CEN/TC 251
Health Informatics
Secretariat: SIS-HSS

TITLE/
SUBJECT: **Short Strategic Study: Enabling Technologies - SGML/XML (Final Report)**

SOURCE: *Andrew Hinchley*

ACTION
REQUIRED: *FOR APPROVAL*

UDC

Descriptors

SGML/XML
in Healthcare

English Version

Health Informatics
Enabling Technologies :SGML/XML

This document is a final report
It is to be considered for acceptance by TC251
It was drafted by Andrew Hinchley

CEN

European Committee for Standardisation

Comité Européen de Normalisation

Europäisches Komitee für Normung

Table of Contents

Summary
This report addressed both SGML and XML, but focuses particularly on XML in the light of the benefits indicated below of using XML, a specification conforming to the SML standard, but of limited complexity.
XML is hardly the first general syntax available for representing data structure. The reasons why TC251 may wish to take a particular interest in it are as follows:-

- It has been designed with a limited number of features in order that it may be supported by software of limited complexity and cost in a great variety of environments;

- It's origins in web-based requirements ensure ubiquitous implementation;

- Web requirements for presentation ensure that solutions in this area can take advantage of these presentation properties;

- It has self-defining qualities which are of potentially of great value in environments where standard data descriptions may need limited enhancement or change to meet local requirements.

- XML parsers output a well-defined data structure (DOM) to applications

For those not familiar with SGML/XML, Annex 1 provides some useful background reading and references. Because of the easy availability of this material, this report does not attempt itself to introduce the features of SGML/XML.

In this report, recommendations on the use of XML by TC251 are placed at the front of the document in Chapter 1.

Chapter 2 introduces SGML/XML.

Chapter 3 looks at specific issues and solutions in relation to use of XML in TC251.

Chapter 4 develops and analyses very specific issues that arise in using XML to support known TC251 requirements in messaging.

Chapter 5 looks at other standards groups' studies of XML.

1. Recommendations

Draft recommendations from this study were discussed at a number of sessions at the Cork TC251 working group meeting of 24-27th May 1998. This final version of the report aligns with the actions taken on reviewing the draft recommendations.

1.1 Recommendation in relation to messaging

Recommendation in relation to messaging
XML should be positively considered as an additional syntax for implementation of TC251 messages.

- *Chapter 4 of the document covers the main issues in using XML to meet messaging requirements. These do not require TC251 to modify its general technology-independent approach to specification.*
- *Developing a solution for messaging will throw light generally on the use of XML for other TC251 requirements.*
- *Given the expected wide availability of XML parsers and supporting software, any limitations of the XML syntax as against use of SGML is a reasonable price to pay for choosing XML.*

This recommendation was accepted at Cork and a cross-WG Task Force was established.

The main task force study areas in relation to messaging were provisionally defined at Cork (see Annex 8 for details) as:-

- XML syntax recommendations
- Application of complex data types
- Use of different element types
- Requirements for a generalised tag repository
- Application of DTDs in messaging
- Validation
- Application to an existing CEN TC251 message

This task force scope includes most of the highlighted technical areas for study described in Section 3 and 4 of this report.

1.2   Action in relation to areas of TC251 output, other than messaging

This report takes the view that, for a syntax-independent body such as TC251, the use of XML needs to be studied against well-defined syntax-independent requirements. However the following considerations are also clear:-

General issues of documents versus messages
XML brings a very practical focus to the question of whether there are any significant differences between documents and messages. Concerns exists that if the use of XML is initially restricted to messages, then the wrong focus may result. This is a view also strongly felt by the HL7 SGML/XML SIG. The task force established at Cork also contains a document-oriented study item.

EPR representation and transfer
TC251 has a number of project teams under way in EPR-related specifications. It is too early to have a complete requirement from these developments, although a preliminary conclusion from PT29 would be that much of the GMD approach to representation in existing TC251 messaging will also be valid for communication elements of an EPR. This confirms that using the current TC251 messaging requirements will be a constructive step in relation to future requirements.

Requirements for image, instrumentation and other areas covered by WGIV
This report was not able to cover any additional requirements of WGIV. This was a resourcing issue in relation to the scope of this strategic study and therefore leaves the question open on WGIV requirements. A current multi-media project team PT34 intends to address the issue of XML to support multi-media reports.

1.3   Other issues recommended for future consideration

This report raised a number of additional issues, most of which could be seen as further steps beyond the scope of the task force established at the Cork TC251 meeting. These are highlighted in sub-sections 1.3 to 1.7 below as worthy of being considered for future action.

1.4   Establishment of a CEN-recommended approach to XML support for messaging

Following its early work on interchange formats, TC251 produced a detailed message development methodology (Annex 2 gives more details), which included an open-ended set of annexes showing how particular syntaxes supported TC251 General Message Descriptions (GMDs). It would be logical to see the XML specification as an additional annex to that report (CR12587). Most communications-related output from TC251 since 1995 is in the form of GMDs which do conform to what is described in CR12587. A list is given in Annex 2.

This report also recommends that consideration should be given to the extension of CR12587 to address the following issues on an implementation independent fashion:

- *A standard method of providing a unique identifier for each HGMD instance (i.e. message sent).*
- *A standard way of identifying and including reference to healthcare message senders and recipients.*
- *A standard way of dating the issuing of a message.*

1.5 Liaison with World-Wide-Web consortium (W3C)

TC251 should be more pro-active in its relations to the syntaxes it uses. As far as XML is concerned, this would mean liaison with the World-Wide-Web consortium.
TC251 should investigate how it might have a liaison/member role with W3C so as to play an active part in voting decisions on any future specification changes to XML.

1.6 Liaison with other organisations implementing XML support for healthcare messaging

Having regard to the commonality of the Tc251 messaging approach, with that of HL7 in its Version 3 developments, Hl7, CEN TC251 should maintain a close relationship with the HL7 study of XML usage, since alignment of solutions may be of considerable benefit from an implementation perspective.

The HL7 Reference Implementation Model (RIM) is a development from the work of CEN TC251 on DIMS and GMDs. The main differences are that the RIM is a collection of all objects within the HL7 scope and also HL7 keep their existing data types. The XML issues arising are however very similar.

1.7 CORBA and Microsoft COM/DCOM

There would clearly be a great benefit if TC251 output expressed in a specific syntax such as XML, could be supported in component technologies without modification. This issue is addressed in detail in a companion report on CORBA and COM/DCOM.

2.

SGML/XML

2.1 Background reading and other references

Annex 1 lists a number of references on SGML/XML. As sources are widely available, this report does not provide a tutorial on SGML/XML. A Powerpoint presentation for TC251 members is however available.

