

# The Online Market Observatory: A Domain Model Approach

Norbert Walchhofer \*      Milan Hronsky \*  
Karl Anton Froeschl †

Preprint of a contribution sent to  
**KSEM2009** - *3rd Int. Conf. on Knowledge Science, Engineering and Management*  
20. October 2009

## Abstract

“Semantic Market Monitoring” (SEMAMO) project aims at the prototypical implementation of a generic online market observatory. SEMAMO is intended to provide a flexible empirical instrument for the continuous collection of data about products and services on offer through WWW portals. Based on a uniform data processing scheme covering all stages from data capture using configurable mediators, through integration of data from multiple sources and persistent storage, up to statistical analyses and reporting functions, SEMAMO delivers a self-contained formal specification framework of market monitoring applications. The formal descriptions of application domains, data integration tasks, and analyses of interest facilitate, by deductive conversions, the arrangement and execution of all internal data and storage structures, observation processes, data transformations, and market analytics, respectively. Specifically, SEMAMO exploits formalised domain structures to adaptively optimise data quality and observation efficiency. The framework is evaluated practically in an application to online tourism.

Keywords: online market monitoring, information extraction, semantic technologies

---

\*EC3—e-commerce competence center, Vorlaufstr. 5/6, 1010 Vienna, Austria

†University of Vienna, Institute of Scientific Computing, Universitaetsstr. 5, 1010 Vienna, Austria

# 1 Introduction

Because of the gradual spread of the World Wide Web into the business domain during the last few decades [1][2] it seems fairly natural to extend traditional methods of market observation [3] to cover online, or electronic, markets as well. While, with an emphasis on single online portals at a time, Web analytics [4] address many of these questions already, there have been few attempts in gathering online market information simultaneously across a variety of online portals, aiming at the statistical integration of collected observations. As a matter of fact, this endeavour links approaches of “business intelligence” [5] to methods of data integration [6], calling for a smart as possible combination of existing proposals and practices from either field of research. This paper describes a generic instrument of observation reflecting the typical conditions of collecting data about online markets. Assuming a reasonably wide class of empirical problems sharing a common structure, the generic observation model is embedded in a configuration space comprising (i) a group of parameters defining a particular observation setting and (ii) another group detailing the processing flow the observations, once gathered, eventually undergo. The ensuing observation instrument is currently implemented as a research prototype named SEMAMO, for “Semantic Market Monitoring”, paying heed to the use of recently so-called *semantic technologies* [7] for the specification of both the observation structure and processing flow of a particular application instance (online tourism markets, this time) of the observatory.

The following sections present first an overview of how to formally characterize the relevant domain structures of an online observation setting, highlighting the generic nature of the devised model using illustrations from online tourism markets [8], before the focus is shifted to the procedural dimension, highlighting the effective utilization of formalized domain relationships in arranging and conducting data pre-processing and data analytical functions of the market observatory: Section 3 sketches a typical task of online observation processing, namely the re-identification of encountered entities as a precondition to linking up successive observations to time series, whereas Section 4 explains the model-driven role of adaptiveness in achieving efficiency of the empirical monitoring process. Section 5, finally, provides a summary and reflects basic assumptions of the SEMAMO approach.

# 2 The Domain Model

From an economic point of view, a market comprises a multitude of different offer channels, some making use of the Internet in terms of online shops or electronic marketplaces. In order to gain insight about market behaviour, systematic surveillance of *many* such channels making up a market is required. In fact, market transactions depend crucially on the provision of information about qualities and prices of things on offer which in case of online channels, such market-specific *digital* information is both easy to capture and fast to process and disseminate, given appropriate resources for data collection and analysis. The proposed empirical instrument of online market monitoring, SEMAMO, revolves around a persistent integrated storage of incrementally accumulated online market data fed forward to multiple, and possibly not entirely anticipated, analytical uses in terms of statistical aggregates and models ready for subsequent decision making or theoretical investigation. SEMAMO implements a three-tier preformed *input>storage>output* approach, gathering data from (predefined) multiple heterogeneous online sources, storing these data in an internal canonical format persistently, and distributing integrated data to various analytical purposes and uses, most of the time comprising “near real-time” statistical aggregation and reporting. More specifically, all of these processing steps are embedded in a periodically repeated *main control loop* in order to achieve an automatic and efficient as possible mode of operation. SEMAMO represents a particular domain of applica-

