Study on Integrating Semantic Applications

Martin Dzbor[1], Philipp Cimiano[2], Victoria Uren[1] and Sam Chapman[3]

[1] Knowledge Media Institute, The Open University, UK
[2] Institute AIFB, University of Karlsruhe, Germany
[3] Natural Language Processing Group, University of Sheffield, UK
{M.Dzbor, V.S.Uren} @open.ac.uk; cimiano@aifb.uni-karlsruhe.de; sam@dcs.shef.ac.uk

Abstract. In this paper we describe two approaches to integrating standalone information processing techniques into a semantic application capable of acquiring and maintaining knowledge. We distinguish between integration through aggregation and through choreographing, and argue that the latter is not only simpler to realize but also provides greater benefits. The benefits were, in our experiment, related to developing a capability of maintaining and validating knowledge through an integration of down- and upstream knowledge processing tools. We describe the principles of integration and relate them to pragmatic challenges for the semantic web and to strategic directions of its evolution.

1 INTRODUCTION

Over the last two years we can observe an emphasis not only on designing and handcrafting ontologies, but also on their automated population. This is arguably the key to bootstrapping the semantic web, thus making it more suitable for practical, larger-scale applications. Current approaches to ontology population are based on various techniques; e.g. automated pattern recognition and extraction on the Web [5, 8] and web mining algorithms relying on information redundancy [2, 3].

Moreover, research into abstract infrastructures for the semantic web and into specialized tools for expert knowledge engineers [9] is complemented by tools for the end users of the semantic web. These emerging tools support end users in annotating documents [11]; in browsing semantically marked up spaces [14], or in integrating web and semantic web resources [7]. They are characterized by *applying* rather than *developing* knowledge models. Their authors stress the importance of achieving rapid reward through good enough semantics rather than global completeness.

These developments may signal that semantic web research is moving from its early vision and early adopter issues [1] to more pragmatic ones [12]. A pragmatic challenge for the semantic web arises from the fact that "…when it comes to promoting the semantic web idea to web users as they have accumulated a 10 year experience with the web, only a truly superior product will win them over" [12].

From a pragmatist's and practitioner's perspective this argument can be translated into *a need to deliver added value to practical applications across a large part of the knowledge processing (or management) chain*. Shortcomings of the current version of the semantic web are remarkably similar to problems with social tools in general [10].

In this paper we address the strategic challenges of automated bootstrapping and supporting both developers and end users. We investigate how information extraction services might interact with the user to provide added value. This interplay of bootstrapping and user interaction issues has not been investigated so far. As a first step towards understanding requirements on integrating different knowledge tools we conducted a case study in which we extended a user-centered tool for browsing the semantic web (Magpie), and integrated it with tools for extracting information from visited pages (PANKOW and Armadillo). Linking end user tools with the knowledge production tools is one approach to tackling the need for the creation of semantic applications addressing a larger part of the knowledge processing chain.

From the architectural point of view, there are several ways to realize such bootstrapping solutions. We consider two approaches to integrating the upstream and downstream knowledge tools: one focusing on data aggregation and another focusing on choreographing independent semantic services into a loose application that is capable of addressing the pragmatic issues of scalability, bootstrapping and simultaneous provision of rewards to both users and developers of the semantic web.

To illustrate the evolution of requirements for the semantic web, consider the evolution of Magpie applications. From a *handcrafted solution* supporting a fixed set of actions [6], the work moved to an *open solution* with dynamically definable actions for navigating the semantic web [7]. However, it fell short of realizing the pragmatic vision of a knowledge management solution introduced earlier. As Magpie, other applications fall short of addressing the knowledge processing chain in its entirety. For instance, Haystack [14] performs well on reuse and sharing but does not help with acquisition and maintenance. Protégé [9] does mostly representation with basic versioning and mapping support. GATE [4] (and its application in KIM) supports discovery and annotation, but there is limited use, reuse and maintenance of the discovered knowledge. Web browsers in general are good at knowledge presentation but not discovery, creation and maintenance.

In order to use the semantic web [7] a user needs a toolkit that efficiently connects semantic and standard (i.e. non-semantic) browsing techniques, e.g. using the automated semantic annotation of web pages. However, the experience with Magpie shows that this approach is often brittle [16]. Magpie recognizes terms in a web page in real time with high precision whilst *within* an ontological domain. When a Magpie lexicon is used outside the intended, well-defined domain, performance rapidly falls.