2.2 SGML and XML

SGML is an international standard approved by ISO. The generality of SGML makes it possible to define a variety of syntaxes, that are all conformant to the SGML specification. In this sense SGML is a meta-syntax, a language for defining other languages. HTML is one of the simplest examples of a conforming syntax. XML is a recent conforming syntax developed by the World-Wide-Web Consortium. It is intended as a replacement for HTML, providing a sounder general-purpose syntax for information description, than HTML which focuses on a limited set of data types, aimed only at simple presentation issues.

There are obviously areas where the extra range and flexibility of SGML may provide a better solution than the more limited XML. In this report, the analysis in Section 4 does not show that that SGML would offer direct improvements over XML in relation to messaging requirements.

Given the benefits arising from the development of easily available XML tools, the direct support in every web browser, together with related developments as indicated below, there are clearly great benefits if XML can be used exclusively

This report therefore proposes XML, rather than SGML, as the appropriate syntax for further study of TC251 requirements.

2.3 W3C architectural views

It is important to understand where XML fits in the structure of specifications being developed by W3C, who have a number of related specifications under development, of which XML is one of the first. There are implications and knock-on effects on longer-term use of XML though these specifications are still under development.

The diagram on the next page shows a number of important additional specifications which will affect the use of XML and the tools that will be available to process it.

RDF
RDF (Resource Descriptor Framework) is a very recent specification aimed at providing a general semantics description language to sit above XML. It is described in the W3C Metadata activity files.

Its relevance in the use of XML in TC251 is its ability to describe contents and relationships between data items defined in XML. The proposal includes the important name space area described in section 3.2 below.

PICS
This is a specific document content description standard allowing specific ratings to be applied to web pages, in terms of suitability for children etc.. As well as providing a standard way in which page designations can be checked by the potential reader, it also provides identity of those who have provided the ratings for this page/document.

PICS illustrates a general point as shown in the diagram. Once RDF is fully in place, then applications such as PICS will sit within the W3C architecture defined, as opposed to their current definition as stand-alone specifications.
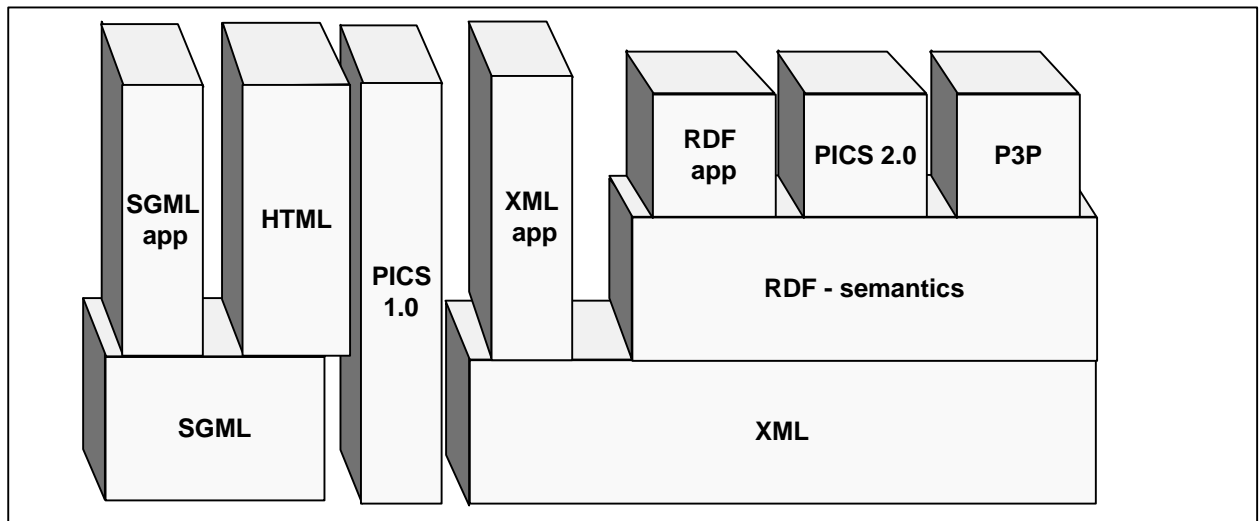
The diagram shows on the left, the early use of the simple HTML specification conforming to SGML. On the right, the future direction is indicated, where XML becomes the base for a number of related standards.

2.4   XML-Data:

This specification does not appear in the diagram above, since it is a proposal initiated mainly by Microsoft. It has the status of a submission to W3C not assigned to a working group. It:-

- introduces a directed graph approach
- provides an enriched set of data types:-number, real, integer and date
- it provides a mechanism to organise element types into a hierarchy
- provides the name space functionality described in Section 3.2

The general policy of W3C is to have only one specification level for XML, so it is likely that these additions would only succeed if there is general agreement on an upgrade and extension to the XML specification.

2.5   Key elements of XML of relevance to TC251 requirements

There are some key features of XML which provide the basis on which it could be use to support TC251 output.

2.5.1   Principles of common parser application

The principle of XML is to define a simple specification such that light-weight parsers can be used on a wide basis, wherever they are needed.

An important part of the approach is for a producer (of an XML instance) to validate that structure as being well-formed, before passing it to a receiver.

In the diagram, the validation of the document is seen happening between the editor and the browser, which in a communications context, would mean validation by the sender and by the receiver.