tion in formal terms as an instantiation of its generic configuration space with respect to (i) a *structural* and (ii) a *processing* dimension, using the structural description of an application domain to determine and govern the actual data processing within the main control loop.

## 2.1 The Structure Model

Any empirical model rests on (i) a sampling population of observation units and (ii) the set of attributes to observe on each population unit [9]. In online market monitoring, well-defined sampling populations can be hardly surveyed directly; rather, they are constructed from a designated set of portals conveying the (quantitative) information of interest: the units of an online population may become accessible through different portals, possibly depending, with at least some of their observable features, on the respective portals. Accordingly, provision has to be made for tracking the appearance of individual population units on different portals. Furthermore, the units of interest may become "clustered" by an intermediating actor, altogether giving rise to (at least) a three-pronged population structure composed of

- a *port* population, designating the portals monitored;
- a *sampler* population, comprising the primary units of observation; linked by
- a *switch* population, the units of which are generally used to represent whole clusters of primary observation (that is, sampler) units made accessible on a subset of portals within a port population.

What is actually observed online, then, always consists of a genuine combination of a port, a switch, and a sampler population, that is, any sampler is accessible at a time in one of the ports, through one of the switches in between. In operative terms, hence,  $\langle \text{port, switch, sampler} \rangle$ -unit triplets represent the units of observation proper, entailing yet another emergent *sensor* population of their own. In turn, sensor populations are naturally stratified by ports and switches, opening a combinatorially rich field of comparative statistics. The resulting SEMAMOGraph is depicted in Fig. 1 (left part), exhibiting a typical (though partial) implementation of the market monitoring schema of the running SEMAMO tourism instance. Note that, for sensor, switch, and sampler populations, the nodes in this schema represent whole populations, respectively, whereas for the port population each node represents a single portal registered.

To further facilitate the formal (schema) definition of populations and population linkages as stated in the SEMAMOGraph, lower-level information structures, such as (structured) value domains, attributes, and the like, are supplied (but not shown in Fig. 1); while these are provided and maintained in supplementary directories and codebooks, *registries* hold the current population frames, that is, the population units encountered during repeated observation processes. A specific registry records information on *sources*, represented by dashed arcs in Fig. 1. Logically, a source delivers data from a port to a defined sensor population, implying generally the configuration, and application, of some mediator [10][11] actually delivering observation data on sensors (cf. Subsection 2.2). The SEMAMOGraph, as shown, can be conceived practically also as a direct graphical manipulation interface governing the addition, or update, of schema elements in a context-dependent mode.

Contrary to the perhaps suggestive case illustrated in Fig. 1, switch and sampler populations might be coupled in different ways (by adding further sensor population elements and linking arcs to the SEMAMOGraph) to introduce sensor populations as intended; doing so is reasonable so long as there are real sources contributing data.

Within an individual observation cycle, SEMAMO applies all defined mediators of the established sources to the respective online portals (cf. Section 4). These sources generate, through their attached wrappers (that is, mediators), temporary raw data termed "canonical wrapper records" (CWR, for short) on sensors. More specifically,

