KPIs from Web Agents for Policies' Impact Analysis and Products' Brand Assessment*

Antonio Candiello and Agostino Cortesi

Dipartimento di Scienze Ambientali, Statistica ed Informatica,
Università Ca' Foscari, Venice, Italy
{candiello,cortesi}@unive.it

Abstract. Both Enterprises and Public Authorities (PAs) need a continuous and updated flux of reliable data in order to select the better choices. Classical Business Intelligence tools fed with internal data could be augmented with new tools able to extract KPIs of interest from the Raw Web made of unstructured HTML pages and from the Deep Web made of online DBs. The continuous growth of data made available on the web increases intrinsically, year by year, the reliability of this approach. A "Web Intelligence" agents-based framework supporting the evaluation of the effective impact of projects and initiatives has been designed and is currently being developed and tested; the system combines up-to-date indicators obtained via a systematic and high frequency staggered data scraping with lower-rate extraction of data from online data sources. The corresponding model for the management, monitoring and assessment of projects implemented by Enterprises and PAs is also presented.

Keywords: Public Authorities, Business Intelligence, Quality Management.

1 Introduction

It is yearly registered a continuous increase of the quantity of digital data produced, that is estimated by IDC [1] at 1.8 Zettabyte for the 2011. Even if only a fraction of this data is made available on the web (and even less data in text, HTML or XML form), the availability of up-to-date web data is improving the quality of indicators that can be extracted from this wide amount of structured and unstructured information. Properly designed web information agents could help both Enterprises and Public Authorities to gather updated and geographically referentiated data related to the impact of their projects and initiatives.

The popular Web 2.0 model is also making available increasing "User Generated Content" (UGC) data in forums, chats, blogs, wikis and Social Networks (SNs). UGC can be a precious source of information for Enterprises wishing to evaluate their brands' esteem in general. Enterprises can also search in UGC the specific attribute that consumers associate to the products, like reliability, cheapness, quality, luxury, etc. Marketing 2.0 [2] is the advanced form of marketing

* Work partially supported by Regione Veneto – Direzione Sistemi Informativi.

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 192–201, 2011.

that makes use of Web 2.0 tools in order to facilitate and improve the relation between customers and Enterprises. Both *active* supporting initiatives (to suggest Enterprises' products and brands in the appropriate contexts) as well as *passive* analysis activities and tools (searching for relevant product/brand feedbacks) are conducted by the more advanced and customer-oriented Enterprises.

On the other hand, a relevant priority for Public Authorities aiming at promoting innovation and supporting social and economic developments, is the combination of eGovernment policies and domain-related support policies (frequently ICT). The effects of correlated innovation projects should be measurable via focused analysis of specific statistical indicators [3,4]. These can be either *direct* eGovernment or domain-related indicators (as is the case, for instance, of portal/online services access, wide band internet coverage of population, number of patents, etc) or *indirect* (impact) socio-economical indicators (as, for instance, average income, local GDP growth, availability of qualified engineers and so on).

Both consumer feedback trails related to products/brands brought by Enterprises (see [5]) and impact of innovation policies implemented by PAs [6] should be evaluated in an adequately wide time-span, month by month, quarter by quarter. The related feedback/impact measurements should follow this pattern, by registering at definite temporal intervals the state of the monitored indicators. A comprehensive strategy for medium-term branding assessment and/or impact measurement could help to outline the searched correlations between the *effects* (brand appreciation for Enterprises, social and economic improvements for PAs) and the *causes* (new products, better marketing, for Enterprises, or specific innovation policies/projects for PAs [7]). We searched, when possible, to assign to indicators a geographical dimension, where the smallest units considered are the municipalities.

The approach requires a thoughtful choice of specific statistical *Key Performance Indicators* (KPIs). These can be obtained via three main sources:

1. by harvesting the "Raw Web" with webbots, data scrapers, crawlers, spiders, searching for specific trails left by the consumers and the citizens;
2. by retrieving (via the so-called "Deep Web" of the online databases) the indicators available from official Institutions like Eurostat, National Governments, local Chambers of Commerce, or from companies specialized in market analysis, ratings and statistical analysis like Nielsen, IDC or Gartner;
3. by addressing consumers or citizens specific surveys.

