Web services communicating in the language of their user community

Kurt Englmeier
1st author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
KurtEnglmeier@computer.org

Pedro Contreras
2nd author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
p.contreras@qub.ac.uk

Steven Johnstone
3rd author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
s.johnstone@qub.ac.uk

ABSTRACT

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 miracle they were once conceived to be. Even though they support interoperability, developing large complex specifications for an application domain is still incredibly complicated.

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, WS-Talk, helps to establish interoperable web services across platforms. Much in the vision of JXTA 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. Our approach fosters quick ad-hoc solutions for a small network of peer-organizations. The advertisements cooperate with the transport level of well-established web service standards. The architecture emerges from lessons learnt in merging heterogeneous data collections during our European research project IRAIA.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval – *information filtering, retrieval models, search process*

H.3.5 [**Information Systems**]: On-line Information Services – *web-based services.*

General Terms

Human Factors, Standardization, Languages.

Keywords

Web services, natural language, text mining, semantic web.

1. INTRODUCTION

Context awareness usually refers to the system's capability of being aware of the user's context and of adapting its behavior accordingly while being minimally intrusive. A user's context can be quite complex, consisting of attributes like current physical position, personal history, and recurring behavioral patterns [2, 6, 13]. An example may be a PDA in a museum that provides context-sensitive information to the user depending on his/her physical location within the building. In the development of our information system IRAIA we addressed issues that were more abstract than physical environments. IRAIA regarded a user, instead, within an information landscape.

Context has to be communicated. Information items such as documents, for instance, have to be mapped into a coherent context landscape that allows, at the same time, to locate the whereabouts of a user. 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. Semantic web technologies and text mining techniques provide for features for the mapping process itself. Developed to represent knowledge and to analyze content they are key enablers when it comes to map information items and user locations into the landscape of an information ambience. Concept hierarchies derived from established and agreed taxonomies endow these analysis and representation techniques with a global concept layer. And, in addition, this layer can be mapped into different natural languages in parallel. A user's background, interests, and preferences can be expressed by semantic profiles which are also derived from the controlled vocabulary which reflects on an abstract level the information landscape and the user acting within it [5].

Of course, we could also describe the user by ontologies reflecting his/her experience or interests. However, whenever the interest changes we have to change the ontology [11]. In addition, developing ontologies for a large user group is an incredibly complicated task. Nevertheless there are common and basic attributes and relationships in such user descriptions that are not too volatile (like name, profession, date of birth, etc.). It is quite helpful to express these attributes in a more or less stable semantic structure that is in addition machine-processable. For the representation of more dynamic attributes we recommend controlled vocabularies and application of text analysis methods.

If we regard now a user in her working environment, then representations contain in addition descriptions of task and of tools that help to accomplish 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.

WWW 2004, May 17-22, 2004, New York, NY USA.
ACM xxx.

There is a lot of sobering truth in this example, and a lot to be learnt about the standards development process. Developing large, complex specifications is not an easy task. Once implementations start, 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. The demise of ebXML, for instance, can be seen in this light: it collapsed under its own weight. [9]

Nevertheless we acknowledge that semantic web standards are very helpful and without them a high level of interoperability would be out of reach in any case. However, in the combination of semantic web standards for the stable part of the service identification, and communication (for the transport level, for instance) with natural-language processing (NLP) mechanisms for the dynamic part, we put the technology users in the position to apply web services to create point-to-point connections. On a corporate or market sector level it is easier to reach an agreement on suitable controlled vocabularies than to design complicated XML models that can not be understood by the majority of the technical users.

The paper is organized as follows: section 2 outlines the rationale of semantic context-awareness. With examples from the IRAIA system we explain structure and function of a semantic co-ordinate system, in section 3, and the information mapping process in section 4. Section 5 presents the transformation of this approach to a network of peer-to-peer services. Section 6 concludes the discussion of this paper.

2. SEMANTIC CONTEXT-AWARENESS

The importance of a solid linguistic basis is often underestimated if not neglected entirely when it comes to the design of context-aware applications [4]. Whenever a machine has to understand or render language in any form, powerful underlying linguistic capabilities are required. Without that, many web intelligence applications addressing information ambiences will never take off.