One way to tackle brittleness is by extending fast annotation techniques with some form of knowledge discovery. This would allow *more robust browsers for the semantic web*. When we compare the brittle Magpie lexicons with those augmented by the information extraction (IE) tool PANKOW [2] (see also [16]), the users' performance in an exploratory task indeed improved when using the augmented lexicons where the domain boundaries were less brittle. However, this achievement came at the price of not being able to generate lexicons in real time – knowledge discovery takes time.

Hence, the idea of robust browsing is clashing with real-time responsiveness. An alternative path draws on the idea of having a 'shell' for building semantic web applications. Such a shell provides generic mechanisms for integrating the ontologies, knowledge bases (KB) and web accessible resources with the hyperlinks, (semantic) web services and tools interacting with them. Rather than producing applications for browsing the semantic web, we propose to view them as the result of *integrated,* *online solutions for knowledge management (on the semantic web)*. This vision is novel, and as such, it is an intersection of three different areas: semantic web, web services and knowledge management.

Here we use a larger-scale integration effort aimed at providing a practical knowledge management solution based on two premises. First, we turn standalone IE tools into services. Second, these independent services are loosely connected in a virtual framework from which a robust solution for implementing pragmatic semantic web applications emerges. Our integration addresses the needs of both the end users (with their interest in browsing, retrieving and annotating), and the developers (who need to create and maintain KBs).

2 MOTIVATION FOR AN INTEGRATED APPLICATION

We believe that adoption of the semantic web depends on satisfying the different needs of different users. Focusing solely on end users is rewarding in the short term, but in addition to the instant, shallow and short-lived rewards, semantic applications should offer sustainable, "delayed gratification" [15].

As a scenario, take a preparation of research review in a field. One starts with a few explicit keywords and domain anchors. A large part of the review is about retrieving additional knowledge about the existing facts. In terms of a Magpie-based semantic application, semantic services may include e.g. "Find papers on theme Y". In addition to recalling and reusing the existing references, the review writing has a hidden but important *exploratory component* whereby we create new knowledge correlated with the existing facts. We want to create knowledge of e.g. "What has argument A in common with person P". The state-of-the-art tools support either knowledge retrieval and *reuse*, or knowledge *creation*, but not both.

2.1 End user's perspective: Towards more robust knowledge navigation

Fig. 1a shows extract of a web page annotated using a user-selected lexicon populated from internal databases of research activities. As can be expected, concepts like "*Magpie*" or "*BuddySpace*" are highlighted because they were explicitly defined in a KB. However, a few potentially relevant concepts are ignored (e.g. "*web services*" or "*human language technology*"). These are closely related to the existing terms but are not in the KB. This is a sign of (i) an incomplete lexicon and (ii) ontology brittleness (i.e. rapid degradation of performance when outside the intended domain).

To overcome brittleness, Fig. 1b shows a collection of additional concepts discovered by an offline IE tool. These relate to the 'discourse' of the user-selected lexicon, yet do not currently exist in the KB, from which annotation lexicons are generated. New concepts are mostly on the level of instances, but the IE tool (which supplies these facts) also proposes a finer-grained classification using discovered classes such as "*Activity*" or "*Technology*". Fig. 1c shows some of the instances (e.g. "*browsing the web*" or "*workflow and agents*") already incorporated into the KB and used for annotation. Importantly, items such as "*workflow and agents*" are not only highlighted as "*Research activities*", but the user can also invoke associated services to obtain additional knowledge about these discoveries; e.g. the semantic menu marked as ❶ in Fig. 1c shows services relevant to the discovered "*workflow and agents*" instance.