As already said, the Raw Web webbot-based strategy is gaining relevance, due to the rapid growth of the quantity and the quality of the information that is available on the web. We expect an increasing trustworthiness of the measurements that will overcome the critical point of this strategy, i.e. the reliability of web-derived indicators data. The *webbots*, said also *data scrapers* when extracting data from a single web source, or *crawlers*, *spiders* when moving between links finding for the needed information, need to be updated when the source web sites change their content or even their appearance (see [8] for an introduction on the

theme; see also [9] for a performance comparison between crawler classes). The research focused on indicators connected to social community-related sources, like blogs, forums and social networks (the Web 2.0) that have the advantage to be continuously up-to-date – and that perhaps will never be closed to the web agents as could be for some internet content [10].

The second approach offers the highest quality data and represents the basic reference thanks to its officiality. The main problem with this approach is that the data officially monitored often do not cover the innovation context to enough detail for PAs, and domain-wide market statistics do not always match Enterprises' needs. Also, these indicators are not always provided in finer territorial detail than the regional one.

Focused surveys offer the opportunity to select specific product or innovation questions, generally not inspected by general statistics. Survey campaigns are generally limited in time (campaigns rarely last for more than a year), in space (provincial or municipal level) or in the context (market sector or consumers segment). Online tools for survey submissions could help in maintaining some efficiency trade-off in the use of this information channel.

The goal of our research is to explicit a comprehensive model for (a) the management of Enterprises' marketing initiatives or PA innovation projects, (b) their systematic monitoring and related impact analysis measurement and to support the model with (c) an integrated *Web Agents Intelligence* information system capable of registering and monitoring such policies, monitor the relative initiatives/projects against their goals, systematically evaluating their impact and finally reviewing the policies themselves on the basis of the resulting analysis.

We conducted the PA-related research in collaboration with our local government partner, Regione Veneto (in northern-east Italy) by extending the QoS monitoring strategy [11] to include impact analysis. The Enterprises-related research was conducted by analysing business-related KPIs on a regional scale and focusing on the products of the local district of sportswear.

The information system developed within this applied research provides public administrators capability to continuously improve their services via an objective evaluation of the resulting impact, keeping their citizens better informed and up-to-date regarding goals set in advance for the policies and the success rate of the local government funded innovation initiatives carried out for the public benefit. Enterprises, on the other hand, are able to evaluate the effective appreciation of their products/brands by the consumers; specific attributes that customers associate to the products (like *performance*, or *durability*, or *comfort*) can also be inspected.

The paper is organized as follows. In Section 2 a model for Enterprises' product initiatives and PAs' services innovation policy management is introduced, and in Section 3 the supporting Web Intelligence framework, its modules and its interaction with the model are presented. Then, in Section 4, some conclusions are drawn.

2 A Model for the Management of Policies for Enterprises and PAs

A comprehensive [12] model for monitoring innovation projects, validating the related policies and evaluating the effective direct and indirect impact on the areas affected could improve the success rate of the Public Authorities innovation initiatives, specifically for ICT [13,14]. Policies and related projects, which we consider mainly relating to eGovernment [4] as well as related to the wider context of ICT infrastructures [15] should be assessed also by the citizens themselves [16]. We adapted the classic Deming plan-do-check-act (PDCA) cycle to the Public Authorities requirements for innovation policies management [17]. This same model can be applied to internal quality management processes of Enterprises for the development of product- and brand-related initiatives. In this last case the impact analysis is substituted by the search for brand-appreciation evidences.

Each policy management PDCA phase is identified by a organizational process and is supported by specific subsystems of the Web Intelligence agents-based framework. The complete model is shown in Fig. 1.

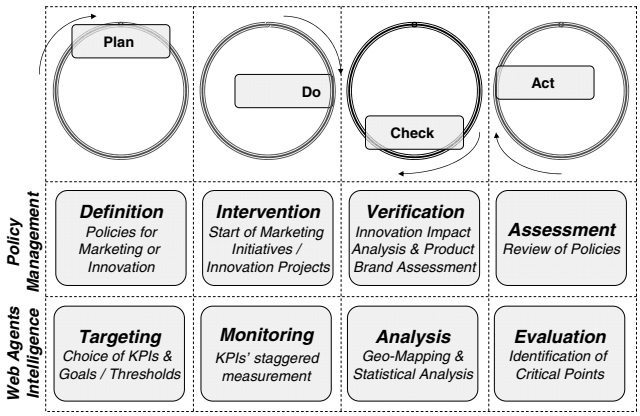

Fig. 1. The comprehensive model for Policy Management, supported by the tools of the Web Intelligence agents-based framework

