

Towards Models for Judging the Maturity of Enterprises for Semantics

Marek Nekvasil, Vojtěch Svátek

Department of Information and Knowledge Engineering, University of Economics, Prague,
Winston Churchill Sq. 4, 130 67, Prague 3, Czech Republic
{nekvasim,svatek}@vse.cz

Abstract: In recent years, semantic technologies have been included in broader and broader areas of application deployment, and their scope has been constantly expanding. The differences amongst them, however, are often vast and the successes of such investments are uncertain. This work provides a possible approach to the categorization of semantic applications and uses it to formulate a set of critical success factors of the deployment of these technologies in a business environment. Finally, it outlines how it is possible to formulate the maturity models of enterprises for preliminary assessment of the investments into semantic applications.

Keywords: semantic technology, critical success factors, process maturity

1. Introduction

Before 2001, the web was regarded by the wider community as a mere conglomeration of static web pages, but then Tim Berners-Lee, former director of the W3 consortium¹, introduced in his most famous article [1] the concept of Semantic Web. In the following period semantic technology became popular and nowadays we find its applications in much broader areas than ever before: from applications that integrate data from different sources, support the search in a diverse range of data, derivate new relationships across heterogeneous databases, including the application support of social networking, management decision-making, annotating and indexing of any content, for up to such different tasks as information extraction from unstructured sources, and even so-called Business Intelligence 2.0 [5].

Such a wide range of semantic (i.e. knowledge-based) technologies is mainly caused by the fact that the general understanding of what can be considered as a semantic application is somewhat loose, thus there are no universally accepted definitions. For example, according to [9] any application that stores data separately from the meaning and content files, and in the same time does not have the meaning hard-wired into the application code, can be called *semantic application*. This concept

¹ <http://www.w3.org/>

includes the use of ontology languages (such as RDF², RDFS³, OWL⁴, etc.) and rule-based systems.

The W3C Semantic Web Education and Outreach (SWEO) Interest Group collected and published case studies of existing applications and potential use cases that take advantage of semantic technologies in praxis [8]. Thanks to this overview it is possible to gain insight into the current state of semantic applications and their usability in the production environment.

In the pursuit of using the semantic applications in commercial sphere it is necessary to justify the respective investments. However there are many views on what the gains of such investments are. Some of these views are clear and straightforward, such as the analysis of financial characteristics and indexes which compare just the costs and revenues. Others are not so clear but are at least equally important; especially in this case where the benefits of the investment can only be quantified with great difficulties and very roughly. Such views include but are not limited to the added value for customers or the productivity increase of employees. Much more essential by the time of assessing the investment is estimating (or defending) its feasibility and determining the necessary conditions under which the whole project will not be loss-making.

However, as we have already mentioned, the notion of semantic application is very diverse from project to project, hence the conditions of feasibility and potential profitability cannot be set generally (yet, some overviews have also been published, see [7]) but it is necessary to identify some *categories of knowledge-based applications* in the first place. Only once these categories were identified, it would be possible to formulate the requirements, because each kind of semantic application can be substantially different.

This work aims at several objectives and identification of the most common categories of semantic applications is only the first one of them. After the applications have been categorized it will be possible to isolate some of the substantial properties according to the categories. While judging the gains on this level would still be very general, we will propose some possible *critical success factors* (CSFs [6]). Therefore the next objective is to establish the most important CSFs of deploying (and developing) the semantic applications.

Finally we will try to outline the manner of how it would be possible to formulate the *maturity models* for deployment of knowledge-based applications (in the sense of the maturity of enterprise processes, according to the original W. Humphrey's work [3]) for some types of such applications based on the aforementioned critical success factors (as a reference model [4]).

2. Categorization of knowledge-based applications

As noted above the knowledge-based applications cannot be considered as a compact area of interest, because indeed they are very heterogeneous uses of the appropriate

² <http://www.w3.org/RDF/>

³ <http://www.w3.org/TR/rdf-schema/>

⁴ <http://www.w3.org/TR/owl-features/>

technologies. The individual applications can differ between each other in many aspects, be it the scale of the used database, number of interested parties, kind of inputs and outputs or the very subject of operation. Because of this the categorization of knowledge-based applications is a multidimensional question.

