Approach for Enterprise Search and Interoperability using Lightweight Semantic

Martin Šeleng*, Michal Laclavík*, Štefan Dlugolinský*, Marek Ciglan*, Martin Tomašek**, Ladislav Hluchý*
* Institute of Informatics, Slovak Academy of Sciences, Dúbravská cesta 9, 845 07 Bratislava, Slovakia
** InterSoft, a.s., Floriánska 19, 040 01 Košice, Slovakia
{martin.seleng, michal.laclavik, stefan.dlugolinsky, marek.ciglan, ladislav.hluchy}@savba.sk ,
martin.tomasek@intersoft.sk

Abstract— **In this paper we want to describe a solution for the enterprise search and interoperability by using the lightweight semantic approach, which is suitable for small and micro enterprises. Our approach is based on discovering and reusing an existing knowledge hidden in enterprise infrastructure ecosystem, like emails and content management systems (documents). Using the lightweight semantic approach our solution is able to support lightweight semantic search and recommendation in order to fulfill interoperability tasks.**

I. INTRODUCTION

Semantic interoperability is a challenge for the enterprises of any size, but it is really hard for SME and micro enterprise. There are lot of different formats and standards, such as EDI, Core Components, ebXML, which are established but are seldom used by SMEs and micro enterprises mainly due to their complexity. In our FP7 project VENIS[1], we are trying to overcame these limitations by providing a new level of interoperability between large and small enterprises based on Virtual Enterprise paradigm. One of the sub-goals of the VENIS project is achieving semantic interoperability through a lightweight semantic (LWS) approach inspired by Web 2.0 applications.

On the other hand the semantic interoperability can also be addressed by standards from semantic web like RDF(S) or OWL. These semantic technologies bring in interesting possibilities, but they have many unresolved problems which prevent them widely used. These problems include the exponential complexity of inference techniques on rich models; unavailability of exhaustive semantic description of the problem area; and the problem of contradictory knowledge.

In contrast, we can create LWS in the form of tags and annotations. Such LWSs may be more appreciated by the end users, since it is easier to manage and can deal with contradictory knowledge, as well as provide scalable inference based on graph algorithms. To process LWS we can also exploit a wide range of graph mining methods developed in recent years to analyze the semantic networks created by the users, as well as a number of interoperability activities or semi-automatic IE (Information Extraction). Main difference between ontologies and LWS is that ontologies uses the top=>down approach and LWSs bottom=>up approach.

[1] http://www.venis-project.eu/

II. LIGHTWEIGHT SEMANTIC APPROACH

Users or enterprises in VENIS concept will be able to attach tags or annotations to their data and documents either manually or automatically via annotation API using information extraction methods. Moreover, semantics can be further enriched when integrated with legacy system such as databases, CMS, ERP, etc. VENIS will support mainly manual creation of tags or annotations, but we will publish REST/SOAP services to allow automatic tagging and automatic access to tags. We will also deliver an automatic solution which can learn from user annotations and user interaction. Tags or annotations can either be hidden or shared to interoperating parties. Large enterprises can have rich semantics inside their repositories, but share with others only a few valuable meta-data that can be perceived, updated and reused by the interoperating SMEs. Semantic tags and annotations will be stored and used in the form of semantic networks of interconnected repository items, entities, tags, annotations, users and events, which can be used for recommendation of resources, information, knowledge and process activities. Thus the user driven LWS will serve as a basis for a simplified semantic interoperability between LEs and small enterprises.

Semantics can be improved also by simple user actions, like merging entities or annotating, deleting entities and we assume that users will be willing to do it if they see benefit in the form of the immediate improvement of recommendation and search functionality. Semantics in VENIS uses free schema based on tags and annotations with user defined types in a similar way as on social networks or Web 2.0 sites. Because of free schema semantic, no interoperability task is done automatically, but it is supported by user, which can fix, resolve inconsistencies and thus system will adapt and learn.

Free schema in VENIS can be based not only on tags and annotations, but also on semantic trees and semantic networks of annotations or events. Both, automatic and user driven annotation with predefined types such as people, organizations, locations, amounts, dates, contact data, etc. will be supported. User defined types based on gazetteers like products, services, projects, etc.

