Grant Agreement Number **ECP 2006 DILI 510049**

ENRICH

Vicodi Ontologies implementation report

Deliverable number	***D-6.4***
Dissemination level	*Public*
Delivery date	*15. 12. 2009*
Status	*Final*
Author(s)	*SYSTRAN S.A., France*

e*Content*plus

This project is funded under the *e*Content*plus* programme, a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

Document Version Control

Version	Date	Change Made (and if appropriate reason for change)	Name and Institution of Commentator(s) or Author(s)
The first draft	November 2009		SYSTRAN S.A., France
Final	15. 12. 2009	Formatting	CCP

Document Review

Reviewer	Institution	Date and result of Review
Jakub Heller	CCP	7.12.2009, Accepted
Zdeněk Uhlíř	NKP	14. 12. 2009, Approved for submission

Document Signature/Approval:

Before the table of contents each document is to contain an approval signoff form.

Approved By (signature)	Date

Accepted by at European Commission (signature)	Date

Structure Summary

The Work package 6 aims to the integration of a tailored multilingual module via a user friendly sophisticated access.

Based on SYSTRAN's machine translation technology, this module will provide also terminology extraction and machine translation customization tools for the construction and retrieval of personalised metadata within the aim to create new multilingual digital documents and multilingual ontologies in Czech, Polish, Spanish, Portuguese, German, Italian, English, French, Danish, Russian and Serbo-Croatian.

This reports aims to present the integration of the Vicodi ontology developed during a 24-month IST project Visual Contextualization of Digital Data http://www.vicodi.org applied to the application http://www.eurohistory.net .

For Enrich translation customization purposes, SYSTRAN has extracted the Vicodi ontology distances to build a Enrich-specific dictionary including all the person names and events hierarchised via ontologies.

Content

I. Introduction

SYSTRAN has extracted the Vicodi Ontology instances referring to historical persons and events of the European history from the RDF-based storage. At the same time NKP provided to SYSTRAN an ontology for Medieval names and locations that were converted to Medieval Do Not Translate Dictionaries composed by Name Entities. The Medieval dictionary as well as the Vicodi ontology dictionary has been associated to the translation profile called via the web service for machine translation customization purposes and name entities recognition.

Following the lines of the Description of Work, ENRICH, Annex 1, the main objectives of the Multilingual and user friendly sophisticated access are described in the Work package 6.

Work package Description

Work package number :	**6**	**Start date:**	**3**	**End date:**	**24**
Work package title:	**Multilingual and user friendly sophisticated access**				

Objectives

This work package aims to the integration of a multilingual module via a user friendly sophisticated access: multilingual search application, multilingual forums, and multilingual ontology editor.
Based on SYSTRAN's machine translation technology, this module will provide also terminology extraction and machine translation customization tools for the construction and retrieval of personalised metadata within the aim to create new multilingual digital documents and multilingual ontologies in Czech, Polish, Spanish, Portuguese, German, Italian, English, French, Danish, Hungarian, Russian and Serbo-Croatian.

Description of work

Work package leader: SYS
Task 6.1 Multilingual access development (m0-m12)
Task leader: SYS
Task participants: NKP, AIP

The project will provide two types of multilingual access:

- via the API integration in the data retrieval interface associated or independent of a multilingual search.
- a dedicated translation interface where ENRICH expert users can fine-tune dynamically the machine translation tools thanks to adapted linguistic tools for terminology extraction and translation post-editing and customization. The parameters and resources constructed will be automatically taken into account by the API in the access presented above

Task 6.2 Translation Stylesheet design and use (m3-m24)
Task leader: SYS
Task participants: NKP, AIP, CCP, NFC, NLF, IMI, ULV, SAM, CSH, DSP, NLI, BNE, BUTE, ULW
Activities:
• analysis of heterogeneity of metadata regarding machine translation.
• implementation of STS exploiting metadata information.

