The Importance of Creating an Ontology-Specific Consensus Before a Markup-Based Specification of Clinical Guidelines

Erez Shalom[1] and **Yuval Shahar**[1] and **Eitan Lunenfeld**[2] and **Meirav Taieb-Maimon**[1] and **Ohad Young**[1] and **Guy Bar**[2] and **Susana B.Martins**[3] and **Laszlo Vaszar**[3] and **Yair Liel**[2] and **Avi Yarkoni**[2] and **Mary K.Goldstein**[3] and **Akiva Leibowitz**[2] and **Tal Marom**[4]

ABSTRACT. We have previously developed the Digital electronic Guideline Library (DeGeL) framework, which includes a methodology for a markup-based, increasingly formal structuring of free-text clinical guidelines (GLs), and tools to support guideline-based application. The methodology includes activities be-fore, during and after the markup process. To reduce the ambiguity of the interpretation of a GL among the Expert Physicians (EPs) who are marking up the GL, and to achieve an interpretation common to the EPs and the knowledge engineers (KEs), an indispensable step before markup is the creation of an Ontology Specific Consensus (OSC) regarding the semantics of the GL. To evaluate the role of the OSC, we created OSCs for three GLs in incremental level of de-tail, using the Asbru GL ontology. The EPs quantified the subjective aspects that most helped them in creating the OSC, while we assessed the clinical and ontological markup errors committed by the EPs. Using medical knowledge and understanding the GL ontology were considered more helpful than understanding the DeGeL tools; and the more detailed the OSC, the less the number of markup errors committed.

1 INTRODUCTION

Care providers, overloaded with information, rarely have the time, or the computational means, to use the valuable knowledge encoded in clinical guidelines (GLs), during treatment. In addition, although there are many text-based GLs, most do not have any automated support, although these GL's could obviously benefit from such support. There is therefore a pressing need to facilitate automated GL specification, dissemination, application, and quality assessment. Furthermore, there is a need for a well-defined methodology that integrates the roles of experts physician (EPs) and knowledge engineers (KEs), and that supports incremental specification and conversion of the GL's free-text into a machine comprehensible representation Thus, we have previously developed a hybrid representation model, which combines in optimal fashion the relative skills of the KE and the EP and which gradually converts the free-text GL into a semi-structured, semi-formal and formal representation.

This hybrid model was implemented in an architecture and set of tools, called the Digital electronic Guideline Library (DeGeL) [1] which was developed to support GL classification, semantic markup, context-sensitive search, browsing, run-time application, and retrospective quality assessment.

1.1 The Need For an Overall Specification Methodology

One of the DeGeL framework tools is the web-based URUZ markup tool. As a web-based tool and part of the DeGeL framework, using the infrastructure of the hybrid guideline representation model, URUZ enables EPs and KEs on different sites to collaborate in the process of GL specification and to mark up the GL at any representation level - semi-structured, semi-formal, and formal representations. However, there is still a need for an overall comprehensive methodology for the GL specification process and for the evaluation of this methodology. Bearing this in mind, the next objective of this study was to define the crucial steps for the specification process. It was known that it should involve both EPs and KEs: Patel et al. [2] have shown in the case of the GLIF GL ontology that joint efforts of EPs and KEs lead to improved results.

This approach is acceptable, although collaboration between EPs and KEs is considered to be the main bottleneck in the specification process, a fact which led to the recent trend of enabling the EP to perform the specification alone [3].In addition, in their recent study, Peleg et al.[4] found that in the process of creating algorithms from narrow guidelines, teamwork is crucial for detecting errors, and that the team should include a KE. Thus, we developed and evaluated a methodology for the overall specification process which is independent of a particular GL specification language, which provides support for all phases before, during, and after markup, and is based on a collaboration between EPs and KEs in all activities of the methodology.