III. LARGE ENTERPRISE TO SMALL ENTERPRISE INTEROPERABILITY AND VICE VERSA

Large enterprises usually handle business document in more formalized way. An ideal case of interoperability between LEs is the exchange of business data via

978-1-4799-4615-0/14/$31.00 ©2014 IEEE

standardized documents (e.g. CoreComponents, EDI, ebXML). However, in the case of SMEs, the situation is very different. SMEs often handle business documents as human readable documents (e.g., doc, PDF).

When LE sends data/business object/business document to SME there are two possibilities:

1. LE can generate human readable business documents by its BPM/ERP/CRM system and sends it via e-mail to SMEs (Fig. 1). In this case, the Data Services of the VSI installed on the SME intercepts it and extracts lightweight semantics both from e-mail text and attachments. Additional business data can be annotated by the SME user, on the received document, which will help the SME to process the document. Business activities such as search for additional related information, processing it into the legacy system or passing it on with semantics to other party will be supported. Fig. 1 describe scenario, where the data is send by legacy system based on a business rule, no request for data needs to be formulated by a SME.
2. The LE is not able to generate human readable documents and send it by e-mail (Fig. 2). In this case the VSI installed on the LE side will be connected via adapters to the legacy applications like BPM, ERP, CRM, Relational Database or Document Repository via one of the supported adapter types (VENIS supports repository adapter FTP/WebDav/BSCW/MS SharePoint; web service adapter; SQL database). By monitoring the legacy application, the adapter will detect that business object/document needs to be sent to the SME and extract data from the legacy system using DB queries, simple file copy or even SQL or EDI template on order to create a simple human readable document. If the legacy application has an EDI/ebXML interface, the adapter will monitor the production of the EDI/ebXML document in order to understand when the data must be sent to the SME.

This is the same work done today by the LE employee, when they take the information from the legacy application and compose an e-mail for the SME employee describing a business interaction (orders, offers, invoices, etc.).

By using the VENIS Services, the Adapter will store this document into the VCR and the VCR will signal the upload of a new file to the Inter-layer event log. Then, the Process Services are triggered by this VCR event and will start the procedure (previous modeled) in order to send an e-mail to the SME employee containing the token for this document. The SME will receive the token and will have access to the business object/document with its lightweight semantics, in order to process it as in the first case.

Fig. 2 depicts scenario, where SME retrieves data from LE's legacy system. This will be done in the following way:

- LE configures a connector to its legacy system, specifying the data view the LE wants to share with a particular SME
- this data view will be periodically refreshed and

