First European Workshop on Computer-based Support for Clinical Guidelines and Protocols
Vojtěch Svátek et al.
IOS Press 2001

Guide-X - a Step-By-Step, Markup-Based Approach to Guideline Formalisation

Vojtěch Svátek, Tomáš Kroupa, and Marek Růžička
Laboratory for Intelligent Systems, Faculty of Informatics and Statistics,
University of Economics,
W.Churchill Sq. 4,
13067 Praha 3, Czech Republic
E-mail: {svatek,xkrot04,xruzm06}@vse.cz

Abstract. The main difficulties of converting the original textual form of medical guidelines to a computer-tractable form are connected both with the ambiguity of the natural language text and with the complexity of the resulting formal (and operational) representation. Proceeding directly from one to the other is thus an extremely demanding task. The proposed Guide-X methodology addresses this problem by breaking the whole process of guideline operationalisation into several steps, each of which requires a different mixture of types (medical, knowledge representation, typographical) and degrees of expertise. The principal technology used is that of XML tagging (using both pre-existing and newly developed languages). The result of each step is connected, element-by-element, to the results of previous steps, thus making the verification and revision of the operationalisation process easier. The methodology is currently being tested in the field of hypertension treatment, within the framework of the Medical Guideline Technology project of the EU Fourth Framework Programme.

1 Introduction

Medical guidelines (MG) are standard means for dissemination of medical knowledge and for setting forth healthcare standards. They are applied in various areas of medicine and recognised (although sometimes criticised) by the medical community. In addition to the textual form of MG's, attention is currently paid to the possibility of their *formalisation* and subsequent *computational processing.* Within the framework of current projects (see sections 2.2 and 4), sophisticated languages for formal and operational representation of guidelines have been developed, which allow for automated processing of several guideline-related tasks - advising, critiquing, comparing etc.; see [24] for an overview.

Comparably less attention has been, however, paid to the problem of *transferring* the knowledge contained in the original guideline *text* into these languages. Most approaches assume that the formalisation is based on extensive interaction with a *domain expert* (here, a medical doctor), who takes full responsibility for the quality of the resulting model and for its adherence to the intentions of the guideline. However, the expert may be prone to enter tacitly pieces of his own (or his clinic's) knowledge. In addition, interactive diagrammatical tools and pre-arranged forms are mostly used for input. Since these tools have limited expressive capability compared to free text, this approach is obviously not suitable for tasks in which the original (consensual) text should play the key role. The formalisation process should then be viewed rather as *expert-assisted knowledge extraction from text*, and the loss of information and addition of external information should be minimised.

Since we are dealing with free-text documents on the one end of the process, and with structured, symbolic knowledge bases on the other end, an extremely useful technology seems to be that of *mark-up languages* such as XML. In the computing world, the applications of XML actually range from adding structural and semantic information to textual documents, to describing the structure of database records and expressing complex knowledge ontologies (e.g. the XOL language, see [14]) and knowledge bases.

In this paper, we first briefly characterize the MGT project, which forms the context of the research described in this paper, with particular attention to one of its novelties, the two-tiered model of MG (section 2). Then we formulate our step-by-step methodology, named GuideX, for converting the textual form of medical guidelines into an operational representation, with the help of XML mark-up and related languages (section 3). Finally, we review some related work (section 4), and outline perspectives for our future work (section 5).

2 The MGT Approach to Guideline Computerisation

2.1 General Characteristics of the MGT Project

The *Medical Guideline Technology* (MGT) project of the Fourth Framework Programme is one of the numerous research projects funded by the EC that focus on medical guidelines[1]. It obviously shares many features with other projects: the need of formal representation of MG's, participation of medical doctors as domain experts, and provision of access to electronic patient records (EPR). Nevertheless, specific emphases have been formulated (both before the start of the project, and throughout its course, in reaction to new findings), which form the distinctive face of MGT. Let us summarise them only briefly:

- Although the guideline models developed should be as generic as possible, the first applications of the approach will be in the field of *cardiology*; the pilot application deals with *hypertension treatment.*

It has been pointed out recently (e.g., in [24]) that the types of MG significantly differ in their properties, and that a single, monolithic model is unlikely to capture all of the important features. Instead, more specialised models or sub-models should be designed. Most of the intended applications of the MGT project outcomes are actually concerned with the *type of guideline* denoted as *guidelines for care of clinical condition* in the IOM categorisation [7]. The scope of applicability of this sort of MG spans over multiple (often, a large number of) encounters, and the effects of treatment are observable only after a certain period. Even more specifically, most of the MG's in question deal with

[1]Other guideline-related EU projects not mentioned further in this paper are e.g. PRESTIGE, PROGUIDE, or PROMPT.

drug (and other forms of conservative) treatment rather than with invasive procedures, and with outpatients rather than with in-patients (in a sense, with GP-oriented tasks).

- The project takes into account an important but somewhat neglected group of MG-related tasks to be potentially performed by computer: the (statistical) *analysis of compliance* between current medical practice and *ideal* practice incorporated in the guidelines. A disclosure of non-compliance may entail, depending on the situation, measures for improving the practice and/or suggested revisions of the guideline itself.

Not only the guideline types differ, so do the *tasks* related to their processing (see, e.g., [24] for a sketch of typology). And, obviously, the role of guidelines themselves varies according to the task. While in the direct guideline-based support of treatment the addition of new (institutional or personal) knowledge is more or less desirable (see, e.g., [10]), in the analysis of guideline compliance we would like to keep the guideline statements clearly separate from other knowledge prior to the phase of result interpretation.