[1]The Medical Informatics Research Center, Ben Gurion University, Beer Sheva, Israel; email: {erezsh, yshahar, ohadyn, , meiravta}@bgu.ac.il; website: http://medinfo.ise.bgu.ac.il/medlab
[2]Soroka Medical Center, Ben Gurion University, Beer Sheva, Israel email: {guybar, yarkoni, liel, akival, lunenfld}@bgu.ac.il
[3]Veterans Administration Palo Alto Heath Care System, Palo Alto,CA email: {Susana.Martins, Mary.Goldstein}@va.gov , laszlo@vassar.info
[4]E.Wolfson Medical Center, Holon, Israel; email: maromtal@013.net.il

The activities-performed *before* markup include: 1) Choosing the specification language 2) Learning the specification language 3) Selecting a GL for specification 4) Creating an Ontology-Specific Consensus 5) Acquiring training in the markup tool; The activities-performed *during* markup include: 1) Classifying the GL according semantic indices 2) Performing the specification process using the tools and consensus; The activities-performed *after* markup include evaluation of the results of the GL specification according to a Gold Standard created by a KE working with a senior EP.

A detailed analysis of the results of the evaluation of this methodology in the case of the Asbru GL specification language [5], in all phases including markup, can be found elsewhere [6], and is outside of the scope of this paper, which focuses mainly on the importance of creating an Ontology Specific Consensus as an indispensable, mandatory step as the main activity to be performed before the markup process.

1.2 The Importance of Using an Ontology-Specific Consensus

Cognitive studies of the InterMed group [7] have found that, given different domain knowledge and strategies, designers and users do not represent the information in the same way, leading to different interpretations and decisions. Our experience has shown that the EPs trying to specify the GL source according to some specification language have different interpretations due to the fact that the GL source is too abstract to support the requirements of automation, because it is not meant to be directly executed by a computer and is not in any formal specification language. Furthermore, since most of the knowledge in the GL source may be implicit rather than explicit, and sometimes the necessary knowledge in the text regarding some patient management is missing, each EP interprets the implicit or the missing content differently, leading to disagreement and a great deal of variability among the EPs. In addition, sometimes the textual sources do not include all the necessary knowledge underlying the GL; therefore, when specifying the GL, each EP expresses his own interpretation, a fact which leads, again, to different approaches of the treatment. In addition, the EPs developed their own interpretation of the medical knowledge represented by the textual GL through the years, based on their own learning and experience; thus, for the same management GL, different EPs have different interpretations

The problem becomes even more complicated when trying to automate the GL according to some specification language: The EPs are not programmers and thus it is hard for them to understand the semantics of the specification language; in contrast, the KEs are not familiar with the clinical semantics. The specification of a GL intended for sharing in different clinical settings, especially a GL that is meant to be automated, requires the EPs to answer explicitly many questions by the KE with which they do not have to contend when they use the GL free text. Therefore, neither of them (the EP nor the KE) can perform the specification alone. This problem still exists when the EP has some experience with the specification language and is assigned the role of KE: some content might still be missed as the specification is then done according to only one opinion (the EP's) who typically thinks only in unilaterally fashion. However, as experienced as the EP can be in the specification language, he usually cannot be more experienced than a KE. Even an EP working together with a KE does not suffice, since there is no collaborative process of thinking about the specification process among all EPs in the clinical setting.

Thus, a uniform interpretation of the GL across the EPs, and between the EPs and the KEs, that is, a *consensus*, is required. A consensus is defined in this study as: "An interpretation of the GL agreed upon by the EP and the KE". Furthermore, for the specification purpose, the consensus is a structural document that describes schematically the clinical directives of the GL and the semantic logic of the GL, in terms of the intended target specification language. (For example, if the specification language includes an eligibility criterion KR, the consensus will refer to that KR, and include its content, although not in a formal fashion.) Thus, the result is an *Ontology Specific Consensus (OSC)*. An OSC is created collaboratively by EPs and KEs, and its creation might involve converting implicit into explicit knowledge, even when it is not included in the GL sources, considering the localization of the GL and other directives that were not a part of the clinical agenda until this step.