At the receiving end, there are two separate pathways, depending on whether simple display is intended, or whether further processing is taking place .
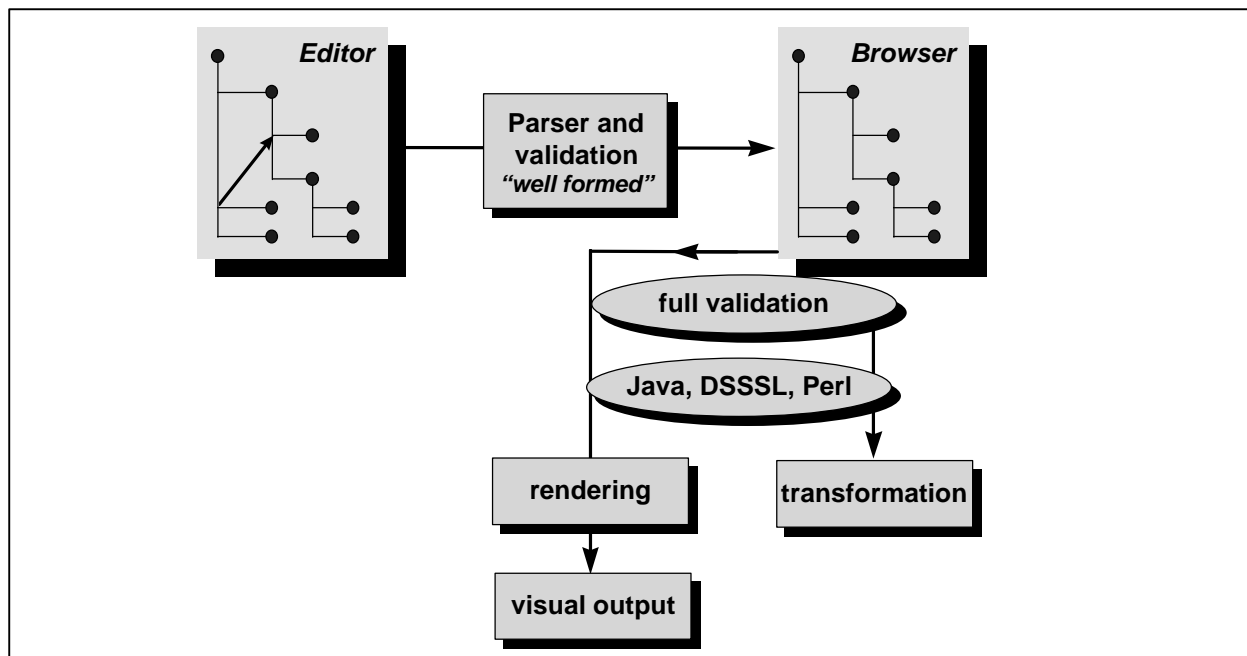
2.5.2   DOM (Document object model)

The Document Object Model is a platform and language-neutral interface that will allow programs and scripts to dynamically access and update the content, structure and style of documents. It's role as a "document-API" is a general one, not limited to its initial role in relation to use in current browsers. The specifications can be found at W3C DOM .

Use of the DOM would be a huge step forward in the implementation and analysis of messaging, something which many current EDI standards have never got a grip of. The DOM is defined in terms of common interface languages such as CORBA's IDL (which is an ISO standard in its own right), Java and C++.

For the application developer, each XML structure appears as a tree, for which the API then provides the ability to "walk" up and down the tree extracting data as and when required.

In addition to this "Level 1" DOM, W3C have also defined a "Level 2" which is concerned more with presentation issues and the associated use of style-sheets.

 2.5.3   Style-sheets

This is an important area for web usage and a number of alternative ways of implementing style sheets with XML have been under consideration. Presentation per se is not a primary requirement of TC251, at least not as currently identified. However some proposals do use style sheets as a method to complete rigorous validation of XML content, of the type that would be required by strong profile validation.

CSS: Cascading style sheets has been an urgent development for HTML in order to provide a fast separation of presentation for existing HTML usage. It's use with XML is non-strategic in view of the developments described in the next sub-sections and therefore it is of no specific interest within this study.

DSSSL: Document Style Semantics and Specification Language, ISO 10179 is a comprehensive specification, used with SGML.

This whole area has only been recently formally accepted by W3C as a work item (in February 1998) in relation to XML, although a proposal specification was published in August 1997 (http://www.w3c.org/TR/NOTE-XSL-970910).

XSL allows formatting information to be associated with elements in a source document to allow the production of a formatted document. A set of construction rules can be defined for a document, together with a set of style rules. XSL is itself expressed in the XML syntax.

Some experts (particularly the XML/EDI group -section 5.2) believe XSL has potential in relation to specification of validation rules. This is of course only true as long as this function really can be separated from display requirements and functionality.

2.6   Standards making structure

W3C is open to any organisation paying fees and currently has several hundred members. Members have a seat on the Advisory Committee, which decide on work items. W3C was founded in Autumn 1994 and has 240+ Members. It has a Director, some permanent staff and is structured as one international consortium, with 3 Hosts:

- MIT in Cambridge, MA
- INRIA in Sophia-Antipolis, France
- Keio in Tokyo, Japan

W3C specifications are output subject to distribution rules set by W3C. Documents are the copyright of W3C or individual authors.

Distribution rights for W3C documents allow free distribution and use and the limitations cover basic requirements, such as a requirement that software or specifications derived from W3C specifications cannot both modify the specification and still use the W3C designation. In relation to approval of W3C specifications, the following text is reproduced from the W3C published procedures (http://www.w3c/Consortium/Process).

> - Integral to the W3C process is the notion of *consensus*.
> - The W3C process requires those who are considering an issue to address all participants' views and objections and strive to resolve them.
> - Consensus is established when substantial agreement has been reached by the participants.
> - Substantial agreement means more than a simple majority, but not necessarily unanimity.
> - In some circumstances, consensus is achieved when the minority no longer wishes to articulate its objections.
> - When disagreement is strong, the opinions of the minority should be recorded alongside those of the majority.

Clearly major sponsors such as Netscape, Microsoft and Sun etc. do have a major influence on specifications, but equally the rules appear to allow as open as possible decision making. There has been a clear decision not to produce different levels of specification for XML. It may be for this reason that W3C do not appear to have particularly backed the EDI/XML development.

2.7   XML tools

Although the XML1.0 specification was only approved earlier in 1998, there are a number of parsers and related tools already available. This is mainly because both Netscape and Microsoft completed much development work prior to the standard being validated. Current browser products such as Internet Explorer 4 have a limited XML capability built-in, to support the new CDF (Channel Definition Format) facility . Future browsers are likely to have a general purpose XML parser included.

The availability of tools which are decoupled from specific browsers obviously starts to point the way to incorporating XML support into additional communications environments. A summary of currently available tools is given in Annex 5.

3. Specific issues and solutions in relation to use of XML in TC251

3.1   DTD registration

XML defines structure in terms of Data Structure Definitions (DTD). TC251 requirements to define structures such as HGMDs (Hierarchical General Message Descriptions) can be defined and managed, if required through an appropriately defined registration procedure.

However DTDs can also be user-defined or existing DTDs can be dynamically modified and extended. User-defined DTDs offer great flexibility where structures cannot be pre-determined. Dynamic modification of DTDs may be useful as a way of supporting message profiles.

As the number of DTDs increase, registration procedures become necessary. This is particularly true where there are overlapping, but separate communities of interests. This is limited however by the following factors:-

- In any one instance, only one DTD applies. Elements can be added/removed only by in-line additions to the DTD
- No appropriate registration authorities for DTDs currently exist.