- The project takes advantage of the results of other projects previously undertaken by the project partners. The most important ones are the *I4C-TripleC* project oriented toward a highly structured EPR (using the ORCA system [19, 25]), and the *MUM* (Managing Uncertainty in Medicine) project, dealing with probabilistic modelling of medical knowledge (see, e.g., [13]).

Interfacing MG's to *complex data structures* of EPR requires a very thorough analysis of concepts dealt with in the guideline text. Similarly, in order to handle *uncertainty* (both in the MG itself and with respect to missing data), symbolic structures capturing the knowledge extracted from the guidelines have to be properly designed.

- The prototype applications will be created with the help of the OCML - Operational Concept Modelling Language [14].

Since OCML is an extremely powerful but also an appropriately complex language, we have to keep track of all the transformations made in the process of formalisation and operationalisation.

The above requirements have led, among other results, to the recent suggestion of a *two-tiered guideline model* [23], and the *formalisation methodology* (main topic of this paper) described in section 3. We will now summarise the essence of the two-tiered model, which is also the starting point of the methodology.

2.2 The Modular-Structural Dilemma and the Two-Tiered Model

The last decade has seen an intense discussion (see e.g. [5, 18]) about the advantages and disadvantages of two styles of guideline modelling:

- Using *independent modules* (Situation-Action Rules, Medical Logic Modules, Arden Rules) operating in forward-chaining (often event-driven) mode [12].

- Using *complex structures* of interconnected elements (branching logic, state transitions, plans) [6, 17, 18, 21]. This approach also favourites the use of *deep* knowledge, such as domain ontologies, tasks and generic problem-solving methods.

For *guidelines for care of clinical condition*, the latter seems to be an obvious choice, since we have to deal with sequences of tasks, temporal relations, evolution of patient states etc. There are, however, two important points that motivate a certain reconsideration of this conclusion:

- As observed in [5], representation with independent modules is *more intuitive* for healthcare providers, who are, in turn, more apt to contribute their knowledge.
- Compact structures of branching logic are difficult to apply in the situation of *missing data*. Moreover, any *deviation from the logic* leads to subsequent inapplicability for the particular patient - in other words, the individual chunks of knowledge cannot be used per se if the plan is inapplicable as a whole.

The *GuiDE* architecture [5] mainly addresses the first objection, via integrating the two paradigms at the level of user interface. As we are interested in automated comparison of formalised MG's and (large numbers of) EPR's, the second objection is more pressing to us. One type of solution has been proposed in the well-known *EON* [17] project, which understands MG's as collections of *skeletal* plans. The plans are specialised and revised for each patient's visit, using additional knowledge. This approach, supported by multiple knowledge-based tools developed at Stanford Medical Informatics, has proven very powerful in assisting the process of guideline-based decision-making, in several medical applications. The closely related *Asgaard* project [21] aims to discover implicit *intentions* of the guideline and of the healthcare provider and to compare them at an abstract level. For this, powerful mechanisms for conceptual as well as temporal abstraction have been devised. Among the projects using the *plan paradigm*, Asgaard is probably the one most striving to eliminate the potential rigidity of the MG (as a plan structure) via resorting to a goal-oriented and declarative view.

Our proposed solution, denoted as *two-tiered model* [23], is comparably lightweight and seems to be better tailored to the *compliance-analysis* task, in which a *large number of EPR's* has to be processed with *limited user interaction*. It is already at the level of the (informal) *representational ontology* of MG (see [22]) that two groups of elements have been identified:

- The first group (i.e. the first tier of the model) contains independent pieces of knowledge: goals, causal relations, concept definitions, and finally, *scenarios*. The scenarios are similar to Arden rules [12] as well as to the concept of scenarios recently proposed within the EON project [24]: they relate a particular set of conditions (pertaining to patient states and history of treatment) to recommendations, i.e. *actions* and/or *decisions* to be *immediately* performed. Long-term consequences are captured both by the second tier and by the *activities* triggered by certain actions.
- The second group (second tier of the model) consists of graph structures, linking the elements of the first group, in particular the scenarios, together, to reflect the *logic of the steps* corresponding to the (long-term aspect of the) guideline.

In the context of the compliance-analysis task, the interpretation of the MG with respect to a particular EPR consists of two steps:

- In the *first step*, the patient's visits are processed one by one. For each visit, the up-to-then available data (visit data plus history of patient states and treatment activities) are compared with the conditions of scenarios, and the relevant subset of scenarios is linked to the visit. Next, the actions taken during the visit are compared with the recommendation part of the selected scenarios. The result of this comparison corresponds to the *base level of compliance* for this particular visit. After having processed all visits, the results are summarised.

- In the *second step*, the graph structure of scenarios extracted from the MG is compared with the linear structure of visits, understood as a hierarchical graph of sets of scenarios. The result of this comparison corresponds to the *upper level of compliance*.

The pragmatics of this approach are to enable the scenarios to be (even partially) matched with the current state of the patient, regardless of the second tier, which will be applied only posterior to the processing of the whole EPR (in a way, this bottom-up approach is inverse to the *skeletal-plan-refinement* operator in EON [17]). In other words, should a portion of cases (i.e. EPR's) *leak out* from the *state-transition* structures implicitly described in the guideline[2] , it does not necessarily get completely beyond its scope, since base-level compliance can still be applied. Minor deviations (e.g. in action sequencing) thus will not obscure serious infringements of the MG.

There is a certain duplication of information, since long-term statements are reflected both in the second tier and in the first tier (references to history and to triggered activities, in the scenarios), special care has thus to be taken to maintain consistency. It is, however, not a total overlap, as the first tier should contain only the *critical requirements of applicability* for each scenario, while the second tier can also reflect *usual* or *implicit* tracks of diagnosis and treatment. Obviously, the distinction of *critical* and *non-critical* requirements can bring back the subjective bias we wanted to eliminate; this undesirable feature is alleviated by the fact that the distinction only influences the level (base vs. upper) on which the non-compliance is detected.