Our initial attempts to create a consensus without a KE being involved in the guideline specification process have led to nonsatisfactory results. It was quite clear that a collaborative standardization of the guideline scenarios should be performed before the markup process and should include all clinical and semantic aspects of the GL in great detail. Furthermore, it is our impression that the creation of a consensus regarding the semantics of the GL is an indispensable, crucial mandatory step before markup, and should be done in the first stage by a group of EPs from the medical field, in collaboration with a KE who is familiar with the specification language. The later stages of creating an OSC can be done by a senior EP who has significant practical knowledge and experience, and a KE. Peleg et al. [4] suggest that to avoid errors in encoding guidelines, it should be verified that: 1) all relevant knowledge of the guideline is carried to the algorithm; 2) all the information necessary for the treatment is provided; and 3) patient scenarios are considered.

This study's objective was to find the aspects (such as use of ontology, understanding of the specification language, etc.) which helped the EPs who created an OSC, and furthermore to find the level of OSC expressiveness which will be the best for the markup process.

2 METHODS

The OSC is created in a iterative fashion by performing the following steps:

1. After selecting the GL sources, the EPs in the clinical institution in which the GL will be applied is performing "brain storming" in order to create a clinical consensus based on the GL sources. The consensus can sometimes be modified with a little refinement of common local clinical settings as well as by adding concepts and directives by the EPs, representing their own knowledge and expertise.

   This clinical consensus should describe, in flow chart fashion, the general clinical steps of the GL; each step is

described in a rectangular task box with a suitable title. Decisions are intuitively described a rhombus. When one step needs to be divided further into sub-steps, an arrow is added, pointing to the next task box. This clinical consensus describes the basic procedural knowledge of the GL, and is free of ontology-specific details. For example, choosing between three antibiotic options mentioned in the PID guideline [8] was explicitly stated as the "inpatient treatment plan" which calls for the use of one of three possible regimens.

2. The KE(s) and the EPs collaborate. The KE, who is an expert in the ontology specification language, together with the senior EP overviews the clinical consensus and contributes his emphasis regarding the structure of the clinical consensus, adding some control structure in the notion of the specification language. For example, he might define the order for a group of steps, or refine the text and the logic of the clinical consensus, according to the answers the EP gives to his questions. Thus, the clinical consensus is refined into an ontology-specific procedural part of the consensus (See Figure 1, and Figure 2 for another example).

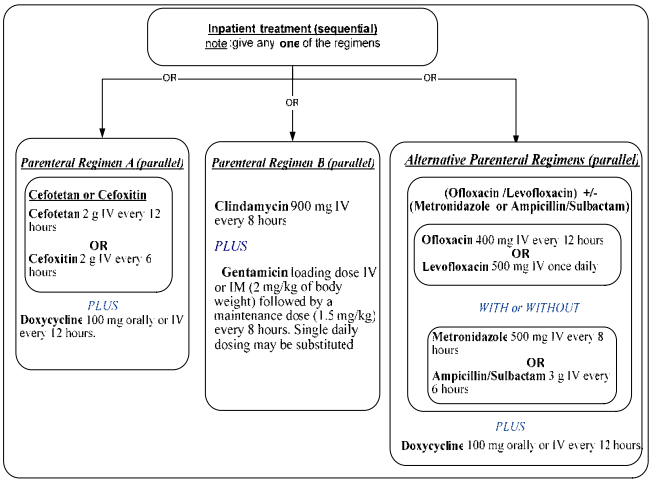

Figure 1. The second stage in forming an ontology specific consensus, in this case, choosing between three regimens for the "Inpatient treatment and evaluation" plan of the "Pelvic Inflammatory Disease" guideline [8]. Each procedural step (regimen) is a task box. Note the semantic order between the regimens ("Or" operator) and the order between the plans in each task box (e.g., parallel or sequential) the knowledge engineer add after defining the clinical consensus. The notations "With or Without" and "Plus" are taken verbatim from the GL source.

3. The EPs and the KE(s) create a table of the relevant knowledge roles (KRs) of the specification language for each step (task box) that they consider as necessary (note that not all the steps will necessarily be detailed here), and describe the content of each KR (e.g., eligibility conditions). The input of the KE is very important: he must revise the semantics of each step, and look for contradictions or special cases where the semantics of the ontology-specific language is not clear to the EP, and is therefore not defined properly. The result is the ontology-specific declarative part pf the consensus (Table 1).