The goals of this model are:

- finding an *objective validation* for the effectiveness of PAs / Enterprises policies,
- *qualifying and quantifying* the effectiveness through appropriate impact statistical indicators,
- *gathering the relevant indicators* via automatic (webbots/scrapers) and semi-automatic (extractors/wrappers), completing the data when needed with focused survey campaigns,

– *representing and mapping the indicators* showing the explicit relation with the affecting innovation projects and the areas involved.

The indicators are classified in three categories: (a) *direct* innovation indicators, mainly the ICT indicators enabling the Information Society, connected to the technology – examples of indicators in this category are population reached by the internet wide band, eGovernment services offered, eGovernment use of services, ratio of PCs/users, ICT education knowledgeability; (b) *indirect* socio-economical indicators related to the resultant impact of the innovation over the local communities, participation, sociality, economy and culture; (c) *specific* product- and brand-related indicators, able to report the evidence of consumers positive or negative opinions regarding specific products or lines of products.

Direct, innovation indicators are easier to manage, as they are strictly related to the innovation projects. For instance, internet wide band penetration ratio, or renewable energy end-user production could be directly related to infrastructure funding and incentives set up by the local governments; similarly, the growth of the number of accesses to eGovernment Portals depends on quality of the offered online services. These indicators require however the setup of specific measurement processes, as innovation evolution is not systematically monitored by the National or Regional Institutions dedicated to statistical data analysis.

Indirect, socio-economical indicators are more easily found in the periodic data reporting produced by National and International statistical or economical Institutions, but these are *second-level correlated* to innovation projects, i.e. there is the need to estimate and then evaluate their effective correlation with the intermediate innovation indicators which can then be put in direct correlation to the monitored projects. For instance, an investment for wide-area internet wide band could in the long-term sustain new business developments, widen social communities opportunities and in general improve the quality of life. The management of indirect socio-economical indicators requires however carefully staggered gathering of statistical data, and the estimation of the correlations between them and the "raw" innovation indicators. In the current phase of the research, we concentrated our efforts in extracting direct ICT innovation and eGovernment indicators and in selecting simple cases for the socio-economical impact indicators without modeling the effective correlations between the two classes of indicators – we are leaving this task for subsequent research phases.

Specific, product- and brand-related indicators are to be actively searched via appropriate techniques for the elaboration of the text found regarding consumers' opinion trails found on the general web. These indicators can be extracted directly (1) from the number of results of simple queries via online search engines as well as (2) from complex lexical semantic analysis of the pages where the brands and products are cited [5]. The first approach was used for the construction of KPIs, leaving the second approach at an experimental stage for the construction of maps similar to the Brand Association Map from Nielsen (see [18,19]), in an effort to deploy effective Decision Support Systems (DSS) for the Enterprises [20] built on the Web of Data.

3 The Web Agents Intelligence Technical Framework

We developed the Web Agents Intelligence framework around the following elements (see Fig. 2):

- the *Policy Manager*, a GUI panel for the management (and review) of policies, projects, the selection of impact indicators and the setting of targets,
- the *Events Scheduler* for the reliable planning of monitoring events with a minimal temporal unit of a day,
- the *Agents Manager* for the management of webbots/data scrapers, wrappers/adapters and for the launch of email/online survey campaigns,
- the *Map Viewer*, a geo-referentiated visualization engine built around the SpagoBI open source platform.

Let us discuss them in detail.