2.1 Content proxies

In the approach outlined here context-aware information provision is constructed around a linguistic interface layer. The layer ensures that the system "understands" a phrase expressed in an ad-hoc way in natural language. "Understanding" means mapping an information item such as speech, a text, or a written or spoken passage 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. To supplement purely semantic analysis of content we have to apply in parallel text mining techniques in cases where semantic analysis is not suitable enough or not applicable. Semantic analysis addresses soft facts contained in an information item whereas data mining looks for facts stored as templated data or meta-data expressed in XML, for instance.

WS-Talk contains thus features that enable web services to communicate in free text based on well-established interface standards. WS-Talk scanning service analyzes the keywords of the above phrase (source language) and translates them into the controlled vocabulary (target language) generating automatically a content proxy for this request. The subsequent matching service 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 count with text mining methods that translate service descriptions into suitable terms of the target language.

The matching processes may also reflect the facets of semantic context-awareness emerging from the combined application of text mining methods and factual data mining methods as they are applied likewise in ontology analysis. The communication process between the services is realized through a combination of different independent services. They are independent in terms of each service being applicable in a different context, too. The scanning service, for instance, can cooperate with a speech processing service that is attached to an automatic telephone switchboard. It is obvious that the messages of the services themselves are wrapped into a standardized XML format to establish the basic layer for the communication.

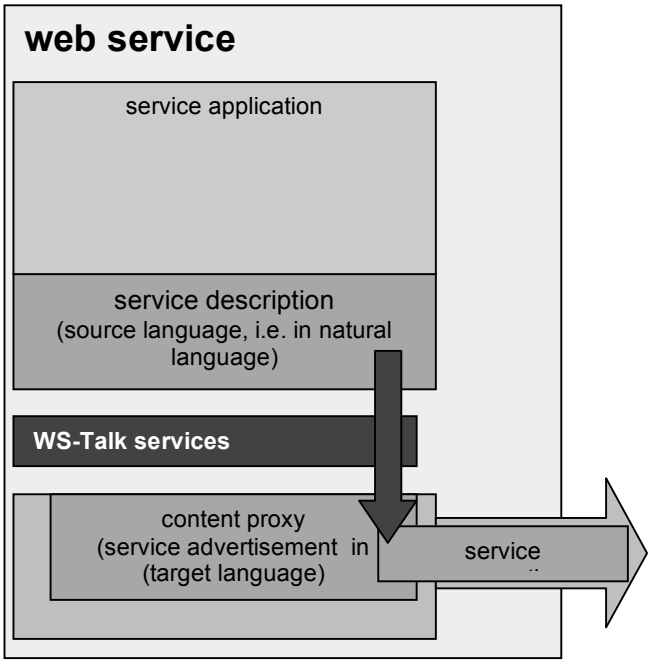

Figure 1. WS-Talk itself consists of a number of embedded services. The communication between services is realized through a content proxy (advertisement files) that are wrapped by an XML and/or JXTA-compatible structure.

2.2 The potential of natural language processing

The past few years have witnessed several major advances in NLP, allowing text and speech processing to make vast inroads in universal access to information. Owing to recent progress in open-domain question answering, multi-document summarization, and information extraction, exploring information posted on the web is becoming faster and more efficient. [3] Moreover, research efforts are enabling natural language processing in a wide range of languages. Users will receive a concise, exact answer to their questions expressed in her own language – independently of the communication to the knowledge base channel. Users communicate in their spoken language with an instant consumer messaging system or with a reservation system either on the phone or using a traditional terminal to a library system. New ways of service-service interaction emerge through the application

of NLP when interfaces "understand" protocols defined in the users' written or spoken language. NLP relies on predefined templates and pattern-based extraction rules to extract meaning.

TFIDF or LSI 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. This is a crucial drawback when topic shifts occur, and specifics for a certain topic are to be identified – a situation occurring similarly in information customization when shifting user interests have to be identified correctly within a dialog or interaction with an application. Content matching, whether for knowledge acquisition purposes or broadcasting purposes between web services, can be improved significantly if content associations can resort to multilingual controlled vocabularies (MCV) since they enhance NLP capabilities. In the context of this approach MCVs are derived from taxonomies (or a representative collection of information items) that are usually available for a huge variety of information domains. They manifest useful vocabularies that help to identify specifics in content without having to resort to huge corpora. Moreover, they are a useful instrument to tackle the problem of ambiguity that is inherent in spoken or written language.