• cross-language validation of STS, optimization of translation parameters.
As far as the Metadata translation module implementation is concerned SYSTRAN will provide a fully customized Translation Stylesheet.
SYSTRAN Translation Stylesheets (STS) use XSLT to drive and control the machine translation of XML documents (native XML document formats or XML representations — such as XLIFF — of other kinds of document formats). STS will provide a simple way to indicate which part of the document text is to be translated, and will enable the fine-tuning of translation, especially by using the structure of the document to help disambiguate natural language semantics and determine proper context. Thanks to STS machine translation is considered as part of the authoring and publishing process: source documents can be annotated with natural language mark-up produced by the author, a mark-up which will be processed by STS to improve the quality of translation, the gateway to the automatic publishing of a multilingual website from a monolingual (annotated) source. The mechanism is implemented through XSLT extension functions for consulting and for setting linguistics options in the translation engine. SYSTRAN will deliver this xslt file in order to fine-tune the system according to the ENRICH xml data elements.

Task 6.3 VICODI implementation (m6-m24)
Task leader: SYS
Task participants: NKP, AIP,

• definition and homogenization of initial ontology applicable for this project
• specification of user-friendly web-interface for visualization of multilingual ontology - special interface for modification
• implementation of the web-interface
Based on previous experience in the visualization and contextualization of digital content (IST project VICODI) SYSTRAN technology has been implemented for the construction of multilingual ontologies. The Research Center for Information Technologies (FZI) constructed multilingual ontologies available under GNU Free Documentation License (FDL) thanks to the EU-funded IST project Vicodi http://www.vicodi.org/). Enrich will implement and use VICODI ontologies for the contextualization of the digital content.

(Inter-) Dependencies, milestones[1] and expected result

Based on the ENRICH Corpus Analysis (Month 6) SYSTRAN will build ENRICH Translation Stylesheet.
After a quality assessment procedure and based on the evaluation results (Month 20) SYSTRAN will proceed to the finalisation of ENRICH Translation Stylesheet (Month 24).
The WP depends mainly on the results of WP3, but is also interrelated with WP4.
The feedback is necessary from WP7.

Deliverables

D 6.1 ENRICH Corpus Analysis report (Month 6), responsible partner: SYS
D 6.2 Personalised Translation Interface delivery report (Month 12), responsible partner: SYS
D 6.3 ENRICH Translation Stylesheet delivery report (Month 24), responsible partner: SYS
D 6.4 Vicodi Ontologies implementation report (Month 24), responsible partner: SYS

II. Vicodi Ontology Presentation

The Management System of the Knowledge space (MSKS) ontology, resource management and search modules were implemented to allow storage and retrieval of the textual VICODI resources and ontology instances. The main Ontology development activities were carried out through the ontology editor, which is a stand alone JAVA GUI application, and may be started using Java Web Start or installed locally on client machines. The MSKS (Management System of the Knowledge space), provided an API to open-source KAON framework to work with VICODI ontology stored within the PostgreSQL.

Development of the VICODI Ontology

Examination and rejection of existing thesauri and glossaries as potential sources of relevant information for VICODI ontology was done. For the purposes of ontology development, competency questions were listed. The creation of initial three mini-worlds was set-out as objective, which was implemented by content partners.

The Ontology design was developed by experimental processing of various instances, concepts/sub-concepts and property relations examples from the miniworlds

The ontology editor has been in use to work on the ontology. It has been customized for VICODI and more customizations are planned. It is based on the general OIModeler of the KAON ontology management framework. The historians use the production ontology, while the developers use copies of the production ontology for testing.

RDF-based storage

After evaluation the open-source KAON framework and found it suitable for the purposes of the RDF-based storage, as it provided many of the required features like RDF export, multilinguality and scalability. We added missing PostgreSQL database support for KAON, as this was the database of choice in the VICODI project.

New KAON query language was implemented to meet the requirement "RDF query language". We installed the KAON framework together with PostgreSQL and a JBoss server. The VICODI ontology is edited online, using the central installed ontology server.

Repositories

The object-relational mapping part of the Expresso framework (www.jcorporate.com) as the basis of the repositories solution was selected. The decision to store the historical articles in (X)HTML files was made. Also we made the decision to provide a central repository with several logical repositories. Logical repositories can be handled separately by the MSKS Server application.

We defined the logical database schema, which is capable to store the required content of the repository (resources, their metadata and content, their annotations and context).

VICODI ontology integration to Enrich

For name entities recognition and tagging as well as machine translation customization, SYSTRAN proceeded to the extraction of the Vicodi ontology instances for name persons and historical events in order to create a Enrich-specific name entities dictionary. With that occasion Medieval ontology instances provide by NKP had equally been integrated to a separated Enrich-specific name entities dictionary.