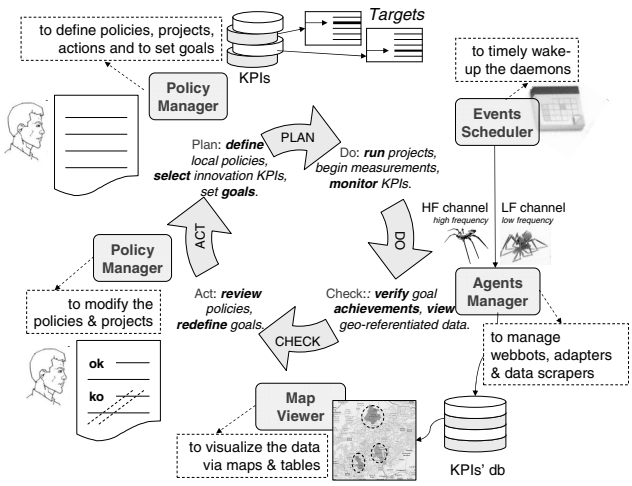

Fig. 2. The agent modules and their relationship with the PDCA cycle

Policies are defined by Enterprises or Public Authorities, and then developed into explicit projects/initiatives. Such projects have definite dates of deployment, they can frequently be geographically localized and they have associated milestones. A *policy manager* has been designed to provide the system all the available data regarding policies, projects/initiatives. The policies, their associated KPIs and targets are defined *ex ante* in the PLAN phase; then, in the ACT phase, the policies are reviewed on the basis of the *ex post* analysis of the associated KPIs.

The *Events Scheduler* manages then (in the DO phase) the execution of the daemons that scan the web, access the online repositories or launch the survey

campaigns via specific web portals. The scheduler uses as the smallest temporal unit the day, and offers the possibility to activate programmed monitoring for timed events ranging from daily intervals, passing through weekly, monthly, quarterly and other sized periodic intervals. The reliability of the scheduler is a central requirement for the framework, as the analysis phases need complete data sets in order to produce trustworthy and significative statistical elaborations. The scheduler fulfills a relevant requirement of our framework: the capability to monitor the trend of the KPIs inspected.

The *Agents Manager* contains the business logic to extract indicators from the web sources. We developed classes of webbots/data scrapers addressing unstructured Raw Web web pages and classes of wrappers/adapters addressing structured Deep Web online databases.

Webbots and data scrapers search for raw information found on the general web. Simple keyword-based searches via major search engines like Google and Yahoo! were experimented. Innovation and ICT-related words were combined to municipality names in order to create geo-referentiated maps; with the same technique products and brands were placed on geographical maps. General Web 2.0 production/consume indicators were also collected by querying Youtube, Flickr and other popular digital repositories. eGovernment indicators were also inspected (indirectly) via Yahoo! Sites, that counts the referring links to the selected web sites. We are currently experimenting the direct extraction of information from blogs- and forum-related web sites. These indicators are mainly used to estimate general innovation and ICT parameters and the brands' popularity via the analysis of the content produced.

The indicators extracted via this channel – webbots and data scrapers for web data harvesting – have a weaker reliability, due to the nature of raw information they are derived from. Also, the agents require a continuous maintenance activity because of the frequent change of the layout of the inspected web sites and/or of their respective APIs used to access the data. On the other way, there is the advantage that this data can be tuned both with respect to the time (as even daily updates can be considered), constituting a *high frequency* channel, and in space (a municipality level is reachable).

As online data source, Eurostat offers the widest option choices for the gathering of structured data, from CSV to XML to several web interfaces and API to manage data: as noted in [21], Eurostat offers more than 4,000 datasets containing roughly 350 million data values/items. Other data sources are more limited in the choices they offer, sometimes even limited to a fixed-format CSV or PDF file. We developed specific extractors for the data reported by the National Statistical Institute collected along with other socio-economical data by the regional Statistical Office managed by our local government partner. Specifically, income, number of inhabitants/families, age distribution data is available year by year at the required territorial resolution of a single municipality. We developed also specific wrappers for common online business directories.

The indicators extracted via this second channel – extractors for official / institutional data retrieval – mainly of socio-economical nature, have the highest

reliability. They also offer the advantage of being completely available at the municipality level. This is a *low frequency* channel, as the updates are typically released with an yearly periodicity. The scheduler has to be instructed to wake up the relative daemons at the right date when the updated data are available.

Webbots/scrapers for Raw Web data and webbots/wrappers for Deep Web data constitute the core of the Agents Manager. We are currently working on the (Semantic Web) extension of the Agents Manager to access the so-called Web of Data [21] or Linked Data [22] in order to extract from these sources higher quality KPIs.

