# SGML and Office Document Management Systems: Tools for Building Code Writers

J. R. Thomas and J. L. Worling[1]

## Abstract

In conjunction with research being conducted into providing computer based tools to support code users the National Research Council of Canada has also undertaken a research and development program to support the Authors of Code documents. Support for the Code Writers has been based upon the adoption of the Standard Generalized Markup Language (SGML) as a mechanism for supporting the management, publication, and information enrichment of the code development and delivery process. All of the current National Building Code documents have been converted to SGML and an Office Document Management System (ODMS) has been implemented to support this process. In addition, a number of prototype authoring tools have been developed to both hide the code authors from the SGML encoding of their text and at the same time ensure that the text of the code articles are consistent with the rest of the document.

## Introduction

In recent years there has been considerable interest and research into providing computer based tools for the support of users of building code documents [Bourdeau91; Vanier89]. In comparison there has been little work carried out into supporting the code writing process which is surprising considering that many of the problems which the developers of code user support tools face could be significantly reduced by providing the right information within the code documents during the authoring process.

The work reported here arouse out of two problems that we were facing. First, the research that we were conducting into developing Code Compliance Checking tools [Cornick90] had, amongst other things, led us to the conclusion that we needed to enrich the basic code documents with information over and above that which is needed in the legal text. The second problem was raised by the code writers themselves, who were beginning to recognize that they were having problems in both maintaining the highly complex documents and in minimizing the work required to publish both paper and electronic versions of their documents.

## What is SGML?

The Standard Generalized Markup Language [ISO 8879-1986] is a meta-language for describing the internal structure of document types and for tagging the structural elements of documents. As such it provides a:

- method of marking up a document for printing;
- vehicle for electronic document interchange
- method for encoding "informational markup" within a document
- basis for "Hypertext" and "Hypermedia" documents

[1] Both Authors are members of the National Research Council of Canada, Institute for Research in Construction, Ottawa, Ontario, Canada.

The biggest impact on the establishment of SGML, has been its required use within the US Department of Defense CALS initiative - (CALS Computer Aided Logistics Support ). It has now being taken up by many of the worlds publishing houses (e.g. Springer-Verlag) and being promoted by their trade associations (e.g. Association of American Publishers). Many national and international bodies are adopting SGML for use in their own publishing activities (e.g.. European Parliament) and it is being adopted by a number of Codes and Standards bodies (e.g. International Standards Organization) for document dissemination and publication.

The stimulus for our own adoption of the Standard Generalized Markup Language (SGML) has been the need to provide the base documents in a format which can be used to support traditional paper-based and electronic publication. The advantages of the SGML approach are that it provides a mechanism for enriching the information content within the document and provides the ability to treat the document as an information database.

Our problem is further complicated by the need to publish our documents in both official languages and the need to maintain consistent representation of accented characters over multiple hardware and software platforms (ease of document interchange).

A typical SGML file consists of four components:

- SGML declaration (Reference Concrete Syntax)
- DTD (Structure Markup, content Markup, External notation etc.)
- Document Instance (Document)
- External files --> Graphs etc.

An example of markup is shown below (Figure 1.) and in Table 1. with explanatory annotations in the right hand column:

```
<PARTID="H1"><NUM>1</NUM><H0T>Application and Definitions </H0T>
<TOC LEVEL="2"><H1ID="H11"><NUM>1.1</NUM><H1T>Application
</H1T>
<H2ID="H111"><NUM>1.1.1.</NUM><H2T>General</H2T><H3ID="H1111"
> <NUM>1.1.1.1.</NUM><H3T> Responsibility </H3T><P> Alternatives to
requirements in this Code may be permitted provided the <TERM REFID="THRT-
HV-">authority having jurisdiction </TERM> is satisfied that the existing fire
protection measures provide an <TERM REFID = "CCPTBL"> acceptable
</TERM> degree of fire safety</P>
```

**Figure 1.** Example of SGML marked-up document

The SGML markup can be interpreted by an appropriate application program and in conjunction with a set of directions for associating display characteristics with the individual markup components of the document a fully formatted version of the document can be produced (Figure 2).

