Combining WordNet and Crosslingual multi-terminology health portal to access health information

S.J. Darmoni [a] , *J. Grosjean* [a], *T. Merabti* [a], *N. Griffon* [a], *B. Dahamna* [a], *D. Dutoit* [a,b]

[a] *CISMeF & TIBS LITIS EA 4108, Rouen University Hospital*, Cour Leschevin, Porte 21, 3ème étage, 1 rue de Germont, F76031 Rouen Cedex, Franc

[b] SensGates, Vigneux, France

Abstract. (à revoir à la fin)

Background:

WordNet® is a large multilingual lexical database in various languages. The European Health Terminology/Ontology Portal (EHTOP) is a repository dedicated to European health professionals and students. Currently, it provides access to thirty two health terminologies and ontology available mainly in French or in English, but also German, Italian and Dutch. EHTOP can be used by humans and by computers via Web services. The main objective of EHTOP is to provide an access to terminologies and ontology, allowing dynamic browsing and navigation.

Methods: To integrate terminologies and ontology into EHTOP, three steps are necessary: (1) designing a meta-model into which each terminology and ontology can be integrated, (2) developing a process to include terminologies into EHTOP, (3) building and integrating existing and new inter & intra-terminology semantic harmonization into EHTOP.

Results: EHTOP is available freely for the MeSH in French (URL: pts.chu-rouen.fr). The access to other terminologies and to other languages is restricted and available only for the scientific community. A total of 32 terminologies are included into EHTOP, with 980,000 concepts, 2,300,000 synonyms, 222,800 definitions and 4,000,000 relations. Twenty one of these terminologies are not included yet in the UMLS among them, some from the World Health Organization. Since January 2011, EHTOP is daily used by CISMeF librarians to index health resources in the CISMeF catalogue in a multi-terminology mode. Currently, 600 unique machines are using the MeSH version of EHTOP, whereas 230 are already registered for its extended version. More recently, the multilingual version of EHTOP is available (URL: http://cispro.chu-rouen.fr/ehtop_site/) and freely provides access to ICD10 in five languages.

Conclusion: Combined with a cross-lingual dictionary, EHTOP is a rich tool, useful for a wide range of applications and users, whatever in education, resources indexing, information retrieval or performing audits in terminology management.

MeSH keywords: Abstracting and indexing; Cataloguing; Controlled Vocabulary; Internet; Database; Dictionaries; Europe; Information Storage and Retrieval; Internet; Subject Headings; Terminology as subject.

Les usages : littérature scientifique
Apprendre l'anglais médical
Serveur Terminologique MeSH préexistant devenu EHTOP a démontré son efficacité à interroger MEDLINE dans la langue maternelle. C'est pourquoi il est autant utilisé ! 600 personnes par jour ouvré avec un profil particulier : librarians, health students and… translators.

Existant : intégration d'Alexandria partially based on Wordnet au sein des applications CISMeF (mettre un transparent)

Introduction

The Internet is currently the major source of scientific and health information and knowledge. If health information and knowledge for health professionals is generally in English, health information for lay people is available in each language. Nonetheless, people around the world is more and more travelling, health information should transcend borders and should become multilingual and based on several health terminologies and ontology. Some institutions are already proving health information in several languages: e.g. MEDLINEplus[1] is providing health information for lay people in English and Spanish, whereas Europe Medicines Agency[2] is providing drug information for health professionals and lay people in each European language.

Health is with law the main scientific fields, where coexist several terminologies and ontology (T/O). For the English languages, over 150 terminologies and classifications are included in

[1] http://www.nlm.nih.gov/medlineplus/

[2] http://www.ema.europa.eu/

the Unified Medical Language System (UMLS) meta-thesaurus developed by the US National Library of Medicine since 1986.

There is an increasing amount of interest today not only in developing and maintaining healthcare T/O but also in making them interoperable within information technology (IT) systems delivering services to applications. Terminology server has been defined as a tool to manage and to give access to various terminologies [3]. Several terminology servers have already been developed, mostly in English. One was recently developed for French terminologies by three partners [4]: the private company Mondeca and two academic medical informatics labs: the LERTIM from Marseilles and the CISMeF team from Rouen.