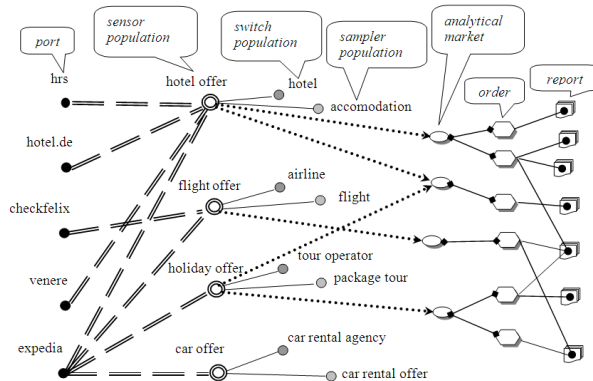


Figure 1: Tourism Application SEMAMOGraph

wrappers implement port-specific query plans covering predefined portions of sensor populations, expressed in terms of domain model structures. Accordingly, the outcome of a wrapper application consists of a set of sensor observations, representing the “state” of these sensors at extraction time. In order to link the “harvested” sensor data to the already registered population units, a re-identification process (cf. Section 3) commences seeking to match CWRs to switch and sampler population units, respectively; in case of success, the harvest (that is, observation) series of a sensor becomes extended by this new observation, otherwise a sensor has been encountered the first time and, hence, population registries have to be updated accordingly. Unit re-identification amounts, of course, to a record linkage problem [12] making use of some distance metrics, again governed by information supplied through the SEMAMO domain model.

From a data model point of view, the interrelation between the different types of data structures representing port, source, switch, and sampler elements revolving around a single sensor population instance could be sketched as shown in Fig. 2. While the top layer of the exhibit indicated the sensor data flow from ports to the staging area pooling the temporary observation CWRs, both master and harvest data areas denote structurally constant SEMAMO storages: obviously, the *master data* area gathers the registries of all populations defined through a SEMAMOGraph, thus depending on a particular SEMAMO application, though reusing the same basic structure elements over and over again, whereas the *harvest data* area—holding the collected harvest series for all sensor populations maintained—resorts to a uniform built-in data representation called “canonical harvest records” (CHR, for short).

Harvest series, built of time-indexed CHRs, are linked to master data through so called “sensor registry records” (SRR, for short) chaining together ports (sources, in fact) with switch and sampler units to individual sensors by means of a simple composite foreign key relationship. Likewise, the SEMAMO domain model establishes the outbound data structures of a SEMAMO application instance. Naturally, the arranged sensor populations of a SEMAMO instance define the interface between storage and analysis areas, respectively, by supplying to the latter the time-dependent harvest series linked to the various available attributes cross-classifying a sensor population. This is to say that, on top of the set of sensor populations in a SEMAMO application, so-called *analytical markets* define the crucial domain elements of the analytical area as depicted in Fig. 1 (right part): each analytical market implements a particular view on the harvest data base of SEMAMO by selecting a targeted subset of gathered sensor (that is, CHR) data.

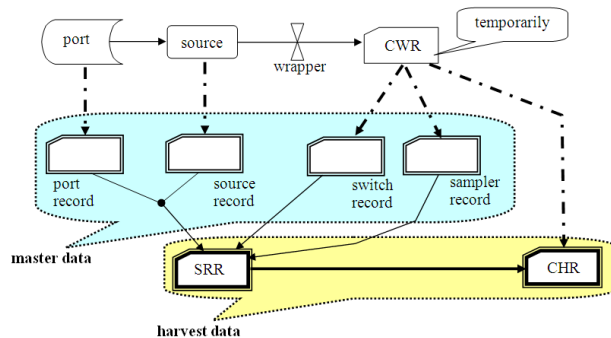


Figure 2: SEMAMO Registry Structures

As a view on harvest data, analytical markets buffer harvest data for subsequent analysis, still referring to individual sensor data so as not to exclude any reasonable type of statistical market analysis. Additionally, for each analytical market arranged, a market pivot table (MPT, for short) is created internally which provides elementary aggregates—or, market indicators—for each of the indicated value dimensions (counts, price, capacity), cross-classified with respect to all market dimensions inherited from the harvest data definition: thus, an MPT in fact represents a basic data warehouse cube [13] in support of all ensuing OLAP. Again, analytical markets are registered in the SEMAMO domain model, guided by information taken from the sensor population registry. With respect to MPTs, also the statistical aggregation functions on value dimensions to be used (e.g., count, average, median, sum of squares, min, max, . . .) are specified here.