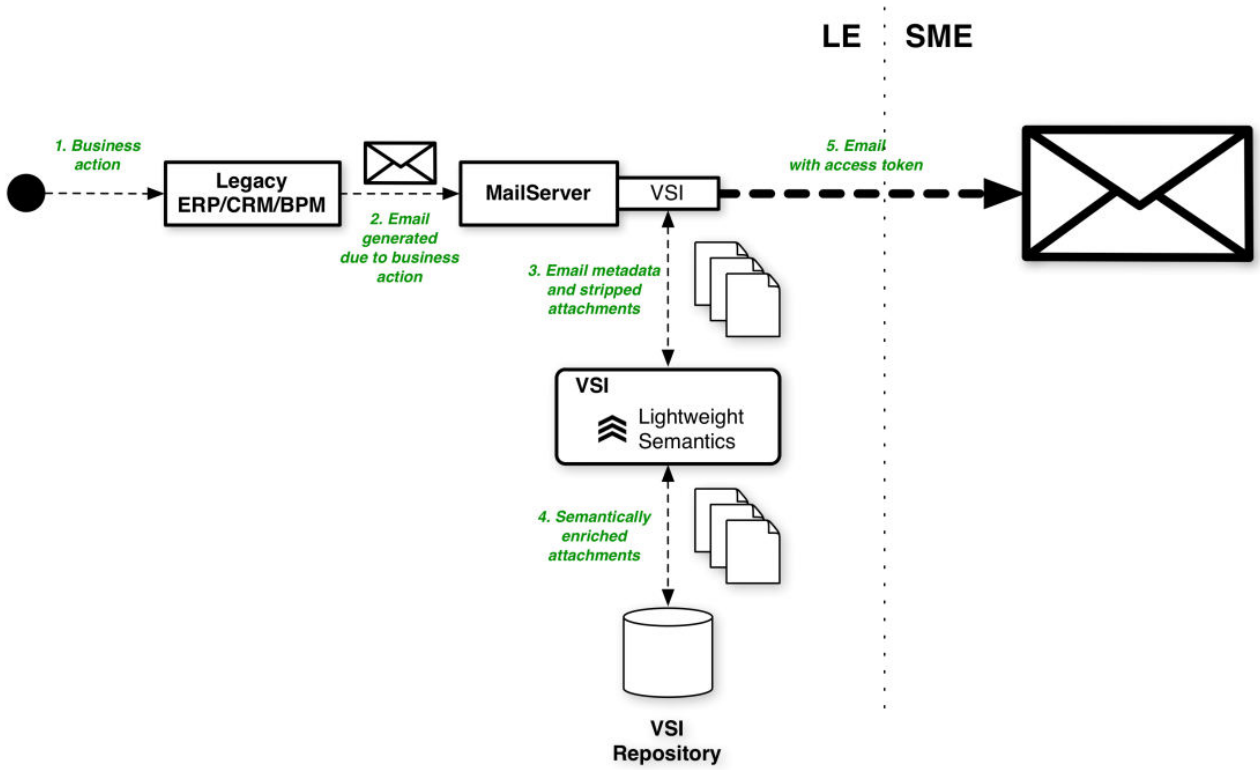

Figure 1 LE generates a human readable file by legacy system and sent to SME

materialized in form of file(s) containing structured data

- when SME needs to access the data of LE's legacy system, SME users can access files containing materialization of those data.

In this way, SME users access LE's structured data in the same way they access other shared files.

When the SME sends a business document to the LE, the SME user fills in a human readable document and sends it to the LE, by e-mail; then there are two possibilities:

1. The LE has a human operator, who handles requests from SMEs by a manual insert of data into the legacy application. The requests will arrive to the LE via e-mail, with the business document stored in the VCR. The LE operator opens the document that has arrived from the SME and the VSI provides lightweight semantics and recommendations, which will help the operator to fill in the manual input of data towards the legacy application.
2. If the LE does not have a human operator who handles requests from SMEs, the adapter of the legacy application will be used to flow data inside the legacy application. When the adapter is installed, a template of the data needed for executing the legacy application inset will be stored in the VCR, to be filled in by the SME when the request is sent to the LE. The VENIS lightweight semantics helps the SME user in the completion of the template and then the completed form is sent to the LE. On the LE side, the adapter takes the completed form from the VCR and transforms it into a formal EDI/ebXML document to be processed by the existing LE legacy system. If the legacy application has no EDI/ ebXML interface, but it has a VENIS adapter, the adapter can be used to insert the data directly to the underlying structured data management system e.g. for relational database adapter exploit SQL updates.

The transformation in a formal EDI/ebXML document can be also done on the SME side, as shown in the Fig. 3; in this case, the adapter handles authorization via VENIS tokens and passes the formal document to the LE legacy application.

IV. Proof of Concept Implementation of Lightweight Semantics

In this chapter we summarize proof of concept implementation of LWS, which was reported in [4].

1) Rule based Named Entity Recognition

IE techniques [1] usually focus on 5 main tasks of IE as defined by the Message Understanding Conferences (MUC):

1. *Named entity recognition* (NE) - Finds and classifies the names, places, etc.;

