Formalizing 'Living Guidelines' using LASSIE: A Multi-step Information Extraction Method

Katharina Kaiser[1] and Silvia Miksch[1,2]

[1] Institute of Software Technology & Interactive Systems
Vienna University of Technology, Vienna, Austria
[2] Department of Information and Knowledge Engineering
Danube University Krems, Krems, Austria
{kaiser, silvia}@ifs.tuwien.ac.at
http://ieg.ifs.tuwien.ac.at

Abstract. Living guidelines are documents presenting up-to-date and state-of-the-art knowledge to practitioners. To have guidelines implemented by computer-support they firstly have to be formalized in a computer-interpretable form. Due to the complexity of such formats the formalization process is challenging, but burdensome and time-consuming.
The LASSIE methodology supports this task by formalizing guidelines in several steps from the textual form to the guideline representation language Asbru using a document-centric approach. LASSIE uses Information Extraction technique to semi-automatically accomplish these steps.
We apply LASSIE to support the implementation of living guidelines. Based on a living guideline published by the Scottish Intercollegiate Guidelines Network (SIGN) we show that adaptations of previously formalized guidelines can be accomplished easily and fast. By using this new approach only new and changed text parts have to be modeled. Furthermore, models can be inherited from previously modeled guideline versions that were added by domain experts.

1 Introduction

The development process for a clinical practice guideline (CPG) takes at least two years. Thus, CPGs can be out of date as soon as they are produced, as new research findings are continuously published. To overcome this problem sometimes the shelf life for a guideline is identified; either by a date (e.g., this guideline will be reviewed in 2 years) or by a statement that the review date will be determined by the availability of new evidence (e.g., this guideline will be considered for review as new evidence becomes available). Alternatively, we can consider a new option – the living guideline. A living guideline is one that remains under review on an ongoing basis, with updates published at set intervals (e.g., annually).

The review of the guideline (i.e., a new article in the specified field is available) may have various characteristics. On the one hand it can add additional evidence and thus alter the evidence level of a recommendation. On the other hand it can lead to a new recommendation or it may change an existing one. However, in the majority of cases only small text parts are changed; often only the reference to the new article is added or to an obsolete article is removed.

Modeling CPGs in a computer-interpretable form is a prerequisite for various computer applications to support their application. However, transforming guidelines in a formal guideline representation is a difficult task. In [1] and [2] we have proposed a semi-automatic methodology called LASSIE to model treatment processes in multiple steps using Information Extraction (IE).

We will now show that we can use LASSIE to support the formalization of living guidelines. Applying this method, which traces both the general formalization steps and the changes to new versions has the potential to reduce the modeling effort. The Scottish Intercollegiate Guidelines Network (SIGN) has already published a living guideline [3]. Based on the documents provided we will show that adaptations of formalized guidelines can be accomplished easily and fast.

In the next section we will discuss some work on guideline formalization tools and guideline versioning methods. Afterwards we will give a short introduction in LASSIE. In Section 4 we describe the adaptation of LASSIE for supporting living guidelines followed by a case study illustrating our methodology. Section 6 summarizes our work and represents our conclusions.

2 Related Work

In this section, we present relevant work describing guideline formalization tools and approaches for guideline versioning.

For formalizing clinical guidelines into a guideline representation language (see [4] for an overview and comparison) various tools exist. We can classify such tools in document-centric and model-centric tools.

2.1 Document-Centric Approaches

Markup-based tools utilize a document-centric approach. Thereby, the original guideline document is systematically marked-up by the user in order to generate a semi-formal model of the marked text part.

The *GEM Cutter* [5] was one of the first exponents of this apporach transforming guideline information into the GEM format [6]. *Stepper* [7] is a tool that formalizes the initial text in multiple user-definable steps corresponding to interactive XML transformations. The *Document Exploration and Linking Tool / Addons (DELT/A)* [8] supports the translation of HTML documents into any XML language. It uses links between the text part in the original document and its corresponding XML model. To generate a specific model user-definable *macros* can be used. *Uruz*, part of the *Digital electronic Guideline Library (Degel)* framework [9], is a web-based markup tool that supports indexing and markup using any hierarchical guideline-representation format. It enables the user to embed in the guideline document terms originating from standard vocabularies.

