WITH: a System to Write Clinical Trials Using XML and RDBMS

Paola Fazi*, Daniela Luzzi#, Mariarosaria Manco^, Fabrizio L. Ricci+, Giovanni Toffoli^, Marco Vignetti*

* Dip. di Biotecnologie Cellulari e Ematologia, Università Roma – La Sapienza, Via Benevento, 6, 00161 Rome, Italy
# ISRDS-CNR, Via C. De Lollis, 12 – 00185 Rome, Italy
^ Link s.r.l., Via Udine, 30 – 00161 Rome, Italy
+ IBFSNC-CNR, Viale R. Margherita, 6 – 95123 Catania, Italy

Abstract
The paper illustrates the system WITH (Write on Internet clinical Trials in Haematology) which supports the writing of a clinical trial (CT) document. The requirements of this system have been defined analysing the writing process of a CT and then modelling the content of its sections together with their logical and temporal relationships. The system WITH allows: a) editing the document text; b) re-using the text, and c) facilitating the cooperation and the collaborative writing. It is based on XML mark-up language, and on a RDBMS. This choice guarantees: a) process standardisation; b) process management; c) efficient delivery of information-based tasks; and d) explicit focus on process design.
Keywords: clinical trial, collaborative writing, XML

1. Introduction
The Italian Group for Haematological Disease of Adults (GIMEMA) involves about 100 Italian centres, participating into studies mainly for acute leukaemia. It joins EORTC (European Organisation for Research and Treatment of Cancer) Leukaemia Group for several international cooperative clinical trials (CTs). GIMEMA defines and manages several CTs active at the same time. It is organising a kind of national register of acute leukaemia and to have data on the ratio of eligible population above the total population referred to a given centre with a given disease. GIMEMA is therefore interested in developing a system which supports the entire life cycle of CTs, helps to define and write a new CT based on standard procedures, and to re-use parts of ongoing or already closed CTs. Another important task of GIMEMA is the CT diffusion.

A CT is defined under two related perspectives: a) a clinical study, and b) a study report. In the first one a CT is “any investigation in human subjects intended to discover or verify the clinical, pharmacological and/or pharmacodynamic effects of a product [...] with the object of ascertaining its safety and/or efficacy” [1]. In the second one it is defined as “a written description of a trial of any therapeutic, prophylactic, or diagnostic agent conducted in human subjects, in which the clinical and statistical description, presentation and analyses are fully integrated into a single report” [1].

Our research project takes both aspects into consideration [2] by integrating the technologies of WfMS and RDBMS with the use of XML. In this paper, however, our attention is focused on the CT document and in particular on the analysis, which has brought us to define a XML document model and to develop the WITH (Write on Internet clinical Trials in Haematology) system which supports the editing of a CT. The aim of the WITH system is to help the end-user to write the CT document guiding him/hem in the preparation of standard and/or mandatory sections, controlling the internal consistency of the text and finally improving its diffusion on the network according to different views.

2 Preliminary considerations
In the literature many studies have drawn attention on clinical guidelines (CGs). Various models have been used to represent CGs in their algorithmic parts, for example using flowchart [3], Arden syntax [4], state transition diagram [5], decision table [6], etc. Currently XML is asserting itself as an exchange format, to represent and disseminate the knowledge contained in a CG (for instance [6,7,8,9]), also because it facilitates the interoperability among different systems.

It is however worth noting that CTs are different from clinical CGs, in particular for:
- their use - a CT has to be rigidly followed because it identifies an uniform clinical behaviour which has to be adopted by the different physicians carrying out the test; while CGs contain recommendation to improve the treatment;
- their aim – in the case of CTs it is necessary to obtain a correct evaluation of the investigation, while CGs are useful to diffuse good medical practice.

The two perspectives mentioned above make it necessary to describe them by means of two classes of formalisms.

We use:
- Considering the clinical study (algorithmic perspective) - a formal description of the clinical activities carried out during the investigation, which is represented by a conceptual workflow (WF) model in order to guarantee the correct execution of
each task performed by the different participating centres.

- Considering the study report (textual perspective) -
  a structural description of the CT document which
  has to respect the rules of "good clinical practice"
  and to bind the behaviour of each one taking part in
  to the test. In this case we have adopted XML.

3 Methodology adopted
In order to define the WITH system requirements we have
analysed the writing process of a CT and then modelled
the content of its sections together with their logical and
temporal relationships [10].