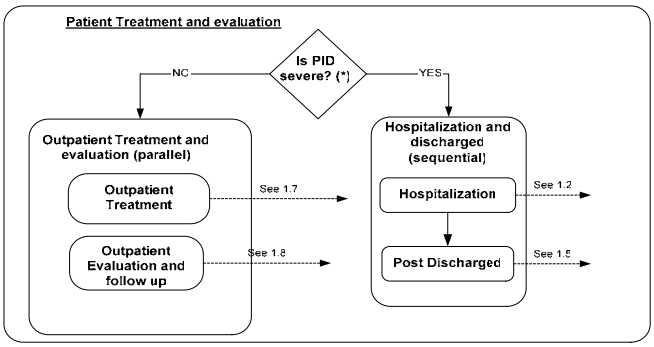

Figure 2. Another example for the second stage in forming an ontology specific consensus, in this case, a condition control structure (if-then-else) for the "patient treatment and evaluation" plan of the "Pelvic Inflammatory Disease" guideline [8]. Each procedural step is a task box. Note the order between the plans in each task box (e.g., parallel, sequential) added by the knowledge engineer after defining the clinical consensus. Note the arrows, pointing to the next sub-steps for further structuring into sub-plans.

Table 1. The declarative part created in third stage in forming ontology specific consensus

Inpatient (IV) treatment Plan - Declarative Content	
Knowledge Role	**Content**
Level of evidence	N.A
Strength of Recommendation	N.A
Actors	Doctors – gynecological ward
Clinical context	Ward
Intention intermediate process	To administer IV drug according to one of the regimens: A,B, Alternative
Intention overall outcome	To achieve substantial clinical improvement (*) **OR** To allow Per Os treatment
Filter Condition	N.A
Set Up Condition	N.A
Abort Condition	N.A
Complete Condition	Substantial clinical improvement (*) **OR** Post IV PO treatment started

Note the * notation which implies a concept (the concept "substantial clinical improvement" in this case) that the editor should elicit as part of the GL knowledge.

4. After some iteration of steps 2 and 3, an OSC is formed, composed from two parts: procedural part and declarative part Table 2 summarizes the steps in the consensus methodology. The N.A stands for non applicable content for its categories KR

As we explained, a methodology for the overall specification process of GLs from a textual representation into a semi-formal representation was developed and evaluated. For the evaluation, three GLs from three different clinical domains were used: Pelvic inflammatory disease (PID) in the Gynecology domain, Chronic Obstructive Pulmonary Disease (COPD) in the Pulmonary domain, and Hypothyroidism in the Endocrinology domain.

Table 2. Summary of steps towards ontology-specific consensus.

Step	The Participants	The process	The Output of this process
1	The EPs of the clinical institution where the GL will be applied in	Describing in flow chart (in task box) fashion the general clinical steps of the GL	Clinical consensus
2	The KEs, who familiar with the specification language together with the EPs	Adding control structure in the notion of the specification language for each step (e.g. parallel, define if the plan is mandatory for application)	The *Procedural* part of the clinical consensus refined with ontology-specific language.
3	The KEs with the EPs	Creating a table of the relevant KRs of the specification language for each defined step (task box) which they decide is necessary	The *Declarative* part of the clinical consensus refined with ontology-specific language.
4	The KEs with the EPs	After some iterations of steps 2 and 3, an Ontology-Specific Consensus (OSC) is formed	Ontology-Specific Consensus (OSC), composed of the two parts: procedural and declerative

For each GL, an OSC was formed in increasing level of detail. for the PID GL the operators between the expressions (such as "And","Or" operators) were excluded, and for the PID and COPD GLs it wasn't denoted explicitly in the OSC which sub-plans are mandatory for execution and which are optional. Thus, we used three OSCs with increasing level of detail: The least detailed OSC for the PID GL, then more detailed OSC for the COPD GL and the most detailed OSC for the HypoThyrd GL. For each GL two markup documents were made by different EPs, after some training. With respect to the size and complexity of the markups: There were 368 KRs which were specified by both EPs in the PID GL; 194 in the COPD; and 96 in the HypoThyrd GL, together, 658 KRs. In addition, Table 3 describes the number of plans and plan types in each GL, categorized by the types of plans in each GL: overall 196 plans were created in the GS, for the three GLs.