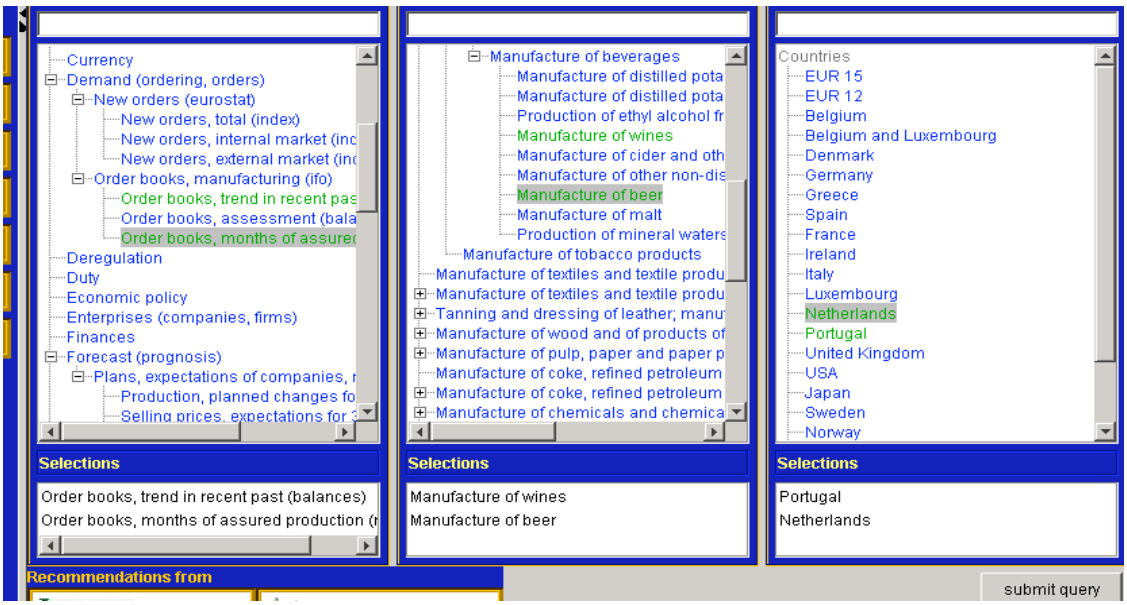

Figure 2. Searching and navigation in a semantic coordinate system for economic information. Selected concepts make up the initial query profile. While realizing their retrieval strategy the users usually perform iterative steps in defining a query and analyzing the retrieved results. In IRAIA, the documents are annotated solely with entries from the hierarchies.

3. A SEMANTIC COORDINATE SYSTEM

The following scenario from the information system IRAIA demonstrates the role and potential of a semantic coordinate system. Hierarchically arranged and grouped along major content facets a controlled vocabulary acts as a stable coordinate system easy to comprehend and memorize[1]. The users are thus much more in the position to localize themselves effortlessly. Successfully searching and navigating now means guided travelling from information to information just by changing the semantic coordinates, i.e. by pointing to relevant concepts. This structure on the other hand enables us to pinpoint the semantic location of any kind of information. Moreover, this structure gives the possibility for a user to modify his query easily by refining or reformulating it.

They are supported in this task by the system displaying correlated terms to the initial ones of the query. Indeed, a query concept is shown within its generic concept and the user can get more documents presented with a more specific or generic view.

For the ambience of economic information we produced a powerful taxonomy that merges two of the most important structures in this field: Eurostat's NACE[2] nomenclature and the industry systematic of the ifo Institute for Economic Research (see Figure 2). The unified taxonomy creates a semantic coordinate system that enables exact and automatic positioning[3] of coherent documents even if they are of different types. It also provides users with the necessary orientation while exploring the information ambience or the actual workspace[4]. Like in using

[1] The coordinate system itself can be presented simultaneously in different languages. This ensures that the domain model mentioned later features multilinguality.