2.2 Model-Centric Approaches

In model-centric approaches a conceptual model is formulated by domain experts. The relationship between the model and the original document is only indirect.

AsbruView [10] uses graphical metaphors to represent Asbru plans. *AREZZO* and *TALLIS* [11] support the translation into PROforma using graphical symbols representing the task types of the language. *Protégé* [12] is a knowledge-acquisition tool that supports the translation into guideline representation languages EON, GLIF, and PROforma. It uses specific ontologies for these languages, whereas parts of the formalization process can be accomplished with predefined graphical symbols. AREZZO, TALLIS, and Protégé offer a flowchart-based representation of the processes.

2.3 Guideline Versioning

Unfortunately, guideline versioning has not been adequately addressed by now. There are two approaches dealing with versioning:

Peleg and Kantor [13] propose a model-centric approach for GLIF. Thereby, the underlying GLIF ontology is extended by version information and a versioning tool was developed that supports the creation of a new CPG model or the modification of an existing one as well as the displaying of versions of a CPG model, highlighting the differences.

Seyfang et al. [14] describe the formalization of 'living guidelines' using a document-centric approach. They start with an HTML version of the guideline and use different intermediate representations to derive a formal model of the guideline. The first intermediate representation is MHB and the DELT/A tool is used to mark-up text chunks. The original marked-up guideline document is then manually updated to the new version by highlighting both newly added and removed text fragments. Using the DELT/A tool the highlighted text fragments are selected to visualize the corresponding MHB chunk in order to make the necessary changes.

But still, using the mentioned tools the modeling process is complex and labor intensive. Methods are needed to automate parts of the modeling task.

3 LASSIE – Modeling Treatment Processes Using Information Extraction

Most guideline representation languages are very powerful and thus very complex. They can present a multitude of different information and data. We apply a multi-step transformation process that facilitates the formalization process by various intermediate representations (IRs) obtained in stepwise procedures.

Our multi-step transformation methodology, called LASSIE[3], supports the document-centric approach by marking the original guideline document and generating the particular models for each marked text part. It is intended to be a semi-automatic approach. This enables the user not only to correct the transformations, but also to augment them by implicit knowledge necessary for a subsequent execution. After each step the user is able to view the results using the DELT/A tool [8].

The benefits of the multi-step approach and in the following of the IRs are that IRs (1) support a concise formalization process, (2) provide different formats and separate

[3] modeLing treAtment proceSSes using Information Extraction

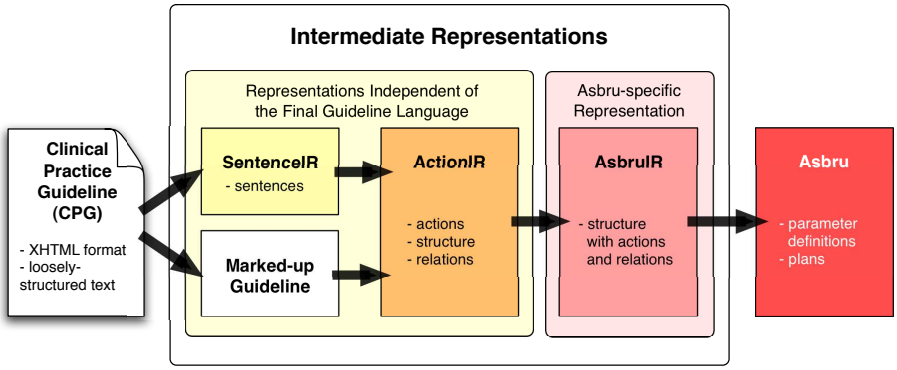

Fig. 1. Steps to (semi-)automatically gain an Asbru representation of CPGs. To gain process information from a CPG the first two steps are accomplished in order to have a representation independent of the final guideline language.

views and procedures for various kinds of information, (3) specific heuristics for each particular kind of information can be applied, and (4) a simpler and more concise evaluation and tracing of each process step is accomplishable. The IRs are specific templates used by IE methods to present the desired information. The IE methods use a terminology based on the Medical Subject Headings (MeSH)[4] [15] and manually generated extraction patterns.

