Submitted version

Adapting the communication capacity of web services to the language of their user community

Kurt Englmeier[1], Josiane Mothe[2], Fionn Murtagh[3], Maximilian Stempfhuber[4]
[1] Fachhochschule Schmalkalden, Fachbereich Informatik, Germany
KurtEnglmeier@computer.org
[2] Université Paul Sabatier, Institut de Recherche en Informatique de Toulouse, France
Mothe@irit.fr
[3] Queen's University Belfast, School of Computer Science,
Belfast, United Kingdom
F.Murtagh@qub.ac.uk
[4]Informationszentrum Sozialwissenschaften,
Lennéstr. 30, 53113 Bonn, Germany
Tel.: +49 228 2281-139 Fax: +49 228 2281-120
st@bonn.iz-soz.de

Abstract. This paper addresses an interface layer as an integral part of web service technologies that takes well-established web service specifications as a wrapper and accepts natural language specifications for the inter-application communication. In a growing complexity of web service networks the architecture of context becomes more and more an issue for both service providers and service requesters. While providers will concentrate more on the basic layers of the activity structure within a service network, the users, i.e. the service requesters, will form ad-hoc collaborations between services for their own specific solutions. We present the Web Service (WS)-Talk layer as a structured-language interface for web services. This "open building block" can be implemented by both the software developer and technology end users who operate on the natural language interface. WS-Talk takes taxonomies for a semantic layer that transforms service descriptions or service requests, expressed in natural language, into specifications necessary for the task-specific composition and use of web services. The objective is to fasten the process of building connected interoperable applications by including more intensively the community of end-users that become an active part in the architecture of web service-based applications. The WS-Talk layer helps web service requesters and providers to create ad-hoc solutions (or rapid prototypes) and it facilitates adoption and management of web services. The paper presents the rationale of WS-Talk and demonstrates its applicability in a prototypical environment providing economic information.

1. Introduction

Years ago the IT community was excited over XML standards that opened avenues for a completely new type of interoperability of software across networks. The dream of each individual using the other's applications to develop new ad-hoc services and appliances seemed to be in reach. In the meantime the sobering truth about semantic web standards is that they are not the silver bullet they were once conceived to be. Even though they support interoperability, developing large complex specifications for an application domain is still incredibly complicated.

Nevertheless, the rapid adoption of web services is spurred on by the benefits they bring to development communities. XML, and related tools and technologies used to set up and maintain web services, help developers create small, task-specific applications with software modules that can be used and re-used. Web services could in fact revolutionize the way we develop applications like the Internet itself changed our life [4].

If we regard now a user in their working environment, then specifications of potential web services address descriptions of tasks and of tools that help to accomplish their tasks. No doubt, tasks and tools can be represented solely by XML. However, if we imagine for a moment a business process in all its ramifications we can easily discern the complexity of the model we have to develop. In addition, it is quite likely that the model requires a lot of adaptation to situations we have neglected or those that have recently come about. Developing large, complex specifications is not an easy task. Once implementation starts, problems are often found in these specifications, but it is very difficult to go back to the specification of the authors in order to put them right.[13]

The demise of ebXML, for instance, can be seen in this light: it collapsed under its own weight. [12]

Nevertheless we acknowledge that semantic web standards and web service specifications are very helpful and without them a high level of interoperability would be out of reach in any case. The context of web services, thus, comprises a stable framework constituting basic principles of the service and details to be specified by a variety of persons and exposed to frequent changes over time. In the combination of well-established web service specifications for the stable part of the service identification and communication with natural-language processing (NLP) mechanisms for the dynamic part, we put the technology users in the position to apply web services to enrich the design of web service networks with their competence and experience. In particular the definition of stateful resources [16] benefit from a more direct involvement of the relevant personnel. On the level of an organization or market sector it is easier to reach an agreement on a suitable vocabulary than to design complicated XML models that can not be understood by the majority of users.

The paper is organized as follows: section 2 outlines the rationale and advantages of extending web service semantics towards natural language (NL). Chapter 3 explains how a context can be specified in natural language that is both machine-processable and understandable to humans. This context reflects a certain domain, the business of an organization or a market sector for instance, including its respective tasks, processes, and operations. Design details of WS-Talk enriched with examples from its first prototypical realization are presented in section 4. Chapter 5 concludes the discussion of this paper.