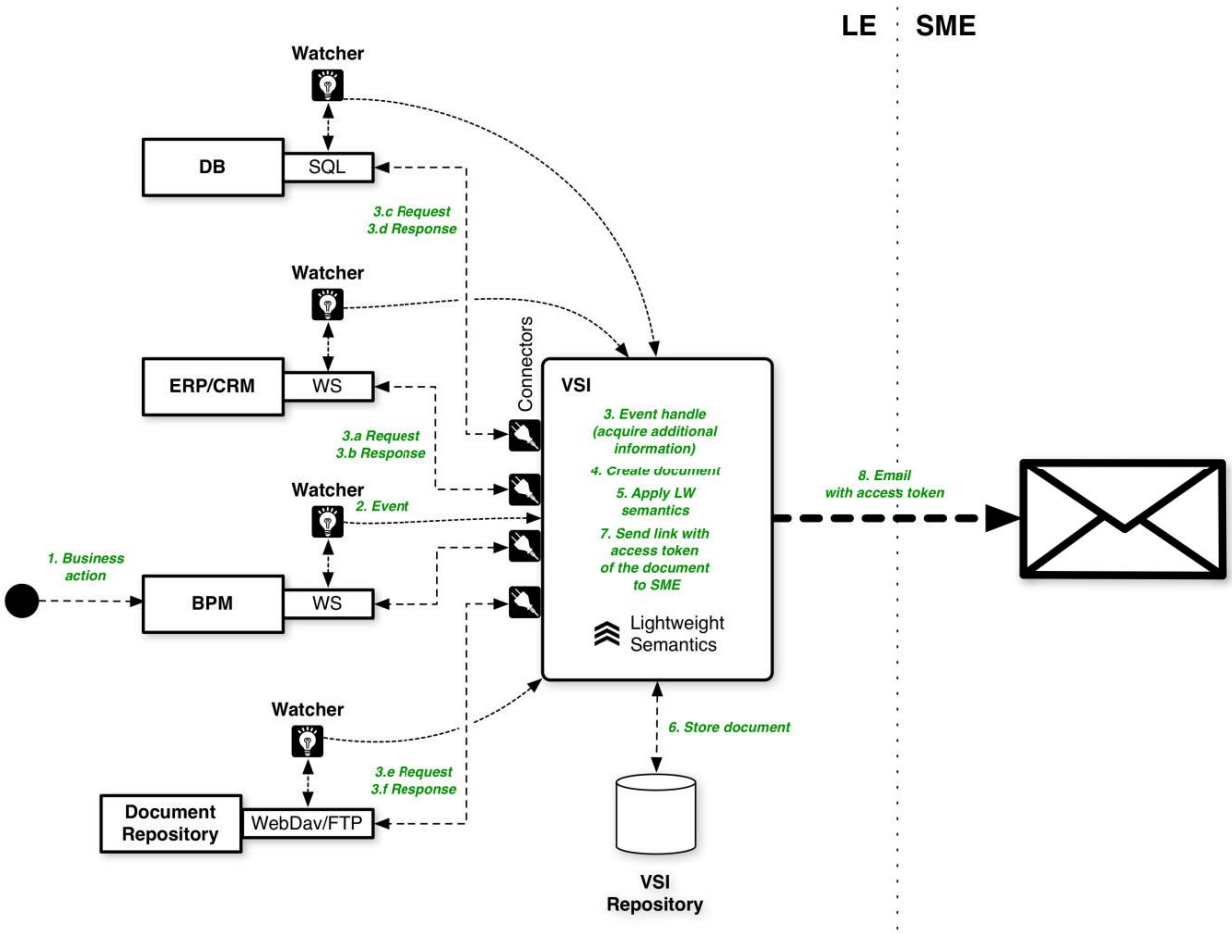

Figure 2 LE's legacy systems are connected to VSI and data are sent to SME

2. *Coreference resolution* (CO) – Finds aliases and pronouns referencing the same entity and discovers the identity relations between entities;

3. *Template element construction* (TE) – Finds properties or attributes of entities and adds descriptive information to NE results (using CO);

4. *Template relation construction* (TR) – Finds relations between NE entities;

5. *Scenario template production* (ST) – Finds events involving entities and fits TE and TR results into specified event scenarios.

For IE we use Ontea IE techniques [3] developed in the FP6 Commius[2] project, but any other IE can be used. Detailed examples can be found in our previous work [2, 3].