3 Overview of the Guide-X methodology

3.1 Motivation for the Step-By-Step Approach

Most approaches to guideline computerisation assume extensive interaction between the domain expert and the knowledge engineer, the formal model of the guideline being created in a way similar to traditional medical knowledge bases. Such a *one-step* formalisation process, however brilliant its output may be, remains quite tedious, and requires a *mixture of expertise*: medical expertise, general knowledge engineering skills, as well as expertise in handling the particular target formalism. In addition, the dominant role of the expert leads to (sometimes uncontrollable) addition of *personal*, extra-guideline knowledge to the model. This may be desirable if the task is to build a pragmatic system for medical decision-making but is completely misleading if the task is, for example, to *compare* the actual medical practice (reflected in EPR's) with the standards set by the guidelines (see section 2.1).

Breaking the formalisation process down into multiple steps can be helpful both in *separating* the different types of expertise required into relatively independent batches, and in making the process more transparent and allowing for easier *verification* and correction of misinterpretations. Knowledge added by the expert can be explicitly pointed out.

[2] This may occur quite frequently when comparing the actual medical practice, in its diversity, with generic guidelines (such as those of the WHO).

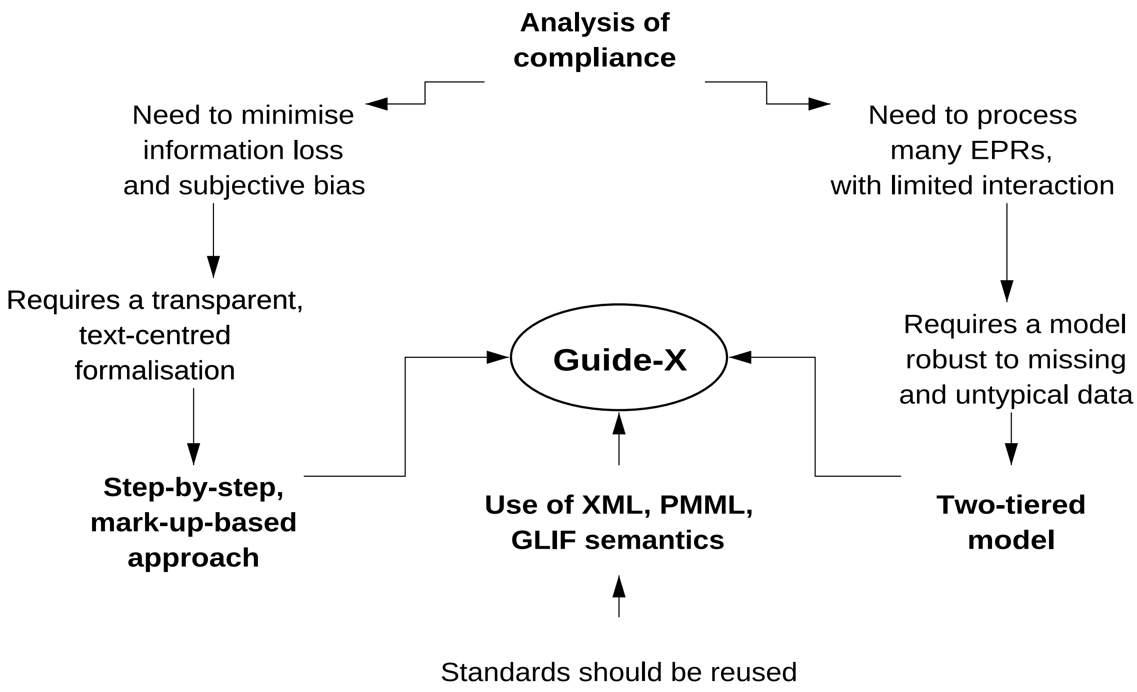

Figure 1: Context of the Guide-X methodology

The formalisation methodology we are currently developing under the name of Guide-X (for Guide*line formalisation based on* X*ML*) consists of five steps, each of which will be described in more detail. Their outputs are, in turn: a well-structured XHTML document, a document with coarse-grained semantic tagging (GLML-S language), a document with rich semantic tagging (GLML-R language), a systematically arranged XML knowledge base (GLKL language), and, finally, the computational representation, which is currently based on OCML (Operational Concept Modelling Language). All but the last are strictly XML-based. It should be noted that, in particular, the latter steps of the methodology are not yet properly supported by complete language definitions and software tools, and can be subject to future modifications and improvements.

The diagram in figure 1 shows the methodology in its context, formed in particular by the compliance-analysis task and by its requirements.

3.2 XHTML as Input Text Format

Since our methodology is XML-based, a natural choice for the initial text format was XHTML[3]: an XML-valid adaptation of the widely used HTML (HyperText Mark-up Language). Its main advantages, within the context of Guide-X, are as follows:

- It provides for straightforward (formatted) *display* with most (up-to-date) web browsers, like ordinary HTML. The documents can thus be stored within ordinary web sites, e.g., those of medical consortia.

[3]See `http://www.w3.org/1999/xhtml`.

- It can be processed by XML *editors* in a way similar to subsequent *semantic* languages.
- It can be *referenced* by these languages using the XLink-XPointer [7, 8] technology.
- The creation of an XHTML document requires *little expertise* beyond the quite common web page design skills.