3.2   Name-spaces

A draft specification for introducing separate (tag) name spaces into XML can be found at http://www.w3c.org/TR/WD-xml-names.

This document is at working document stage only. These functions could however be useful in any context where well-defined and specified DTDs, such as HGMDs are supplemented by user-defined structured additional material. This could be, for instance, to support clinical headings material, for which fully agreed coding did not exist, or ad hoc locally structured.
The specification does not however remove the need to consider a tag repository (next sub-section).

3.3   Uniqueness of tags in relation to other support XML requirements

There is a requirement coming not from the specific requirement to support TC251 messaging, but from other areas to ensure tag uniqueness. This is illustrated in the diagram below.

There are two main areas where this may come from:-

- A requirement for structured text sections within a GMD structure, where user-defined XML tags may be encountered.
- Support structures such as security envelopes which may need to be defined such that the structure can be within the DTD and able to be analysed as a single tree by a parser.

These particular requirements could be met by the XML namespace proposals. These appear both in the XML-Data drafts and in the RDF drafts and therefore would appear to have some chance of successful addition to the XML specification.
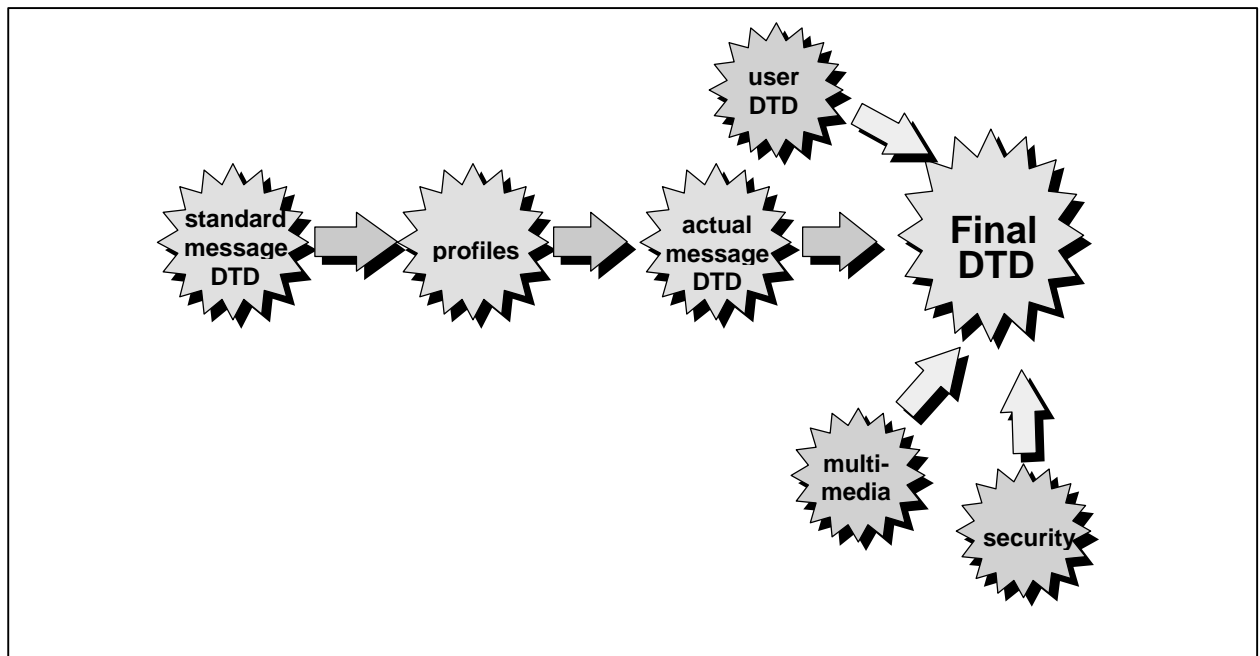
4. TC251 messaging support from XML

This section details the issues in providing XML support for TC251 messaging. It does not represent a specification for implementation, but does offer a way forward towards developing such a specification.

It is a broad conclusion of this study that there should be a single XML specification for TC251 messaging support. This obviously implies that there must be a single solution to the issues mentioned, such that there is an unambiguous recommendation for implementors.

The main elements of XML support are indicated in the sub-sections below. The requirements are comparable to those of HL7 in supporting their Reference Information Model (RIM), which of course is based heavily on CEN TC251 experience and specifications. For this reason the work of HL7 (Section 5.1) provides much of relevance and applicability to the sections below.

As already indicated the base TC251 requirement is to support the Tc251 General message Descriptions (GMDs) developed following the methods described in the TC251 report CR12587, which includes a definition of how two syntaxes (EDIFACT and ASN.1) can support HGMDs. XML support, at minimum, therefore requires an equivalent annex to be written.

The main elements and issues of XML support are covered under the following headings.

4.1 A DTD for every HGMD

The XML Data Structure Definition (DTD) approach is appropriate and suitable for defining an HGMD. Each HGMD can be expressed as a single DTD. A registration process may be appropriate to manage the ongoing use of such DTDs.

4.2 Objects and attributes

Objects and attributes could be supported, for example as XML parameter entity declarations and XML tags need to be allocated corresponding to each object and attribute. Further study is required of the available alternatives.

4.3 Achieving a unique set of tags

There is a real incentive in XML to use short tags, since XML instances use explicit tagging and therefore over-long tags greatly increase the length of the message to be transmitted. A method is needed to create short tags from the object and attribute definitions of any GMD.

Current TC251 output makes each set of GMDs consistent and unique within their scope, but does not require that objects and attributes are consistent in meaning across the entire range of GMDs produced.

A wider requirement for uniqueness across all TC251 output would result from viewing things in relation to the support of a common API to clinical messages received, however this is outside the TC251 messaging scope to date.

There could be several reasons to attempt to use a unique identifier scheme:

- The simplest additional requirement would be for arbitrary XML structures inserted by users within a standard message structure .

- A more complex requirement would be to correlate across multiple syntaxes, such as proposed by the XML/EDI group.

XML allows any structure to carry an identifier of unspecified type and structure. To operate a unique scheme, a registration schema would need to be identified, together with a tool which could assist the use and implementation f a unique identifier.

4.4

Common attribute groups

Common attribute groups are structured and used in a similar way to attributes which belong to a specific object. There is no particular problem therefore in representing them in the same way as attributed that belong to a specific object.

4.5   Support for TC251 data types

XML is a weakly typed language and therefore it is not possible to easily support the main basic types of INTEGER, REAL BOOLEAN and STRING. These are most easily mapped to a simple string type, removing the chance of ever validating those data types, from the type property per se. Further review is needed as to the effect of this loss. There is in fact in most GMDs, very limited use of basic types other than string.

