Grant Agreement Number: 257528

KHRESMOI

www.khresmoi.eu

Final report on KHRESMOI contribution to standards

Deliverable number	*D12.4.2*
Dissemination level	*Public*
Delivery data	*August 2014*
Status	*Final*
Authors	*Célia Boyer, Henning Müller, Georg Langs, Wim Peters, Angus Roberts, Thomas Schlegl, Veronika Stefanov*

This project is supported by the European Commission under the Information and Communication Technologies (ICT) Theme of the 7th Framework Programme for Research and Technological Development.

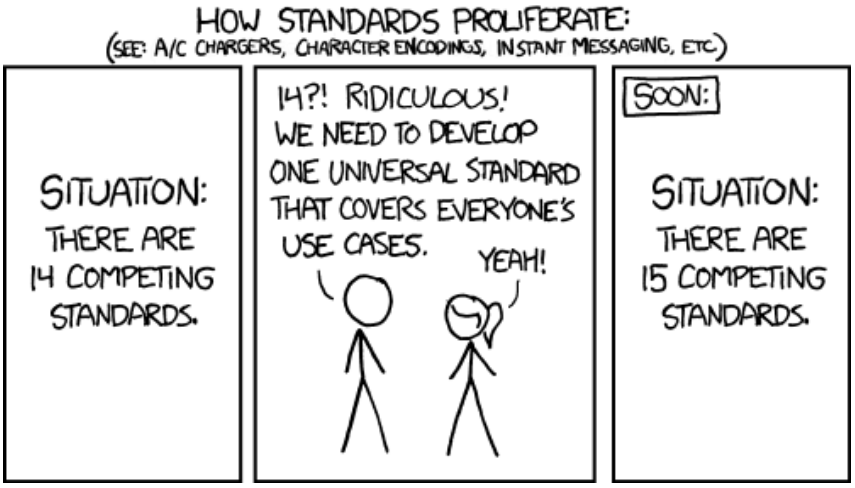

Figure 1: How standards proliferate [http://xkcd.com/927/]

Executive Summary

This document gives an overview of the KHRESMOI project's contribution to standards. Contributions were made in the areas of radiology, language resources and health informatics.

With regard to standards in radiology, we report that RadLex was adapted to the specific local environment, specifically that the German version of RadLex has been extended and improved, and made available to the community. A translation to Spanish is ongoing, and a French version is planned.

The annotated text corpora produced within KHRESMOI were made public in GATE's markup format, for which tools are available. Results from KHRESMOI are compliant to standard representation formats for exchange and reuse of multilingual language resources. KHRESMOI consortium members concerned with language resources are involved in standardization bodies, projects and initiatives for this and other issues such as time representation, ontology interoperability, and the standardization of part-of-speech tag representation.

Safeguards for the health-related Internet domain names and the new dot health domain are part of the relations by HON with the World Health Organization.

This document is the follow-up to D12.4.1 ("Analysis of standards and potential contribution to standards", 2011)[13], which gave an overview of standards potentially used or contributed to by the KHRESMOI project.

Table of Contents

List of Figures

Abbreviations

BioASQ	Biomedical Question Answering
CES	Corpus Encoding Standard
CLARIN	Common Language Resources and Technologies Infrastructure
DICOM	Digital Imaging and Communications in Medicine
GATE	General Architecture for Text Engineering
GBM	Glioblastoma multiforme
HON	Health On the Net (KHRESMOI consortium partner)
HONcode	Health On the Net Code of Conduct
ICANN	Internet Corporation for Assigned Names and Numbers
ISO	International Organization for Standardization
KHRESMOI	Knowledge Helper for Medical and Other Information users
LIRICS	Linguistic Infrastructure for Interoperable Resources and Systems
MeSH	Medical Subject Headings
MLi	MultiLingual Data & Services Infrastructure
MUW	Medical University of Vienna (KHRESMOI consortium partner)
RadLex	Radiology Lexicon
RSNA	Radiological Society of North America
SC	Subcommittee
SGML	Standard Generalized Markup Language
TC	Technical committee
TEI	Text Encoding and Interchange
TimeML	Markup Language for Temporal and Event Expressions
TLD	Top Level Domain
uComp	Embedded Human Computation for Knowledge Extraction and Evaluation
USFD	University of Sheffield (KHRESMOI consortium partner)
W3C	World Wide Web Consortium
WHO	World Health Organization
XCES	XML Corpus Encoding Standard
XML	Extensible Markup Language