Depending on the initial form of the guideline document in question, the *0th* step of the formalisation process thus amounts to either:

- Typewriting the XHTML document from scratch (for paper documents - unless OCR can be applied).
- Addition of XHTML tags to plain text documents.
- *Canonisation* of older HTML documents: this task can be partially automated using software tools (such as the publicly available HTMLTidy program[4] that can to a certain extent resolve crossed tags and other ambiguities), the result should, however, still be verified by a human.

To guarantee *uniform output formatting* for medical guideline documents encoded in XHTML, a dedicated Cascading Style Sheet (CSS) has also been created.

3.3 GLML-S: Coarse-Grained Semantic Element Mark-Up

The XHTML document resulting from the previous step is well structured - not only in the sense of XML validation, but also in the common sense of formal document structuring. It, however, typically contains parts that are not likely to be exploited in the operational code, such as results of clinical studies (giving empirical support to important assertions), or even ad hoc illustrations. We assume that, if the first addition of *semantic information* into the guideline document is carried out carefully enough, it can be done even by persons without (deep) medical expertise. Real medical experts can thus be relieved of scrutinising long documents with little tangible effect, and their involvement will be postponed to later phases.

According to the *representational ontology* developed within the MGT project (see [22]), we distinguish among four types of elements:

- Definitional statements, i.e. definitions of *concepts* dealt with in the text.
- Purely descriptive statements in the form of *causal relationships* among various clinical parameters, events etc.
- Intentional statements, i.e. *goals* to be achieved (or states to be maintained or avoided). Some are related to the guideline (-based treatment) as a whole, some to particular actions or activities within.
- *Procedural* statements: what should be done under which conditions, i.e. parts of text likely to produce scenarios. Since the scenarios are the essential *knowledge fuel* for guideline-processing software, procedural statements have to be given particular attention.

[4] At `http://www.w3.org/Status.html#TIDY`.

We assume, also by comparison with other approaches and methodologies, that these four rather general types cover more or less any important statement that occurs in guidelines for *long-term care for a clinical condition.* For example, even the (recently introduced) Unified Service Action Model [20], which is significantly more ambitious than our model, and aims to cover, in addition to MG, also the statements from EPR and workflow, is essentially based on the notions of event, definition, intent, goal and criterion (the so-called *action moods*).

It should be noted that the boundary between the element types is not always clear: concept definitions can have procedural aspects, while some causal relationships and procedural statements can themselves be understood as concepts etc. These problems have, however, little to do with medical expertise, and should be systematically solved within the knowledge-engineering endeavour.

For this *crude* semantic annotation, we have developed an XML-based mark-up language named GLML-S (Guideline Mark-Up Language - Simple). Its DTD defines five top-level elements: `goal`, `causal-rel`, `concept-def`, `procedural` and `other`. Pieces of text conforming to any of the ontological notions are enclosed in the respective tags and inserted into the GLML-S document, while the rest of the text is left out. Minor reformulation is sometimes needed in order to pick up relevant phrases consistently, out of a complex sentence. In figure 2, you can see how two instances of basic GLML-S elements are extracted from a single paragraph of the XHTML document.

The `other` element is destined to cover pieces of text that can be potentially formalised and processed by computer, but do not conform to the currently used guideline model. In the pilot application of this methodology, we have come across two situations of this kind:

- the statement does not *conceptually* conform to any of the predefined notions: e.g. meta-statements about the guideline itself;
- the statement is conceptually related to some of the notions but at a different level of *granularity*: e.g. detailed instructions for performing procedures and measurements, which (though being *procedural statements*) cannot be handled in a model with *visit* as the smallest time unit.

The effect of using `other` (instead of leaving such text out) is to minimise the text-analysis effort incurred by submitting the same guideline document into a different (or, enriched) model. Some new elements could then be extracted directly from GLML-S, without resorting to the full text.

For all basic elements, the *XPointer* language [8] embedded in *XLink* [7] references is used to indicate the location of the original text; see again, figure 2. Here, the location is defined as "the fifth paragraph following the element (a second-level heading) with attribute *id* set to *chp1*." Since the sectioning is done with header elements in the XHTML file, the paragraphs within a chapter are counted from the heading forward; naturally, the paragraphs could have labels as well.

In addition to concept definitions, GLML-S includes also *concept references*: any piece of text within basic elements can be marked with the `con` tag to indicate that the text *may* refer to a concept. At this stage, not all concepts referred to also have to be defined; the reference is, in principle, based on the embedded text, and the inclusion of an attribute pointing directly at the ID of the concept definition is merely optional.

In our experiments with the 1999 WHO-ISH Hypertension Guidelines [1], the first conversion step (XHTML to GLML-S) has shrunk the original document to less than

```
...
<h2 id="chp1"> What is high blood pressure and hypertension </h2>
...
<p>
In some patients, office (or clinic) BP is persistently elevated
whereas daytime BP outside the clinic environment is normal.
There is continuing debate as to whether {\em isolated} office hypertension
({\em white coat hypertension}) is an innocent phenomenon or whether it
carries an increased burden of cardiovascular risk.
</p>
...
```

```
...
<concept-def
id="cd4"
xlink:href="who99mid.xhtml#xpointer(id('chp1')/following::p[5])"
xlink:role="src">
"Isolated" office <con>hypertension</con>
("white coat <con>hypertension</con>"):
in some patients, office (or clinic) <con>BP</con> is persistently
elevated whereas daytime <con>BP</con> outside the clinic environment
is normal.
</concept-def>

<causal-rel
id="cr3"
xlink:href="who99mid.xhtml#xpointer(id('chp1')/following::p[5])"
xlink:role="src">
There is continuing debate as to whether "isolated" office
<con>hypertension<con> ("white coat <con>hypertension</con>")
is an innocent phenomenon or whether it carries an increased burden
of <con>cardiovascular risk</con>.
</causal-rel>
...
```

Figure 2: Conversion from XHTML to GLML-S

50% of the size. We assume that this is characteristic for *long-term patient management guidelines* (cf. [23]), since long-term effects of treatment are less obvious and have to be documented with additional (esp. statistical) material. Since we do not have to persuade the computer to trust the guidelines with the help of statistical figures, these parts of the text can be safely omitted.