2. Rationale and advantages of a web service understanding natural language

A web service context defines basic information about the activity structure within a network of services. Web service specifications raise the capabilities of services to enter relationships between them by sharing common attributes as an externally modeled entity. For the time being the architecture of context is exclusively in the hands of service providers that resort to web service technology, that in turn enables them to develop useful service networks. However, the more the development leaves the transport layer, the basic communication layer between the services, and enters the application domain layer the more providers get trapped in the specification problem mentioned above.

Instead of jumping on the desperate bandwagon of standardization we propose a web service wrapper that enhances XML-based interfaces with features for free-text interchange. The resulting wrapper, Web Service (WS)-Talk, helps to establish interoperable web services across platforms. Much in the vision of JXTA[1] a web service announces its service context through advertisements in domain talk defined by its user community. The non-standardized part enables a very flexible way to define a web service and its usage. To ensure that the advertisements are machine-processable we apply robust text mining methods that map advertisement content into a suitable controlled vocabulary. From a different angle our solution resembles the current successful application of natural language processing in speech recognition systems where command languages are combined with NLP interfaces to provide for robustness.

Each company, and each industrial sector, has its specific domain talk used to describe products, services, objectives, processes, tasks, and so on. WS-Talk distils from this domain talk a structured representation and takes the resulting condensed enterprise talk as quasi-standard for inter-application communication. Quite often domain talk is represented in taxonomies. More and more organizations are making a comprehensive taxonomy of the organization's content a high priority. Taxonomies as well as controlled vocabularies organize content and context of the organization's subject by grouping similar items into broad categories which themselves can be grouped onto ever-broader concept hierarchies. In essence, taxonomies provide a degree of structure to the organization's unstructured content. Because of a taxonomy reflecting the most important business categories, organizations that carry out the same business activities tend to have similar taxonomies. This fact and the rising importance of taxonomies to organizations brought us to the idea to harness this structured information as quasi-standard for the context architecture of a web service network applied within an organization's scope, and beyond if similar taxonomies allow this expansion.

The WS-Talk service then consists of a transport layer representing a typical web service interface, and a context layer resorting to WS-Talk features for capturing the enterprise talk. Through its structured-language interface WS-Talk services can be implemented by both the software developer **and** technology end users who operate on the natural language interface.

The **objectives and advantages** of this wrapper are

1. a high availability for ad-hoc solutions for small technology user communities and
2. a high flexibility in responding to the dynamics of the community's environment to which it is applied.
3. It fosters the proliferation of web services as flexible means to construct complex and dynamic applications and

[1] JXTA is an open network computing platform launched by Sun Microsystems, Inc., USA. It is designed for peer-to-peer computing. Its goal is to develop basic building blocks and services to enable innovative applications for peer groups. The JXTA protocols enable developers to build and deploy interoperable web services and applications. The protocols are designed to be independent of programming languages, and independent of transport protocols. More information is available at http://www.jxta.org.

4. passes the design competence more to the technology users that are better aware of the application area than purely tech people.

Natural language itself is not appropriate for programming, but NL referentiality is a key factor for the design of program instructions that are both machine-processable and understandable to humans[7]. For information retrieval and speech recognition systems[2] NL referentiality is essential and aspect-oriented programming shows how powerful program organizations can be realized based on instructions and structures expressed in NL [6, 9, 11].

3. Communicating context

Context awareness in human-computer interaction usually refers to the system's capability of being aware of the user's physical or – more abstract – "semantic" position and of adapting its behavior accordingly while being minimally intrusive. A user's context can be quite complex, consisting of attributes like current position, personal history, and recurring behavioral patterns (see examples in [1], [5], [15]).
The quality of an architecture of context depends on the capability of its underlying approach to communicate the context of a specific domain [8]. In a web service environment, the service is propagated and identified through its description. To avoid ambiguity the service descriptions, for instance, have to be mapped correctly into a coherent context map that allows, at the same time, to locate the whereabouts of a service. Semantic coordinates – i.e. controlled vocabularies derived from taxonomies and structured according to concept hierarchies – are elements of orientation that can be communicated to service-providing machines. Concept hierarchies derived from established and agreed taxonomies endow the context representation techniques with a global concept layer of the respective domain. And, in addition, this layer can be mapped into different natural languages in parallel [3].