The particular dimensions (i.e. categorization criteria) however had to be identified. This is where we started the analysis of the mentioned case studies published by the W3C interest group [8] (by the time of publishing this paper 20 were taken into account). We split the individual case studies amongst 7 workers from Department of Information and Knowledge engineering (every one of them interested in semantic technologies) and went through them in detail. Afterwards, every case study has been discussed in particular by the whole team. Thanks to comparing the individual cases the following aspects of differentiation of the semantic applications emerged (not all however have a direct impact on the forming of critical success factor – this will be considered in part 3). Although none of the analysts had a personal experience with the case considered and only a description was available, the results are credible because of the fact that the SWEO catalogue gathers together cases that represent more than single software a distinctive kind of applications. The categorization criteria we found are these:

- **Information sources.** The semantic character of considered applications directly implies that at least one knowledge model (ontology or taxonomy) has to be used. Some applications also use other knowledge models or even expect a variable knowledge base. Apart from that the applications can of course also use other data of various kinds. Knowledge-based applications can be divided according to whether they process structured knowledge, structured data or unstructured data.
- **Data source provenance.** Semantic applications can be distinguished according to whether the information they are working with arise in other systems (or are already available in a structured form) or whether they are created specifically for this system. If the data are created exclusively for the semantic system we can further distinguish the cases where this is done manually, automatically from other sources or as a side effect of other activities (such as normal user behavior).
- **Accuracy of inputs and outputs.** Considering the semantic applications we find different approaches of transforming inputs to outputs. Here the applications can be divided e.g. into those firmly relying on full precision of data, applications that expect that the data may be incomplete but do not expect them to be inconsistent and do not work with uncertainty, and finally, applications that include treatment of uncertainty.
- **Domain-specificity and reusability of applications.** Because of the separation of data from their meaning the semantic applications should be much less domain-dependent than conventional solutions, but even here there are exceptions, which include, for example, specific interfaces tailored to a specific domain or particular treatment of data on the application level.
- **Number and kind of users.** Users of semantic applications may constitute of unprofessional individual users, professional users (domain experts), knowledge

experts and management. Applications can also be distinguished according to whether they are intended for individuals, working groups or thousands of users in social networks.

- **User × provider relationship.** Here we managed to identify several options for operating the applications: the user is an individual and operates the application for his/her own use; there are a few users and they are more or less equal subjects or form a social network and the operation is granted commercially, by the community or non-profitably; the users are the customers of the provider; and finally the users are the employees of the provider. For the last two possibilities we can distinguish cases where the operation of the application is the core business of the company and where it is only a supporting process and can therefore be considered as a possible target for outsourcing. The cases when the operation is ensured by the community can be broken down by whether the operation is centralized or decentralized.
- **Frequency of access to the application and its availability.** Applications may be used continuously (24/7), at random, regularly or by a single opportunity. Furthermore, a distinction must be also made by the availability of such applications: either the application must be available constantly, in defined intervals or on demand (e.g., the reactive manual start of the application).
- **Subject of operation.** From the analysis of case studies we managed to identify several main types of activities of semantic applications. These are data indexing, data integration and reasoning. These activities are, however, in most cases the means rather than the purpose of the activity (the exception is the integration of heterogeneous data). From these, we can derive several other activities which support the main purpose of the application, for example, they are enabling better searching capabilities (indexing + integration), heterogeneous database browsing and navigation in the domain (integration + indexation), recommending new relations among entities (reasoning) and allowing the adaptability to change the systems' data structures (data integration).

By sorting the considered case studies we can find out how often some specific values of the proposed criteria are seen in the real-world applications. (The sorting was done by filling a prepared form by the responsible analysts in the first place and the results were then reviewed and normalized by one of the authors.) The relative frequencies of these values are shown below on Fig. 1.

Of course, one can imagine a semantic application that is classified by the mentioned aspects more or less arbitrarily, but in the presented case studies certain coincidences and clusters can be identified amongst the various aspects. This sorting thus enables us to and to identify and name some basic archetypes of semantic applications, based on examples clustering:

- **“Improved search engine”.** These applications focus on indexing the data, often associated with integrating data from various other systems where the data are generated automatically. These automatically acquired data are also often accompanied by manual annotation. Applications of this type work with both structured data and unstructured data (using automatic filters and wrappers).