1 Introduction

From the beginning of the KHRESMOI project it was clear that the success of the project would depend on the semantic and technological interoperability of the system, and that building on relevant standards would be important for achieving this. In this context, possibilities for the contribution of the KHRESMOI project to ongoing standardization initiatives were identified [13].

Now at the end of the project we can report that KHRESMOI's results in several fields of research either already have been added to existing standards, as is the case with the German RadLex improvements created in the context of the 3D Radiology prototype for KHRESMOI, or are in the process of slowly having an effect through the project members' numerous involvements in standardization bodies, projects and committees.

2 Contributions to Standards

2.1 Standards in Radiology

In the context of clinical imaging data mining, the KHRESMOI project relies on standards such as DICOM and established terminologies such as MeSH and RadLex. In the course of the project we have contributed to the adaptation of RadLex to specific local environments (extensions of the vocabulary and mapping of frequently used terms). In the following we provide a brief summary of the basis, and our contribution.

RadLex [6] is an ontology for the radiology domain. RadLex is developed by the Radiological Society of North America (RSNA). It provides an unified language for reporting, organizing and retrieving images and medical reports.

The current version of RadLex (version 3.11) includes more than 68.000 terms. Each term has a corresponding unique identifier (RadLex ID). Terms are pooled together in distinct groups, for instance in terms which are related to anatomy, pathology, imaging modality or image observations [12]. Each group in turn is structured hierarchically, so that navigation between terms of the same hierarchy is possible. Various relations exist between terms of the same hierarchy (synonym, Is_a, subClassOf, Part_Of, Has_Part, Contained_In, ...) and between hierarchies as well.

A German version of RadLex is also available [7]. The German version of RadLex contains fewer terms and has a different hierarchical structure than the English version as it was based on an older RadLex version. In KHRESMOI the German version was linked where possible to the newer English version so links between the hierarchies become usable as well. The German and English version were also made available in various formats to the RadLex community by KHRESMOI. KHRESMOI also initiated the manual translation of a Spanish version that is currently underway, and plans to work on a French version.

In the 3D Radiology Prototype of the KHRESMOI project RadLex was utilized to extract standardized terms from radiological reports. This prototype was developed in Vienna, where clinical images and corresponding radiological reports from the Vienna General Hospital were used. Standing to reason the German version of RadLex was used. Because of the following reasons, the German RadLex had to be extended:

- Not every term of the English version of RadLex has a corresponding German translation.
- Radiological reports show a large diversity of used terms. The English version of RadLex contains English synonyms of terms but does not contain additional German synonyms.
- Abbreviations (e.g. GBM for Glioblastoma multiforme) are frequently used. Abbreviations are neither contained in the English nor in the German version of RadLex.
- Even the English version of RadLex does not cover completely the language used in radiological reports. For example, "Nierenbeckenkelchsystem" is a term which is frequently used in German reports. The English version of RadLex does not contain a suitable term.

The extensions have been implemented in the framework of the clinical radiology prototype, and additional term mappings were performed by MUW. The term extraction pipeline was developed by the University of Sheffield.

2.2 Language Resource Standards

In Khresmoi deliverable D12.4.1 [13], we discussed that fact that linguistic and terminological knowledge is expressed in various ways within language resources, including lexicons, machine readable dictionaries, and term banks, and how this knowledge covers linguistic description at various levels of granularity, including the syntactic, morphological, orthographic and semantic levels. The differences in language resources are driven by a number of factors, including application requirements and end user [8]. Divergence in these resources have in turn driven the need to harmonize resources, in order to enable reusability and interoperability [13, 8].