3.1 The definition process of a CT
The test of a new clinical or diagnostic procedure is carried
out only when an ad hoc committee has given its approval.
This is based on the evaluation of a document, that is the
CT, designed by a group of physicians/researchers who
suggest a plan of the test. Both the procedure of the test
approval and the CT content follow nowadays standardised
rules which facilitate the evaluation procedure and
describe in detail the activities to be carried out during the
test. The writing of a CT is therefore a complex activity in
itself. It is accomplished through the collaboration of a
team, which participates in the writing committee,
generally working under the supervision of someone in
charge of planning and co-ordinating the writing of the
different sections of the CT. The physician responsible
for the CT proposal has to control that the text is coherent and
consistent in its parts also according to the rules of
document presentation mentioned above. We consider the
writing of a CT as a process in itself which can be
described through a Workflow, where tasks and their
interrelations are identified together with the different
actors, who have to describe specific sections of the CT
according to their competency and responsibility [10].

3.2 The CT schema draft
We have analysed the content of several CTs, the logical
and temporal relations between the different sections, the
importance which some sections have on the evaluation of
the entire document and we have modelled the CT content
according to well defined relationships [10].

The relationship “bound” represents a connection of logical-semantic coherence which indicates that the
different sections are closely interdependent. This means
that between these sections there must be a consistency of
both data and knowledge.

The relationship “knows” represents a connection of both
logical-semantic and temporal coherence. The sections
connected by this relationship have to be written in
sequential, according to the order of the edge direction.
This means that different single writers may write these
sections.

The relationship “equivalence” represents a loose
connection of logical-semantic coherence, characterised by
the presence of the same data, which has to be reported in
different sections of the document. However there is a
different information view of the same data reported in the
document.

Sections 1, 2 and 3 (fig.1) describe the fundamental
information of the CT and therefore they have to be
elaborated together, as they are closely interdependent.
These sections are the starting point for the drafting of
section 4 and 5. These last ones have to be written at the
same time, as they are strictly interdependent. For the
description of the trial design the use of WF libraries can
help this task. The information content of these elements
has to be reported also in section 6, but with a different
information view (relationship “equivalence”).

The fundamental sections of a CT are now written, the
other ones are drafted in an autonomous way, generally
using standard descriptions related to message exchange
(section 10) or to forms (sections 2 and 15). The authors
of the CT have to know the content of the common sections,
while a dictionary available to everyone will help to use
the same terminology.
4 The system requirements
For the development of the WITH system we have taken into account the requirements necessary to: a) editing the document text; b) re-using the text (protocol libraries, data dictionary, etc.); and c) facilitating the cooperation and the collaborative writing.

4.1 Requirements for the editing of the CT document
The standardisation of CTs, considered as a minimum set of information mandatory for the description of the clinical and diagnostic investigation as well as for its approval, is a process in an advanced phase, but still not yet completed. The standardisation is limited by the differences existing among CT typologies, which present their own characteristics according to the type of investigation phase (I, II, III and IV steps) and to the type of study (randomised or not randomised). Besides these objective differences, there are other variations, which depend on the writing process of any type of text, which has its own feature, is subjected to customisations and variants related to a specific medical school, or are the result of the collaboration between different competencies (i.e. physician and statistician) or between different healthcare unites. Finally a CT is also subjected to variations resulting from emerging needs, such as the ethic issue, which in the last years is getting an increasing importance.

These considerations, which derive from the analysis of a meaningful number of CTs already developed by GIMEMA, have pointed out the need of developing a system, which identifies and suggests a standard structure of a CT document, but which is at the same time flexible enough to allow the user to change the structure, add new paragraphs, modify their order or change their titles, etc. The analysis of CTs has also brought to identify the 'key information', which can be contained in specific sections. The use of key information within the different sections guarantees the uniformity and the consistency of the text. These consistencies represent the links "knows" and "equivalence" of the schema shown in fig. 1.

4.2 Requirements for the text re-use
The editing of a new CT is embedded within the medical activities conceptually related to the development of ongoing or already closed CTs, to guidelines used in a specific clinical praxis, to the results of laboratory and pharmaceutical research, which need a wider investigation. The complex process of writing a new CT can be facilitated by accessing libraries of CTs, and guidelines, data dictionary and databases of scientific papers, by the retrieval of information which can be stored and modified according to the specific requirements of the new CT.

The analysis of the CTs has also drawn attention on parts of standard texts (represented by the link "uses" of fig. 1), which are automatically shown to the user, who can leave them unchanged or modify them.

4.3 Requirements for the cooperative writing of the CTs
The writing process of a CT is carried out in collaboration with a team of physicians/researchers often geographically distributed and with different roles and competencies. A system which supports a collaborative writing has to make it possible to:
- Manage the authority access identifying the users qualified to access CTs in a only reading mode, or enabling him/her to modify the text or making it available outside the writing committee (Internet publication);
- Manage the different versions of the CT, which represent the 'history of the document' and report the author’s modified sections.