Based on analytical markets, order and report structures are added in a straightforward way; orders simply encode particular types of statistical analyses applied to analytical markets (typically, in a periodical fashion) in order to gain information on market monitoring or decision making levels, triggered by reports defining analytical schedules by grouping sets of analyses composed of tabular and graphical representations of analysis outputs obtained from order processing. Eventually, reports encode certain information interests of decision makers or market researchers (in fact, reports may exhibit a tree-like composite structure) and, in doing so, provide an interface to external SEMAMO service customers. To this end, both reports and orders are recorded in appropriate domain model registries. In particular, each order specifies both (i) the type of analysis alongside with any method-specific parameters, and (ii) its linkage to the analytical market it depends on. Conversely, report schedules determine the due dates for order processing which, in turn, trigger the replenishment of the data buffers (and MPTs) associated with analytical markets. Finally, the SEMAMO domain model provides auxiliary “semantic” information structures, mostly on the level of value domains, such as synonym tables, domain vocabularies/thesauri, or coincidence/correspondence data relating different semantically overlapping classifications or meronomic subdivisions [14] useful in various SEMAMO processing stages.

## 2.2 The Process Model

The SEMAMO structure model provides formal representations of all the different objects and established object relationships characterizing a particular application, as defined through a SEMAMOGraph instance. In addition to using this formal frame of reference, the actual operation of the online market observatory calls for a minute specification of all processing steps within the generic processing flow. On top, the operation of SEMAMO is controlled by two main processing loops, viz.

- a *harvest cycle* accessing all defined sources attached to Web portals registered for monitoring through the arranged wrappers executing predefined (but adaptively instantiated, depending on different internal harvest parameters; cf. Section 4) query plans;
- an *order cycle* responsible for executing all registered statistical analyses as mandated by report schedules.

Indirectly, both loops interact since the harvest cycle writes periodically to the harvest data base while the order cycle reads from it, although this interaction need not be in sync. As to the job specification, the harvest cycle is by far the more complex one and, contrary to the order cycle, only few processing steps can be composed simply from toolbox functions: in spite of the repetitive nature of Web extraction, data cleansing (or, more generally, so-called ETL processing), and unit re-identification tasks, many transformation functions critically depend on local circumstances and, hence, need careful tuning and testing prior to deployment—activities hardly amenable to automation. Still, there are better defined processing steps such as intermediate schema mappings [15] (notably, the conversion of query plans to parameters controlling wrapper application to portals through binding patterns [16] in Web extraction jobs, or the definition of forward mappings used to integrate sources [17] at sensor population level), for which function libraries can indeed be provided in terms of configurable tool sets.

In the SEMAMO prototype, data integration tasks are located in two places: first, in combining different sources to sensor populations (dashed arcs in Fig. 1) and, secondly and technically much simpler to accomplish, combining different sensor populations to analytical markets (dotted arrows in Fig. 1). While, at both stages, a “global-as-view” integration applies, data requests actually move upstream from reports compiling orders which, implicitly through the analytical markets they refer to, demand the provision of queries to data from some subset of portals. Thus, eventually, queries expressed in terms of analytical market dimensions have to be re-translated into wrapper binding patterns of query plans which harvest portal data to be converted into sensor attributes. In fact, SEMAMO makes use of a weighting scheme reflecting the (sometimes changing) data demands of the order cycle which is then passed on to the harvest cycle.