[2] Nomenclature des Activités dans la Communauté européenne - systematic of the economic activities of the European Union

[3] Automatic positioning means annotating a document with entries from the concept hierarchies. During the automatic process the terms of a document are matched with those of the hierarchies and their synonyms. The hierarchy entries which get the highest frequency of matches are chosen for the annotation. (See below for details.)

[4] Time series are usually related to a certain taxonomy, a hierarchical description of all the themes covered by these documents. IRAIA benefits from these taxonomy structures by employing them as concept hierarchies. Conversely, access keys of documents in time series bases are linked closely to these

languages it helps users as a passive vocabulary to identify the topics of their information problem.

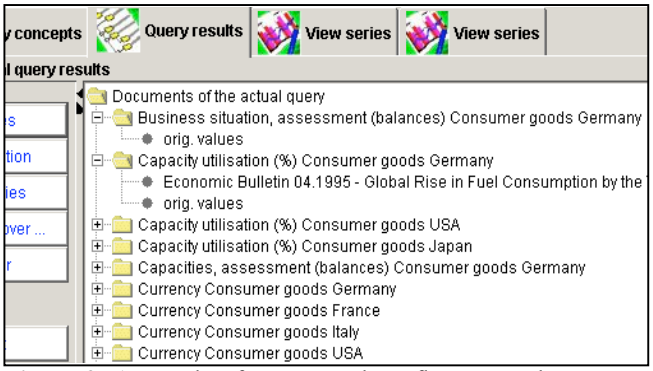

Figure 3. A sample of query results reflects a workspace. Its corresponding documents are presented together independently from the type of data and even the underlying query language.

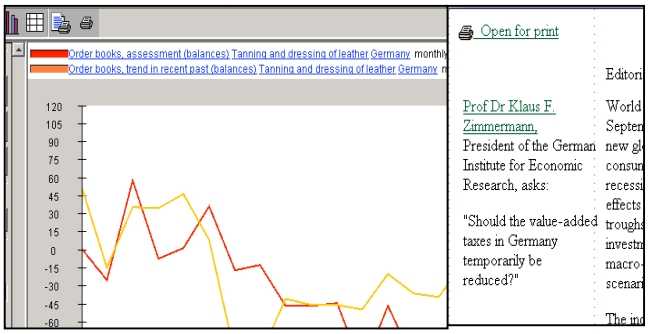

Figure 4. In the case of economic information for instance, time series and texts are presented together. Text may explain the numerical data or give additional information the user should know.

Grouping large data samples along easily discernible aspects of an information ambience tackles the problem of ambiguity which the traditional search engines cannot master. This means that at each navigation step a user gets only a workspace, precisely tailored to the selected entries in the respective concept hierarchy, and taking into account multimedia documents actually retrieved. This also means that the sample of retrieved document (i.e. the workspace) form a contextually coherent group and are presented as such.

The example shows how a semantic coordinate system works to determine the correct location of a document or a user request within an information system. This positioning process is tantamount to correctly matching requests and information items.

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:

- 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.
- 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.
- Instructions how to handle data that are passed from one service to another are also expressed in terms of the controlled vocabulary.

Providing the controlled vocabulary in different languages in parallel has the advantage of exchanging service descriptions even across language boundaries, and this avoids the final users having to apply a language they are not familiar with or having to be trained in a formal description language.

4. INFORMATION MAPPING

In the end, every design model for finding the adequate web service goes back to the good old quest to find an efficient matching of representations of services. The matching process itself can be regarded as a process that compares service representations (descriptors) based on different forms of linguistic and/or statistical evidence. Different representations of services (words, features, phrases, additionally assigned words and structured combinations of them) are matched in this process. [1]

4.1 Content Matching

The task to be carried out corresponds to document categorization. While parsing the concept hierarchies the entries of their nodes are treated as profiles. These profiles are composed of phrases because the entries of the concept hierarchies usually consist of a couple of words describing a concept. From the text analysis results an index list of words that is in this form a parsed representation of the controlled vocabulary. Alternatively, a profile is composed of a set of phrases automatically extracted during a learning stage. This latter method is used when a training set of service descriptions is available, that is to say when the system is provided with descriptions and their associated correct entries. The salient difference is that, like in text parsing, the underlying morphology of the concept trees is preserved. This allows the expansion of the meaning of a node by propagating the content from its ancestors. Thus, a node together with its ancestors can be regarded as a pseudo-document that may be helpful in the further matching process.