As a producer of the most widely used open source language engineering toolkit in the world, Khresmoi partner USFD have continued to be involved in standards bodies, expert committees, and standards driven project initiatives. This includes continued involvement in the International Organization for Standardization (ISO[1]) expert committees 37/SC 4 on Language resource management. This has created standard representation formats that enable the exchange and reuse of multilingual language resources. We have ensured that all GATE work carried out within Khresmoi is compliant with these standards, through use parts of the reference implementation of TC37/SC2 standards.

Corpora produced within Khresmoi are made public in GATE's stand-off markup format. This is compatible with the XCES standard, an XML [5] based corpus encoding standard that builds on the Corpus Encoding Standard (CES)[2], itself an application of SGML1 (ISO 8879:1986, Information Processing–Text and Office Systems–Standard Generalized Markup Language). XCES is a corpus annotation standard that conforms to the TEI Guidelines for Electronic Text Encoding and Interchange[3][2]. For use of the Khresmoi corpora in a standards compliant environment, a suite of GATE based XCES tools are available from the American National Corpus.

USFD has continued to be involved in standards work through various projects. A major part of this input has been through our role in the MLi support action[4] which is defining the strategic

[1] http://www.iso.org
[2] http://www.cs.vassar.edu/CES/
[3] http://www.tei-c.org/index.xml
[4] http://mli-project.eu/

vision and specifications for Europe's language data and services infrastructure. Several Khresmoi partners have also been involved on the steering committee of the BioASQ project[5], which develops community challenge efforts in semantic indexing. As mentioned in [13], we have also been involved in standards efforts through the LIRICS [6] project, which provided ISO-ratified standards for language technology, and CLARIN [7] which established interoperable language resources research infrastructure [4, 14].

More recently, we have begun to explore standards on time representation in natural language, as expressed in biomedical documents and clinical records. This builds on work carried out at USFD and by members of the GATE team on the time markup standard, TimeML[8].

USFD has also joined the Ontology-Lexica Community Group of the World Wide Web Consortium (WC3)[9], addressing the issues of ontology interoperability in [9]. USFD has also been involved in standardisation of Part Of Speech tag representation via the UComp project[10], and TEI compliance of various annotated records.

2.3 Health Informatics Standards

There is still a critical need to ensure appropriate safeguards for the health-related Internet domain names and to protect the citizen from biased manipulated health related online information. KHRESMOI had an important role to show to the community and in particular to the ICANN and ISO TC 215 that it is possible to guide citizens toward trustworthy health related information. HON and KHRESMOI are active to promote safeguards for a dot health management and has contributed to the Safeguards for Health Domains, World Health Organization Proposals on Principles and Governance [11] [3, 10, 11, 1].

With its active collaboration with WHO via several contributions such as the HONcode of Conduct and the KHRESMOI project, Health On the Net is now a non-profit organization which is in official relations with the World Health Organization since February 2014, which is an acknowledgment of the service provided also by KHRESMOI to the health internet.

The HONcode principles will be shortly included in the repository of the Joint Initiative for Global Standards Harmonization support by ISO TC 215 [12].

2.4 Other

Regarding MT, there are no current standards for data (text), except for them to be plain text. Moses MT system[13] uses its internal formats and structure, which is primarily not meant for data interchange.

[5] http://www.bioasq.org/
[6] http://lirics.loria.fr/
[7] http://www.clarin.eu/external/
[8] http://www.timeml.org/
[9] http://www.w3.org/community/ontolex/
[10] http://www.ucomp.eu/
[11] http://www.who.int/ehealth/programmes/governance/en/index3.html
[12] http://www.skmtglossary.org/AddEditDocument.aspx
[13] http://www.statmt.org/moses/

3 Conclusion

The KHRESMOI project has not only benefited from using existing standards for its work. KHRESMOI project members have collaborated in various roles within standardization committees, groups, projects, and other bodies to create, improve or extend existing standards and formats.