Useful aids to collaborative writing are also the scheduling of a workplan (workflow) of the CT compilation, where the tasks of each editor can be reported together with the scheduled deadlines, as well as the management of e-mails exchanged by the different text authors.

5 System implementation
The choice of integrating XML with a RDBMS is mainly connected with the need of an efficient management of CT libraries in terms of efficient storage, fast access, data integrity, security, etc.

The system is composed by function libraries, developed using Microsoft, ADOT™ e XML-DOM™ components. The content of a generic CT is represented using a XML schema, often developed as an alternative of DTD (Document Type Definition). This schema constitutes an extension of DTD and makes it possible to define a more precise document model. We have adopted the language XDR (XML Data Reduced), a standard suggested by a Working Group of W3C [11] and supported by Microsoft. Our CT model is composed by about 150 element types.

The system proposes a document structure subdivided into sections, each one corresponding to a specific XML element. Some of these sections have a predefined structure (paragraphs and subparagraphs). The user can choose to use the proposed structure or he/she can change it inserting titles and texts in a number of non predefined subparagraphs. In some sections recursive structures are also foreseen, for instance in the case of "patient registration", where the number of subparagraphs is function of the number of the randomisation of the investigation (XML example 1).

```
<elementType
  name="PatientRegistration"
  content="altOnly">
  .......
  <element type="text" />
  <element type="nestedpara"
    minOccurs="0" maxOccurs="*" />
</elementType>
```

```
<elementType
  name="nestedpara"
  content="altOnly">
  .......
  <element type="title" />
```
The key information is proposed by the system as an XML element to be inserted within the text of an identified section, using the attribute type "macros" and the content "mixed" (XML example 2). The tagged key information represents a guide in the composition of a 'well formed' paragraph also under the perspective of the completeness and consistency of its content. For instance, the section "title" is meaningful when it contains the name of the drug to be tested, the pathology and the indication of the cycle of treatment considered for the investigation (XML example 2). This set of key information is recorded by the user in the "general information" section. They contribute to define a data dictionary and in particular they represent the metadata needed to identify the CT in a univocal way.

Besides representing a generic CT model, the XML schema contains attributes used to:

- Mark up special types of entities (such as pathology, drug, etc.) in order to relate them with other information sources, which can represent a legacy. This facilitates the retrieval and comparison between CTs. See how we use the attribute type "reference" (the name of legacy) in XML example 3.
- Map the XML document structure and the relational database schema. See how we use the attribute type "column" (the name of relational attribute) in XML example 3.
- Develop interactive layouts for the composition and editing of the document. See how we use the attribute type "en" (in order to display the name of document section) in XML example 3.

The re-use and modify function is developed by the system through both the XML marked up document and the insertion of these text sections into a relational database. This facilitates the retrieval of the information previously described by XML elements. The relational schema describes: a) a CT by means of a relation, where each attribute corresponds to a document section; and b) each entity external to CTs (legacy) by means of a relation.

6 The WITH sessions

The WITH system is available on the GIMEMA Intranet by Explorer™. Fig. 2 shows the display used to define search criteria. The result is a CT or a CT section. In the search mode it is possible to combine free text key-words (within single sections) with key information by means of AND and OR operators.

Fig. 2: The search mode

Fig. 3 shows the layout of a CT. The text is browsable by hypertext links. Some of them are based on the "equivalence" relationship of the schema shown in fig. 1.

Fig. 3: The layout mode

Fig. 4 shows the layout used to insert new drugs in a controlled dictionary, which is one of the means for guaranteeing text consistency. From this dictionary the user gets the value of the key information of the CT document as shown in fig. 5.
draft [10]. There are systems which support the draft writing using a decision support system, for example DaT [12]. Our major focus is posed on a system design developed for an expert user, who takes advantages of consistency constrains, text re-use and collaboration. Therefore the system guarantees: a) the process standardisation; b) the process management; and c) the efficient delivery of information-based tasks.

We are now working on the extension of WITH functionalities improving in particular the WF editor to describe the "trial design" section. We have already developed at a prototype level a system which permits the description of a WF using a WF library, automates the procedures for the management of case record forms, patient registration and eligibility [2]. WITH is now in a β-test phase carried out by a selected group of GIMEMA; the first results are encouraging.

References

7 Discussion and Conclusions
The WITH system makes it possible to write the complete CT document, where the writing activity represents the final step of a complex process, started with the development of a