WordNet® is a large lexical database of English [5], which is also available in various languages, including French for this work: the SenseGates Web site is providing this translation [7]. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept that should be linked to EHTOP. The main relation among words in WordNet is synonymy [5], where in EHTOP the main relation is the hierarchy among descriptors, although synonymy is also an important relation (e.g. CISMeF has manually added 16,000 French synonyms to the MeSH thesaurus in the last 16 years).

Cooperation between the CISMeF team and SenseGates (previously Memodata) began in 2004 with the Vodel project with the same objective "ontological valorization of electronic dictionaries" following in 2007 by the InterSTIS project[3] [4] (Semantic Interoperability between T/O), both funded by the French National Research Agency. Bilingual dictionaries based on WordNet (Alexandria tool) were integrated in various CISMeF tools, in particular the CISMeF quality-controlled health gateway (Catalog and Index of Health Resources in French-Fr) [URL: http://www.cismef.org] [1] and French MeSH Browser [2].

The principal aim of this work was to create a Health Multi-Terminology Cross-Lingual Portal mainly based on European languages (EHTOP) and combining EHTOP to WordNet®. The primary goal of EHTOP was to search concepts among all the health terminologies available in French (or in English and translated in French) included in this portal and to browse it dynamically. The ultimate goal was to use this search: (a) to index resources manually or automatically in the CISMeF quality-controlled health gateway (Catalog and

[3] http://www.interstis.org/about-interstis/

Index of Health Resources in French-Fr) [URL: http://www.cismef.org] [1]; (b) to allow multi-terminology automatic indexing and information retrieval; (c) to evaluate the integrity of terminological data (audit); (d) to provide a new tool to train health students.

Material and methods

A placer éventuellement :

WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms—strings of letters—but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity.

List of terminologies included in EHTOP

Thirty two T/O were included in the CISMeF Information System (n=32), and therefore in EHTOP (see Figure 1). Some T/O are included in the UMLS meta-thesaurus (n=11) but most are not (n=21).

Among these 32 T/0:

- ✓ MeSH thesaurus [9], including the MeSH Supplementary Concepts (MeSH SC), including the translation of 15,300 MeSH SC in French and the add-on of over 16,000 synonyms to MeSH terms; the MeSH thesaurus is the pivotal thesaurus used to index health resources in the CISMeF catalogue. During the first 10 years of existence, CISMeF used only one thesaurus to index Internet resources as only MeSH is used in the MEDLINE bibliographic database. Because Internet resources are more diverse than scientific articles, the CISMeF team has undergone a major strategic shift in 2005: switching from a mono-terminological world to a multi-terminology universe for the overall CISMeF information system, which includes multi-terminology automatic indexing, multi-terminology information retrieval and integration of several terminologies in the CISMeF information system, then allowing the creation of the EHTOP portal.

- ✓ CISMeF thesaurus, which is an extension to the MeSH thesaurus, including 130 metaterms (super-concepts to unify MeSH terms of the same medical discipline), 300 resource types (adaptation to the Internet of the publication types), over 200

predefined queries and the translation of 12,000 MeSH Scope Notes (8,000 manually and the rest semi-automatically); this semi-automatic translation was performed by Sensegates tools based on WordNet.

- ✓ SNOMED CT & SNOMED International (French version) to describe electronic health records [8],
- ✓ And all T/O developed by the World Health Organization (WHO), in particular ICD10 (International Classification of Diseases, 10th revision) [10].

Some other terminologies and ontology will be integrated in the coming months, in particular US National Cancer Institute Terminology. Most of these T/O (those freely available), in particular all the MeSH add-ons were then included in Sensegates tools to improve its dictionaries with health and scientific terns and its definitions.

Integration of the terminologies

To integrate the terminologies in the CISMeF Information System (Oracle 11.1g database), three steps are necessary:

- to design a terminology generic model into which each terminology model can be integrated,
- to design a process capable of integrating terminologies into the EHTOP,
- to build and integrate intra & inter-terminology semantic harmonization into EHTOP. Two inter-terminology mappings were performed: one based on UMLS concepts (and one based on NLP tools developed by the CISMeF team.

A generic model was designed for the database in order to fit all the terminologies into one global structure (see Figure 2): this database is the CISMeF Information System. Then, a model of each terminology was designed as a specialization of the meta-model. Many conceptual and technical issues have been encountered, especially in the model creation for several terminologies (MedDRA model, FMA ontology to terminology). It was necessary to understand the whole structure and the functional purpose of each terminology to propose a good representation for human. Another problem with terminologies is the space complexity when data is very large (e.g. SNOMED international with more than 80,000 terms and 62,000 relations). We had to adapt our tools to allow integration in short time while keeping a control on data. For every new terminology integrated in the CISMeF Information System, we learnt more and more about structure and data to be able to integrate all kinds of terminology in our system.