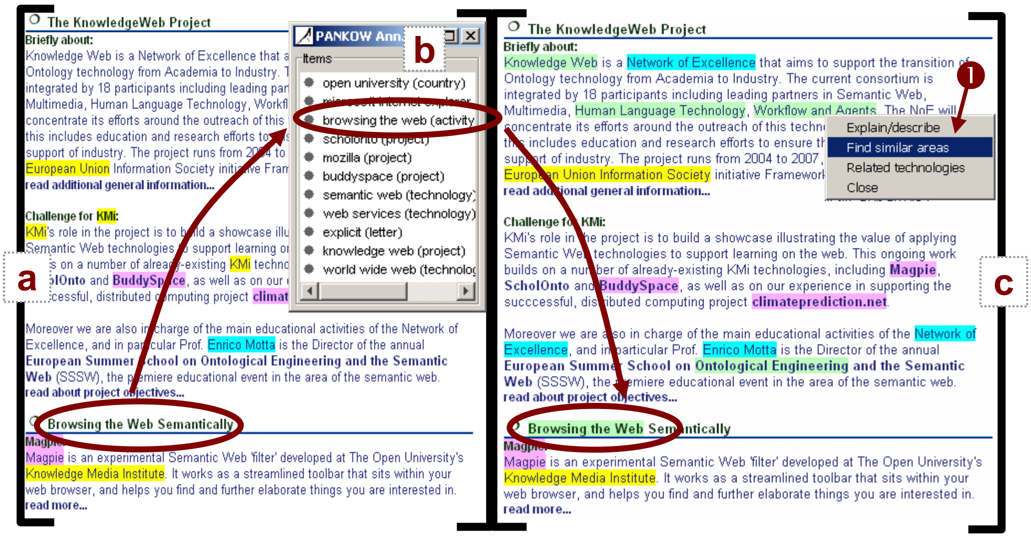
2.2 Engineer's perspective: Towards automated knowledge maintenance

The challenge from section 2.1 is to avoid having the user manually entering and committing the discoveries into a KB. Having a lexicon/KB that evolves and exhibits learning and adaptation would obviously justify using the adjective "semantic". From engineer's perspective, the downside is the manual filtering, validation and inclusion of the suggestions from IE into the KB. In our experiments using IE tool PANKOW, the existing lexicon of 1,800 concepts was extended by additional 273 organizations, 80 events, 176 technologies, etc. The engineer had to manually adjust classification for 21-35% items. Linking PANKOW to another (validating) IE technique reduced the need for adjustment (e.g. for events) to 8%. Thus, by integrating the pragmatic IE techniques and by making them accessible from a semantic browser we achieved benefits in terms of evolving knowledge and assuring its quality.

Benefits for the end user include more robust support for browsing, but also create opportunity to personalize a generic, handcrafted KB. While the personalization was not the focus of our case study, it is an important side-effect of our strategic aim to develop a semantic application addressing a larger part of the knowledge processing chain. Personalization, in turn, may lead to creating tools that are aware of such pragmatic issues as trust or provenance [12].

3 TWO APPROACHES TO APPLICATION INTEGRATION

Architecturally, there are several ways to realize the strategic and pragmatic aims. Two such approaches described below integrate different techniques, originally de-



The KnowledgeWeb Project

Briefly about:

Knowledge Web is a Network of Excellence that a
Ontology technology from Academia to Industry. T
integrated by 18 participants including leading par
Multimedia, Human Language Technology, Workfl
concentrate its efforts around the outreach of this
this includes education and research efforts to
support of industry. The project runs from 2004 to
European Union Information Society initiative Fram
read additional general information...

Challenge for KMi:

KMi's role in the project is to build a showcase illu
Semantic Web technologies to support learning o
s on a number of already-existing KMi techno
olOnto and BuddySpace, as well as on our
cessful, distributed computing project climat

Moreover we are also in charge of the main educational activities of the Network of Excellence, and in particular Prof. Enrico Motta is the Director of the annual European Summer School on Ontological Engineering and the Semantic Web (SSSW), the premiere educational event in the area of the semantic web.
read about project objectives...

Browsing the Web Semantically

Magpie:

Magpie is an experimental Semantic Web 'filter' developed at The Open University's Knowledge Media Institute. It works as a streamlined toolbar that sits within your web browser, and helps you find and further elaborate things you are interested in.
read more...

The KnowledgeWeb Project

Briefly about:

Knowledge Web is a Network of Excellence that aims to support the transition of Ontology technology from Academia to Industry. The current consortium is integrated by 18 participants including leading partners in Semantic Web, Multimedia, Human Language Technology, Workflow and Agents. The NoE will concentrate its efforts around the outreach of this techn
this includes education and research efforts to ensure th
support of industry. The project runs from 2004 to 2007,
European Union Information Society initiative Framework
read additional general information...

Challenge for KMi:

KMi's role in the project is to build a showcase illustrating the value of applying Semantic Web technologies to support learning on the web. This ongoing work builds on a number of already-existing KMi technologies, including Magpie, ScholOnto and BuddySpace, as well as on our experience in supporting the succcessful, distributed computing project climateprediction.net.

Moreover we are also in charge of the main educational activities of the Network of Excellence, and in particular Prof. Enrico Motta is the Director of the annual European Summer School on Ontological Engineering and the Semantic Web (SSSW), the premiere educational event in the area of the semantic web.
read about project objectives...