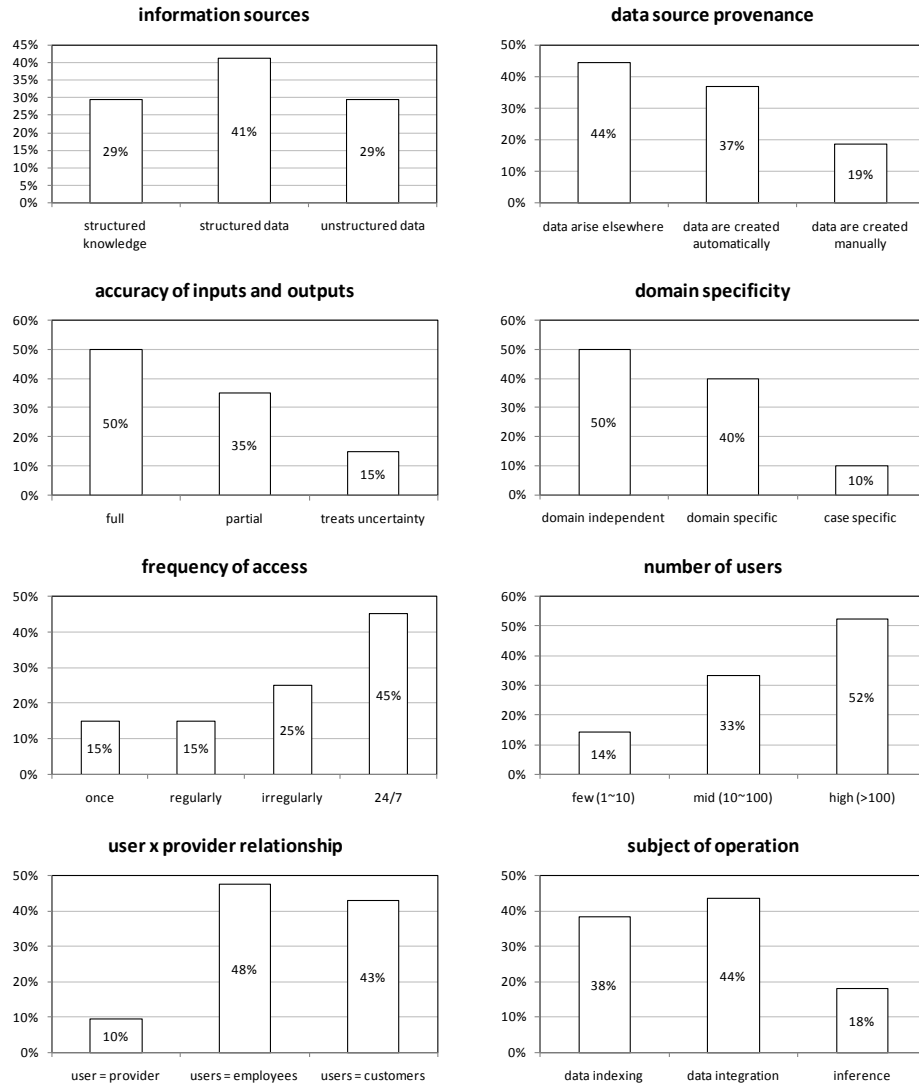


Figure 1 - Relative frequencies of certain criteria values

The main benefit of these applications is to enable searching in heterogeneous data base and the creation of complex queries without the need for a priori knowledge of data structures. In other aspects, however, they can vary greatly, so such different applications as support for annotating and searching of files on a personal computer⁵, a public portal for searching for findings of Chinese medicine⁶ and management of sound recordings archives by a Norwegian radio station⁷ can be classified here.

- **“Data-browsing interface”**. These applications follow the abilities of the previous archetype, but enhance not only the possibility of displaying diverse content (videos, articles, chemical formulas, etc.), but also the possibility of visual data browsing, regardless of their structure. These applications are mainly focused on the use by professionals and are operated either commercially for internal use or non-profitably to support a professional community (and simultaneously promoting the technology). Examples of this archetype can be e.g. systems for the aggregation of medical data, whether in order to facilitate the treatment of patients⁸ or achieving savings in the development of new drugs⁹ or portal for the association of programming knowledge by Oracle¹⁰.
- **“Recommending system”**. The nature of these applications is the derivation of new relationships between entities. Moreover, apart from all other types of source data these applications often utilize data that are automatically generated as a side effect of normal user activity, which enables, inter alia, to propose new relationships on the basis of the current users’ context. The user is often the customer of the provider, be it either as a paid service, public service (e.g. designing of individual city tours in Zaragoza¹¹) or a commercial way to personalize advertisement targeting together with the provision of services (such as a system for recommending services to users of mobile devices¹²). Applications in this category often work with uncertainty, thus one can also include a variety of expert systems.
- **“Data interchange framework”**. Operations of applications in this category (because of their nature) are distributed, thus these knowledge-based systems “only” allow to unify the structure of data exchanged between the participants, regardless of their content. This content can evolve over time and be adapted to the needs of a particular bilateral exchange relationship and yet be transmitted in a standardized format. An example is the initiative for the establishment of semantic data interchange in the oil and gas industry¹³.