SYSTRAN had extracted from a Vicodi ontology metadata description like

```
</ontology:Person>
<ontology:Person rdf:about="&ontology;AEthelbert-II-of-Kent"
        rdfs:label="AEthelbert-II-of-Kent">
       <rdfs:comment>Author=GP. Mass Upload sheet Persons.xls. Roles=King Links=Kent
Websites= Trans= Remarks=AEthelbert II King of Kent King of Kent</rdfs:comment>
       <ontology:hasCategory rdf:resource="&ontology;Politics"/>
       <ontology:related rdf:resource="&ontology;Politics"/>
       <ontology:exists rdf:resource="&ontology;i-1090590633795-1600645804"/>
       <ontology:related rdf:resource="&ontology;i-1090590633795-1600645804"/>
</ontology:Person>
```

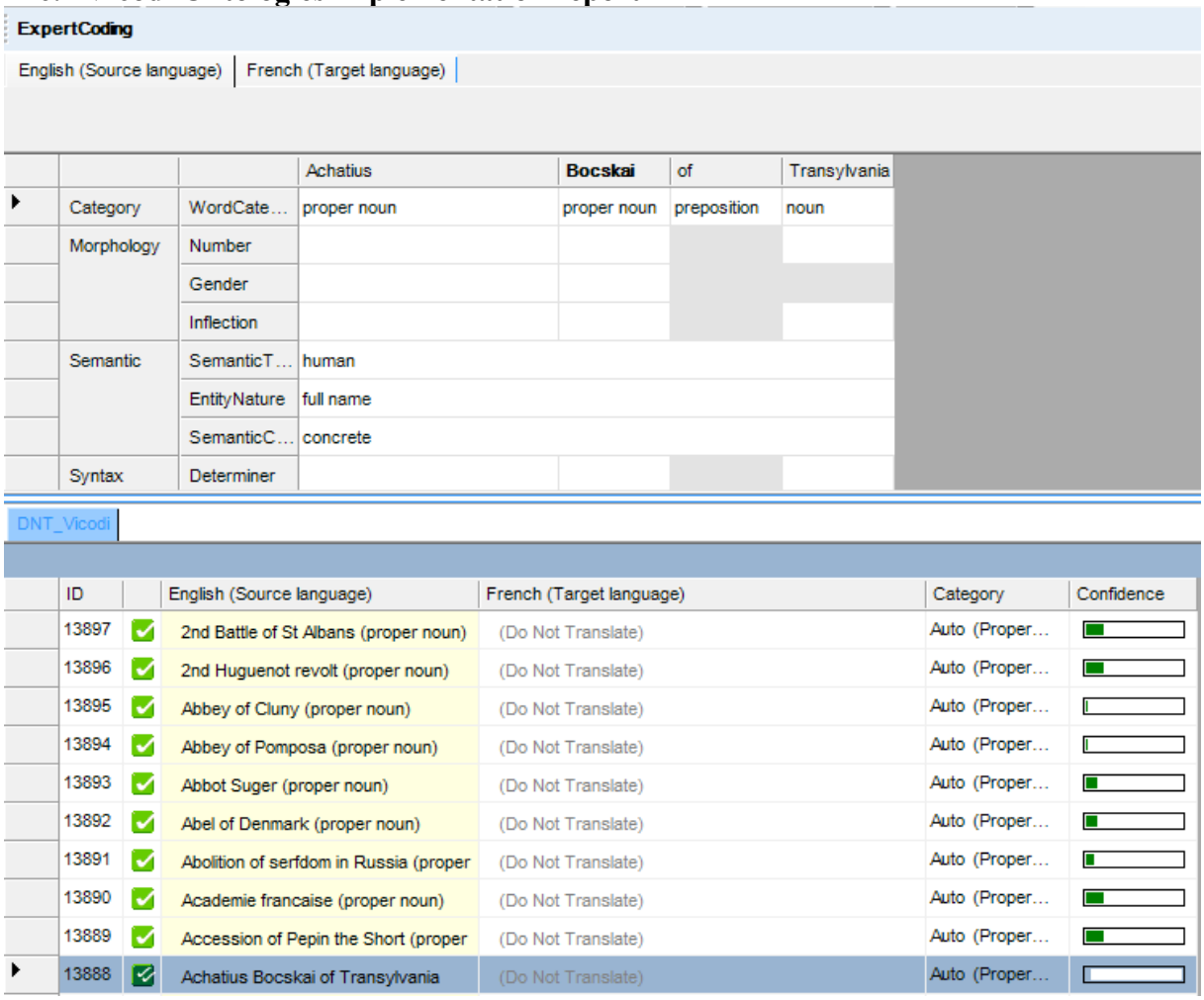
the rdfs label item. A Do Not Translate Proper Noun dictionary had been created composed by 7.043 entries.