CPGs present effective treatment processes. One challenge when authoring CPGs is the detection of individual processes and their relations and dependencies. We can generate simple representations of treatment instructions (i.e., actions), which are independent from the final guideline representation language. Based on this independent representation we can transform the information in further steps into the guideline languages. In [1] and [2] we have demonstrated that it is possible to formalize processes using IE for modeling guidelines in Asbru (see Fig. 1).

4 Adaptation of LASSIE for 'Living Guidelines'

Using LASSIE a unique identifier (i.e., the DELT/A link) marks information transformed from one step to the next. We now apply LASSIE to support the formalization of living guidelines. The document provide us the information that has changed: Adaptations of every new revision are marked by arrows and highlighted in terms of color (or in different gray scales) (see Fig. 2).

We now propose a new method utilizing this information. Thereby, the new guideline is not going to be modeled from scratch, but already modeled parts from previous versions are inherited. Thus, only new text parts have to be modeled (see Fig. 3).

[4] http://www.nlm.nih.gov/mesh/

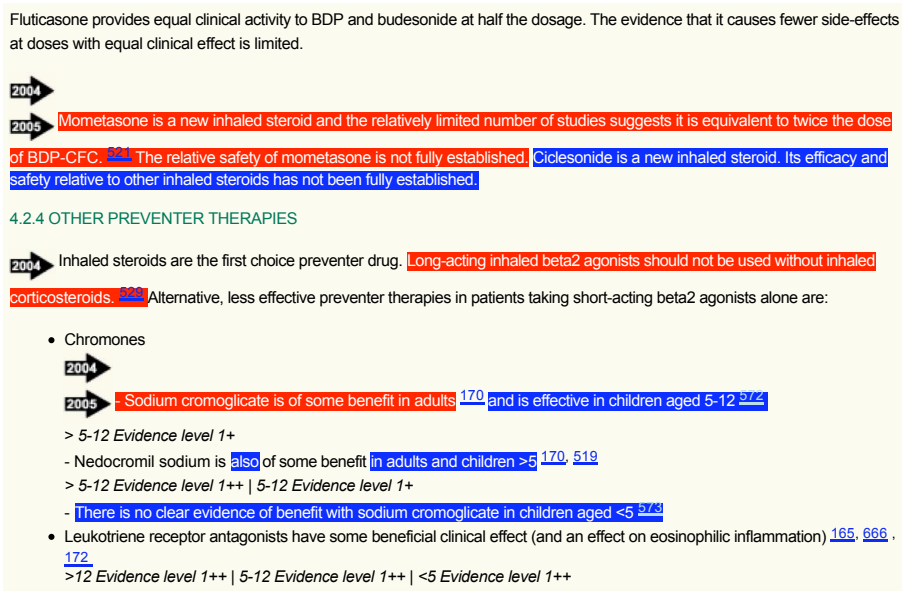

Fig. 2. Excerpt of the 2005 version of the ”living guideline” [3]. Adaptations of every new revision are marked by arrows and highlighted in terms of color (or in different gray-scales).

As LASSIE is a multi-step methodology, we have to satisfy each step for the living guideline’s formalization.

4.1 Pre-processing

As the input of LASSIE’s first step is the XHTML-conform guideline document, we have to preprocess the document to get a unified document format. We accomplish this by XSLT scripts, HTML Tidy[5], and manual post-processing in order to obtain not only a well-formed but also a hierarchically well-structured XHTML document.

4.2 Marking-up the New Guideline Version

LASSIE’s first step is to detect relevant sentences and text parts in the guideline document. Text parts are thereby list entries that may not be complete sentences, but are referred to as sentences in the remaining paper.

The output of LASSIE’s first step are two files: (1) the marked-up guideline document, where relevant sentences are marked and tagged by a DELT/A link, and (2) a file containing all relevant sentences and their corresponding DELT/A links.