Browsing the Web Semantically

Magpie:

Magpie is an experimental Semantic Web 'filter' developed at The Open University's Knowledge Media Institute. It works as a streamlined toolbar that sits within your web browser, and helps you find and further elaborate things you are interested in.
read more...

Fig. 1. Extract from a web page annotated with a brittle KB and lexicon (a), a list of KB extensions proposed by PANKOW and visualized in Magpie collector (b), and the same page annotated by an extended KB/lexicon as learned by PANKOW and validated by Armadillo (c).

veloped as standalone applications. Both approaches use the existing open semantic services framework from Magpie [7]. From knowledge management perspective the two integration approaches can be seen as interactive learning loops comprising knowledge acquisition, extension, validation, presentation and reuse.

As shown in Fig. 2, one loop (starting by action 1 and ending with 8a) is shorter and shallower. It acts as a channel for delivering newly acquired and potentially relevant facts to the user. We call this *integration through aggregation*, and discuss it in section 4.1. The other loop (from 1 to 15, but replacing action 8a with 8b) is more complex but it already addresses the issue of knowledge validity not only relevance. This strategy facilitates *integration through choreographing*; details in section 4.2.

3.1 Knowledge reuse: Integration through aggregating applications

Information aggregation is "a service that gathers relevant information from multiple sources" and "adds value by analyzing the aggregated information for specific objectives" [17]. Key capabilities of an aggregating service are: (i) to use multiple information sources, repositories or providers, and (ii) to act as a central hub in the information exchange. Multiple providers of partial knowledge or information are typically unaware of each other and rely on the central hub – the aggregator – to mediate between them. Aggregation, however, has several limitations.

First, knowledge maintenance is an asynchronous interaction among services. The aggregator is likely to become a bottleneck. Also a tension arises between maintaining the real-time responsiveness of the semantic application [6] and the quality of discovered knowledge. Second, the interaction that emerges from aggregating the discovered instances and their subsequent presentation is user-driven. As shown in Fig. 2, this part of the integrative framework delivers rather trivial discoveries to the user. In practice, knowledge delivery uses visual user interfaces of the Magpie plug-in. Magpie's functionality has been described extensively e.g. in the context of supporting organizational memory [6] or online undergraduate education [7].

Trivial knowledge maintenance occurs by employing an IE service, which can be plugged into Magpie using its trigger services facility [6]. The IE service is informed about the web page the user visited by means of an asynchronous message. If new information is extracted, the user is alerted automatically via a trigger service interface, where s/he can interact with the discovered instances. The interaction may comprise the invocation of semantic menus, if there are any defined. However, most IE tools only suggest a class membership of a new instance, no other properties.

The IE engine that has been published as a trigger service for Magpie and tested in this particular role was PANKOW [2]. The aggregation of PANKOW discoveries in Magpie's collectors added value to the original Magpie's capabilities, however, this was limited due to *mistakes in classifying* discoveries and *transient nature* of the IE service responses. Particularly the last factor is a major shortcoming of using an aggregation strategy. This design constraint allowed PANKOW to inform the user instantly, but the discoveries were discarded when the browsing session ended and Magpie's interfaces were reset. Knowledge that was not persistently stored in lexicons (i.e. serialized KBs) was lost for any future reuse. We should take in account not only the amount of discovered knowledge but also its quality and extensibility.

3.2 Knowledge creation: Integration through choreographing applications

The second approach to integrating semantic applications assumes a reflection on results acquired by one application in another one. Instead of instant knowledge presentation emphasized by aggregation, the emphasis shifts to *knowledge validation*. To validate knowledge we need to go beyond mere discovery of additional facts in a corpus and their shallow classification [4]. In order to *create* new knowledge we need two consecutive steps: (i) an instantiation of the discovered concepts to appropriate classes, and (ii) shallow facts to be supported by relational knowledge.

Instead an instant user notification, the IE tool for validation is catching the output of the initial fact discovery step. Thus, we do not aggregate additional services with Magpie. Instead we create a choreographed sequence of standalone services. A choreographed process is defined by "the roles each party may play in the interaction; it tracks the sequence of messages that may involve multiple parties and multiple sources" [13]. While aggregation typically assumes the information exchange between services is controlled by a single party, choreography is different. It only tracks the message sequence, where no party owns the conversation [13].