⁵ <http://nepomuk.semanticdesktop.org>

⁶ <http://www.cintcm.com>

⁷ <http://www.nrk.no/>

⁸ <http://www.pharmasurveyor.com/>

⁹ <http://www.lilly.com>

¹⁰ <http://otnsemanticweb.oracle.com/>

¹¹ http://www.zaragoza.es/turruta/Turruta/en/index_Ruta

¹² <http://www.w3.org/2001/sw/sweo/public/UseCases/SaltLux-KTF/KTF.pdf>

¹³ <http://www.w3.org/2001/sw/sweo/public/UseCases/Chevron/>

Surely it would be possible to discover other archetypes of semantic applications; however, we consider these four to be the most usual. Of course there are also applications that cannot be assigned to any of these archetypes, as well as others which, on the contrary, lie in between two or more.

3. Critical Success Factors

As already indicated in the section 2, by the synthesis of the risks mentioned in the individual case studies [8] it is possible to outline the most frequent critical success factors in the development of the semantic applications and their deployment into the production environment. These factors are not universal, but each only applies to a particular group of applications given by the aspects of their categorization, referred to above. Critical factors for success of semantic applications identified so far are:

- **Correctness of the core ontology/taxonomy.** This factor holds for all knowledge-based applications, and the more complicated and less volatile the used model is, the more crucial is its correctness. Achieving this success factor entails the need for recruitment of high-quality analysts and knowledge engineers in the phase of development and deployment of the application, which involves considerable costs. The quality and reach of the used ontologies is not limitless; apart from the costs of creation it also has other more structural constraints (see [2]).
- **Sufficiently steep learning curve of end-users.** This applies to applications that have individual end users. Semantics used in this type of applications entails quite atypical method of control compared to standard applications and the learning curve is rising very slowly. Not only a comprehensive and intuitive user interface of the system is a must, but also clarity and accuracy of the outputs and results is vital for the users' work.
- **The potential of possible benefits to compensate the temporary reduction in productivity during implementation and learning** (as well as operating costs). The benefits of the applications are very diverse and often very vague (in contrast with conventional solutions) and thus can be hardly estimated (and quantified) at the time of the deployment of the system. Operating costs are mostly comparable to conventional applications, but in the phase of deployment, it is necessary to count with temporary decrease of productivity of the users (see previous item). For an application to be successful, this temporary decrease should not be so serious that it overshadowed its potential benefits.
- **Will and discipline of all parties to use the same knowledge model.** In case the operation of application is distributed, it is necessary that all interested parties use a central shared knowledge model. There is therefore a potential risk in terms of the need to negotiate on its form and content.
- **Synchronized distribution of central ontology.** Gradually, there may be modifications of the central knowledge model that arise subsequently and if the operation is distributed, it is necessary that these changes are properly disseminated amongst interested parties, or else this could lead to some

inconsistencies. While these changes and modifications take place the previous item still holds.

- **Sufficient number of users.** If the respective semantic applications is based on social data, its success is conditioned by the existence of a large enough number of users that produce this data. The risk in this case occurs in the form of necessary expenses for the promotion of an emerging system.
- **Users' motivation.** This critical factor occurs at two levels. The first is in the time of the introduction of a new application while the user experiences a negative stimulation in the form of the slow rise of the learning curve. The user thus lacks the motivation to learn to deal with the system in the first moment. Moreover the user that does the work is not always the one who benefits from it (discussed in [2]) which can be of a further burden. The second is in the actual phase of operation; a common source of data for the semantic systems of all sizes is manual annotation, whose creation is up to a certain point very labor-intensive for the users. Partial source of motivation may be a potential benefit of the better results or a facilitation of work in the future. In addition the user can be motivated by the possibility of using the experience gained elsewhere, while even the most different semantic applications use similar technologies (e.g. SPARQL querying).
- **Sufficient supply of data.** For applications that use some reasoning having a sufficient data source is very essential for providing beneficial results (i.e. utilizing the added value of semantics). Even for applications based on data indexing, having enough data is critical to the success, because for small volumes of data they give similar results as traditional methods but with higher initial costs.
- **Diversity of sources and forms of data.** The greater the richness of the knowledge-modelling language (namely, its part actually used in the application), the more beneficial results can be produced by applications based on the derivation of new relationships. Likewise, the greater is the diversity of data content the more useful are the results given by applications performing semantic integration. The use of semantic technologies on trivial systems will therefore likely not pay off.
- **Maintaining at least the same accuracy of results as the sub-systems.** Applications that integrate data of some source systems are at risk of finding an inconsistency in the aggregated results. The functionality of semantic applications itself is not subject to consistent data, but the possible inconsistency should be expected in the design. Good estimation of the reliability of data sources is thus crucial at this point.
- **Reliability of parsers and wrappers.** If the application handles unstructured data, it is dependent on the output of parsers and wrappers of various content and, where appropriate, the natural language processing systems. Here again the same applies as in the previous paragraph, namely that it is necessary to correctly estimate the reliability of the information obtained in such a way.