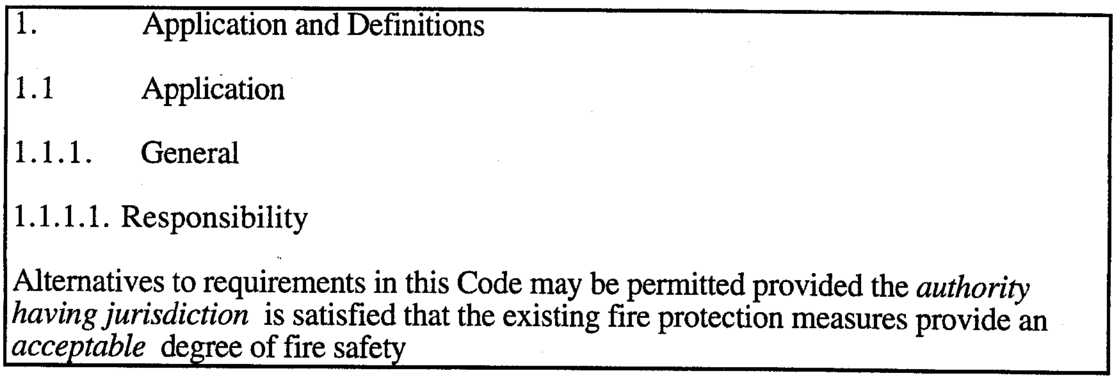
The Document Type Definition (DTD) provides a mechanism for describing both the content and structure of a single or class of document. It also provides for the encoding of additional information within the text which can be used by programs that are designed to make use of such information. One such program is the MiniCode Generator [Thomas92] which is able to make use of multiple attribute - value pairs which are tagged in association with each of the articles of the Code document.

| | |
|---|---|
| <<PARTID="H1"> | H1 is the top level of the document hierarchy. Its id is "H1" |
| <NUM>1</NUM> | the printed section number is "1" |
| <H0T>Application and Definitions </H0T> | the heading of this section is "Application and Definitions" |
| <TOC LEVEL="2"> | level 2 in Table of Contents |
| <H1ID="H11"> | second level with an id of "H11" |
| <NUM>1.1</NUM> | the printed sub-section number is "1.1" |
| <H1T> Application</H1T> | this sub-level heading is "Application" |
| <H2ID="H111"> | third level with an id of "H111" |
| <NUM> 1.1.1.</NUM> | the printed sub-section number is "1.1.1." |
| <H2T>General</H2T> | his sub-level heading is "General" |
| <H3ID="H1111"> | fourth sub-level with an id of "H1111" |
| <NUM>1.1.1.1.</NUM> | the printed sub-section number is "1.1.1.1." |
| <H3T> Responsibility </H3T> | his sub-level heading is "Responsibility" |
| <P> Alternatives to requirements in this Code may be permitted provided the <TERM REFID="THRT-HV-">authority having jurisdiction </TERM> is satisfied that the existing fire protection measures provide an <TERM REFID = "CCPTBL"> acceptable </TERM> degree of fire safety</P> | This is the contents of a provision<br>This is a definition with an id "THRT-HV-"<br>This is a definition with an id "CCPTBL" |

**Table 1.** Example of Standard Generalized Markup Language.

SGML allows the data suppliers to markup their information according to the structure of the documents as opposed to encoding the output format. It provides a method for the interchange of document instances independent of both hardware and software platforms. For example, the International Standards Organization (ISO) has developed a DTD for publishing their own standards which has been adopted by both the Canadian Standards Association and various other national organizations. This standardization will greatly assist the flow of accurate and timely information not only between Standards Organizations but also to users.

A number of the projects currently underway at IRC involve either adding additional information, such as classification information for specific clauses of documents, or inter-linking various sections of the document with hypertext capabilities. For example, using the functionality provided by SGML, NRC is able to encode the Provincial variants of the prototype national code documents within a single document structure.



1. Application and Definitions

1.1 Application

1.1.1. General

1.1.1.1. Responsibility

Alternatives to requirements in this Code may be permitted provided the *authority having jurisdiction* is satisfied that the existing fire protection measures provide an *acceptable* degree of fire safety

**Figure 2.** Example of displayed text

## Office Document Management System

We have undertaken to develop new ways of working with our Codes documents and we needed to implement a system which would integrate the whole process. This includes not only maintaining and developing code documents but also the considerable correspondence and committee documents that are part of the process. We needed to implement a system which provided:

- version control;
- timeliness of information (e.g. referencing the most current version of a standard);
- assurance that all "public review" comments are processed to completion;
- tracking status of tasks and documents (e.g. graphics produced by outside consultants) and;
- the integration of database techniques into the document itself (e.g. all references to a standard in the document are pointers to a database entry).