The least standardized processing stage of SEMAMO concerns cleansing and rectification of portal wrapper data. Generally, wrappers have to be hand-crafted [18] (often supported by visual tools, such as the Lixto Visual Wrapper [19]) and data cleansing operations custom-tailored to the peculiarities and anomalies of individual data sources. However, as data quality is often checked against statistical criteria, many parameters for data validation and editing derive from internal statistical analyses applied to the harvest data base and, thus, are provided in SEMAMO routinely for inclusion in data cleansing functions. The explicit specification of work flows in SEMAMO transformation and schema mapping steps amounts to the (serial) attachment of function modules to SEMAMOGraph arcs much in the same way popular data mining tools [20][21] support GUI-based process specifications. Despite these indispensable customisation tasks, the specification of many core and auxiliary processing steps of SEMAMO can be deduced directly from an application’s domain model (cf. Subsection 2.1). This includes ancillary service processes (such as the update of population registries in case of sensors encountered for the first time, or the empirical induction of coincidences between terms of different classification hierarchies, the continuous update of data validation statistics, etc.) as well as more fundamental building blocks of the main harvest cycle of SEMAMO such as the periodic computation of the adaptive harvest schedule actually governing wrapper application (cf. Sec. 4).

### 3 Unit Re-Identification

Careful maintenance of population registries is a salient prerequisite of good data quality [22]. In SEMAMO, the reference to population units is established through the observation process: by scanning online portals, units are registered and, in case of repeated observation, have to be re-identified properly in order to assign time-dependent observations correctly to (re-)identified population units. For the sake of distinction, let *canonical unit* denote the rectified representation of a real entity encountered in an online observation process whereas a *harvest unit* (represented in a CHR; cf. Subsection 2.1) refers to an observation not yet assigned to a canonical unit. Contrary to canonical units bearing unique registry entries by definition, harvest units may feature ambiguous representations, possibly depending on the source they originate from. Figuring out this assignment constitutes the critical task of unit re-identification. The SEMAMO domain model defines type and structure of population units used in an application. In particular, monitored portals contribute switch and sampler units to the population registries such that, in general, portal sub-populations overlap. Accordingly, population registries are induced by sifting out matching harvest units and maintaining the unique ones as canonical population units in a rectified representation (cf. Fig. 2). Apparently, this matching rests on probabilistic decisions, with a re-identification process composed of two stages, viz. (i) *pair-wise record matching* and (ii) a ‘*link analysis*’ chaining pair-wise matching instances to the identified canonical unit. However, contrary to stage (i), stage (ii) does not depend on the domain model of an application and, hence, is not dealt with any further in what follows. Pair-wise record matching in SEMAMO is a combinatorial process based on some population schema  $\vec{S}(A_i, \dots, A_k)$  comprising the  $k > 0$  attributes  $A_i, \dots, A_k$  defined in a SEMAMO domain model. The extension of schema  $\vec{S}$  is the relation  $R(A_i, \dots, A_k) = \{r(a_1, \dots, a_k) | a_j \in \text{dom}(A_j), 1 \leq j \leq k\}$ . Let  $r(B)$  denote the projection of a tuple  $r \in R(A_i, \dots, A_k)$  over a non-empty attribute subset  $B \subseteq \{A_i, \dots, A_k\}$ . Furthermore, let  $r' \equiv r''$  express the fact that  $r', r'' \in R(A_i, \dots, A_k)$  represent the very same canonical unit (equivalence) whether or not  $r' = r''$  (syntactical equality). To facilitate unit re-identification, SEMAMO resorts to two kinds of semantic rules. First, value ambiguities are resolved using edit rules covering the case of ‘near functional’ attribute dependencies: for some  $J \subseteq \{A_i, \dots, A_k\}$  provided that  $r'(J) = r''(J)$  for a record pair  $r', r'' \in R(A_i, \dots, A_k)$ ,  $r' \equiv r''$  can be assumed. Edit rules are justified on statistical grounds that  $J$  is chosen such that the entailed rate of false positive matches remains very low (although, strictly speaking,  $r' \equiv r''$  is logically not warranted even in case  $r' = r''$  because of undetected homonymy). For instance, letting the attributes  $A_1$  = ‘hotel name’,  $A_2$  = ‘street address’,  $A_3$  = ‘ZIP code’,  $A_4$  = ‘city name’, and  $A_5$  = ‘country code’ in a hotel population schema, it is fairly safe to expect that  $J = \{A_1, A_2, A_3, A_5\}$  already match a pair of hotel harvest records even if city names in the records compared differ:  $r'(\{A_4\}) \neq r''(\{A_4\})$ . Viewed another way, given such an edit rule, the disparate values of attributes in set  $M = \{A_1, \dots, A_k\} - J$  represent *synonyms* relative to the matching values of  $J$  (known synonyms can also be tried, of course, as a replacement of originally non-matching attributes in  $J$ ). Quite frequently, a set  $E \subseteq R(A_1, \dots, A_k)$  such that, for  $r', r'' \in E$ ,  $r'(J_E) = r''(J_E)$  for some non-empty  $J_E \subseteq \{A_1, \dots, A_k\}$  may not justify to conclude  $r' \equiv r''$  from statistical evidence. In those cases, *discriminant rules* are applied to decide whether or not two records match. To this end, letting  $M_E = \{B_1, \dots, B_m\} = \{A_1, \dots, A_k\} - J_E$ , a proximity vector function