Of course, these critical factors will be weighted differently in the scope of different applications. If, for example, the source application collects data automatically and passes the outputs to the user in almost natural language and in an appropriate context, we can expect a relatively steep learning curve, so that the period of reduced productivity is quite minimal and as a result it will be compensated enough even by minor benefits. These universal critical success factors can only be taken as starting points when considering a particular case.

4. Future Work – Maturity Models

Maturity models [3] have developed over the past two decades in order to enable to assess the readiness of enterprise to implement some kind of structural investment. Most commonly they are used in the deployment of any IT applications such as CRM systems, ERP and Business Intelligence. In our opinion it should be possible on the basis of the above aspects of categorization and the associated critical success factors to establish enterprise maturity models for the deployment of a certain type of semantic technologies. Although these models would be without factual content and without the target statement, they could be formulated by concrete examples and at least for each archetype would thus make it possible to set a certain level of requirements for an enterprise, which should be met in order to consider the feasibility of the solution.

An example maturity requirement for the archetypal semantic search engine could look like this: *If an enterprise uses a single source of data and a proprietary data structure, then it is unprepared for the introduction of this kind of system. If it is using multiple systems with heterogeneous data structure, the introduction of search engines with semantic indexing can bring some improvements to the search results. The enterprise achieves next level of readiness if it uses more systems with a standardized data structure; in such case it can start thinking about the integration of these systems with a semantic data exchange, etc.*

Before formulating such exemplar maturity models the critical success factors need to be evaluated because their validity is vital. We plan on performing a detailed survey of the successful semantic technology projects (based on the SWEO catalogue) and test whether the critical factors hold. As a side effect such a survey will help quantify exact boundaries and limits in our dimensional categorization approach. Only after verifying the CSFs we can move onto formulating the maturity models.

5. Conclusion

This work outlines a dimensional method of categorization of semantic applications, and with the help of case studies published by the W3C interest group then identifies some of the basic archetypes of semantic applications. Given the proposed categorization we then formulated the most critical success factors for the deployment of semantic applications in a business environment, together with the validity limits of these factors.

Finally we also outlined how the proposed critical factors and the categorization can be used in the process of finding valid models for maturity of enterprises for semantics. The formulation of such specific models is the subject of our future work.

Acknowledgement

We wish to express our sincere thanks to our department colleagues Tomáš Kliegr, Jan Nemrava, Ondřej Šváb-Zamazal and Jan Zemánek, who helped us a lot with the survey of SWEO case studies, and, in particular, to Ota Novotný from the Department of Information Technology, who offered his valuable consultations on CSFs and maturity models.

References¹⁴

- [1] Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web, *Scientific American*, May 2001. <http://www.sciam.com/article.cfm?id=the-semantic-web>
- [2] Hepp, M.: *Possible Ontologies: How Reality Constrains the Development of Relevant Ontologies*, in: IEEE Internet Computing, Vol. 11, No. 1, pp. 90-96, Jan-Feb 2007 (invited), doi:10.1109/MIC.2007.20
- [3] Humphrey, W.: *Managing the Software Process*, Addison-Wesley Professional, Massachusetts, USA, 1989
- [4] Novotný, O.: *IS/ICT Management Reference Model*, Revista de Engenharia de Computacao e Sistemas Digitais, 2007, y. 3, n. 3, p. 53–61. ISSN 1678-8435
- [5] Raden, N.: Business Intelligence 2.0: Simpler, More Accessible, Inevitable, *Intelligent Enterprise*, <http://www.intelligententerprise.com/showArticle.jhtml?articleID=197002610>
- [6] Rockart, J. F.: *Critical Success Factors*, Harvard Business Review, 1979, p. 81-91
- [7] Sauermann, L.: *Benefits of Semantic Web .. for you*, DFKI GmbH, 2008 <http://www.dfki.uni-kl.de/~sauermann/2008/04/benefits/>
- [8] Semantic Web Education and Outreach (SWEO) Interest Group: *Semantic Web Case Studies and Use Cases*, <http://www.w3.org/2001/sw/sweo/public/UseCases/>
- [9] Staab, S., Studer, R.: Handbook on Ontologies - Preface. *International Handbooks on Information Systems*, Springer Verlag, 2004

¹⁴ The hypertext links are valid on the date of February 1st, 2009