Initially we evaluated a commercial ODMS system but rapidly discovered that it had never been loaded with the number of documents that we had on the system at that time with the result that we exceeded many inherent system limits. Once we had managed to have the system limits adjusted we discovered that running the system with even a limited number of user resulted in a totally unacceptable response time from the system. As a result we embarked upon the development of a custom designed solution using outside consultants to put together an initial prototype system. Although the prototype system did provide an initial system that met many of our requirements it required in excess of one additional person year of effort to re-code much of the original system in addition to adding new functionality.

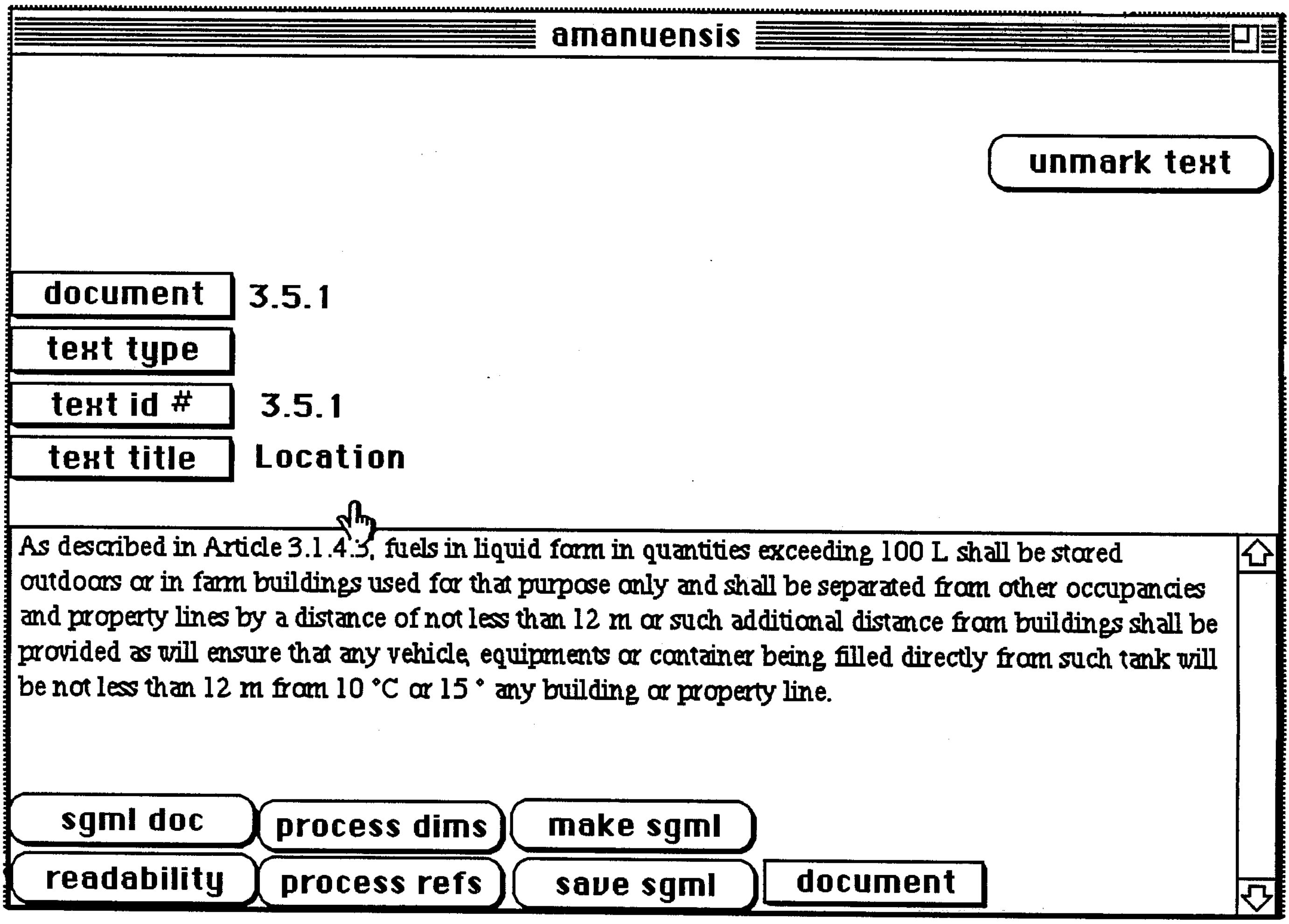
One of the key aspects of implementing the ODMS was the management of the process of introducing the system to the users. Rather than introducing them to a complete system at the start, the system was introduced over a period of time and involved phasing in the users access to new functionality. In part this was necessary as a considerable number of the users were changing from a PC platform to a Macintosh platform at the same time as having to adapt to the new document management system.