$$\vec{p}_E = (p_1, \dots, p_m) : \times_{j=1}^m (B_j \times B_j) \rightarrow [0, 1]^m \quad (1)$$

is arranged, composed of component proximity functions  $p_j$  depending on the data type of attribute  $B_j \in M_E$  (for instance, string comparison metrics [23] for alphanumeric attributes). Each classifier  $D_E(r'(M_E), r''(M_E))$  has to be trained, of course, using

a representative sample of application-specific data, but may be as simple as a linear (perceptron-type) normalised separator criterion  $\vec{p}_E \cdot \hat{w}_E$  with estimated weights  $\hat{w}_E > \vec{0}$ , yielding acceptable results in the tourism application case, that is, in detecting matches between hotel units harvested from four different online hotel booking portals. Albeit expressed in numerical terms, the trained classifiers encode salient knowledge about the application domain.

## 4 Adaptive Online Observation

A pivotal function of any empirical instrument consists in generating observations. In case of online markets, observations originate from taking “snapshots” of Web portals using mediators wrapping portal content or, more specifically, capturing attribute values of defined population units in a systematic fashion. Moreover, monitoring online markets entails the repeated observation of population units in order to capture market dynamics appropriately, suggesting a cyclic organisation of the observation process (cf. Subsection 2.2). Taking observations online is subject to quite specific “experimental” conditions, invalidating many commonly presupposed sampling conditions:

- the conceived online populations can rarely be enumerated; in other words, well-defined sampling frames and, thus, selection probabilities are lacking;
- selecting individual population units at random within portals may not be feasible; rather, because of technical access properties, bulks of population units can be sampled only, which may impede representativeness and coverage of the sample data gathered;
- the number of observations taken within a period of time happens to be bounded because, otherwise, the virtually unrestricted gathering of observations for purely observational purposes interferes with the normal business mode of portal operation [24], possibly provoking access denials;
- observation attempts simply may fail (that is, resulting in unit or item non-response), introducing a further kind of non-sampling error into the data collection process.

To embrace a large as possible class of monitoring scenarios, SEMAMO implements a fairly generic observation model which attaches a harvest series of successive measurements of a quantitative variable (price, in general) to a single sensor (unit of observation as defined in Subsection 2.1). Thus, essentially, each sensor keeps track of the prices for a product or service on offer (sampler unit) accessible online through a portal and, inside the portal, commercially through a vendor or market intermediary (switch unit). The set of sensors in an observed population is broken down multi-dimensionally in terms of recorded offer features as defined by virtue of the SEMAMO application domain model, in addition to portals and switch units providing further stratifications. Through repeated observation, harvest series per sensor are generated incrementally and stored persistently in the SEMAMO harvest data base for later use in statistical analyses. The dynamics of the observed quantitative variables may be modelled in different ways. In the tourism application, for instance, price trajectories are conceived as a step function, that is, prices are assumed to change discretely once in a while, after an average (but unknown) interval of time. This way the temporal behaviour of a sensor mirrors the updating of online content in terms of a change rate model [25][26]. Generally speaking, for sampling purposes the sensors’ change rates are estimated from observations recorded in the harvest data base, taking into account also failed observation attempts, and converted into adaptive harvest weights based on a heuristic which maps (price) change dynamics to observation frequencies [27]: basically, this is to say that the sampling probability (or, frequency) of a sensor is made dependent adaptively on the evident variability of previous measurements. In each