We use these files of the previous guideline version to detect unchanged relevant sentences in the new guideline version. We parse the new guideline document and

[5] http://tidy.sourceforge.net

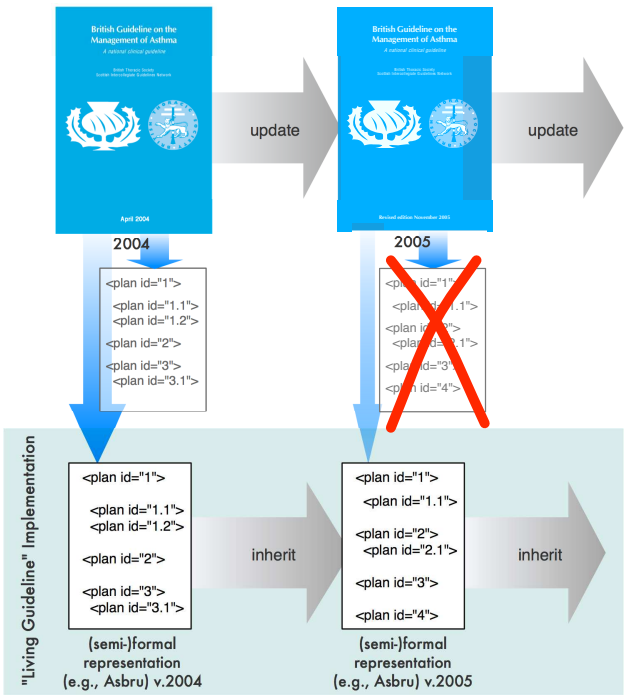

Fig. 3. Formalizing a living guideline using LASSIE. The documents provide us the information that has changed. After comparing the new documents with the previous ones we are able to adapt the former formalized documents using LASSIE.

search for each sentence marked-up in the previous version. Thereby, we have to consider not only equal sentences but also equal contexts of them. This is necessary as a marked-up sentence can appear repeatedly in the document and we have to assign the correct DELT/A link in the new document. For each sentence in the new guideline that is marked as updated as a part or whole we apply step 1 of the LASSIE methodology (see [2] for details) in order to detect relevant sentences for further processing. Relevant sentences of the old guideline version that are not found in the new document can be seen as removed.

For each sentence of the new guideline that has been marked as relevant by the LASSIE methodology we assign also a version id. Furthermore, we have to be aware to not assign an obsolete DELT/A link to a new sentence.

Thus, we obtain the new marked-up guideline version and are now able to extract the processes in order to gain a representation independent of the final guideline language. After this step the user is also able to view the resulting files with the DELT/A tool and make corrections.

4.3 Further Transformation of the Extracted Information

After obtaining the new marked-up guideline document we can proceed with the subsequent steps coming up with LASSIE. That means, we can inherit models of subsequent representations that correspond with text parts that were not changed in the new guideline version. For new or changed models the particular processing step of LASSIE is applied. For instance, to detect processes we proceed as following:

Within the next step of the LASSIE methodology relevant sentences are structured and relationships between sentences are found. The output of this step is a representation (*ActionIR*) containing actions, relations controlling the process flow between these actions, and the structure illustrating the hierarchy and nesting of groups of actions.

An *action* contains the action sentence, possible assigned annotation sentences, treatment instruments, information about the dosage, duration or iteration of a drug administration, and conditions. If the action is part of a selection, it is given a selection id. DELT/A links are inherited from the *SentenceIR* representation in order to provide the traceability of the process.

In order to obtain actions from our new version of the marked-up guideline, we can inherit action and annotation sentences from the previous *ActionIR* version. Furthermore, new relevant sentences of the current guideline version are classified in action and annotation sentences. When sentences are classified as annotation they must be assigned an *action* sentence. If an action and its assigned annotation sentences were not changed in the new version, the complete action node is inherited to the new *ActionIR* representation. Otherwise, the action node and its additional information has to be generated by LASSIE. Additionally, a version id is assigned for these new nodes. Likewise, we are able to inherit relations between actions nodes if none of the both action nodes has changed. Otherwise, we have to detect new relations using LASSIE. The third part of the *ActionIR* representation, the structure of the actions, is then generated by LASSIE.

The output of this step is then a new version of the *ActionIR* representation, which can be viewed with the DELT/A tool. Changed information is identifiable by the version id. The user may then make corrections or add new information to the representation.