3.4 GLML-R: Breaking the Elements Down Further

As soon as we have identified the key elements in the text (thanks to GLML-S), we can proceed to their further elaboration. Basic elements are refined to *sub-elements*, e.g. the `causal-rel` element is refined to sub-elements specifying the cause and the effect, and optionally the context (conditions of validity) and time specification (e.g., delay of effect). The respective declaration in the GLMLR DTD looks as follows:

```
<!ELEMENT causal-rel (context?, cause, effect, time?)>
```

The `concept-def` element is broken out as follows:

```
<!ELEMENT concept-def (name, sub, alias*, spec?, def)>
```

The sub-elements correspond to a (chosen) basic name of the concept, a reference to the concept the current concept is subordinated to, an alternative name/s, a semantic specification (for polysemous terms), and a textual definition of the concept.

The `procedural` elements are replaced with proto-scenarios. The structure of the proto-scenarios (already denoted by the `scenario` element) essentially corresponds to that of fully developed scenarios. However, some more complex proto-scenarios (corresponding to long-term sequences of actions and states) are bound to be decomposed in the next step of conversion.

Important components of many scenarios are complex decision structures. So far, we have considered the two that are most common in the medical environment: *decision graphs* and *heuristic choice*. For the former, we partially reuse the existing definition of *decision tree* embedded in the *Predictive Model Mark-up Language* (PMML) [2]. From this language, we also borrow the notion of *Data Dictionary* that lists all variables and their values required for decisions. Our Data Dictionary, as part of the GLML-R document, serves as one of the starting points for interfacing to the EPR stored in a database. Since PMML is well on its way to becoming the world-wide representational standard for interchange of *data-mining results*, our adoption of PMML-compatible mark-up seems to be favourable for our future efforts toward EPR mining.

Fragments of the 1999 WHO hypertension guidelines expressed in GLML-R are shown in figure 3, 4 and 5.

Figure 3 shows three *concept definitions*: two definitions of the concept of *hypertension* and the refined definition of the *isolated office hypertension* concept we have already seen for GLML-S. Obviously, the two definitions of *hypertension* are due to the polysemy of the term: increased *value* of blood pressure obtained within a measurement, vs. persistent *diagnosis* that can be assigned to a patient. The concept references are marked more thoroughly: for example, the terms *elevated* and *normal* are identified as fragments of *elevated blood pressure* and *normal blood pressure*, in the first definition.

Figure 4 shows the complex structure of a *goal*: the complexity is partly due to the limited ontological expressiveness of DTDs (see [11] for an analysis of XML as a

```
<concept-def added="parts" id="cd1" type="syndrome"
xlink:href="#xpointer(concept-def[@id='cd1'])" xlink:role="src">
  <name>hypertension</name>
  <alias>chronic elevation of <con>blood pressure</con></alias>
  <spec id="diag">hypertension as diagnosis</spec>
  <def><con sem="value">hypertension</con> obtained several times on
    several separate occasions</def>
</concept-def>
...
<concept-def added="parts" id="cd2" type="finding"
xlink:href="#xpointer(concept-def[@id='cd2'])" xlink:role="src">
  <name>hypertension</name>
  <sub>blood pressure</sub>
  <alias>elevation of <con>blood pressure</con></alias>
  <spec id="value">hypertension as result of <con>blood pressure
        measurement</con></spec>
  <def><con>systolic blood pressure</con> greater or equal to 140 mmHg
    or <con>diastolic blood pressure</con> greater or equal to 90 mmHg
  </def>
</concept-def>
...
<concept-def id="cd4" type="syndrome"
xlink:href="#xpointer(concept-def[@id='cd4'])" xlink:role="src">
  <name>isolated office hypertension</name>
  <sub><con sem="diag">hypertension</con></sub>
  <alias>white-coat hypertension</alias>
  <def>In some patients, office (or clinic) <con>blood pressure</con>
    is persistently <con frag="yes">elevated</con> whereas daytime
    <con>blood pressure</con> outside the clinic environment is
    <con frag="yes">normal</con>. </def>
</concept-def>
```

Figure 3: Concept definitions from a GLML-R document

```
<goal overall="no" id="g5"
xlink:href="#xpointer(goal[@id='g5'])" xlink:role="src">
 <goal-of><action>
   <data-coll><con>clinic and laboratory evaluation</con></data-coll>
 </action></goal-of>
 <is-goal><state>
   <parameter>secondary causes of <con sem="diag">hypertension</con>
   </parameter>
   <value>excluded or identified </value>
 </state></is-goal>
</goal>
```

Figure 4: Goal from a GLML-R document

knowledge representation language). The goal relates the intention (state to be reached: *secondary causes of hypertension* diagnosis *have either been excluded or identified*) to a particular action (*clinic and laboratory evaluation*).

Finally, figure 5 shows a relatively simple *scenario* for deciding about the suitability of starting the (long-term) *activity* of ambulatory BP monitoring. The start itself is viewed as (one shot) *action* - this distinction of actions or activities is similar to the one recently introduced in the EON project [24]. One of the important conditions is the occurrence of *isolated office hypertension* defined above.