2.1 Evaluation Design

In order to evaluate the importance of an OSC, two types of measures were used: Subjective measures and Objective measure:

Table 3. Number of plans and plan-types in each guideline.

	Guidelines			
Plan Type	**PID**	**COPD**	**HypoThyrd**	**Total**
Cyclic (Periodic)	24	2	2	28
If-Then-Else	16	4	6	26
Plan Reference	9	NA	3	12
Parallel Plan	11	8	2	21
Sequential Plan	14	3	4	21
Simple Plan	28	36	14	78
To be Defined	4	6	NA	10
Sum of Plans	106	59	31	196

- **Subjective Measures:** A subjective Questionnaire was administered to find essential aspects which helped the EPs in creating the OSC. Fourteen different aspects, which were considered to be relevant to the process of creating the OSC and the markup, such as the overall process methodology and specification language, were listed. For each concept; a description and an example were also given. The EPs who participated in the process of creating the OSC were asked to score the contribution of each aspect on a scale of -3 [interfered with creating the OSC] to +3 [very helpful].

- **Objective Measures**: To objectively evaluate the role of OSC in the markup specification process, we have classified the markup errors made by the EPs into two categories: (1) ontological errors committed in applying the Asbru syntax and semantics incorrectly, and (2) Clinical errors regarding the medical content. Both measures were in comparison to a Gold Standard (GS) markup prepared by a KE and a senior EP in each GL domain.

In addition, we defined a Markup Error Rate (MER) measure (see formula 1): The Mean number of markup errors committed per KR, per EP, given a particular GLi (MERi). The MER can be used for either ontological or clinical errors as well as for the total number of errors. In addition, to find whether the MER is significantly different among different GLs, we used the proportion test1 [10] for the MER (proportion of errors) between each pair of GLs. (The division by a factor of 2 assumes that two EPs marked-up each GL; in general, the factor of 2 can be replaced by the number of markup editors).

$$\frac{No.of\ errors\ in\ GL_i}{No.of\ KRs\ in\ GL_i * 2} = MERi \qquad (1)$$

[1] The proportion test considers the problem of testing the hypothesis that the proportion of success in a binomial experiment of two populations is equal.

Table 4. The subjective assessment by the EPs of the relevance of various aspects potentially useful for creating the OSC, sorted by level of importance.

Aspect ID	Aspect Name	Aspect Description	EP1	EP2	EP3	EP4	Mean	STDEV
1	Having Medical expertise	All your expertise knowledge regarding the guideline	2	3	3	3	2.75	0.50
2	Reading the guideline sources before making ont. Consensus	Reading the textual content of the guideline	2	2	3	3	2.50	0.58
3	Knowing the multiple representation level model	The Hybrid Representation Model	3	0	3	3	2.25	1.50
4	Understanding Asbru Krs – Procedural part	The clinical pathway described as flow chart	2	2	3	2	2.25	0.50
5	Understanding Asbru Krs - Declarative part	Each plan includes some declerative KRs such as Filter Conditions KR	3	2	2	2	2.25	0.50
6	Using of Ontology	Key concepts, properties and their relations	1	0	2	3	1.50	1.29
7	Having more than one source	The guideline has more then one source	3	3	0	0	1.50	1.73
8	Using DeGeL	The Digital electronic Guideline Library	0	0	3	2	1.25	1.50
9	Using URUZ –main interface	The tool for guideline markup	1	0	0	1	0.50	0.58
10	Using Plan-body wizard	The tool for guideline structuring into tree of plans	1	0	1	0	0.50	0.58
11	Using IndexiGuide	The tool for indexing the guideline	0	0	0	2	0.50	1.00
12	Using Vaidurya	The tool for searching the guideline	0	0	0	1	0.25	0.50
13	Using Vocabulary server	The tool for finding terms In standard vocabularies such as Loinc, CPT	0	1	0	0	0.25	0.50
14	Using Spock	The tool for Applying the guidelines	0	0	0	0	0.00	0.00
		Mean Score	**1.29**	**0.93**	**1.43**	**1.57**	**1.30**	**0.28**