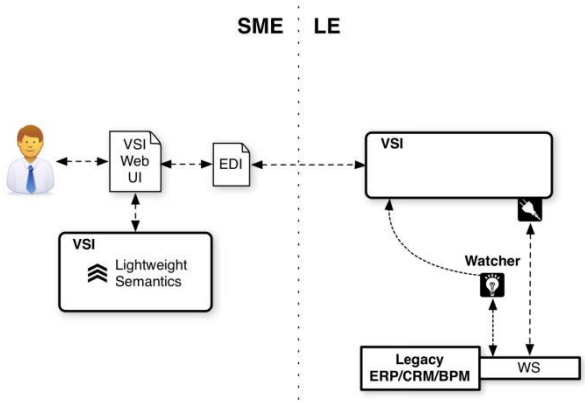

Figure 3 SME sends data to LE and VSI Adapter helps to flow data

2) Semantic Trees and Graphs

Annotations (key-value pairs) representing entities form trees and graphs/networks which can be built from any text collection. A document is then represented by a document node, its paragraph and sentences nodes, and such documents are interconnected by the nodes representing the entities present in multiple documents. Entities found in multiple documents occur only once in the graph (they are unique) and they connect by edges to their respective sentences, paragraphs and documents. In the future we would also like to add activity graph to this structure, e.g. when and by whom the document was updated, downloaded, sent, re-shared. The underlying idea is that any other internal or external data can be included, and the semantic inference and relation discovery will work on such networks. For the relation discovery we are using spreading activation algorithm described in [2], with one example of enterprise search also described in chapter 4 below.

3) Email parsing

The implemented prototype is able to access several kinds of remote and local email collections. It uses Java Mail API to access email collections by IMAP, SMTP and POP3 protocols. During the FP6 Commius project we have developed an email processing framework ACoMA [5]. As this framework is already described, we will not go in to the details. Email text part is parsed with the help of Apache Tika[3], which can recognize several file formats and parse file metadata as well as its textual content. All extracted email parts are then processed by the Ontea tool. Based on the extracted entities Ontea tool adds additional LWS information into the message. After the metadata is enhanced by Ontea, the email represented by the parsed parts is ready to be indexed.

4) Indexing

We are using an Apache Lucene[4] based indexer in our prototype. These indexed documents are later sent to the Apache Solr[5] interface. We are indexing not only email textual parts, but also their metadata and documents extracted from the attachments and connected ERP, CRM, CMS, etc. systems.

V. USE CASE AND PROTOTYPE

The Intersoft use case faces the supply chain aspects related to the software development and maintenance.

In this frame, flexible "Search and Manage" capabilities are needed to reach the expected level of collaboration in the distributed design & development of software and during the post-sales customer care. When a customer needs to implement a complex software project and the LE provider does not have enough or required development resources available, it will look for a suitable set of suppliers and involve them in the collaboration so as to complete the project for the customer. On the Fig. 4 is the overall description of the Intersoft scenario, with data and process flows and an actor description as well.

In the folowing section we discuss two innovative semantic interoperability features in VENIS: (1) Enterprise Search; (2) search and recommendation for relevant semantics in concrete interoperability tasks.

In Fig. 5 we can see an early implementation of search functionality. There is a search field on the top of the screen, which can be used to search for documents or emails using full-text search. Search results can be filtered by clicking on facets on the left. Several facets are predefined like msg_part_type, msg_from_name, msg_to_name, msg_cc_name, msg_bcc_name and filename. The msg_part_type can be either message or attachment.

Full-text search is integrated with an entity search, where the related entities are displayed by clicking on Enty Srch link.

After clicking on a document listed in the search results, the user can access the document or do any other operation with it like share it, add semantic tags and/or share them or even invoke the user interface that exploits the LWSs for system-to-system interoperability with the user in the loop.

The email interface (Fig. 6) will show email semantics, i.e. the detected entities and recommended entities for the email with the possibility to click on a desired entity (tag, person, invoice ID, customer or product) and perform entity search (as seen in Fig. 5 in the two front windows).

The entity search and the recommendation have features similar to those used in the Email Social Network Search developed in the Commius project, which is being further extended in VENIS [2]. In the Fig. 5 the front most shows the skills detected in the development requirement email; the one behind it the list of companies relevant for

[2] http://cordis.europa.eu/projects/rcn/85234_en.html
[3] https://tika.apache.org/
[4] http://lucene.apache.org/core/
[5] https://lucene.apache.org/solr/

the skills. The idea is to return the relevant entities for one or more selected elements (i.e. context) and thus to deliver needed information for the interoperability task.

As we can see in Fig. 6, the email user interface should show the email with the detected context displayed in the email text. The context is also displayed as a set of items which can be modified. Based on the context, VENIS will deliver relevant recommendations. Recommendations can contain information, inferred knowledge or process and activity information relevant for the email and the business activity covered by the email.