The CISMeF Information System

This system was established around the "Descriptor" which is the central concept of the terminologies (aka "keyword"). Each descriptor is labelled and may be defined, linked to other descriptors (such as Related-Term relation) and involved in a son-father type of hierarchy (BT-NT for Broader Term – Narrower Term), which is the main relation of EHTOP. A descriptor may also contain specific attributes and synonyms (which is the main relation of WordNet), abbreviations etc.

It was also necessary to work on the terminologies modeling (OWL format) in order to fit it into the global database structure and to standardize the data in a well known and shared format. That is why the RDF (Resource Description Framework) syntax was chosen with the OWL (Ontology Web Language).

Creation of the EHTOP

The EHTOP was designed as a graphic interface of a Web Service, entirely dedicated to information retrieval and associations between terms of several terminologies. Thus, the main objective was to dissociate the substance from the form, in particular the interface.

Results

This project is a completely innovative work human-oriented to deal with terminologies; this is the main difference between ontology and terminologies and as data is more important for user (structure for computer), we decided to oversimplify ontology and terminologies (if necessary) integrated in EHTOP but we are caring more to data and its representation for human beings.

Overall, in July 2011, the time spent to build and maintain this health multi-terminology multilingual portal is approximately nine man-years. Currently, the CISMeF team is using one junior engineer (JG) to integrate new terminologies (e.g. SNOMED CT) and one post-doc (TM) to perform semantic harmonization on each terminology to another (more than 900 alignments were already performed -32^2-) using CISMeF NLP tools.

Currently, two versions of EHTOP exist: (a) the first version is mainly bilingual (French and English) specifically devoted to French users. This version is available at the following URL: http://pts.chu-rouen.fr/. Only MeSH and CISMeF terminologies are freely available. We provide a restricted access only for the scientific community (click on “Subscribe”); (b) the

second version is multilingual. Because the right to access terminologies varies among countries (e.g. the MeSH is freely available in English and French, but it is not the case in other European languages), the EHTOP multilingual version provides a free access to ICD10 in five European languages (English, French, German, Dutch, Greek).

A total of 32 terminologies are included into EHTOP, with 980,000 concepts, 2,300,000 synonyms, 222,800 definitions and 4,000,000 relations. Twenty one of these terminologies are not included yet in the UMLS among them, some from the World Health Organization. Due to various optimizations, the average response time for one concurrent user takes less than 500 milliseconds. Since July 2011, EHTOP is daily used by CISMeF librarians to index in multi-terminology mode. Since March 2010, the bilingual version is daily used by 600 unique machines, mainly to query MeSH in their native language. Two hundred thirty people have already registered to access other T/O, mainly physicians, health students, librarians and translators. For the 11 T/O included in UMLS, we performed a contextual link to BioPortal.

Via its human interface, EHTOP is daily used by the CISMeF team to index resources. It is also used by various CISMeF academic partners in different French and European projects, which was necessary to develop the EHTOP. EHTOP was qualitatively evaluated in 2010 by 25 medical students from the Rouen Medical School and gave 58% satisfaction for its user interface and 76% for its functionalities and content.

Discussion

The Health Multi-Terminology Multilingual Portal (EHTOP) is used daily by various CISMeF academic partners in different French and European projects. In 2011, the Top 3 EHTOP targets are librarians, health professionals and health students to learn how to manipulate health terminologies (e.g. about rare disease with Orphanet thesaurus or anatomy with the FMA ontology) and to extract knowledge from it, in particular from hierarchies and relations (e.g. various siblings of a rare disease, symptoms of this rare disease or to obtain all the muscles of the forearm in one click). Combining EHTOP to WordNet will improve both tools: allowing for EHTOP to obtain definitions and rich relations from WordNet and for WordNet to obtain more health terms, including definitions and even richer relations (e.g. relations for one rare disease to the genes linked to it).

If French MeSH Browser was also heavily used (500 unique users per working day) mainly to access MeSH using queries in French, EHTOP will allow to access MeSH in 10 other native

languages. Via its Web services, EHTOP may also be used by several interactive applications. The other targeted users include the entire range of medical IT players (e.g. institutions, hospitals, software publishers, information portals) and, through them all those involved in the healthcare sector, in particular healthcare professionals and patients.