At the same time SYSTRAN validated the Enrich-specific dictionaries for translation customization and name entities recognition extracted from Manuscriptorium ontology instances

6.759 entries for Medieval persons
1.060 entries for Medieval locations
2.016 entries for Medieval sequences
1.741 entries for Medieval general instances

Enrich-specific Name Entities Dictionary

The list of the entries had been coded with the necessary linguistic information so as the system applies into context the given terms followed by the accurate context analysis.
Those dictionaries can be maintained by the content partners and authorised users in order to add further proper nouns entries. A specific linguistic tool SYSTRAN Dictionary Manager accessible to the Enrich consortium had been utilized for the linguistic coding. Please find below a sample of the created dictionary as displayed via the Graphical User Interface of SYSTRAN Dictionary Manager.



Regarding open standards the dictionary can be exported to an OLIFF-compliant xml structure for integration to third applications

```
</mono>
  </entry>
  <entry type="dnt" ID="-13888" pos="proper_noun" status="validated">
   <mono lng="EN" cat="NP" confidence="147">
    <syn>
     <lemma cat="NP">
      <word type="guessed" code="1">Achatius</word>
     </lemma>
     <lemma cat="NP" headword="head">
      <word type="guessed" code="1">Bocskai</word>
     </lemma>
    </syn>
    <lemma cat="PREP">
     <word>of</word>
    </lemma>
    <lemma cat="N">
     <word type="guessed" code="1">Transylvania</word>
    </lemma>
    <syntax>
     <type>HU,CON</type>
    </syntax>
    <source>Achatius Bocskai of Transylvania (proper noun)</source>
    <natural_lemma>Achatius Bocskai of Transylvania</natural_lemma>
    <headword>Bocskai</headword>
   </mono>
   <mono lng="FR" cat="NP" confidence="360">
    <lemma cat="NP" headword="head">
     <word>Achatius Bocskai of Transylvania</word>
    </lemma>
    <source>Achatius Bocskai of Transylvania (proper noun)</source>
    <natural_lemma>Achatius Bocskai of Transylvania</natural_lemma>
    <headword>Achatius Bocskai of Transylvania</headword>
    </mono>
   <semcat>HUMANS</semcat>
```

```
</entry>
```

Name Entities Recognition

Using the Translate function of the SOAP API several options can be added to control entity recognition and fields association like variants.

The entities recognized in the request source text are annotated in the response source or translated text. (to get annotated source text in SOAP response, the option "NOTRAN" needs to be used, as defined below).

The option DNT_ENTITY_XXX_MARKER defines the markup for the XXX entity, with XXX being an entity name defined in the entity recognition rules.

PERSON and EVENT are the two entities defined for Enrich. DNT_ENTITY_PERSON defines the markup applied for historical persons, and DNT_ENTITY_EVENT defines the marker applied for historical events.

In the option value, the special character "#" will be substituted by the ID of the recognized entity as defined in the entity dictionary.

Option Name	Description
NOTRAN	Defines request output – when returned, source will include spell check and entity annotation, translation will include entity annotation 0/1/2 – Default 0 0: only translation is returned 1: only source is returned 2: both source and translation are returned

III Conclusions

The aim of the Vicodi Ontology instances extraction and the enrichment effectuated adding the Medieval Ontology Instances was to customize the machine translation result by rendering more accurate the source analysis of the metadata description so as the translation is of the best possible quality. In addition to this name entities recognition for information retrieval purposes between manuscripts can be achieved via the tags that can be attributed if needed next to each name entity.The necessary linguistic tools and applications are at the disposal of the Enrich Consortium in order to maintain or enrich those dictionaries following the same methodology. The dictionaries are called via the translation profile activated via the call of the web service for the multilingual search and the metadata description translation.