Since the transition from GLML-S to GLML-R is quite sophisticated, it can rarely be reliably accomplished without (although limited) participation of a *medical expert*. The Guide-X methodology assumes that the expert inspects the proposed GLML-R structures in interaction both with the original XHTML (and, possibly, GLML-S) document, and with the knowledge engineer responsible for the transition. In order to focus the inspection on the most problematic parts of the model, all basic elements can be characterised according to the amount of *external knowledge* added. The attribute `added` can utilize the values `no` (text without modification), `interp` (text has been reformulated using the most likely linguistic interpretation), `parts` (part of the text has been added) or `whole` (the whole element has been added).

The elements of GLML-R *refer*, via XPointer-XLink references[5], to the corresponding elements in GLML-S.

3.5 GLKL: Systematic Reordering and Normalisation

While both GLML languages are, at least to a certain extent, document-markup languages in the sense of tagging textual fragments in a document (though throwing parts of it away), the transition to the *Guideline Knowledge Language* (GLKL) completely abandons the document structure in favour of systematic ordering of knowledge. This process is likely to be done more-or-less automatically in the future.

In addition to reordering, references will be verified, updated if necessary, and ambiguities resolved. In concept references, *aliases* of the concept will be replaced with

[5]For the sake of brevity, we have included only local pointers; the identity of the source document can be declared in the header of the GLML-R document.

```
<scenario id="s2"
xlink:href="#xpointer(procedural[@id='p1'])" xlink:role="src">
<descr> When ambulatory BP monitoring should be considered  </descr>
 <cond> <state>
  <compound-predicate bool-op="or">
   <predicate attribute="BP variability" op="eq" value="yes" />
   <compound-predicate bool-op="and">
    <predicate attribute="Isolated office hypertension" op="eq"
    value="yes" />
    <predicate attribute="Risk level" op="eq" value="low" />
   </compound-predicate>
   <predicate attribute="Hypotensive episodes" op="eq" value="yes" />
   <predicate attribute="Drug resistant hypertension" op="eq"
    value="yes" />
  </compound-predicate>
 </state> </cond>
 <recom> <action type="start">
  <activity type="monitoring">ambulatory BP monitoring </activity>
 </action> </recom>
</scenario>
```

Figure 5: Scenario from a GLML-R document

basic concept names (according to the concept definitions). References will also be defragmented, even at the cost of reformulation, as shown in figure 6 (again, for the example of isolated office hypertension). Finally, complex *proto-scenarios* with long-term reach are broken down to visit-level scenarios (first tier), and their long-term streams are filtered out into graph structures (second tier). Some of these activities will require substantial involvement of the medical expert.

The structure of GLKL DTD has not been completed yet; it will be largely based on the GLML-R DTD.

3.6 From GLKL to Operational Code

The last step of the methodology consists of two parts.

```
<concept-def id="cd4" type="syndrome">
  <name>isolated office hypertension</name>
  <sub><con sem="diag">hypertension</con></sub>
  <alias>white-coat hypertension</alias>
  <def>Persistent <con id="cd2">hypertension</con> in the office
  or clinic, and <con id="cd3">normal blood pressure</con>
  outside the office or clinic.</def>
</concept-def>
```

Figure 6: Concept definition from a GLKL document

First, the GLKL knowledge base will be *syntactically* converted from the XML format to the format of the OCML conceptual modelling language [15], and references to GLKL elements will be encapsulated in specific slots of OCML classes. This can be done fully automatically; we are currently testing whether the declarative apparatus of XSL style sheets is sufficient for this purpose, without the need of developing dedicated software.

Second, many sub-elements of the GLKL code (now converted to arguments of OCML relations) still contain natural language text, which has to be interpreted and operationalised by a knowledge engineer, with respect to available database (Electronic Patient Record) items, and possibly, again, with the assistance of a medical expert.

The most sophisticated aspect of the conversion will be the transformation of tree structures of XML to OCML classes and relations. We assume that *concept definitions* will be transformed directly to OCML classes, while other elements will be broken down to several OCML relations connected via foreign keys. An example of an operationalised concept definition is in figure 7, again for the concept of isolated office hypertension.

The first part of the definition contains the *sufficient* condition for a patient having isolated office hypertension: all measurements of blood pressure in the office result in hypertension, while all home measurements of blood pressure result in normal blood pressure (this is of course an idealistic condition, which should be further refined). The second part conserves the values of sub-elements, which have been marked-up in previous phases.

3.7 Graphical Summary of the Process

The step-by-step procedure for guideline operationalisation is depicted in figure 8. Each of the *formats* is connected to its *definitions* (pre-existing ones in the upper part, newly created ones in the lower part), and each of the transformation steps is labelled with the symbolic figures of persons involved (medical expert, knowledge engineer, document designer). Processes and definitions that are still the subject of intensive research are distinguished by dashed lines.

4 Related Work on Medical Guideline Mark-up

Since we have briefly reviewed the state of the art in computerised MG processing (with emphasis on the *modular and structural* issue) already in section 2.2, we will now limit ourselves to projects and methods relying on MG document and data *mark-up*.