Knowledge persistence in an aggregating service is not a necessity, but in service choreographing it is critical. Component services of a choreographed process need to be aware of the states knowledge goes through – e.g. candidate, valid and extended knowledge are different states. Simple knowledge retrieval and aggregation services, such as described in [6] or in section 3.1, are state-less. The interaction is controlled by the aggregator. By removing the aggregator as the owner of a conversation, the coordination among choreographed services needs to share knowledge states.

PANKOW was our first-pass IE service for discovering facts in a web page, and delegated the results of the initial discovery to the Armadillo [3] service for maintenance. Armadillo was published into Magpie's open services framework as a special type of service that could be triggered by other services rather than the user's interaction with a web page. It maintained a persistent record of discovered facts in a local RDF store. The candidate facts from the store were checked against additional sources on the Web using the principle of information redundancy [3]:

- o By proving the existence of certain 'expected' relationships we were able to *create new knowledge about the discovered facts*.

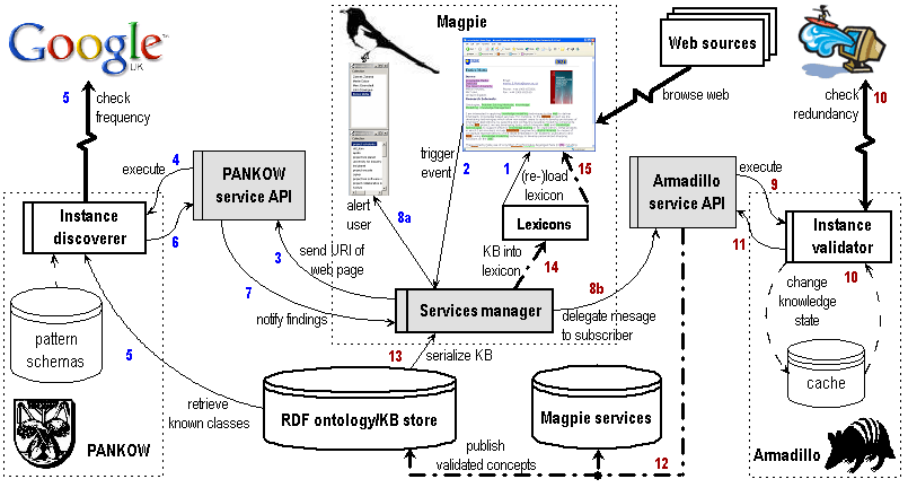


Fig. 2. Magpie – PANKOW – Armadillo interactions: *aggregation* (1-8a) and *choreography* (1-8b-15)

- The existence of the relationships using and extending the proposed class membership of an instance could be seen as *a form of IE validation*.

A discovery is validated when a new relational knowledge can be obtained using the ontological commitments made at an earlier stage. Once a fact is validated, it changes its state from a candidate to knowledge. It can be committed to the public repository containing the KB the original Magpie lexicon for annotating web pages was generated from. Magpie's lexicons are explicit serializations of KBs, and obviously, a different lexicon will be created from a changed KB. Magpie's support for loading a lexicon from a remote URI is beneficial for maintaining knowledge. When Armadillo validates PANKOW proposals and persistently publishes them in the RDF data store, the changes are propagated to the user next time a lexicon is requested from the KB.

4 DISCUSSION

In our opinion, integration based on loose choreography of the independent services into a larger-scale and robust semantic application is the way for meeting pragmatic challenges for the semantic web. It draws on the pluralistic notion of the semantic web being an open knowledge space with a variety of resources, which always mutually interact. If we want to highlight the primary benefit of choreography over aggregation, it is the aspect of *ownership*, *scalability* and an opportunity to bring in *social trust*. Because there is no need to design a service overseeing the interaction of the component tools, this simplifies the application development. Significant time can thus be saved by the developers, who would otherwise need to re-design all tools involved (i.e. PANKOW, Armadillo and Magpie).

What the user receives in our integrated effort are candidate instances that pass several validating tests together with some of their properties and relations to other instances. In respect to knowledge validity, our choreographed integration improves on the individual component tools – mainly observable on the level of knowledge maintenance. While there are many techniques and strategies for knowledge discovery, representation and reuse, the maintenance of knowledge is in its infancy. *We are not aware of any major research into robust and scalable knowledge maintenance*.

Another important aspect is the capability of replicating creation of knowledge from both, direct experiences and social interactions. In our experiments, the direct experience is attributed to C-PANKOW drawing on the constructed hypotheses. The social creation came in a shape of Armadillo looking into trusted and established information sources for obtaining additional evidence for creating new knowledge. To generalize, this might be an approach to delivering a version of the semantic web, which is both *a formally and a socially constructed space*, and which contrasts with the current version, which mostly emphasizes formal, experience-based constructions.