5 Case Study

We tested the applicability of our method to a real *living guideline*. Based on the *British guideline on the management of asthma* [3] from SIGN in its version of 2005 we generated the previous guideline versions (i.e., from 2004) due to the non-availability of the old documents[6]. This was possible because SIGN offers a document which clearly describes every adaptation (i.e., change, adding, removal) of the text. For evaluating the method we only used Section 4 (*Pharmacological Management*) of the guideline. It describes an important part of the asthma treatment and contains also updated text parts.

[6] We were not able to receive the older guideline versions from SIGN.

5.1 Formalizing the Original Guideline Version

We preprocessed the old guideline document to comply our unified document format. Starting with the old guideline document we used LASSIE to generate the particular models necessary for formalization in Asbru. We automatically generated the intermediate representations and adapted them according to our needs. The document consists of 509 sentences. 139 of them were classified as relevant for further processing.

5.2 Formalizing the New Guideline Version

The next step was to model the new guideline version using our new method. Therefore, we prototypically implemented our method to automate this task and adapted our implementation of LASSIE to enable the processing of *living guidelines*.

Preprocessing. We preprocessed the new guideline document in order to gain a unified document format complying the XHTML format.

Markup of new guideline version. Afterwards, we automatically searched for unchanged sentences that were marked in the previous guideline version and added the corresponding DELT/A links into the current document. Now, we were able to have LASSIE check the adapted sentences for relevancy. The new version of Section 4 consists of 515 sentences. We were able to inherit 133 sentences of the old version, which means that six relevant sentences were either changed or removed in the new version. 13 updated or new sentences were found and checked with LASSIE, which classified ten as relevant. The new relevant sentences were marked and assigned a new DELT/A link as well as a version id.

Action generation and further transformations. Within the next step the new sentences were classified in action or annotation sentences. The latter are then assigned to action sentences. We received five action sentences and five annotation sentences. Four of the annotation sentences were assigned to two previously available action sentences; one to a new action sentence. Thus, the remaining unchanged action models were inherited from the previous version.

The same procedure is done for all subsequent steps in an analogous manner.

5.3 Discussion

Our study shows that using a document-centric approach – LASSIE with the DELT/A tool – offers distinct benefits in modeling *living guidelines*. A fast adaptation of the new document is possible. As in *living guidelines* there will not be radical changes from one version to the succeeding version, inheriting of previous models is a simple, time-saving, but effective method for modeling computer-supported guidelines. Also, in the intermediate representations the new models are marked by their version ids to

enable a prompt identification. Thus, the user is able to perform adaptations quickly and conveniently.

A limitation of our methodology is that minor changes in the text may result in applying a new relevance check, sentence classification, action generation, and so on, which will require an evaluation by a human afterwards. In methods described in Section 2.3 such minor changes may be checked and accomplished by a human user more efficiently.

Furthermore, we have to mention that the IRs do not contain the models of all versions, only the actual ones. Thus, it is not possible to have one file for all versions, but one file for each version of a representation.

6 Conclusion

Living guidelines are documents presenting up-to-date and state-of-the-art knowledge to practitioners. To support their application they have to be brought in a computer-interpretable form, which is a difficult task.

We propose a method applicable on documents previously being formalized using a document-centric approach. Thereby, the guideline document is marked-up and corresponding formal models are generated. Our method utilizes these links between the textual document and the formal models. It inherits formalized models of the previous guideline version by re-linking them to their corresponding text parts in the new guideline version. Only changed or added texts have to be analyzed and modeled. The formalization task is thereby done using the LASSIE methodology. It is a semi-automatic approach using IE and various intermediate representations to model different kinds of information in various granularities. Our case study showed that the modelling effort can be reduced considerably by applying our LASSIE methodology.

By re-using previously formalized models of guidelines we are able to quickly and effectively formalize new guideline versions.

Acknowledgements. This work is supported by ”Fonds zur Förderung der wissenschaftlichen Forschung FWF” (Austrian Science Fund), grant L290-N04.