More complex data types in TC251 are based completely on the simple types. They can be supported in XML as entity parameter declarations themselves and therefore do not represent a direct problem in support.

If it was considered necessary to carry the real representation fully, this might be done by :-

- a general method of carrying a type descriptor as part of each instance. This could then be applied to other data types, although there are candidates in TC251 methodology, apart from boolean which is used rarely.

- A specific method would perhaps carry reals as exponent/value pairs.

4.6   Optionality

The TC251 methodology specifically supports profile definitions, being a revised GMD eliminating optional elements which are unnecessary.

In XML there are two different ways of approaching this:-

- In the first approach, the elements which are not going to be used need to be designated as such. Presumable this would need the introduction of additional entity declarations which would over-ride those already defined.

- An alternative approach would be some other way of indicating a change in properties.

4.7   Value and/or range definitions

XML allows value specification so these could be added by using revised additional parameter entity definitions as part of the same approach as described in 4.6 for dealing with optional elements. In any situation where a value was being defined, it would require that a revised ENTITY definition was introduced.

4.8   Cardinalities

TC251 defines clear sets of cardinalities, which are all supportable in XML through existing facilities in parameter lists.

4.9   Choice definitions

Choice is directly supportable in XML.

4.10   Message identifier

A standard message identifier should be considered so that key aspects of message management can be properly handled. The main required fields are:-

- a unique message identifier
- a sender identifier
- a recipient identifier

Uniqueness for a message identifier can clearly only be in relation to a message sender.

World-wide registration scheme for sender/recipient already exist (such as the ERSC/EDIRA scheme, supported by major registration authorities such as Dun & Bradstreet).

4.11   Full element specifications (lengths etc)

HGMDS do not specify lengths of data elements. In the case of both EDIFACT and ASN.1 appropriate lengths are selected by the implementor, limited in the case of EDIFACT by the relevant data element definitions. XML does not limit field lengths. The implications of unbounded fields may require further examination.

4.12   Specification and validation of conditionality, dependencies, code list selection,   range checks

Traditionally in messaging little attention is paid to this area, with disagreement on whether validations should/could be defined as part of the general message processing. Few syntaxes have adequately addressed this area, X12 being an exception. It is an issue which has only recently come to the fore in HL7.

4.13　　　Profiling

Profiling is a critical part of implementing TC251 specifications, although very limited thought has been put into this process at the TC251 level, leaving it entirely to the requirements of a specific implementation.

The requirements in CR12587 are limited and do not make any specific demands on XML. However those with most experience in EDI, such as the ANSI X12 activity do have very well-defined profiling techniques and specifications, developed as a result of considerable implementation experience.

In healthcare, DICOM and the HP-led Andover activity have been the first to address these issues. This report cannot both define a possible TC251 requirement against which to assess a XML solution, but the following issues arise:-

- static and dynamic profiles, should be applicable to version 2 and 3
- a registration process is needed to identify profiles, such that a profile identifier can be carried in the message. Static profile identifier tree can then be created.
- optionality needs to be constrained to indicate whether required, and whether any dependencies apply.

5. Other standards groups studies of XML

5.1   HL7

5.1.1   KONA

KONA is the output of a group who met in 1997 to consider generally the effect of using SGML/XML to support healthcare data structuring and communications. The group is now part of the HL7 Special Interest Group on SGML/XML HL7 who are studying the use of SGML/XML for structured clinical documents. The KONA proposal is not yet a position statement of the SIG as a whole. The SIG is working on a rigorous definition of the differences between messaging and documents and the relationship between the two in a standards context. This area is also addressed in Annex 3.

The KONA approach to healthcare data is somewhat like a microscope, where the right degree of focus is applied to the particular requirements in hand. The benefits are that the same instrument i.e. XML can be used at different levels of magnification, potentially providing a unifying approach to what is currently a deep divide between structured and unstructured.

The KONA group identify four levels of granularity. The KONA group are currently engaged into understanding where these levels correspond to HL7's future reference information model, and a set of comparable arguments can be seen in relation to TC251, particularly as far as Level 2 is concerned.

5.1.2   HL7 syntax version 2

The work on Version 2 is of great interest as quantitative measurements are being done in this area by Bob Dolin and his Kaiser Permanente. The parser used was, not a validating parser, able to confirm XML conformance.The reference is:-

SGML/XML as an Interchange Format for HL7 V2.3 Messages (Revised April '98)

Message size
Initial findings by Bob Dolin is that 2-300% more space is required for these messages, as compared with existing HL7 Version 2 implementations. In some cases, the expansion was considerably more. The reason for this is that in XML tags need to be explicitly defined in all instances.

There are several areas where Bob Dolin highlights where the mapping to XML may affect these sizings:-

- Tag lengths are critical, the algorithm by which tag names are derive is therefore important
- Message complexity has an effect on length

Speed
A processing speed 5 times slower that regular HL7 interface engines, however these only partly interpret a message and the comparison is not like-for-like. It would be reasonable to place too much weight on these preliminary results.

5.1.3   HL7 syntax version 3

The work in relation to Version 3 relates very closely to TC251 interests as the HL7 RIM (Reference Information Model) is substantially based on TC251 DIM/GMD approach.

Version 3 comprises a RIM (Reference Information Model) which is a single reference collection of objects and attributes, as compared with TC251 which defines objects and attributes only within a particular scope. In other respects, the approach is similar, with GMDs derived from the reference model as required.

Only one Message Decsription so far has been define as an example by HL7. This limits experimentation in the short-term. However, it raises issues such as whether CMEDs (HL7's equivalent of common attribute groups) are supported as defined in the reference model, or variably represented in a message or a message profile, depending on the specific sub-set of attributes in the attribute group actually being used?

SGML/XML as an Interchange Format for HL7 V3.0 Messages (Fully revised document issued June 1998)

The work by Bob Dolin has focused on supporting GMDs, although there is some feeling that the full power of SGML/XML as a "variable focus", by revisiting the RIM methodology and specification. There is a belief that there are areas which can be supported which are outside the current capabilities of the RIM to model. Another way of expressing this is that XML can support rougher granularity than the objects defined, where necessary.

HL7 have invested greatly in UML tools and there is clearly scope to create a DTD for any message directly from the HGMD representation in the UML tool.

5.2   XML/EDI

The XML/EDI proposal was first published in September 1997 and a revised document was produced in January 1998. Revision of the draft and further development towards a full specification is expected to continue for at least another year.