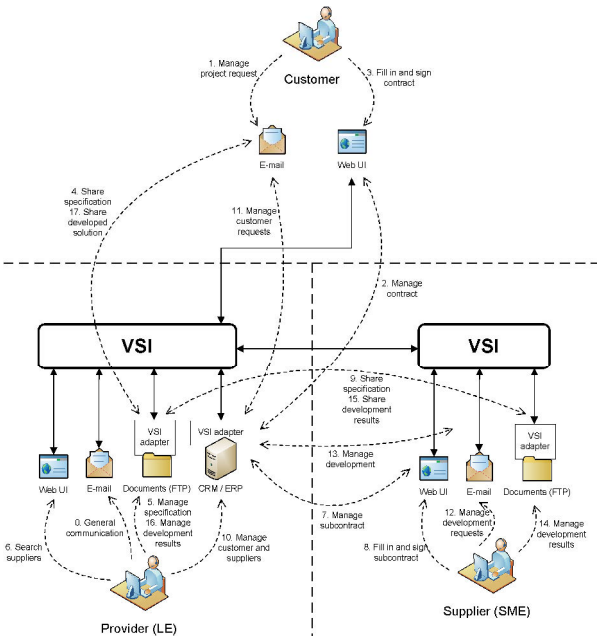

Figure 4 Data flow, actors and services in Intersoft scenario

The recommendation will be provided by the same engine as the already implemented entity search (Fig. 5 front screens), i.e. it will reuse the underlying LWSs in the form of a graph, but instead of user query, the context is detected directly from the email and its attachments. Based on this, other relevant entities (next activity, related people, products, services, or skills) will be recommended in a way similar to the multiple entity search showed in Fig. 5. Additional templates for recommendation can be created and/or adjusted even at a later stage. For example, if a request for outsourcing developers is detected, the system should restrict its recommendation only to people and skills, and omit irrelevant data.

Detected or recommended entities can be selected (as seen in Fig. 6) and further processed. In this way, a system-to-system interoperability activity can be initiated by the user. For example, contact information can be stored to a database or an invoice can be submitted to and processed by a legacy system.

Users can also exploit recommendations for further entity searches. For example, if a user wants to see the skills or contact details of developer Stefan, he/she can click on the recommended item (Person Stefan in this case) and the relevant information will be displayed.