The EHTOP presented here has the main functionalities of any terminology server, except the extensive management of terminologies (e.g. adding a new hierarchy). To the best of our knowledge, EHTOP is the first of its kind to allow cross-lingual navigation. The main added value of EHTOP when compared to any UMLS browser [13] is the possibility to access the main health terminologies in French or the multi-lingual terminologies and classification coming from WHO, which are not (yet) included in the UMLS (e.g. ATC for drugs or ICPS for patient safety), as demonstrated in accessing ICD10 in five languages. Currently, the EHTOP is a necessary basic tool to index any document in a multi-terminology multilingual mode.

Other portals propose to search and navigate T/O such as NCBO Bioportal [14] and the EBI Ontology Lookup Service [15-16]. Those tools are also very user-friendly but do not allow users to navigate through terms or search among synonyms in different languages. They are also not adapted to a daily use to index or to present the FMA to medical students.

In the near future, we have planned to integrate multi-lingual dictionaries (Wordnet) via an applet to the new multi-lingual health search engine Doc'CISMeF.

Conclusion

A health cross-lingual multi-terminology portal connected to a cross-lingual dictionary is a valuable tool to help to index and retrieve resources from a quality-controlled health gateway. It can also be very useful for teaching or performing audits in terminology management.

Acknowledgements

EHTOP was partially funded by PlaIR project, funded by FEDER; URL: http://www.plair.org

The authors thank Richard Medeiros for his advice in the editing of this manuscript and the eight students of the INSA Rouen Engineering School that partially developed the multi-terminology portal.

References

1. Darmoni SJ, Leroy JP, Baudic F, Douyère M, Piot J, Thirion B: **CISMeF: a structured Health resource guide.** *Methods Inf Med* 2000, **39**(1):30-35.
2. Thirion, B; Pereira, S; Névéol, A; Dahamna, B & Darmoni, SJ.. French MeSH Browser: a cross-language tool to access MEDLINE/PubMed.. AMIA symp., 2007: 1132.
3. Burgun A, Denier P, Bodenreider O, Botti G, Delamarre D, Pouliquen B, Oberlin P, Lévéque JM, Lukacs B, Kohler F, Fieschi M, Le Beux P: **A Web terminology server using UMLS for the description of medical procedures.** *J Am Med Inform Assoc* 1997 Sep-Oct: **4**(5):356-363.
4. Darmoni SJ; Joubert M; Dahamna B; Delahousse J, Fieschi M. **a French Health Multi-terminology Server**. *AMIA symp* 2009, 808.
5. About WordNet [URL: http://wordnet.princeton.edu/; accessed July, 22, 2011]
6. Fellbaum C. WordNet and wordnets. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 2005: 665-670.
7. Sensegates [URL: http://www.sensegates.com/; accessed July, 22, 2011]
8. Lussier YA, Rothwell DJ, Côté RA: **The SNOMED model: a knowledge source for the controlled terminology of the computerized patient record.** *Methods Inf Med* 1998, **37**(2): 161-164.
9. **National Library of Medicine. Medical Subject Headings** [http://www.nlm.nih.gov/mesh/]
10. **World Health Organization. International Classification of Diseases, 10th revision.** [http://www.who.int/classifications/icd/en/index.html]
11. Merabti T, Massari P, Joubert M, Sadou E, Lecroq T, Abdoune H, Rodrigues JM, Darmoni SJ: **An Automated Approach to map a French terminology to UMLS**. Stud Health Technol Inform 2010, 1040-1044.
12. Zeng K, Bodenreider O. **Integrating the UMLS into an RDF-Based Biomedical Knowledge Repository.** *AMIA Annu Symp Proc*. 2007 Oct 11:1170.
13. McCray AT, Razi A. **The UMLS Knowledge Source server.** *Medinfo*. 1995;8 Pt 1:144-7.
14. N.F. Noy, N.H. Shah, P.L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D.L. Rubin, M.-A. Storey, C.G. Chute, and M.A. Musen. **BioPortal: ontology and integrated data resources at the click of a mouse**. *Nucleic Acids Research*; Web Server Issue **10**. 2009