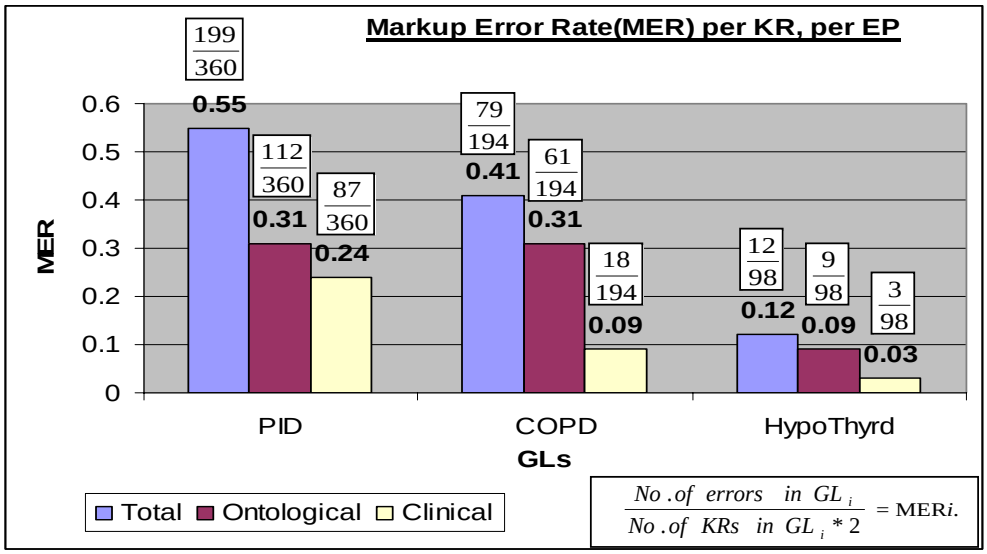

Figure 3. The mean markup error rate (MER) during markup, per KR, per EP, separated into the Clinical, Ontological (Asbru semantics), and overall categories, for each guideline.

3 RESULTS

Using their medical knowledge and their understanding of the specification language (Asbru, in our case) were considered as more helpful by the EPs for creation of an OSC, than understanding the different specification framework tools (DeGeL, in our case) (Table 4). Smaller variability among the EPs was found regarding subjective assessment of aspects that are not helpful (such as comprehension of the DeGeL tools in our case). This can be explained by the fact that the EPs involved in creating the OSC are usually senior EPs with significant clinical knowledge, but with less computer orientation required for

markup. A significant correlation (P<0.05) was found among most of the EPs, supporting this interpretation.

As for the objective measures, we found that the MER of total markup errors committed by the EPs was highest (MER=0.55) in the PID GL (see Figure 3), probably due to its less detailed OSC. The MER decreases as the OSC becomes more detailed, as can be seen in the COPD GL(MER=0.41), reaching a low of 0.12 mean number of errors per KR (per EP) in the HypoThyrd GL, which used the most detailed OSC. The differences in the total MER between the three GLs were highly significant in a proportion test (P<0.001). The difference in the ontological (Asbru semantics) MER between the PID and the COPD GLs was not significant, neither was the difference between the clinical MER of the COPD and the HypoThyrd GLs. The results confirm our hypothesis regarding the importance of the OSC: the more detailed and structured the OSC was, the lower the total number of errors committed by the EPs for each KR in their markupss.

4 DISCUSSION

This study has clearly shown that creating an OSC is an essential step before the GL specification process. Before using the detailed OSC, we first tried using an informal consensus; but made only little progress in the markup phase, due to the multiple ambiguities and missing information inherent in most GLs. Creating an OSC forced the EPs to disambiguate many of the core consensus issues. It was also obvious that without a consensus, each markup editor would create a different version of the same GL, and the comparison to a GS would not be meaningful. Thus, in this research, three different OSCs (one OSC for each GL) at an increasing level of detail were created. The subjective assessments by the EPs marking up the GLs emphasized that using their medical knowledge and their understanding of the GL specification language were considered as more helpful than understanding the specification tools. When assessing objectively and quantitatively the result of the markup, compared to the previously prepared gold standard, we found that the more detailed and structured the OSC was, the less the total (ontological and clinical) error rate per KR (P<0.001).