This process compares, roughly speaking, the content of the concept hierarchies with those of the service descriptions and decides which tree nodes are the most prominent ones for this document. The entries of each hierarchy are ranked according to the weight that those are relevant to the document. This annotation process serves to create references from the service to the hierarchies of one or more thematic domains.

A threshold ensures that only those entries are selected that contribute to an appropriate abstraction of the service description. It's obvious that the ancestors of an entry get less weight than one of its subordinal concepts (in terms of specialization). This is based on the assumption that only the most specific terms are in the position to capture the specifics of a document's content and these terms can be found towards the respective ends in the ramification of a hierarchy. References from a service to the most specific entries of a hierarchy lead to higher precision. (For detailed information on the matching process see [8]).

---

structures enabling comprehensive deduction of a series' meaning from its relationships to the respective hierarchy concepts. Applying these taxonomies both to texts and to numerical data enables an integrated view on both types of information.

4.2 Evaluation and Annotation

The following evaluation calculates the weight (bel) in a profile (a node in a concept hierarchy) due to the occurrence of a concept c_i of a service description d:

$$bel_{Eh}(d) = \sum_{c_i} \left(\frac{tf_i}{s(d)} \cdot \frac{s(h)}{tfE_i} \right) \cdot e^{\frac{s(Eh \cap d)}{s(Eh)}}$$

where

Eh= an entry or node from the concept hierarchy h, represented by a profile,

C_i = a concept from the profile Eh,

tf_i = frequency of the concept C_i in the service description d,

tfE_i = frequency of the concept C_i in the node Eh,

s(d) = size of the description d (number of concepts),

s(h)= size of the concept hierarchy,

s(Eh) = size of the node Eh,

$s(Eh \cap d)$ = number of terms from Eh that occur in d.

$\frac{tf_i}{s(d)}$ measures the importance of the concept in the service description,

$\frac{s(h)}{tfE_i}$ measures the importance of the concept in the concept hierarchy and

$\frac{s(Eh \cap d)}{s(Eh)}$ measures the rate of occurrence of the concepts from the concept hierarchy node in the description (coverage).

This formula is derived from the well-known *tf.idf* weighting function. The values are normalized and remain between 0 and 1. Concept phrases for a given document are ranked by this function. A concept phrase is selected for annotation if its value is above a certain significance threshold (let's say 0.4, for instance). Due to practical reasons only the three entries highest in ranking as well as above the threshold are chosen.

$$ann(d) = \bigcup_h \begin{cases} bel_{Eh}(d) \text{ with } bel_{Eh}(d) >= \delta_0 \\ \{\} \text{ otherwise} \end{cases}$$

where

δ_0 = *significance threshold*

h = concept hierarchy

Calculations are performed for all concept hierarchies. We emphasize once again that the matching process is based exclusively on concept phrases. These entries of concept hierarchies (as well as the controlled vocabulary derived) define the semantic context of the specific domain-related web service space.

5. PEER TO PEER WRAPPER AND SERVICES

A major characteristic of the peer-to-peer (P2P) network model is its independence from any central organization. This autonomy allows for the possibility that any peer can potentially act as a requester or provider of services, depending on the task for which it was defined. JXTA [7] is an open source project which defines a set of six language independent protocols [10] - peer discovery protocol, peer resolver protocol, peer information protocol, pipe binding protocol, rendezvous protocol and end point routing protocol. These protocols define the necessary building blocks with which to build a P2P infrastructure. They set up a ubiquitous, secure and pervasive virtual network [12] that may sit on top of any transport protocol.

We propose to create a JXTA service for WS-Talk that will generate content proxies (XML advertisements) describing a service and that can receive and respond to other XML-wrapped content proxies. The service providers of a specific domain remotely publish their services allowing them to be discovered by other service requesters. Services can be identified using a unique id (UUID) and their content proxies. Once the service has been discovered it can be accessed in a standard way using a JXTA pipe and any result may be returned via HTTP, Java application or a Java Web start method.