XML/EDI incorporates a common set of messaging conventions applicable potentially to all EDI messaging requirements. The strength of such a set of recommendations would be that common tool sets focused on EDI messaging might then be developed, and that both old and new EDI syntaxes could co-habit within a single framework.

The XML/EDI group is not yet officially recognised by the World-Wide Web Consortium (W3C), but is supported by the Graphical Communication Association) GCA, the organisation which acts as a secretariat for a number of SGML activities.

The work is at too early a stage to fully validate whether XML/EDI could meet the requirements of all existing EDI standards groups, including the requirements of TC251, but the development is worthy of detailed further study as and when further specifications emerge. The XML/EDI Group are very active participants in the ANSI X12 studies, which are summarised in the following paragraphs. Annex 6 provides a more detailed analysis of the XML/EDI guidelines draft document. It is clear that XML/EDI will arrive at a specific solution for most, if not all, of the TC251 messaging requirements as analysed in Section 4.

5.3   ANSI X12

ANSI X12 is a widely used EDI syntax in the United States, used by all sectors including non-clinical healthcare applications.

X12 has an initiated an on-going study of a replacement syntax as part of its X12C sub-group. The group is focusing on the use of XML. The group has the formal active input of the XML/EDI group and also has support from Commercenet, business standards consortium for electonic commerce.

The additional resources that X12 can bring to the issues addressed by XML/EDI has accelerated study of a number of the technical areas listed in Section 6. Annex 7 gives further details.

Annex 1
References on SGML/XML

The official web site is at: SGML and XML: Structured Document Interchange
(http://www.w3c.org/XML)

A simple FAQ for XML can be found at:- The XML FAQ (http://www.ucc.ie/xml)

An excellent compact description of XML can found in *"A Guide to XML"*, one of the chapters of XML
principles, Tools and Techniques (published by W3C.$29.95). Details can be found at:
http://www.w3c.com/xml/

"Presenting XML" Richard Light (Sams Books $24.99 ISBN 1-57521-334-6).
This is a good introduction to XML. It's length (at 394 pages) makes it sensible to read it together with a
more compact reference summary, such as "A Guide to XML" .

A full specification of XML can be found at :- http://www.w3c.org/TR/PR-xml-971208

The HL7 SGML/XML group has done much work on applications in healthcare.
The KONA architecture document can be found at
http://www.mcis.duke.edu/standards/HL7/committees/sgml/WhitePapers/KONA

Annex 2
Communication requirements already defined by TC251

Specifications have been defined and used in the great majority of TC251's existing specifications and those currently under development. The following sub-sections summarise these requirements.

Communication and related representation requirements are defined for TC251 in:-
CR 12587 :1996   CEN Report: Medical informatics - Methodology for the development of healthcare messages. The approach follows earlier work in TC251 which created a separation between specification of application information and the technologies which were used to support the transfer of that information.

The approach is summarised in the diagram below, reproduced from CR12587.
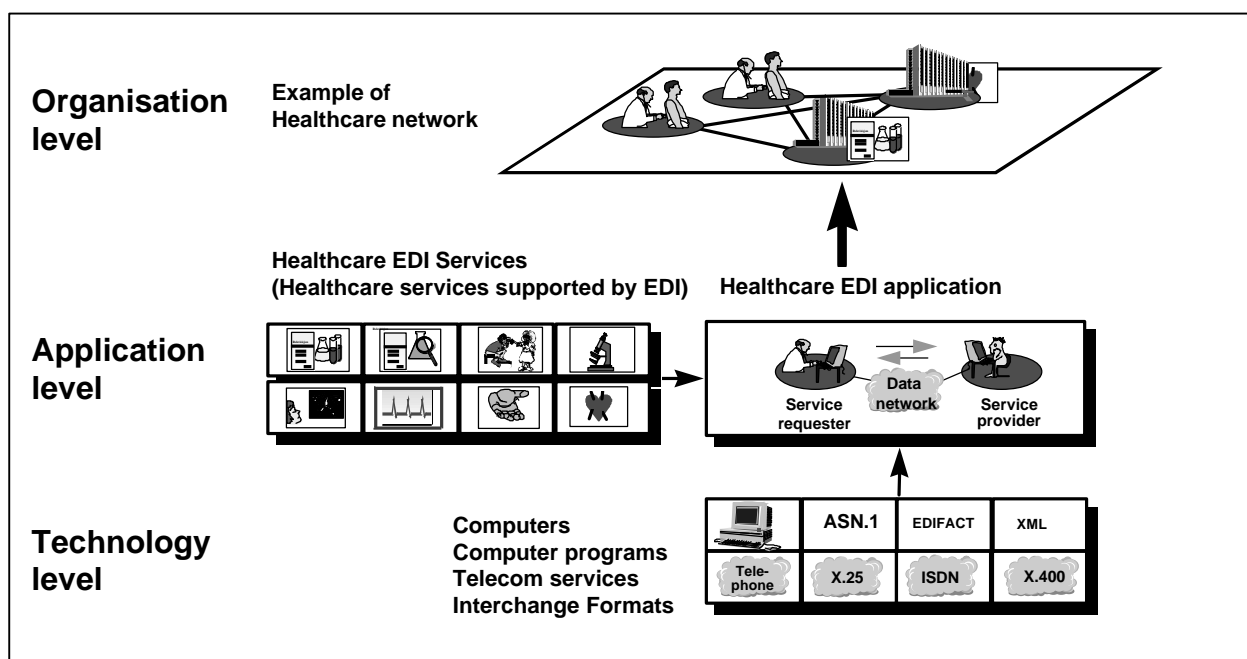
With a small number of exceptions, most TC251 recommendations which include a communications in scope, conform to CR12587. This includes not only the four ENVs covering pathology services, diagnostic services, GP communications, and administrative messages, but also vital signs. For on-going projects all PTs making recommendations in WGI and WGIV are believed to wish to conform to CR12587. A full list is included below.