Thus, our study suggests a general insight regarding the set of tools and frameworks needed to support specification and application of GLs: An OSC is definitely needed; and the OSC should be made as detailed as possible, including all relevant procedural and declarative concepts. This approach has a potential drawback: at the extremism of this approach, the OSC will become very close to the markup itself. However, we believe that in most cases, the OSC will only serve the editors as a very rough draft representing the clinical semantics of the "spirit" of the GL specification in an ontology-specific fashion.

Given our qualitative and quantitative results, we propose, that the OSC should be created initially by a group of EPs of the local medical setting, in collaboration with a KE who is familiar with the specification language. The KE should teach the EPs the core semantics of the specification language in order to create "brain storming" not only between the KEs, but also between the EPs and the KEs. Further research on this front might involve psychologists, to explore the influence of "group thinking," the style of discussion, and the existence of a moderator, to optimize the interaction.

The OSC is independent of the specification tool; thus, the OSC can be used for markup with different specification tools for the same ontology. Thus, an OSC such as was created here can be used when working with another specification tool that can support perhaps the Asbru ontology, such as the GESHER tool [9]. Another future option is to use graphical tools to facilitate OSC creation and specification in multiple GL-specification languages (for example, it might be interesting to create a PID GL OSC in the GLIF ontology).

Finally, re-use and sharing of OSCs among EPs and clinical settings should be supported by saving the OSCs in an appropriate digital library. For example, the DeGeL framework, already used for storage and retrieval of hybrid GL representations at multiple levels of detail, might well be an appropriate medium for that task, perhaps as the next representation level below the free-text guideline source. .

ACKNOLEGMENTS

This research was supported in part by NIH award LM-06806. We also want to thank Dr. L. Basso, and Dr. H. Kaizer for their efforts and contribution for this research, in early evaluations of the markups tools.

REFERENCES

[1] Shahar Y, Young O., Shalom E., Mayaffit A., Moskovitch R., Hessing A., and Galperin M. (2004). A hybrid, multiple-ontology framework for specification and retrieval of clinical guidelines. The Journal of Biomedical Informatics 37(5), 325-344.

[2] Patel, V., Allen V., Arocha J. and Shortliffe E. (1997). "Representing Clinical Guidelines in GLIF: Individual and Collaborative Expertise." J Am Med Inform Assoc.. 5(5): 467-83.

[3] Van Bemmel, J. H. and Musen M. A. (1997). Strategies for Medical Knowledge Acquisition, the HandBook of medical informatics, Heidelberg:Springer.

[4] Peleg, M., Gutnik L., Snow V. and Patel, V. L. (2005). "Interpreting procedures from descriptive guidelines." Journal of Biomedical Informatics: in press.

[5] Shahar, Y., Miksch S. and Johnson P. (1998). "The Asgaard project: A task-specific framework for the application and critiquing of time-oriented clinical guidelines." Artificial Intelligence in Medicine **14**: 29-51

[6] Shalom E. An Evaluation of a methodology for Specification of Clinical Guidelines at Multiple Representation Levels. Master Thesis (2006), Dep. of Information Systems Engineering Ben Gurion University, Beer Sheva, Israel

[7] Patel, V. L., Branch T., Wang D., Peleg M. and Boxwala A. A. (2002). "Analysis of the Process of Encoding Guidelines :An Evaluation of GLIF3." Methods of Information in Medicine 41(2): 102-113

[8] PID-CDC (2002). "Centers for Disease Control and Prevention." http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5106a1.htm.

[9] Shalom, E. and Shahar Y. (2005). A Graphical Framework for Specification of Clinical Guidelines at Multiple Representation Levels. AMIA Annual Fall Symposium, Washington DC, USA.

[10] Walpole, R. E. and H.Myers R. (1978). Probability and Statistics for Engineers and Scientists, Macmillan publishing Co., Inc. New York