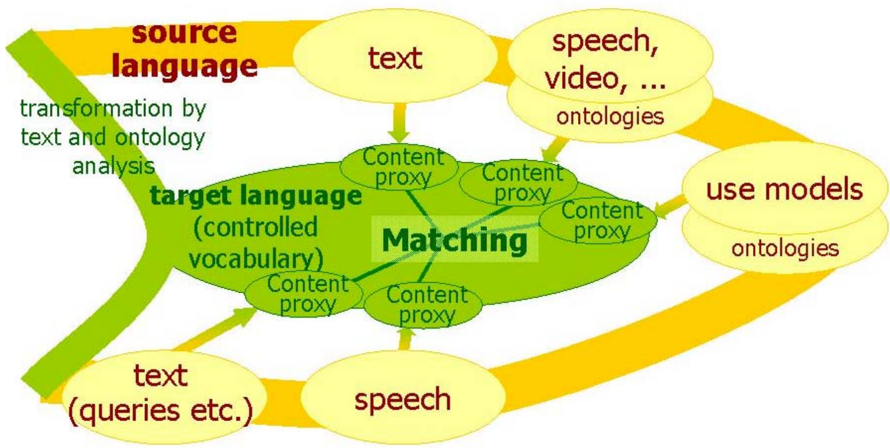

Fig. 1. Structure of an interface layer for semantic context-awareness

In the approach outlined here a linguistic interface layer ensures that the system "understands" a phrase expressed in an ad-hoc way in natural language. "Understanding" means mapping an information item like a web service description into an appropriate content proxy that is developed by terms of the target language, the controlled vocabulary derived from taxonomies, and structured along concept hierarchies. WS-Talk contains thus features to "translate" a service request or a service description (source language) into the controlled vocabulary (target language) generating automatically a content proxy for this request. The source language can be a mixture of natural language and command language (see example below). The subsequent matching process looks for content proxies that are in close semantic proximity to the proxy of the query. Each of these proxies may represent a single service or a set of services.
The "translation" of the essential content of a web service – the proxy generation – therefore has to contend with text mining methods that translate unambiguously service descriptions into suitable terms of the target language. A robust text analysis method has to rely on predefined templates and pattern-based extraction rules to extract meaning. TFIDF (weighting-based) or LSI (latent semantic indexing) are the most prominent methods for the classification of content in unstructured textual data. The problem with such methods is that they can estimate the importance of the relevance of a term only in the shadow of global information about a larger corpus. And in addition, they cannot solve sufficiently the problem of ambiguity. This is a crucial drawback when the relevance

estimation cannot or should not resort to a larger corpus because of performance purposes. To circumvent this problem in WS-Talk, we resort to the "vector voting method"[10] developed for and successfully applied in IRAIA[2]. In principle, each term in a text "votes" for a concept of the concept hierarchy if it contains a matching term. The more "votes" a concept receives, the stronger is the link between the text and that concept. In fact, the method is more complex as concept terms are not only those from the concept itself, but also those from related synonyms and terms associated with the concept ancestors. In the end, the technique uses the whole corresponding concept as index term.

4. Design of the WS-Talk wrapper

The objective of representing a domain context in natural language is to enable a flexible and easy management of **enterprise "standards"** necessary for domain-specific web service communication. Such inner-enterprise "standards" have their origin in the enterprise talk, in its traditional way to represent business processes, products and resources. Sometimes the enterprise talk is reflected down to the labeling of functions and parameters in their software codes. Enterprise talk includes also descriptions of products and services. For external observers, it is sometimes hard to catch up on a discussion by car manufacturers, while the latter have problems to follow food producers, and so on. The issue becomes worse when it comes in combination with production details or IT-details.

Proprietary message-oriented middleware is currently the backbone of enterprise interoperability. Its interfaces and protocols reflect IT-specifics within the enterprise talk. The enormous amount of specific detail makes creating even intra-enterprise standards extremely complicated using a high level specification of a web service standard. And in turn, this fact sheds light on the extreme complexity of establishing industry-wide standards. The situation looks equally grim on the side of industry-wide specifications. The problem of standardized specifications in capturing the complexity of the enterprise's reality is a severe technical limitation that hampers the broad proliferation of a technology that in fact could revolutionize our way of building applications. Facilitating the development of an architecture of context through sophisticated design tools eventually mitigates this problem, but cannot tackle it completely.