TC251 ENV or CR		Year of availability	Conformance to CR 12587	Title	Responsible WG
ENV	1064	1993	NO	Medical informatics - Standard communication protocol - Computer-assisted electrocardiography	IV
ENV	1613	1995	YES	Medical informatics - Messages for exhange of laboratory information	I
ENV	12018	1997	NO	Identification, administrative, and common clinical data structure for Intermittently Connected Devices used in healthcare (including machine readable cards)	I
ENV	12052	1997	NO	Medical Informatics - Medical Imaging Communication	IV
prENV	12443		YES	Medical informatics - Medical informatics healthcare information framework	I
ENV	12537-1	1997	YES	Medical informatics - Registration of information objects used for EDI in healthcare - Part 1: The Register	I
ENV	12537-2	1997	YES	Medical informatics - Registration of information objects used for EDI in healthcare - Part 2: Procedures for the registration of information objects used for electronic data interchange (EDI) in healthcare	I
ENV	12538	1997	YES	Medical informatics - Messages for patient referral and discharge	I
ENV	12539	1997	YES	Medical Informatics - Request and report messages for diagnostic service departments	I
CR	12587	1996	YES	CEN Report: Medical Informatics - Methodology for the development of healthcare messages	I
ENV	12612	1997	YES	Medical Informatics - Messages for the exchange of healthcare administrative information	I
ENV	12623	1997	NO	Medical Informatics - Media Interchange in Medical Imaging Communications	IV
CR	12700	1997	YES	CEN Report: Supporting document to ENV 1613:1994 - Messages for Exchange of Laboratory Information	I
ENV	12967-1	1998	NO	Medical Informatics - Healthcare Information Systems Architecture - Part 1:Healthcare Middleware Layer	
		In progress	YES	Vital signs	IV
		In progress	YES	Multi-media report (yes-with extensions??)	
		In progress	YES	PT29(with a number of extensions!)	
		In progress	YES	PT33 Exchange of coding info.	III

Annex 3
Properties of messages and documents

Properties of messages
None of the technologies under examination is message-based in its origins. It is therefore important to capture the properties of messages, which are often left as implicit, when working, as TC251 has been to date, in an environment where message-based technologies were assumed to be the implementation target and therefore these matters did not need further consideration.

In relation to the technologies under study there are two aspects of the question "What is a message?" which need to be examined. These are:-

- The properties of a messages in relation to use of document technologies
- The properties of message in relation to use of object/component technologies

Messages and Documents
As far as this study is concerned, the issue arises because SGML/XML is defined for structuring documents rather than messages. Applying SGML/XML to messages therefore raises the issues of whether there are underlying differences, which need to be exposed in examining issues arising from use of SGML/XML.

The issues are however would be the same whatever document standards was being evaluated It is unclear that there are differences other than those, which are fairly intuitive. However some of the key aspects are listed below, drawn from a recent discussion in the SGML/XML SIG of HL7 on this subject. There are:-

- no differences based on content alone
- strong and useful distinctions based on intended use
- features fit into "defining" and "characteristic"

The defining aspects of a document are:-

- human and machine readable
- suitable for permanent archival report
- potential for unbounded complexity
- integrity requires inclusion of whole and whole contains sufficient context for human comprehension

The characteristic aspects of a document are:-

- it contains narrative
- it can carry attestation as part of the clinical record
- it's primary use within a local environment
- there is no immediate action is expected on receipt of a document

The defining aspects of a message are:-

- the message has very specific context based on such things as the preceding messages, the sender, the recipient and the point of issuing the message
- the message is processable by software application
- there is bounded complexity, with previously agreed structure
- it does not carry attestation

The characteristic aspects of a message are:-

- it may be cached, but is not always archived
- it can be divided between loss of integrity, if metadata supplied, but require external documentation for human comprehension
- an immediate action expected on receipt

Annex 4
Technology independence - the TC251 view

The TC251 technology independent approach:-

- Reduces (at least in some areas) the size of the specification required and the resource to develop it
- Allows choice of technology by user
- Lengthens the period for which specifications are able to be used

However this very technology independence:-
- Leaves many issues to be solved by each implementation
- Stands in the way of a tool-based approach where the much of the implementation effort can be automated
- Increases development time generally
- Reduces interoperability
- Reduces reuse of components

In general therefore, the questions that need to be borne in mind when evaluating any technology are:-

- Is there any element of the TC251 methodology which would preclude optimal use of a technology?

- Are there additional specifications that would be needed to optimally use a particular technology?

- Can successful interoperability only be achieved by more actively promoting particular technologies?

Supporting a messaging paradigm
The communications approach favoured by TC251 is message-based, yet the technologies reviewed in this report are definitely not message-based.

SGML/XML is document-based, and in XML is oriented specifically to the web as a communications framework.

CORBA and DCOM/COM are object-based component technologies.

However it is clear that the ability to managing and manipulating XML structures can and will extend from current web/browser environment into a much wider set of environments.

It is also clear that both CORBA is supporting messaging ((through healthcare work in the Andover project and OMG's own work on MOM-Message Oriented Middleware) while HL7 and the Microsoft Health User Group (MSHUG) have implemented a messaging environment for COM/DCOM.. This area is covered in more detail in an accompanying report on COM/DCOM and CORBA.

Annex 5
Available software

Company	Product	Capabilities

webMethods	Web Automation Platform:
	• Web Automation Toolkit
	• Web Automation Server
		The Platform is an API for automating access to all Web addresses, it combines rich HTML Parsing, Text Pattern matching and a Web Interface Definition Language (WIDL) that binds data objects to program variables, with HTTP Protocol handling, SSL Security, and an object Repository which supports complex queries across multiple data sources.
		∗ HTML/XML Parsing
		  • Single pass Parser
		  • Fully multithreaded
		  • Document Object Model
		  • Client/Server
		  • Supports HTML 3.2, including Netscape and IE extensions
		  • Fully extensible
		  • Supports XML
		∗ Object Repository - a Document Object Model is used to access objects, object attributes, and data structures in the Object Repository. Queries can be made across multiple documents

DataChannel	XML Development Kit	∗ DXP - DataChannel XML parser ∗ XML generator ∗ DOM builder

	RIO	∗ Save to the Web™ feature - instant web publishing without a Webmaster.
		∗ Extensive management controls for Intranet publishing
		∗ Personalised information delivery to the desktop
		∗ Built in XML technology
		  • New Active Content technology - allows data to be reused across multiple applications
		  • used in the DataChannel client where information is distributed as pointers or metacontent, saving network bandwidth
		  • Open API allows developers to integrate RIO's database-driven user profiling and authorisation systems, into existing applications
		  • XML-based and desktop content-profiling system in RIOs built-in database allows companies to integrate any form of

		data or any type of application
Chrystal	Astoria 3.0	XML based content management solution for producing documents. ∗ Organise documents using familiar metaphors such as file cabinets, folders, and documents ∗ Preserve document integrity with access controls for check-out, editing, and revision tracking ∗ Access and track revisions of documents and structured document components ∗ Reuse information already authored and graphics already created and stored ∗ Search on content in documents throughout the repository and structured document components by content, attributes or structure ∗ Edit documents and document components using existing authoring tools ∗ Manage access and revision tracking using a comprehensive set of system management utilities ∗ Integrate with existing applications or build new ones using the Software Developers Kit (SDK)
ArborText	Document Architecture ADEPT Publisher and ADEPT Editor	∗ Allows developers to write, compile, and test SGML DTDs and stylesheets. These can then be installed as "doctypes" (SGML applications) for use by ADEPT Publisher and ADEPT Editor
POET	POET Content Management Suite	∗ Object-oriented repository for SGML/XML documents ∗ User-defined component granularity ∗ Check-in/check-out and versioning support ∗ Entity and link management

Annex 6
XML/EDI

This is an activity which attempts to formalise those elements of XML and associated standards which could form a broad base of support for all EDI-like applications of XML.