5 ACKNOWLEDGMENTS

The integration effort has been supported by the Dot.Kom, KnowledgeWeb and Advanced Knowledge Technologies (AKT) projects. Dot.Kom and KnowledgeWeb are sponsored by the European Commission by grants no. IST-2001-34038 and FP6-507482 respectively. AKT is an Interdisciplinary Research Collaboration (IRC) sponsored by the UK Engineering and Physical Sciences Research Council by grant no. GR/N15764/01. The AKT IRC comprises the Universities of Aberdeen, Edinburgh, Sheffield, Southampton and The Open University.

6 REFERENCES

[1] Berners-Lee, T., Hendler, J., and Lassila, O., *The Semantic Web.* Scientific American, 2001. **279**(5): p. 34-43.

[2] Cimiano, P., Ladwig, G., and Staab, S. *Gimme' the Context: Context-driven Automatic Semantic Annotation with C-PANKOW*. In *14th Intl. WWW Conf.* 2005. Japan.

[3] Ciravegna, F., Dingli, A., Guthrie, D., *et al. Integrating Information to Bootstrap Information Extraction from Web Sites*. In *IJCAI Workshop on Information Integration on the Web*. 2003. Mexico.

[4] Cunningham, H., Maynard, D., Bontcheva, K., *et al. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications.* In *40th Anniversary Meeting of the Association for Computational Linguistics*. 2002. Pennsylvania, US.

[5] Dill, S., Eiron, N., Gibson, D., *et al. SemTag and Seeker: bootstrapping the semantic web via automated semantic annotation*. In *Proc. of the 12th Intl. WWW Conf.* 2003. Hungary: ACM Press. p. 178-186.

[6] Dzbor, M., Domingue, J., and Motta, E. *Magpie: Towards a Semantic Web Browser*. In *Proc. of the 2nd Intl. Semantic Web Conf.* 2003. Florida, USA. p. 690-705.

[7] Dzbor, M., Motta, E., and Domingue, J. *Opening Up Magpie via Semantic Services*. In *Proc. of the 3rd Intl. Semantic Web Conf.* 2004. Japan. p. 635-649.

[8] Etzioni, O., Cafarella, M., Downey, D., *et al. Methods for domain-independent information extraction from the web: An experimental comparison*. In *Proc. of the 19th AAAI Conf.* 2004. California, US. p. 391-398.

[9] Gennari, J., Musen, M.A., Fergerson, R., *et al.*, *The evolution of Protege-2000: An environment for knowledge-based systems development.* Intl. Journal of Human-Computer Studies, 2003. **58**(1): p. 89-123.

[10] Grudin, J., *Groupware and Social Dynamics: Eight Challenges for Developers.* Communications of the ACM, 1994. **37**(1): p. 92-105.

[11] Kahan, J., Koivunen, M.-R., Prud'Hommeaux, E., *et al. Annotea: An Open RDF Infrastructure for Shared Web Annotations*. In *10th Intl. WWW Conf.* 2001. Hong-Kong.

[12] Kalfoglou, Y., Alani, H., Schorlemmer, M., *et al. On the Emergent Semantic Web and Overlooked Issues*. In *Proc. of the 3rd Intl. Semantic Web Conf.* 2004. Japan. p. 576-590.

[13] Peltz, C., *Web Services Orchestration and Choreography.* Web Services J., 2003. **3**(7).

[14] Quan, D., Huynh, D., and Karger, D.R. *Haystack: A Platform for Authoring End User Semantic Web Applications*. In *Proc. of the 2nd Intl. Semantic Web Conf.* 2003. Florida, USA. p. 738-753.

[15] Takeda, H. and Ohmukai, I. *Building semantic web applications as information/knowledge sharing systems*. In *UserSWeb: Wksp. on User Aspects of the Semantic Web*. 2005. Crete.

[16] Uren, V.S., Cimiano, P., Motta, E., *et al. Browsing for Information by Highlighting Automatically Generated Annotations: User Study and Evaluation*. In *Proc.of the 3rd Knowledge Capture Conf.* 2005. Canada. p. (submitted).

[17] Zhu, H., Siegel, M.D., and Madnick, S.E. *Information Aggregation – A Value-added e-Service*. In *Proc. of the Intl. Conference on Technology, Policy and Innovation*. 2001. The Netherlands.