harvest cycle, by grouping sensors according to hitherto observed change dynamics, a stratified *harvest schedule* is derived which, then, is mapped onto available query plans for execution through the pre-configured portal wrappers. In a sense, schedule strata segment markets reflecting differences in observed dynamics. In general, query plans intersect with schedule strata to varying degrees, so that, in order to strike an optimal balance between the scheduled sample of sensors determined for observation and the tolerated access quotas of the portals monitored, the selection of query plans (resource allocation) in a harvest schedule is accomplished by applying a meta-heuristic [28].

The derivation of adaptive harvest schedules is built into the main harvest cycle of SEMAMO. This feedback harvesting scheme implements a self-calibrating model of online market monitoring, determined and governed through a range of parameters provided through the SEMAMO domain model, such as (i) the subject-matter defining both sensor and constituent populations, (ii) the quantitative variables of interest, (iii) the choice of a model underlying the dynamics of observed quantitative variables, (iv) the query plans available to schedule implementation, (v) a set of numerical parameters and threshold values controlling the heuristics involved, and (vi) access quotas and cost functions per portal relevant for schedule optimisation. Hence, by arranging (or changing) the domain model of an application as outlined in Section 2, this observation instrument is attuned, though indirectly, to the respective experimental set-up of a market monitoring application, provided that the required data sources (wrapper configurations) have been arranged appropriately.

## 5 Conclusions

Amongst other effects, the “digital turn” of commerce tremendously increases the pace of all business processes, bearing competitive forces and responses. In particular, online markets offer the access to a wealth of business-related information on market sizes, market dynamics, and competitive market behaviour at large. Certainly, early trend anticipation and quick reaction to market changes become even more of crucial importance to business success once major fractions of markets have gone online. In some sectors, such as the tourism industry, this situation is no doubt established already. Besides the individual market participant’s view, of course, a surveillance of online markets is of considerable macro-economic interest as well, be it, say, for estimating the growth in sales volumes, or with respect to assessing concentration tendencies possibly calling for antitrust or other regulatory measures.

All of this suggests the use of powerful and flexible means of online market monitoring, exploiting the abundance of publicly available data jointly mirroring an important part of today’s economic reality. Presuming a general structure of empirical approaches to online market dynamics, the presented SEMAMO approach prototypically implements a fairly generic strategy of online market monitoring, defining the empirical process in principle, and leaving the detailed specification of the various structural and procedural parameters of the instrument to the formal-symbolic representation of each application domain from which the internal gears and operations of the instrument are deduced mechanically as far as possible. This way, the use of the instrument itself is economised in terms of a specialised toolbox facilitating “online market statistics on demand” across a wide range of application domains. It remains to be seen, mainly through practical experimenting, whether the SEMAMO approach in fact meets its expectations and, in particular, if the concepts developed may extend to a still more general view of a “Web Observatory”.

### Acknowledgments.

The authors gratefully mention the valuable discussions with Wilfried Grossmann, Marcus Hudec, Wolfgang Jank, Patrick Mair, and Kurt Hornik on statistical issues of

SEMAMO. The project is supported by grant Fit-IT “Semantic Systems and Services” No. 815.135 of the Austrian Research Promotion Agency (FFG).