Much of this work will continue to have effects in the future. Standardization is a slow process. Final acceptance of new developments will take years.

4 References

[1] Célia Boyer and Ljiljana Dolamic. Feasibility of automated detection of HONcode conformity for health related websites. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 5(3), 2014.

[2] Lou Burnard and C.M. Sperberg-McQueen. The Design of the TEI Encoding Scheme. *Computers and the Humanities*, 29(1):17–39, 1995.

[3] Ljiljana Dolamic and Célia Boyer. Detection of pages respecting quality standards: Character n-gram tokenization in automatic detection of web-page trust. In *Med-e-Tel 2014, International eHEALTH, Telemedicine and Health ICT Forum for Education, Networking and Business*, 2014.

[4] A. Funk, I. Roberts, and W. Peters. Implementing a variety of linguistic annotations through a common web-service interface. In *"Language Resource and Language Technology Standards" workshop at LREC, Malta*, May 2010.

[5] N. Ide, P. Bonhomme, and L. Romary. XCES: An XML-based Standard for Linguistic Corpora. In *Proceedings of the Second Language Resources and Evaluation Conference (LREC), Athens, Greece*, 2000.

[6] C. P. Langlotz. RadLex: A new method for indexing online educational materials. *Radiographics*, (6):1595–1597, 2006.

[7] D. Marwede, P. Daumke, K. Marko, D. Lobsien, S. Schulz, and T. Kahn. RadLex-German version: a radiological lexicon for indexing image and report information. *RoFo: Fortschritte auf dem Gebiete der Rontgenstrahlen und der Nuklearmedizin*, 181(1):38–44, 2009.

[8] E. Montiel-Ponsoda, Aguado de Cea, A. G., Gómez-Pérez, and W. Peters. Modelling multilinguality in ontologies. In *Coling 2008: Companion volume - Posters and Demonstrations, Manchester, UK*, 2008.

[9] W. Peters. Establishing interoperability between linguistic and terminological ontologies. In A. Oltramari, P. Vossen, L. Qin, and E. Hovy, editors, *New Trends of Research in Ontologies and Lexical Resources*. Springer, 2013.

[10] Natalia Pletneva, Rafael Ruiz de Castaneda, Frederic Baroz, and Célia Boyer. General vs health specialized search engine: a blind comparative evaluation of top search results. In *25th European Medical Informatics Conference, MIE*, 2014.

[11] Natalia Pletneva, Zdenka Uresova, Jean-Jacques Altman, Nicolas Postel-Vinay, Patrice Degoulet, Jan Hajic, and Célia Boyer. Observations and lessons learnt from non health professionals evaluating a health search engine. In *25th European Medical Informatics Conference, MIE*, 2014.

[12] M. W. Shore, D. L. Rubin, and C. E. Kahn Jr. Integration of imaging signs into RadLex. *Journal of digital imaging*, 25(1):50–55, 2012.

[13] Veronika Stefanov, Matthias Samwald, Célia Boyer, Allan Hanbury, Blanca Jordan, Henning Müller, Georg Langs, Jungyeul Park, Wim Peters, Angus Roberts, and Patrick Ruch. D12.4.1 – Analysis of standards and potential contribution to standards, Confidential report (only for members of the Khresmoi Consortium (including the Commission Services)), 2011.

[14] P. Wittenburg, N. Bel, L. Borin, G. Budin, N. Calzolari, E. Hajicova, K. Koskenniemi, Lemnitzer, Maegaard L., M. B., Piasecki, J.M. Pierrel, S. Piperidis, I. Skadina, D. Tufis, R. Veenendaal, T. Vradi, and M. Wynne. Resource and service centres as the backbone for a sustainable service infrastructure. In *Proceedings of LREC2010*, 2010.