As a third, complementary, input channel, we integrated in the agents manager the eGif engine [23]. The indicators obtained via this channel – mainly citizen feedback regarding eGovernment services and impact of innovation policies – are costly for the effort required in the survey campaign management, but can be useful to complete the statistical analysis for specific themes/areas. Other common online survey solutions can be integrated, as alternatives to eGif, for consumer surveys addressing the needs of Enterprises.

The impact analysis of gathered data is then (in the CHECK phase) managed by the *Map Viewer* with the support of the open source visualization engine SpagoBI engine [24]. The Map Viewer exposes the indicators data over the local territories. It allows to localize the impact of the Enterprises' and Public Authorities' policies. In order to be represented on maps, the KPIs have to be evaluated against a geographical dimension. As the smallest geographical units, we selected the municipalities (corresponding to LAU2 in the Eurostat regional classification).

We are currently experimenting extensions of the SpagoBI platform in order to be able to use also multi-dimensional geo-referentiated data patterns, as the *travelling time distance grid* research case that we tested for mountain-area municipalities and for more abstract patterns appropriate for attributes-brands correlation maps.

4 Conclusions

In this paper a comprehensive policy management model and its supporting Web Agents Intelligence framework has been presented; the model, drawn from quality management methodologies, offers the capability to measure the local impact of innovation policies brought forward by Public Authorities and to reveal the customer opinions regarding Enterprises' specific products and brands. The policy management model and the coupled Web Intelligence framework should help both in reviewing and improving their projects/initiatives by inspecting the resulting impact in detail.

The main features of the model are: (a) the qualification of the policies/projects and the definition of innovation targets, (b) a systematic and staggered measurement of the relevant innovation, economic, social and marketing indicators at the needed scale, (c) a detailed, geo-referentiated analysis of the evolution patterns of the indicators and the relation of products/brands with

specific attributes, (d) the re-assessment of policies and related projects/initiatives against the results obtained.

A first set of core indicators relevant for Public Authorities (ranging from socio-economical data like population, income, business presence, to ICT-related data related to education, eGovernment usage, user-produced content, wide band infrastructures), was extracted from official Institutions and raw web sources. A complete data set of the indicators has been created for all of the regional municipalities; the results, reported on the SpagoBI-powered maps, are currently discussed with regional government staff and the relations with the local ICT innovation initiatives were analyzed.

References

1. Gantz, J., Reinsel, D.: Extracting value from Chaos. Technical report, IDC (2011)
2. Consoli, D., Musso, F.: Marketing 2.0: A new marketing strategy. Journal of International Scientific Publications: Economy & Business 4(2), 315–325 (2010)
3. Janssen, M.: Measuring and Benchmarking the Back-End of E-Government: A Participative Self-Assessment Approach. In: Wimmer, M.A., Chappelet, J.-L., Janssen, M., Scholl, H.J. (eds.) EGOV 2010. LNCS, vol. 6228, pp. 156–167. Springer, Heidelberg (2010)
4. Neuroni, A., Rascon, A., Spichiger, A., Riedl, R.: Assessing and evaluating value and cost effectiveness of e-government initiatives: Focusing the step of the financial evaluation. In: Chun, S.A., Sandoval, R., Philpot, A. (eds.) dg.o 2010. ACM Digital Library, Digital Government Society (2010)
5. Aggarwal, P., Vaidyanathan, R., Venkatesh, A.: Using lexical semantic analysis to derive online brand positions: An application to retail marketing research. Journal of Retailing 85(2), 145–158 (2009)
6. Bernroider, E.W.N., Koch, S., Stix, V.: Elements of Comprehensive Assessments of IT Infrastructure Projects in the Austrian Ministry of Finance. In: Andersen, K.N., Francesconi, E., Grönlund, Å., van Engers, T.M. (eds.) EGOVIS 2010. LNCS, vol. 6267, pp. 84–91. Springer, Heidelberg (2010)
7. Misuraca, G., Ferro, E., Caroleo, B.: Assessing Emerging ICT-Enabled Governance Models in European Cities: Results from a Mapping Survey. In: Wimmer, M.A., Chappelet, J.-L., Janssen, M., Scholl, H.J. (eds.) EGOV 2010. LNCS, vol. 6228, pp. 168–179. Springer, Heidelberg (2010)
8. Schrenk, M.: Webbots, Spiders, and Screen Scrapers: A Guide to Developing Internet Agents with PHP/CURL. No Starch Press (2007)
9. Batsakis, S., Petrakis, E., Milios, E.: Improving the performance of focused web crawlers. Data and Knowledge Engineering 68(10), 1001–1013 (2009)
10. Jennings, F., Yates, J.: Scrapping over data: are the data scrapers days numbered? Journal of Intellectual Property Law & Practice 4(2), 120–129 (2009)
11. Candiello, A., Albarelli, A., Cortesi, A.: Three-layered qos for egovernment web services. In: Chun, S.A., Sandoval, R., Philpot, A. (eds.) dg.o 2010. ACM Digital Library, Digital Government Society (2010)
12. Ojo, A., Janowski, T.: A whole-of-government approach to information technology strategy management. In: Chun, S.A., Sandoval, R., Philpot, A. (eds.) dg.o 2010. ACM Digital Library, Digital Government Society (2010)