The reflection on this enterprise reality brought us to the idea of taking the enterprise talk as quasi-standard for inter-application communication. However, it is obvious that the approach can be applied as well in a context that is different from that of an enterprise.

The definition of an open building block then consists of a transport layer representing a typical web service interface and a context layer resorting to WS-Talk features for capturing the enterprise talk. The possibility to express a service description in natural language offers enormous flexibility.

A semantic coordinate system helps to determine the correct location of an information item like a service description or a user request within a domain-specific context. This positioning process is tantamount to correctly matching requests and information items as is realized by IRAIA or similar search engines. If we now reflect this approach in the context of web services we implement the same features as presented above for service description and identification:

1. Like a user defines a query using the concepts of the controlled vocabulary, she describes in the same way a web service. This description is in any case easier and faster to realize than developing a corresponding XML or WSDL structure expressing the functionality of the service.
2. The matching mechanism to find a suitable service operates the same way as the process that locates a document at the semantic coordinates that match with a user query or a similar request from an intermediary service.
3. Instructions how to handle data that are passed from one service to another are also expressed in terms of the controlled vocabulary and may include command language components.
4. The NLP part of the web service hides unnecessary implementation details from service consumers, the non-tech end users.
5. Providing the controlled vocabulary in different languages in parallel has the advantage of exchanging service descriptions even across language boundaries, and this avoids the users having to apply a language they are not familiar with or having to be trained in a formal description language.

[2] IRAIA is a service provider platform designed, in particular, for large and complex information spaces. The platform emerged from a research project under the 5th Framework Programme of the European Commission. It was applied in the areas of economic information (iraia.diw.de) and cultural heritage related to puppetry (www.epuppetry.com).