## References

- [1] Shaw, M.; Blanning, R.; Strader, T.; Whinston, A. (Eds.) Handbook on Electronic Commerce. Springer, Berlin, (1999)
- [2] Romm C.T.; Sudweeks, F. (Eds.) Doing Business Electronically - A Global Perspective of Electronic Commerce. Springer, London et al., (1998)
- [3] Kotler, Ph.; Armstrong, G. Principles of Marketing. Pearson, Prentice Hall, N.J., 2007
- [4] Dhyani, D.; Ng, W.K.; Bhowmick, S.S. A Survey of Web Metrics. ACM Computing Surveys 34 (4), December 2002: 469-503.
- [5] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy R. (Eds.) Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, Menlo Park, Ca., 1996.
- [6] Lenzerini, M. Data Integration: A Theoretical Perspective. PODS 2002: 233-246.
- [7] Davies, J.; Studer, R.; Warren, P. Semantic Web Technologies: Trends and Research in Ontology-based Systems. Wiley, New York, 2006.
- [8] Werthner, H.; Klein St. Information Technology and Tourism, A Challenging Relationship. Springer, Wien, 2008.
- [9] Levy , P.S.; Lemenshow, S. Sampling of Populations: Methods and Applications. Wiley, New York, 1999.
- [10] Wiederhold, G.; Genesereth, M. The Conceptual Basis for Mediation Services. IEEE Expert 12 (5), Sept./Oct. 1997: 38-47.
- [11] Baumgartner, R.; Frölich, O.; Gottlob, G. The Lixto Systems Applications in Business Intelligence and Semantic Web. ESWC 2007: 16-26
- [12] Fellegi, I; Sunter, A. A Theory for Record Linkage. JASA 64 (328), December 1969: 1183-1210.
- [13] Kimball, R.; Ross, M. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. Wiley, New York, 2001.
- [14] Visser, U. Intelligent Information Integration for the Semantic Web. Springer (LNAI #3159), Berlin et al., 2004.
- [15] Batini, C.; Lenzerini, M.; Navate, Sh.B. A Comparative Analysis of Methodologies for Database Schema Integration. ACM Computing Surveys 18 (4), 1986: 323-364.
- [16] Rajaraman, A.; Sagiv, Y.; Ullman, J.D. Answering Queries Using Templates with Binding Patterns. PODS 1995: 105-112.
- [17] Halevy, A.Y. Answering Queries Using Views: A Survey. VLDB Journal 10 (4), 2001: 270-294.
- [18] Chidlovskii, B.; Borghoff, U.; Chevalier, P. Towards Sophisticated Wrapping of Web-based Information Repositories. RIAO 1997 (Montreal): 123-155.
- [19] LiXto Visual Wrapper; <http://www.lixto.com/li/liview/action/display/frmLiID/12/> (last accessed June 24, 2009)
- [20] SAS Enterprise Miner; <http://support.sas.com/documentation/onlinedoc/miner/index.html> (last accessed June 24, 2009)
- [21] SPSS Modeler (previously: Clementine) <http://www.spss.com/software/modeling/modeler/> (last accessed June 24, 2009)

- [22] Herzog, Th.N.; Scheuren, F.J.; Winkler, W.E. *Data Quality and Record Linkage Techniques*. Springer, New York, 2007.
- [23] Winkler, W.E. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *ASA Proc. of the Sect. on Survey Res. Methods*, 1990: 354-359.
- [24] Hess, C.; Ostrom, E. Ideas, Artifacts, and Facilities: Information as a Common-pool Resource. *Law and Contemporary Problems* 66, 2003: 111-145.
- [25] Cho, J; Garcia-Molina, H. Estimating frequency of change. *ACM Transactions on Internet Technology* 3 (3), 2003: 256-290.
- [26] Grimes, C; Ford, D. Estimation of web page change rates. *JSM: 3968-3973*. 2008 (Denver, Co.)
- [27] Walchhofer, N.; Froeschl, K.A.; Hronsky, M.; Hornik, K. Adaptive Web Harvesting in Online Market Monitoring. submitted to *Journal for Advanced Data Analysis and Classification*. (2009)
- [28] Kellerer, H; Pferschy, H.U.; Pisinger, D. *Knapsack Problems*. Springer, Berlin (1983)