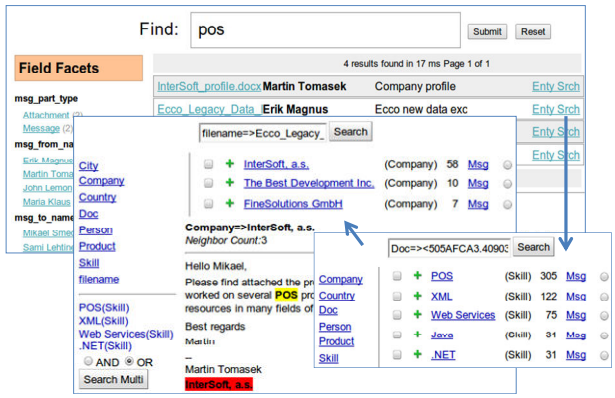

Figure 5 Prototype enterprise search user interface

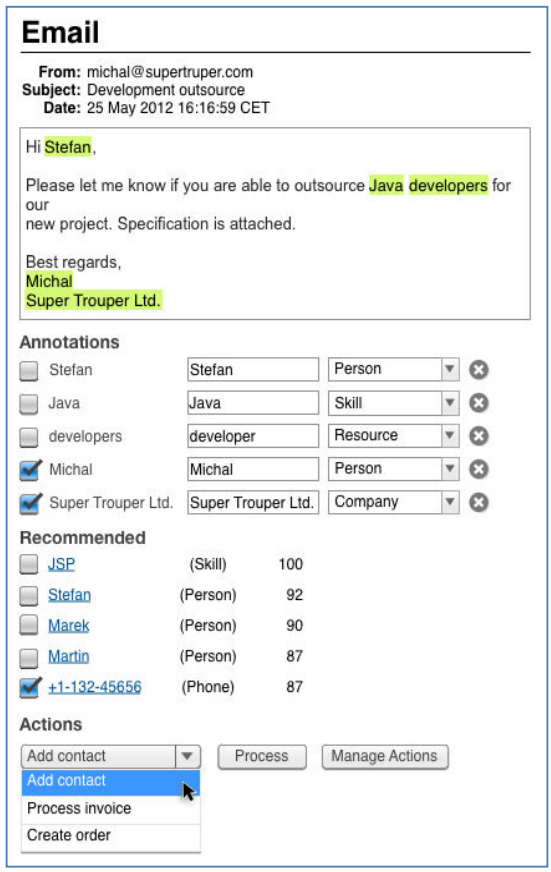

Figure 6 User interface with email and context recommendations

VI. Conclusion and Future Work

In this paper we summarize our work in progress on semantic interoperability and semantic search using LWS networks/graphs. We provide a proof of concept implementation of the semantic enterprise search, which exploits the LWS graph and can help with interoperability tasks communicated by email and shared documents. In our architecture we also provide a way how-to reuse the LWS networks for system-to-system interoperability, but always with users in the loop. Our next work will be focused on providing better tools for users to interact with LWSs, enabling features such as sharing of semantics, adjusting, deleting or annotating tasks.

ACKNOWLEDGMENT

This work is supported by projects VENIS FP7-284984, CLAN APVV-0809-11 and VEGA 2/0184/10.

REFERENCES

[1] Cunningham H. (2006), *Information Extraction, Automatic*. In: Encyclopedia of Language & Linguistics, Second Edition, volume 5, pp. 665-677. Oxford: Elsevier

[2] Laclavík M, Ciglan M, Dlugolinský Š, Šeleng M, Hluchý L. *Emails as Graph: Relation Discovery in Email Archive*. In Email2012 workshop, WWW 2012, April 16–20, 2012, Lyon, France, pages 841-846, 2012

[3] Laclavik M, Dlugolinsky S, Seleng M, Kvassay M, Gatial E, Balogh Z, Hluchy L. *Email Analysis and Information Extraction for Enterprise Benefit*. In Computing and informatics, 2011, vol. 30, no. 1, p. 57-87. ISSN 1335-9150

[4] Laclavík M, Dlugolinský Š, Šeleng M, Ciglan M, Tomašek M, Kvassay M, Hluchý L. *Lightweight semantic approach for enterprise search and interoperability*. In CEUR Workshop Proceedings : INVIT 2012. - CEUR, 2012, p. 35-42. ISSN 1613-0073

[5] Laclavík M, Šeleng M, Dlugolinský Š, Gatial E, Hluchý L. *Tools for email based recommendation in enterprise*. In ENTERprise Information Systems : CENTERIS 2010. Eds J.E.Q. Varajão, M.M. Cruz-Cunha, G.D. Putnik, A. Trigo. - Berlin: Springer, 2010, part I, p. 209-218. ISBN 978-3-642-16401-9. ISSN 1865-0929

[6] Marin C. A, Carpenter M, Wajid U, Mehandjiev N. Devolved Ontology in Practice for a Seamless Semantic Alignment within Dynamic Collaboration Networks of SMEs. In Computing and Informatics, Vol. 30, 2011, No. 1

[7] Dlugolinský Š, Ciglan M, Laclavík M. Evaluation of named entity recognition tools on microposts. In INES 2013, 17th IEEE International Conference on Intelligent Engineering Systems 2013. - Budapest : IEEE Industrial Electronic Society, 2013, p. 197-202. ISBN 978-1-4799-0830-1

[8] Ciglan M, Laclavík M, Nørvåg K. On community detection in real-world networks and the importance of degree assortativity. In KDD´13 Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Eds. Grossman, R.L., Uthurusamy, R., Dhillon, I., Koren, Y. - New York : ACM, 2013, p. 1007-1015. ISBN 978-1-4503-2174-7

[9] Dlugolinský Š, Nguyen T G, Laclavík M, Šeleng M. Character gazetteer for named entity recognition with linear matching complexity. In Proceedings of the 2013 World Congress on Information and Communication Technologies : WICT 2013. Eds. Ngo, L.T. et al. - IEEE Systems Man and Cybernetics Society, Spain Chapter, 2013, p. 364-368. ISBN 978-1-4799-3230-6

[10] Obrst L. Ontologies for semantically interoperable systems. In Proceedings of the twelfth international conference on Information and knowledge management (CIKM '03). ACM, New York, NY, USA, 366-369. DOI=10.1145/956863.956932 http://doi.acm.org/10.1145/956863.956932

[11] Enterprise Interoperability Research Roadmap, ftp://ftp.cordis.europa.eu/pub/fp7/ict/docs/enet/ei-research-roadmap-v5-final_en.pdf