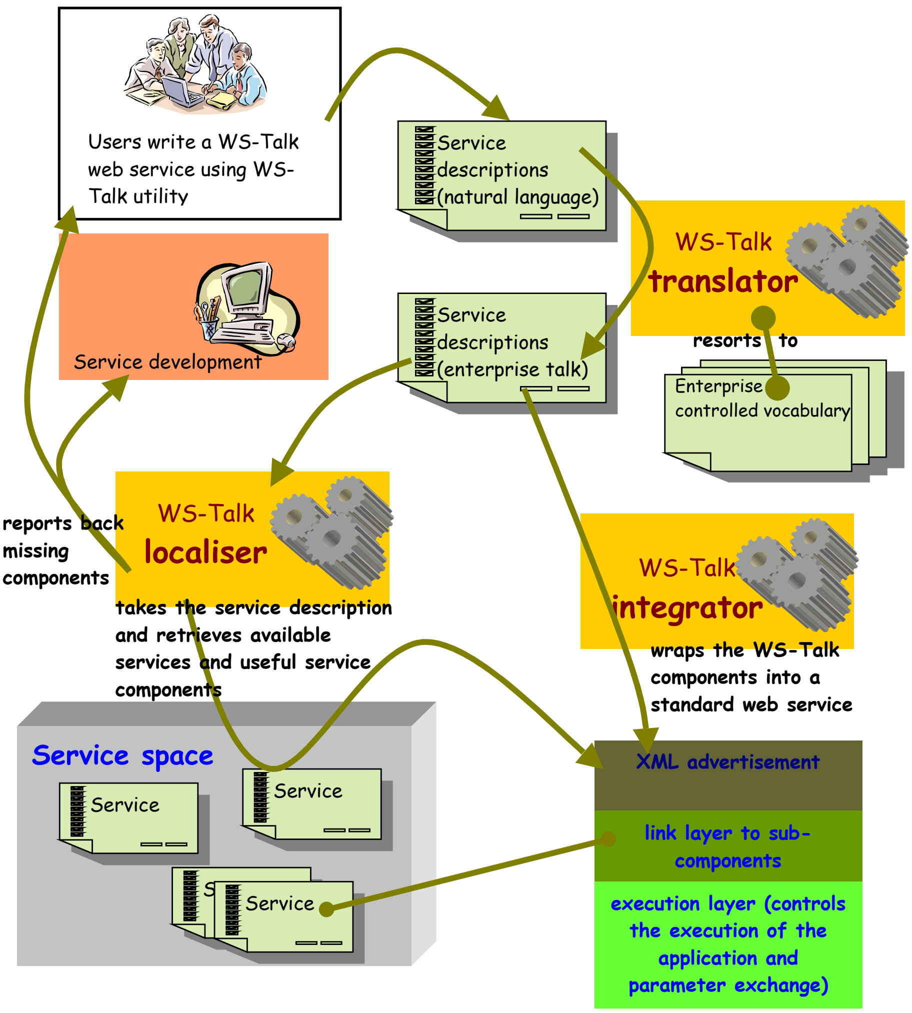

Fig. 2. The schema shows how the WS-Talk components operate within a network of web services

The following **example scenario** addresses a data mining environment supporting the economic analysis of an enterprise. It illustrates in a living example how WS-Talk-specific components as shown in the schema above cooperate in a network of web services.
An analyst may be interested in creating a web service that searches time series related to economic activities (like the amount of incoming orders, productivity, exports, etc.) of all economic sectors that show a certain impact on a specific sector. The principal idea behind this analysis is that present activities in one industrial sector have an impact on the future economic situation of other sectors. In economic analysis, usually such an analysis helps to determine forecast indicators.

The analyst now creates a Web service with the following specification: "Search for time series in our own and the OECD databases that represent all economic activities of all industrial sectors that have an impact on the products of the lubricants section of our company." This service doesn't exist so far. The service definition tool of WS-Talk "translates" the description into key words of the company's enterprise talk, i.e. transforms the original phrase in source language into terms of the target language, the company's controlled vocabulary. The "translation" process may include term replacements or query expansions. The concept of "products of the lubricants section of our company" may be replaced by a list of concrete product definitions. The subsequent localization process identifies the semantic coordinates corresponding to the key words resulting from the "translation" and searches for one or more existing services that have the same semantic coordinates. It is quite plausible that the required web service has to be constructed from a number of services that exist within the company or come from outside. Let's assume that in our example there is a service available at OECD that enables the access to its time series database, a further one of the company that presents these time series in business charts, and a third one that identifies impact relationships between time series using time-warping as known from speech recognition and recently also applied in economic analysis. The latter service may come from the company's consultant. If the requested service is available, it will be propagated when its description is fully matched by the semantic coordinates of the existing services. Missing pieces are reported back to the service requester. She or he may then re-formulate the description or ask the software department to provide for the missing component(s).

The services themselves have some coordination capabilities. From the localization they know which is the superior process they are related to and know, thus, the components they are cooperating with. The time-warping tool knows that it requires input from the OECD tool, and the OECD tool may at first confirm if this web service of this specific company is authorized to retrieve OECD data. The exchange of data may be based on XML-specifications and further metadata descriptions.

The sidebar (see below) extends the example by more implementation details from the actual prototype of WS-Talk, how the semantic coordination system is instrumented and how interfacing is realized.

An example from the WS-Talk prototype

The technology for semantic coordinate systems applied in WS-Talk is based on IRAIA, the European FP5-project for advanced information provision in the area of economic and statistical information. The thematic domain is reflected by concept hierarchies that are derived from existing taxonomies. For the ambience of economic information, for instance, in IRAIA a powerful taxonomy has been produced that merges two of the most important structures in this field: Eurostat's NACE (Nomenclature des Activités dans la Communauté Européenne -systematic of the economic activities of the European Union) and the industry systematic of the ifo Institute for Economic Research.