It was formed in 1997 and to date has an outline document (dated January 1998) in the public domain: *Guidelines for using XML for Electronic Data Interchange.*

The group is not within the W3C structure, but has an affiliation to GCA, the organisation that organises SGML conferences and support. Despite being a small group, they have embarked on an ambitious solution intended to be universally applicable to all EDI-like problems. Much of the current development work is being done in tandem with the ANSI X12 work described in Annex 7.

Important areas being address in XML/EDI include:-

- integration of web-based messaging with conventional EDI messaging
- global tag repository
- sophisticated message validation
- work flow support by defining specific business functions able to be triggered on message receipt

Naming/global repository

Work of the group in this area has accelerated due to two inputs: study with the ANSI X12 group (see Annex 6), as well as discussions with those who have previously worked on global repositories such as the BSR (Basic Semantic Repository) project.

This path has a number of drawbacks:-

- There is a need to support existing EDI standards. However, these have a number of differing requirements, met by different rules for DTD definition, message validation etc. There do not appear to be commonalities which could be brought into a single specification.

- XML/EDI attempts to also address global repository problems. This holy grail (such as attempted by the BSR) is difficult to achieve, and in the first instance only of interest to very large organisations. It is unclear that this is a priority for many users.

- XML/EDI also attempts to address central message repositories, again a holy grail area, of which the Federal Government repository is the best example to date, while both MSHUG and Andover have a more limited concept of message factories. Again this is an area, which to date has only appealed to very large communities.

-

The XML/EDI community argue that the following are binding useful elements:-

- a tag respository, based possibly on the BSR -Business Semantic Register, an ISO project, with XM/EDI envisaging a repository of all elements used and DTDs etc.

Three specifications to be published shortly

- position document on repositories (with input from NIST and BSR). A first draft of this report has recently been made available by the XML/EDI group.
- templates (for conformance/validation)
- flow engine logic concept (for work flow)

A real issue remains as to whether TC251 requirements could be successfully met by XML/EDI.

On the one hand it will impose some constraints on the way HGMDs are represented and there may even need to be some specific support elements in XML/EDI to support the HGMD requirement, or indeed other requirements of TC251.

As a benefit, it may provide a homogeneous environment in which software vendors can develop applications able to flexibly support a number of EDI syntaxes.

There is a plan to commence a European pilot of XML/EDI and an inaugural meeting was held in June under the auspices of the CEN ISSS electronic commerce workshop. It may be worth maintaining close touch with any developments. A final project timescale has yet to be established.

Annex 7
X12 work

Study of comparable issues in the ANSI X12 community is well advanced. The status of the work can be found at: http://www.commercenet.co/projects/x12-xml

There are some interesting components to the work, which as a starting point, has again elements in common with the TC251 requirement, as also evidenced in the work of HL7. The work is being carried out with the involvement of members of the XML/EDI group and therefore there is an intention to use the XML/EDI approach wherever possible *(and presumably extend XML/EDI in cases where X12 demands it).*

Typing issues: They have suggested addressing the basic typing problems by attaching a fixed attribute to an element definition, which is its specific type. This would allow type checking by a type checking analyser. The disadvantage of this approach would be that an additional phase is introduced after the parsing of the document/message and before any more specific validation through an XSL approach etc.

X12 have also looked at the ISO 10744 extension to SGML, which provides a formal method to define lexical types.

Naming uniqueness: X12 have looked at the use of ISO/ITU OID (Object identifiers), where tags would consist of an ordered sequence of digits indicating the place in the hierarchy. (i.e. x12.3.1.850.111.100 would be X12 version 3, release 1,850 being a message type and then followed by the detailed element(s). Such long identifiers would however have a significant effect on the volume of data to be carried.

Annex 8
Task force preliminary plan

Preliminary Working Program
XML has some chances to become the generalised format for representing documents in networks. It is therefore necessary to investigate the new possibilities, chances and also drawbacks which ca occur when XML is also used as an interchange format for communication standards.
The task force should come out with recommendations concerning the future application of XML for this purpose. In the preliminary working program the following issues will be included and discussed within the task force.

1. Syntactic Recommendations
XML provides a large variety of possibilities for representing data which is appreciated in publishing. For using XML as an interchange format some recommendation have to be established to guarantee a more standardised approach. The task force have to find out where such rules are required and has to propose recommendations for an effective use of XML as an interchange format.

2. Application of complex Data Types
The currently used communication standards include a larger number of complex data types to represent relationships between data in an unambiguous way. It has to be investigated to what extend these data types are still required using XML resp. In what way they can be represented

3. Use of different element types
XML provides seven different element types including processing instructions, images, links to internet documents. It has to be investigated to which extend these element types can be used to enrich the application of communication standards

4. Requirements for a generalised tag repository
Tags become an important issue when using XML as an interchange format. They can be defined locally. But it seems to be recommendable to establish a common, generally applicable repository for tag data which could be available in the Internet. It has to be investigated to what extent the already delivered standard ENV 12537- `Registration of information objects used for EDI in
healthcare' can be applied for this task.

5. Application of DTD's in messaging
Documents in XML can be defined by Document Type Definitions (DTD). They describe the schematic composition of the document including element, attribute and entity definitions. XML documents (and also messages) can be validated by these DTD's. But XML allows also the definition of documents without providing a DTD. It has to be investigated to what extent DTD's will be required for messages, where the advantages are and where they are of no additional
value.

6. Validation of messages and its impact
One important advantage of XML is the validation process carried out automatically by XML browsers. The impact of this feature has to be considered

7. Impact of the application of XML in messaging for document handling.
The application of XML in messaging can influence the handling of documents to a certain extent. The task force has to consider these implications in all of its recommendations.

8. Application of the developed recommendation to one already available CEN Standard
The developed recommendations have to be applied to a CEN communication standard which is already published or. in the status of development. The applicability of XML has to be shown at least with one set of messages.