A JXTA-enabled WS-Talk service is one that is made available through the publication of a Module Specification advertisement that contains in addition one or more WS-Talk content proxies. This advertisement gives details of a specification for a service of a certain class of functionality. Its main purpose is to provide references for documentation on how to implement a service of this type. It includes a pipe advertisement that contains the PipeID of the pipe that can be used to communicate with a running service. An implementation of the service specification publishes a Module Implementation advertisement, which contains details of a specific implementation of the service. A WS-Talk service transforms in addition a content proxy's information into additional implementation specifications. The ability to have different Module Implementation advertisements relating to one specification allows the service to be language/platform independent. For example the same service could be implemented in Java or C++. Through the possibility to represent the WS-Talk controlled vocabulary in different languages in parallel the service providers can even write their implementation specifications in their own language.

In order to simplify the proposed model we have separated the client and server tasks onto two different peers. This is shown in Figure 5.

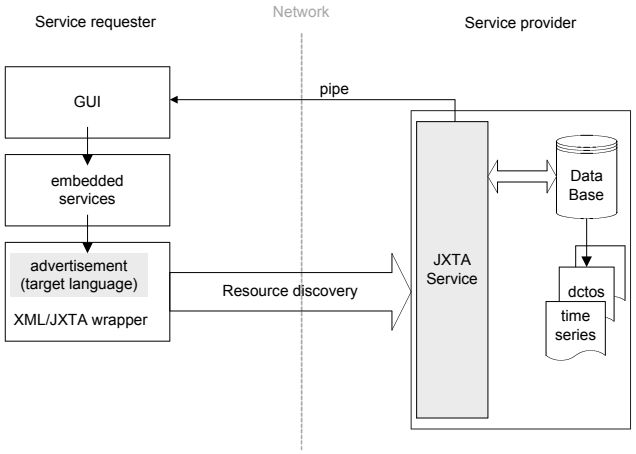

Figure 5. JXTA wrapper and services architecture

At a basic level a wrapper is an XML document that contains a JXTA compatible structure while embedded content proxy can be written in the user's language. The wrapper can be disseminated by a peer through the Peer Discovery Protocol. For instance, a provider could include an instruction how to query data from his own database in a free-text, natural language description. Once the description has been processed and wrapped it can be sent to peers in the peer group that are providing a service that can handle and understand this type of description.

6. CONCLUSIONS

Despite advances in the development of semantic web standards, NLP methods still form part of many techniques for enabling computers to understand and engage in human communication. 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 new generations of semantic web technologies. 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 services.

XML and other semantic web standards help a lot in matching the users' interests with a distributed application domain. However to make this work, all parties – the application designer as well as the application users – have to agree on common domain representation in semantic web standards. It cannot be expected that such a consensus can be reached for the vast majority of situations.

What is required is an automatic context-aware transformer of a web service description into the catalog logic of the respective application domain. This transformer architecture resides on the user's side. They are in the position to describe the services they require in their ambience. The underlying design rationale of WS-Talk is defined by dynamic and flexible customization of the service description. It can be crucial in the design of future web services to adapt to the user community settings instead of requesting the user community to adapt to the services' logic.

7. ACKNOWLEDGMENTS

Research outlined in this paper is part of the project IRAIA that was supported by the European Commission under the Fifth Framework Programme (IST-1999-10602). However views expressed herein are ours and do not necessarily correspond to the IRAIA consortium.

8. REFERENCES

[1] Callan, J.P.; Croft, W.B.; Broglio, J. TREC and TIPSTER experiments with INQUERY. Information Processing and Management, 31, 1995, 327-332.

[2] 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.

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

[4] Davies, N. and Gellersen, H.-W. (2002). Beyond Prototypes: Challenges in Deploying Ubiquitous Systems. IEEE Pervasive Computing, 1 (1), 26-35.

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

[6] 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.

[7] JXTA, http://www.jxta.org

[8] 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.

[9] Newcomer, E. The Web Services Standards Mess, http://www.webservices.org/index.php/article/view/1202/, October 12, 2003.

[10] Protocols, http://spec.jxta.org/v1.0/docbook/JXTAProtocols.html.

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

[12] Traversat et all 2003, http://www.jxta.org/project/www/docs/JXTA2.0protocols1.pdf.

[13] 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.