References

1. Kaiser, K., Akkaya, C., Miksch, S.: How can information extraction ease formalizing treatment processes in clinical practice guidelines? A method and its evaluation. Artificial Intelligence in Medicine **39**(2) (2007) 151–163
2. Kaiser, K., Miksch, S.: Modeling treatment processes using information extraction. In Yoshida, H., Jain, A., Ichalkaranje, A., Jain, L.C., Ichalkaranje, N., eds.: Advanced Computational Intelligence Paradigms in Healthcare – 1. Volume 48 of Studies in Computational Intelligence (SCI). Springer Verlag (2007) 189–224
3. Scottish Intercollegiate Guidelines Network (SIGN), British Thoracic Society: British guideline on the management of asthma. a clinical national guideline. Scottish Intercollegiate Guidelines Network (SIGN) (November 2005)

4. Peleg, M., Tu, S.W., Bury, J., Ciccarese, P., Fox, J., Greenes, R.A., Hall, R., Johnson, P.D., Jones, N., Kumar, A., Miksch, S., Quaglini, S., Seyfang, A., Shortliffe, E.H., Stefanelli, M.: Comparing computer-interpretable guideline models: A case-study approach. Journal of the American Medical Informatics Association (JAMIA) **10**(1) (Jan-Feb 2003) 52–68
5. Polvani, K.A., Agrawal, A., Karras, B., Deshpande, A., Shiffman, R.: GEM Cutter Manual. Yale Center for Medical Informatics, New Haven, CT. (2000)
6. Shiffman, R.N., Karras, B.T., Agrawal, A., Chen, R., Marenco, L., Nath, S.: GEM: a proposal for a more comprehensive guideline document model using XML. Journal of the American Medical Informatics Association (JAMIA) **7**(5) (2000) 488–498
7. Růžička, M., Svátek, V.: Mark-up based analysis of narrative guidelines with the Stepper tool. In Kaiser, K., Miksch, S., Tu, S.W., eds.: Computer-based Support for Clinical Guidelines and Protocols. Proceedings of the Symposium on Computerized Guidelines and Protocols (CGP 2004). Volume 101 of Studies in Health Technology and Informatics., Amsterdam, NL, IOS Press (2004) 132–136
8. Votruba, P., Miksch, S., Kosara, R.: Facilitating knowledge maintenance of clinical guidelines and protocols. In Fieschi, M., Coiera, E., Li, Y.C.J., eds.: Proceedings from the Medinfo 2004 World Congress on Medical Informatics, AMIA, IOS Press (2004) 57–61
9. Shahar, Y., Young, O., Shalom, E., Mayaffit, A., Moskovitch, R., Hessing, A., Galperin, M.: DEGEL: A hybrid, multiple-ontology framework for specification and retrieval of clinical guidelines. In Dojat, M., Keravnou, E., Barahona, P., eds.: Proceedings of the 9th Conference on Artificial Intelligence in Medicine in Europe, AIME 2003. Volume 2780 of LNAI., Protaras, Cyprus, Springer Verlag (2003) 122–131
10. Kosara, R., Miksch, S.: Metaphors of Movement: A Visualization and User Interface for Time-Oriented, Skeletal Plans. Artificial Intelligence in Medicine, Special Issue: Information Visualization in Medicine **22**(2) (May 2001) 111–131
11. Steele, R., Fox, J.: Tallis PROforma Primer – Introduction to PROforma Language and Software with Worked Examples. Technical report, Advanced Computation Laboratory, Cancer Research, London, UK (2002)
12. Gennari, J.H., Musen, M.A., Fergerson, R.W., Grosso, W.E., Crubézy, M., Eriksson, H., Noy, N.F., Tu, S.W.: The Evolution of Protégé: An Environment for Knowledge-based Systems Development. International Journal of Human Computer Studies **58**(1) (2003) 89–123
13. Peleg, M., Kantor, R.: Approaches for guideline versioning using GLIF. In Musen, M.A., ed.: Proceedings of the 2003 American Medical Informatics Association (AMIA) Annual Symposium, Washington, DC, American Medical Informatics Association (Nov. 2003) 509–513
14. Seyfang, A., Martinez-Salvador, B., Serban, R., Wittenberg, J., Rosenbrand, K., ten Teije, A., van Harmelen, F., Marcos, M., Miksch, S.: Maintaining formal models of living guidelines efficiently. In Bellazzi, R., Abu-Hanna, A., Hunter, J., eds.: Proc. of the 11th Conference on Artificial Intelligence in Medicine (AIME'07), Springer Verlag (2007)
15. National Library of Medicine: Medical Subject Headings. The Library (updated annually)