15. Cote RG, Jones P, Apweiler R, Hermjakob H.**The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries**. *BMC Bioinformatics*. 2006 Feb 28;**7(1)**:97 PMID: 16507094
16. Cote RG, Jones P, Martens L, Apweiler R, Hermjakob H. **The Ontology Lookup Service: more data and better tools for controlled vocabulary queries**. *Nucleic Acids Res*. 2008 Jul 1;**36(**Web Server issue):W372-6. Epub 2008 May 8. PMID: 18467421

Figures

Figure 1 - Interelationship between CISMeF Information System and EHTOP

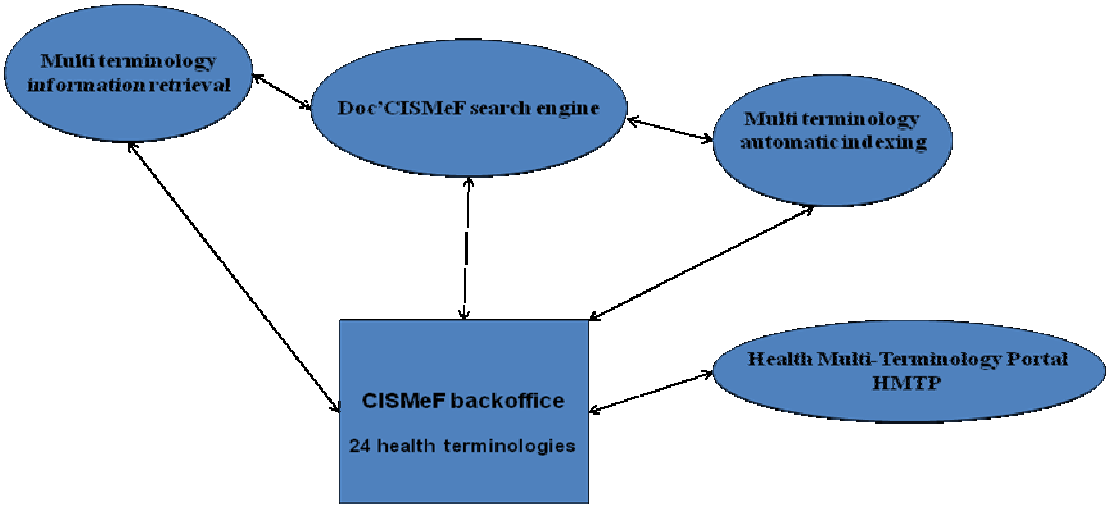
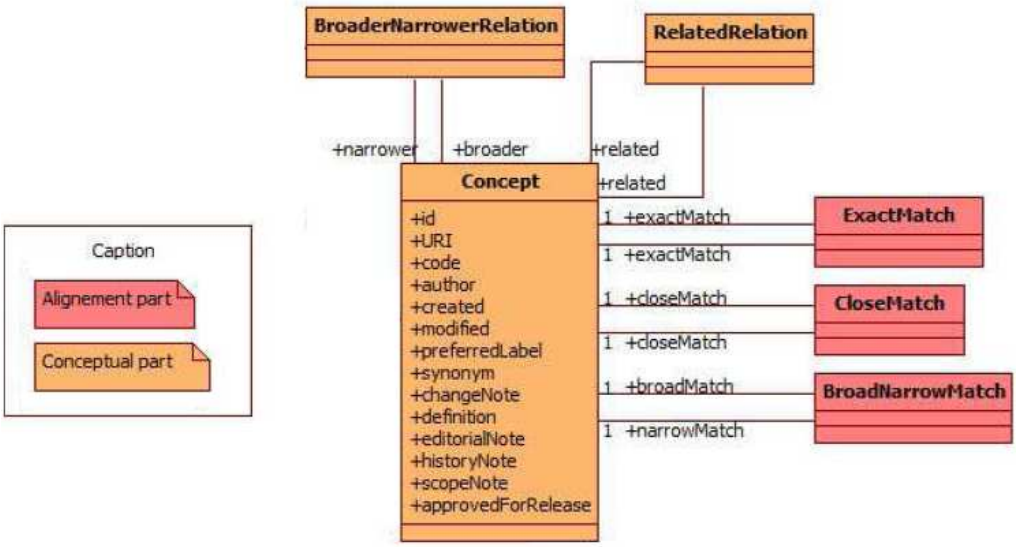


Figure 2 - The CISMeF BackOffice database conceptual structure



When	Jan 9, 2012 - Jan 13, 2012
Where	Matsue, Japan

Submission Deadline	Sep 1, 2011

Call For Papers

First Call for papers 6th International Global WordNet Conference (GWC2012)

CALL FOR PAPERS

Announcement
6th International Global WordNet Conference,
Matsue, Japan
January, 9-13, 2012

First Call for papers 6th International Global WordNet Conference.