The *art* of mapping fragments of free text to formal knowledge elements (originally for the purpose of session transcript analysis) has a long tradition in *knowledge acquisition*, and has been more recently *rediscovered* in *knowledge management*, to capture corporate knowledge encapsulated in documents. In the *PatMan* (Patient Workflow Management) project [1], mapping is created between medical domain ontologies[6] and the guideline text. Unlike our project, that approach is mainly top-down, i.e., formal knowledge elements are populated with appropriate fragments found in the text. In our opinion, this is not suitable for extraction of *all available knowledge* from text. It should, however, be noted that the aim of the mapping is to *verify* the presence of

[6]The fact that PatMan uses OCML [15] as the target modelling language is another common point with the MGT project.

```
(def-class has-isolated-office-hypertension (patient) ?p
  :sufficient
  (and (forall (?m) (=>
                   (and
                    (measurement ?m)
                    (of-patient ?m ?p)
                    (of-parameter ?m bp)
                    (or
                     (in-location ?m office)
                     (in-location ?m clinic))
                    (has-value (?sbp ?dbp))
                    )
                   (hypertension ?bp sbp ?sbp dbp ?dbp)))
       (forall (?m) (=>
                   (and
                    (measurement ?m)
                    (of-patient ?m ?p)
                    (of-parameter ?m bp)
                    (not (in-location ?m office))
                    (not (in-location ?m clinic))
                    (has-value (?sbp ?dbp))
                    )
                   (normal-blood-pressure ?bp sbp ?sbp dbp ?dbp)))
  )
  (
   (has-element-status :value concept-def)
   (has-addition-status :value interp)
   (has-name :value "isolated office hypertension")
   (has-sub :value "hypertension")
   (has-type :value syndrome)
   (has-alias :value "white-coat hypertension")
   (has-def :value
            "hypertension in the office or clinic and
             normal blood pressure outside the office or clinic")
   )

)
```

Figure 7: Example of a final OCML code

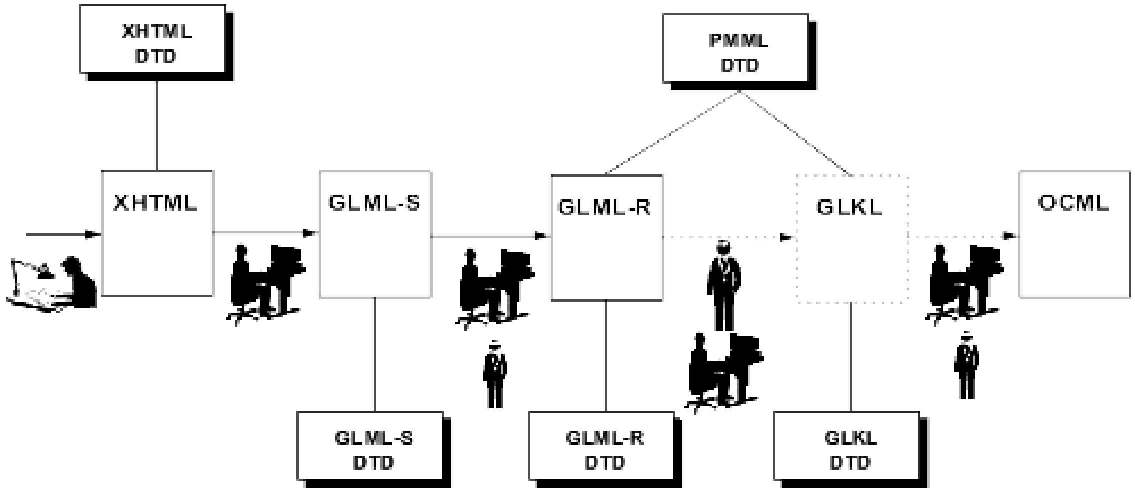

Figure 8: Steps of the Guide-X methodology

ontological concepts in the text rather than to *convert* the textual information into a knowledge model (see [16]).

XML, which is becoming a worldwide standard, also appears more and more frequently as a data format. The *Guideline Interchange Format* (GLIF) [18] developed within the *InterMed* collaborative project of four American universities, uses the XML format as an alternative to the original ODIF format. Also the above-mentioned PatMan project expects to adopt XML in the modelling process, in the future [1]. In our project, we use XML not only as a data representation (and interchange) format but attempt to use systematically other features of XML - document tagging, linking, and styled output.

5 Conclusions and Future Work

In the paper, we have described a methodology for converting medical guideline documents from *textual form* into a *formal representation* in several, relatively independent steps. The key technology in the conversion process is XML based languages.

The main advantage of the step-by-step approach is *better control* over the conversion process, and *distribution of expertise* required in each step. The bottom-up, text-focused (rather than expert-focused) process should render the target representation of the guideline to a form favourable for the *analysis of compliance* task, which would be difficult to achieve by other state-of-the-art approaches.

The methodology is so far supported only by declarative tools - Document Type Definitions (DTDs) and styles. Although general XML editors are well suitable for the knowledge-engineering part of guideline encoding, they are not ideal for medical experts. Further work should thus include the development of *dedicated interfaces* for creation and inspection of annotated documents. Improved use of the methodology by medical experts will then provide systematic feedback for further improvements of the formalisms.

In order to fill the gaps (due to uncertainty and vagueness) in the textual form of guidelines, consensual *medical background knowledge* rather than private knowledge of a single expert has to be elicited. In case the consensus cannot be fully reached, we intend to provide the possibility of recording *fuzzy* values for various parameters. This will naturally impose new requirements on the computational apparatus applied for operational MG processing.

The work on XML mark-up will also converge with the work on *database* (EPR) *interface* and on the OCML *library of medical guideline concepts* [26] within the MGT project. The methodology should incorporate the know-how of matching the concepts referenced or defined in the text with library concepts and fields in the structured EPR.

Acknowledgement

The work on this paper has been partially supported by the MGT (Medical Guidelines Technology) Project No. ICT15-CT98-0315 of the 4th Framework Program of the EC, by the grant no. 201/00/D045 (Knowledge model construction in connection with text documents) of the Grant Agency of the Czech Republic, and by the grant no. VS96008 (Laboratory for Intelligent Systems) of the Czech Ministry of Education.