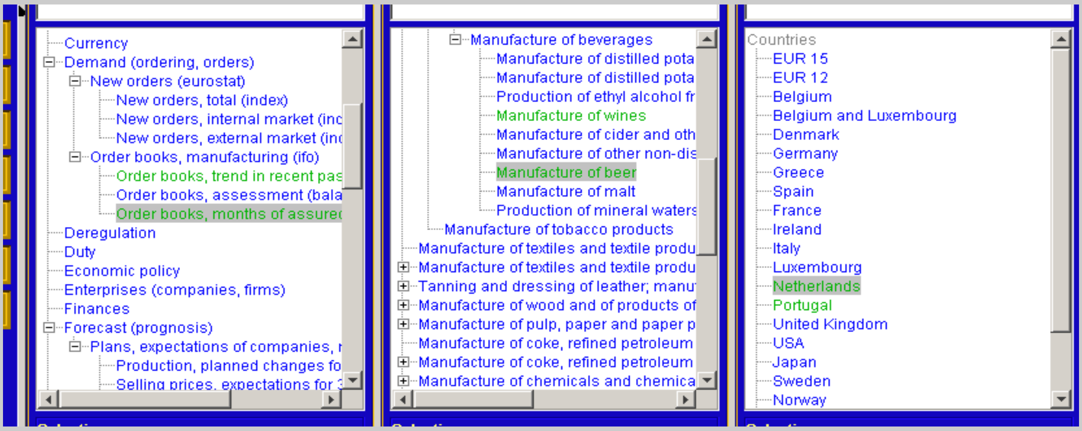

Fig. 3. The unified taxonomy creates a semantic coordinate system that enables exact and automatic positioning of coherent information items like documents or service descriptions. The concept hierarchies presented comprise the domain-talk of economic analysts.

The concept hierarchies above can be regarded as a controlled vocabulary of the thematic area a European economic research institute is dealing with. They are constantly checked for completeness against the content covered by the text and time series documents reflecting comprehensively and in detail the organization's business. This approach can be extended easily towards descriptions of tasks that are related to economic analysis and to web services performing or supporting these tasks.

The following example shows a part of a description of a JXTA service as being used for accessing and processing time series. A well-formulated description usually reflects precisely the nature of the essentials of the service and the way how to use it.

Tab. 1. Description section of a (JXTA) web service that retrieves and processes time series.

```
<Svc>
  ...
  <Desc>
    TimeSeries service. Retrieves time series from a specified database and
    provides processing methods.

    Parameters:
    Types of processing methods:
    moving averages, smoothing::SMOOTH::int
    period-on-period::PERIOD-ON-PERIOD
    ...
    Time scope, recent n periods::SCOPE::int
    Time scope, from a to b::SCOPE::Calendar::Calendar
    Time scope, from a on::SCOPE::Calendar
  </Desc>
</Svc>
```

A concept hierarchy containing an arrangement of task concepts and commands can then be used to analyse a service description and to represent its purpose by its most significant concepts. Commands are added to the title

of the respective function like synonyms. The same holds for database specific codes of thematic concepts. In our example here “Chemical Industry” is represented by the code “b25000” of the Eurostat database.
In the same way a request expressed in natural language can be distilled to its most significant concepts derived from the same controlled vocabulary. This “translation process” is a preparatory step for the subsequent matching process that identifies in the end the service(s) necessary and suitable for the request. The following screen shot shows a user’s request of a service and the machine-processable “translation” of this request as feedback from the WS-Talk system.

Fig. 4. Screen shot from the current WS-Talk prototype showing the “translation” of a user request into the target language manageable by the web services.

The proposed web service architecture is also applicable in scenarios where heterogeneous data like statistical time series and text documents has to be integrated at the end user’s side to satisfy complex information needs. Here, WS-Talk can transform the more formal and restricted source language from official statistics to the much richer target language of newspaper or journal articles or economic reports. In the project ELVIRA [14], an ontology has been developed that currently connects around 20 international trade nomenclatures and two thesauri from the electrical engineering and electronics industry with cross-concordances and transfer relations based in statistical corpus analysis and co-occurrence analysis. ELVIRA makes use of an already established network of web services to scientific information in Germany, Vascoda[3], as a test bed for the architecture outlined here.

[3] http://www.vascoda.de

5. Conclusions

Despite advances in the development of semantic web standards, natural-language processing methods still form part of many techniques for enabling computers to understand and engage in human communication [2]. This is not at all a paradox since much information is passed or stored in natural language. Semantic web standards do not replace NLP and vice versa, and the synthesis of both disciplines offers an enormous potential for a new generation of web service technology. WS-Talk emerges from the cross-fertilization between semantic web technologies and text mining methods. The idea behind this approach is to enable small communities to set up ad-hoc web services using their own language instead of resorting to tools that too complicated for them to define or modify their own context architecture for their services.