13. De', R., Sarkar, S.: Rituals in E-Government Implementation: An Analysis of Failure. In: Wimmer, M.A., Chappelet, J.-L., Janssen, M., Scholl, H.J. (eds.) EGOV 2010. LNCS, vol. 6228, pp. 226–237. Springer, Heidelberg (2010)
14. Janssen, M., Klievink, B.: Ict-project failure in public administration: The need to include risk management in enterprise architectures. In: Chun, S.A., Sandoval, R., Philpot, A. (eds.) dg.o 2010. ACM Digital Library, Digital Government Society (2010)
15. Lampathaki, F., Charalabidis, Y., Passas, S., Osimo, D., Bicking, M., Wimmer, M.A., Askounis, D.: Defining a Taxonomy for Research Areas on ICT for Governance and Policy Modelling. In: Wimmer, M.A., Chappelet, J.-L., Janssen, M., Scholl, H.J. (eds.) EGOV 2010. LNCS, vol. 6228, pp. 61–72. Springer, Heidelberg (2010)
16. Tomkins, A.J., PytlikZilllig, L.M., Herian, M.N., Abdel-Monem, T., Hamm, J.A.: Public input for municipal policymaking: Engagement methods and their impact on trust and confidence. In: Chun, S.A., Sandoval, R., Philpot, A. (eds.) dg.o 2010. ACM Digital Library, Digital Government Society (2010)
17. Candiello, A., Cortesi, A.: KPI-Supported PDCA Model for Innovation Policy Management in Local government. In: Janssen, M., Scholl, H.J., Wimmer, M.A., Tan, Y.-H. (eds.) EGOV 2011. LNCS, vol. 6846, pp. 320–331. Springer, Heidelberg (2011)
18. Akiva, N., Greitzer, E., Krichman, Y., Schler, J.: Mining and Visualizing Online Web Content Using BAM: Brand Association Map. In: Proceedings of the Second International Conference on Weblogs and Social Media, pp. 170–171. Association for the Advancement of Artificial Intelligence (2008)
19. Till, B.D., Baack, D., Waterman, B.: Strategic brand association maps: developing brand insight. Journal of Product & Brand Management 20(2), 92–100 (2009)
20. Dai, Y., Kakkonen, T., Sutinen, E.: MinEDec: a Decision-Support Model That Combines Text-Mining Technologies with Two Competitive Intelligence Analysis Methods. International Journal of Computer Information Systems and Industrial Management Applications 3, 165–173 (2011)
21. Hausenblas, M., Halb, W., Raimond, Y., Feigenbaum, L., Ayers, D.: SCOVO: Using Statistics on the Web of Data. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E.P.B. (eds.) ESWC 2009. LNCS, vol. 5554, pp. 708–722. Springer, Heidelberg (2009)
22. Bizer, C.: The Emerging Web of Linked Data. IEEE Intelligent Systems 24(5), 87–92 (2009)
23. Candiello, A., Albarelli, A., Cortesi, A.: An ontology-based inquiry framework. In: Gangemi, A., Keizer, J., Presutti, V., Stoermer, H. (eds.) DEGAS 2009. CEUR Workshop Proceedings, vol. 426 (2008)
24. Golfarelli, M.: Open Source BI Platforms: A Functional and Architectural Comparison. In: Pedersen, T.B., Mohania, M.K., Tjoa, A.M. (eds.) DaWaK 2009. LNCS, vol. 5691, pp. 287–297. Springer, Heidelberg (2009)