The authors would like to express their thanks to their colleagues from the MGT project - R. Jiroušek, A. Říha, T. Zíka, and J. Zvárová, for the feedback on ideas expressed in this paper, as well as to J. Kosek for his assistance in getting acquainted with the intricacies of XML.

References

[1] HC 4017 PatMan. Final Report.
`http://aim.unipv.it/projects/patman/ExecSumV5.pdf`

[2] PMML 1.1 – Predictive Model Markup Language. Online at
`http://www.dmg.org/html/pmml_v1_1.html`

[3] WHO/ISH Hypertension Guidelines for the Management of Hypertension. *Journal of Hypertension*, 17, 1999, 151-183.

[4] Clark J.: XSL Transformations (XSLT) Version 1.0. W3C, 1999.
`http://www.w3.org/TR/xslt`

[5] de Clercq P. A., Blom J.A., Hasman A., Korsten H.H.M. Task-Specific Models for the Acquisition and Execution of Clinical Guidelines in High-Dependency Environments. Submitted to *Medical Informatics*.

[6] Dazzi L., Fassino C., Saracco R., Quaglini S., Stefanelli M.: A Patient Workflow Management System Built on Guidelines. Proceedings of AMIA'97, Nashville 1997, 146-150.

[7] DeRose S., Maler E., Orchard D., Trafford B.: XML Linking Language (XLink) – W3C Working Draft. W3C, 1999. `http://www.w3.org/TR/xlink`

[8] DeRose S., Daniel R., Maler E.: XML Pointer Language (XPointer) – W3C Working Draft. W3C, 1999. `http://www.w3.org/TR/xptr`

[9] Field M. J., Lohr K. N. (eds.): Guidelines for Clinical Practice: From Development to Use. Washington, DC: National Academy Press, 1992.

[10] Fridsma D.B., Gennari J.H., Musen M.A.: Making Generic Guidelines Site-Specific. In: Proc. 1996 AMIA Annual Fall Symp. (Cimino J.J., ed.), 597-601.

[11] van Harmelen F., Fensel D.: Practical Knowledge Representation for the Web. In: Proc. IJCAI-99 Workshop on Intelligent Information Integration.

[12] Hripcsak G.: Rationale for the Arden Syntax. Comput.Biomed.Res. 1992, 25, 435-467.

[13] Jiroušek R., Kushmerick N.: Constructing probabilistic models. International Journal of Medical Informatics. Vol.45, Nos.1,2. pp. 9-18.

[14] Karp P. D., Chaudhri V. K., Thomere J.: XOL: An XML-Based Ontology Exchange Language. In: Bio-Ontologies'99 Meeting, 1999.
`ftp://smi.stanford.edu/pub/bio-ontology/`

[15] Motta E.: Reusable Components for Knowledge Modelling: Principles and Case Studies in Parametric Design. IOS Press, Amsterdam, 1999.

[16] Motta E., Buckingham-Shum S., Domingue J.: Ontology-Driven Document Enrichment: Principles and Case Studies. In Proceedings of the 12th Workshop on Knowledge Acquisition, Modeling and Management (KAW '99). Banff, Canada, 1999.

[17] Musen M.A., Tu S.W., Das A.K., Shahar Y.: EON: A Component-Based Approach to Automation of Protocol-Directed Therapy. JAMIA 3:367-388, 1996.

[18] Ohno-Machado L., Gennari J.H., Murphy S., Jain N.L., Tu S.W., Oliver D.E., Pattison-Gordon E., Greenes R.A., Shortliffe E.H., Barnett G.O.: The GuideLine Interchange Format: A Model for Representing Guidelines. Journal of the Americal Medical Informatics Association 1998;5(4):357-72.

[19] Pierik F.H., van Ginneken A.M., Timmers T., Stam H., Weber R.F., Restructuring Routinely Collected Patient Data: ORCA Applied to Andrology. Methods Inf Med 36, 1997 184-190.

[20] Schadow G., Russler D. C., McDonald C. J.: Conceptual Integration of Guidelines and Workflow into the Electronic Health Record. In: MIE2000/GMDS2000, Medical Infobahn for Europe, IOS Press, 807-811.

[21] Shahar Y., Miksch S., Johnson P.: The Asgaard Project: A Task-Specific Framework for the Application and Critiquing of Time-Oriented Clinical Guidelines. *Artif. Intelligence in Medicine* 14:29-51, 1998.

[22] Svátek V., Říha A., Zíka T., Zvárová J., Jiroušek R., Zdráhal Z.: Informal, Formal and Operational Modelling of Medical Guidelines. In: HRUKA, Tom, HASHIMOTO, Masa-aki (ed.). Knowledge-Based Software Engineering. Amsterdam: IOS Press, 2000, 9-16.

[23] Svátek V., Zvárová J., Jiroušek R.: A Two-Tiered Model of Medical Guideline. In: MIE2000/GMDS2000, Medical Infobahn for Europe. (CD-ROM). Berlin: Quintessenz Verlag, 2000.

[24] Tu S. W., Musen M. A.: A Flexible Approach to Guideline Modeling. 1999, Tech.Rep. SMI-1999-0789.

[25] Zvárová, P. Hanzlíček and V. Přibík, Application of ORCA multimedia EPR in Czech hospitals. In: Proceedings of Third Eur.Conf. on Electronic Health Care Records, Sadiel, Sevilla 1999, 160-165.

[26] Zvárová J., Doležel T., Peleška J., Říha A., Svátek V., Jiroušek R., Kroupa T., Růžička M., Štochl J.: Medical Guidelines on Hypertension: Formalization, Interpretation and Execution. Submitted to MEDINFO2001, The 10^{th} World Congress on Medical Informatics, London 2001.