A crucial component of such a design environment is the automatic and precise transformer of a web service description into the catalog logic of the respective application domain. This transformer resorts to methods supporting semantic context-awareness. It resides on the users' side putting them in the position to describe the services they require in their ambience.

Allowing software developers and service consumers to write an interface specification in their own words (i.e. allowing all to become context architectures), to avoid the complicated process of reaching an enterprise-wide agreement on this interface standard, sounds first of all like shutting a Pandora's box in favor of opening another one. However, as we harness only the most stable part of web service specifications we also embark only on text analysis and information mapping methods that are in the position to tackle adequately the problem of ambiguity in interpreting service request and correctly locating the adequate services.

Future applications will emerge from networks of web services that are defined within an organization and are used across the organizations' boundaries and the architects will be service providers and service requesters. The example of WS-Talk shows that, beside semantic web specifications, natural language processing has a powerful role in these networks as it allows also non-tech people to set up and maintain essential parts of a company's palette of services.

References

[1] Cheverest, K., Mitchell, K., and Davies, N. *The Role of Adaptive Hypermedia in A Context-Aware Tourist Guide*. Communications of the ACM, 45 (May 2002), 47-51.

[2] Ciravegna, F., Harabagiu, S. *Recent Advances in Natural Language Processing*. IEEE Intelligent Systems, 18 (January/February 2003), 12-13.

[3] Englmeier, K., and Mothe, J. *Natural language meets semantic web*, http://www.ktweb.org/doc/Englmeier-NLP-SW.pdf (July 2003).

[4] Filman, R.E., *Semantic Services*. IEEE Internet Computing, July-August 2003.

[5] Fleck, M., Frid, M., Kindberg, T., O'Brien-Strain, E., Rajani, R., and Spasojevic, M. *From Informing to Remembering: Ubiquitous Systems in Interactive Museums*. IEEE Pervasive Computing, 1 (January/February 2002), 13-21.

[6] Kiczales, G., Lamping, J., Mendhekar, M., Maeda, C., Lopes, C., Loingtier, J.-M., Irwin, J. *Aspect-Oriented Programming.* In proc. of the European Conference on Object-Oriented Programming (ECOOP'97). 1997.

[7] Lopes, C.V., Dourish, P., Lorenz, D.H., Lieberherr, K. *Beyond AOP: toward naturalistic programming*. ACM SIGPLAN Notices, 38 (December 2003), 34-43.

[8] Meisel, J.B. and Sullivan, T.S.: *Portals: the new media companies*. Info – The Journal of Policy, Regulation and Strategy for Telecommunications Information and Media. 5(2), 2000.

[9] Miller, L.A., *Natural Language Programming: Styles, Strategies, and Contrasts.* IBM Systems Journal, 20 (2), 184-215. 1981.

[10] Mothe, J., Chrisment, C., Dousset, B., Alaux, J. DocCube: *Multi-Dimensional Visualisation and Exploration of Large Document Sets*. Journal of the American Society for Information Science and Technology, JASIST, Special topic section: Web Retrieval and Mining, 54 (March 2003), 650-659.

[11] Nardi, B. *A Small Matter of Programming: Perspectives on End User Computing.* MIT Press, Cambridge, USA. 1993.

[12] Newcomer, E. *The Web Services Standards Mess*, http://www.webservices.org/index.php/article/view/1202/.

[13] Shirky, C. *Web Services and Context Horizons*. IEEE Computer, 35 (September 2002), 98-100.

[14] Stempfhuber, M., Hellweg, H., Schaefer, A. ELVIRA: User Friendly Retrieval of Heterogeneous Data in Market Research, p. 299 - 304. In: Callaos, Nagib; Harnandez-Encinas, Luis; Yetim, Fahri (eds.): SCI 2002:

The 6th World Multiconference on Systemics, Cybernetics and Informatics; July 14-18, 2002, Orlando, USA; Proceedings, Vol. I: Information Systems Development I.

[15] Voida, S., Mynatt E.D., MacIntyre, B., and Corso, G.M. *Integrating Virtual and Physical Context to Support Knowledge Workers*. IEEE Pervasive Computing, 1 (May/June 2002), 73-79.

[16] WebServices.Org *New Specifications Intended to Harmonzie Grid and Web Service Standards*. http://www.webservices.org/ index.php/article/view/1314/, January 20, 2004.