The Global WordNet Association is pleased to announce the 6th International Global WordNet Conference (GWC2012).
The conference will be held at Kunibiki Messe (www.kunibikimesse.jp/en/index.html), Matsue, Japan, in January 2012. It is organized by the Global WordNet Association, the Toyohashi University of Technology and the National Institute of Japanese Language and Linguistics. Matsue is located in the west part of Japan, about 1000km west from Tokyo.

Local organization: Hitoshi Isahara (Toyohashi University of Technology) and Kyoko Kanzaki (National Institute of Japanese Language and Linguistics).

Details about the Association and the full announcement for the conference can be found on the GWA website (www.globalwordnet.org) and on the conference website (to be launched).

We invite papers addressing the topics listed below. Proposals for tutorials on building wordnets and demonstrations of wordnet databases, and wordnet-based software are welcome too.

1. Linguistics and lexical semantics, including:
* In depth analysis of Semantic Relations,
* Determining and representing word meanings (definitions, relations, semantic components, co-occurrence statistics, etc.)
* Necessity and Completeness issues.
* Ontologies and wordnets.
* The lexicon and wordnets.
2. Architecture of lexical databases, including:
* Language independent and language dependent components
3. Tools and Methods for WordNet Development, including:
* User and Data entry interface, organization,
* Extending and enriching wordnets
4. WordNet as a lexical resource and component of NLP and MT, including:
* Word sense disambiguation using wordnet,
* Ontologies and WordNet,
* The Lexicon and WordNet
5. Applications of WordNet, including:
* Information Extraction and Retrieval,
* Document Structuring and Categorization,
* Automatic Hyperlinking
* Language Teaching,
* Psycholinguistic Applications
6. Standardization, distribution and availability of wordnets and wordnet tools.

Presentations will fall into one of the following categories:

* long papers (30 mins)
* short papers (15 mins)
* project reports (10 mins)
* demonstrations (20 mins)

Submissions will have to state one of the preferred categories. Acceptance may be subject to changes in the category of the presentation, e.g. a long paper submission may be accepted as a short paper.

Final papers should be submitted in electronic form (PDF only):

* Long papers should contain approximately 4,000 words (~ 8 pages),
* Demonstration papers must be at most 5 pages text (2,500 words long) and can have additional 3 pages screen dumps or images;
* Short papers and project descriptions should be approx. 2,500 words long (~ 5 pages).

Papers need to be submitted to the EasyChair website:

GWA 2012 Easy Chair site: https://www.easychair.org/conferences/?conf=gwa2012

The format of the paper is in ACL format (PDF):

ACL 2010 paper formats: http://acl2010.org/authors_final.html

The conference program will include oral presentations and demonstration sessions with sufficient time for discussions of the issues raised.

The deadline for submissions is September 1st, 2011. Decisions regarding acceptance will be announced to the authors by end of September. Final papers are due at October 10th.

Important note: Inclusion of accepted submissions into the final program and the proceedings is contingent upon at least one author's registration. Late registration and on site registration for participants is possible without inclusion of the paper and without presentation.

Important dates:

1. September-1-2011 Deadline for paper submission
2. September-30-2011 Acceptance of papers
3. October-10-2011 Final Papers
4. October-15-2011 Registration is open
5. November-15-2011 Registration closes for author(s) to be included in the proceedings
6. January 9-13, 2011 Conference

Please note that traveling to Japan may require a Visa. Please, consult the Japan National Tourist Organization: http://www.jnto.go.jp/eng/arrange/essential/visa.html

Also visit the conference website (to be launched) for relevant information.

Conference Chairs:

Christiane Fellbaum
Fellbaum@clarity.Princeton.edu

Piek Vossen
P.Vossen@let.vu.nl

Local Organizing Chairs:

Hitoshi Isahara
Toyohashi University of Technology
Email: isahara@tut.jp

Kyoko Kanzaki
National Institute of Japanese Language and Linguistics
Email: kanzaki@ninjal.ac.jp