## Tools for Authors and Users

Although the ODMS helps support the management of the documents that are involved in the code development process it was evident that there was a need for additional tools to support the authoring process itself. These tools needed to provide:

- an SGML "Hiding" environment;
- consistency checking of language use;
- reference checking (e.g. recognized Standards);
- text "style" enforcement (e.g. we use italics to indicate a defined term);
- grammar checking and "reading level" checking;
- cross reference linking (e.g. checking the target reference exists).

Two prototype authoring tools were developed which provided most of the above functionality. These were developed using Hypercard on the Macintosh along with some additional functions programmed in "C". The initial system, called amanuensis, provided the author with a form filling environment in which the user entered the necessary details to identify the location of the article within the code document (article number and title) as well as the text for the article (see figure 2). This functionality was further extended by a later system called cobre.



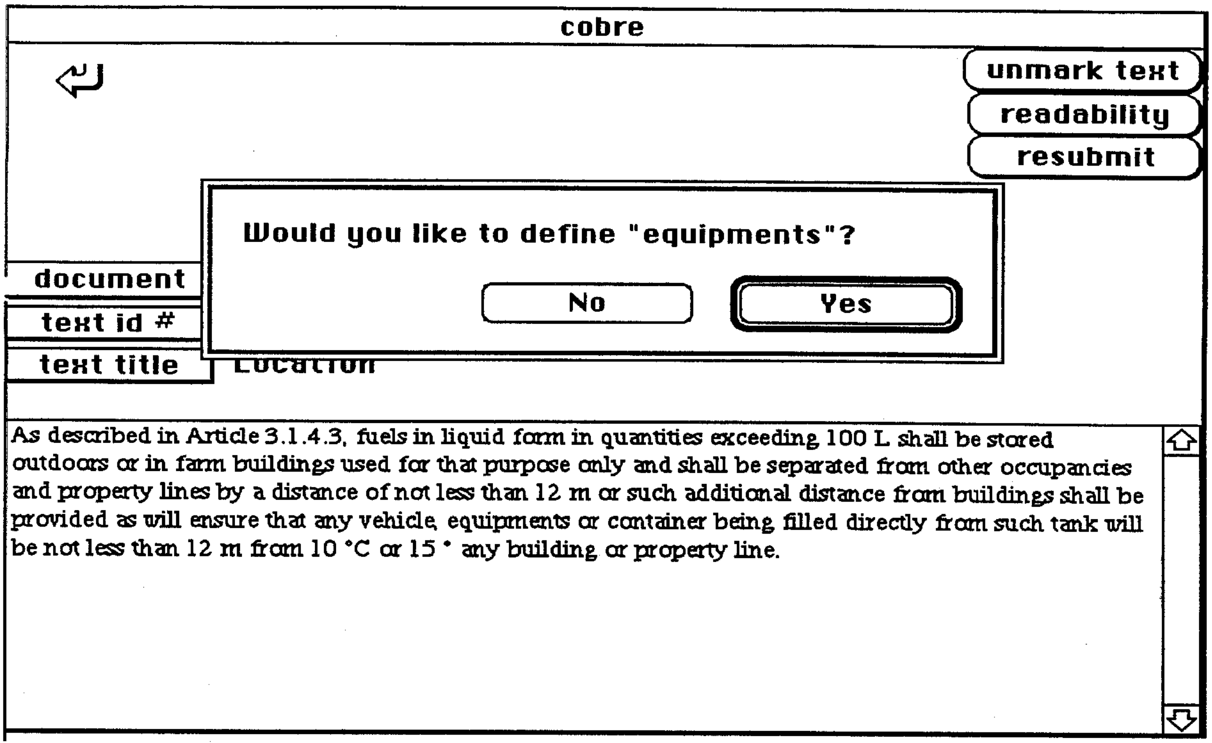
document 3.5.1

text id # 3.5.1

text title Location

As described in Article 3.1.4.3, fuels in liquid form in quantities exceeding 100 L shall be stored outdoors or in farm buildings used for that purpose only and shall be separated from other occupancies and property lines by a distance of not less than 12 m or such additional distance from buildings shall be provided as will ensure that any vehicle, equipments or container being filled directly from such tank will be not less than 12 m from 10 °C or 15 ° any building or property line.

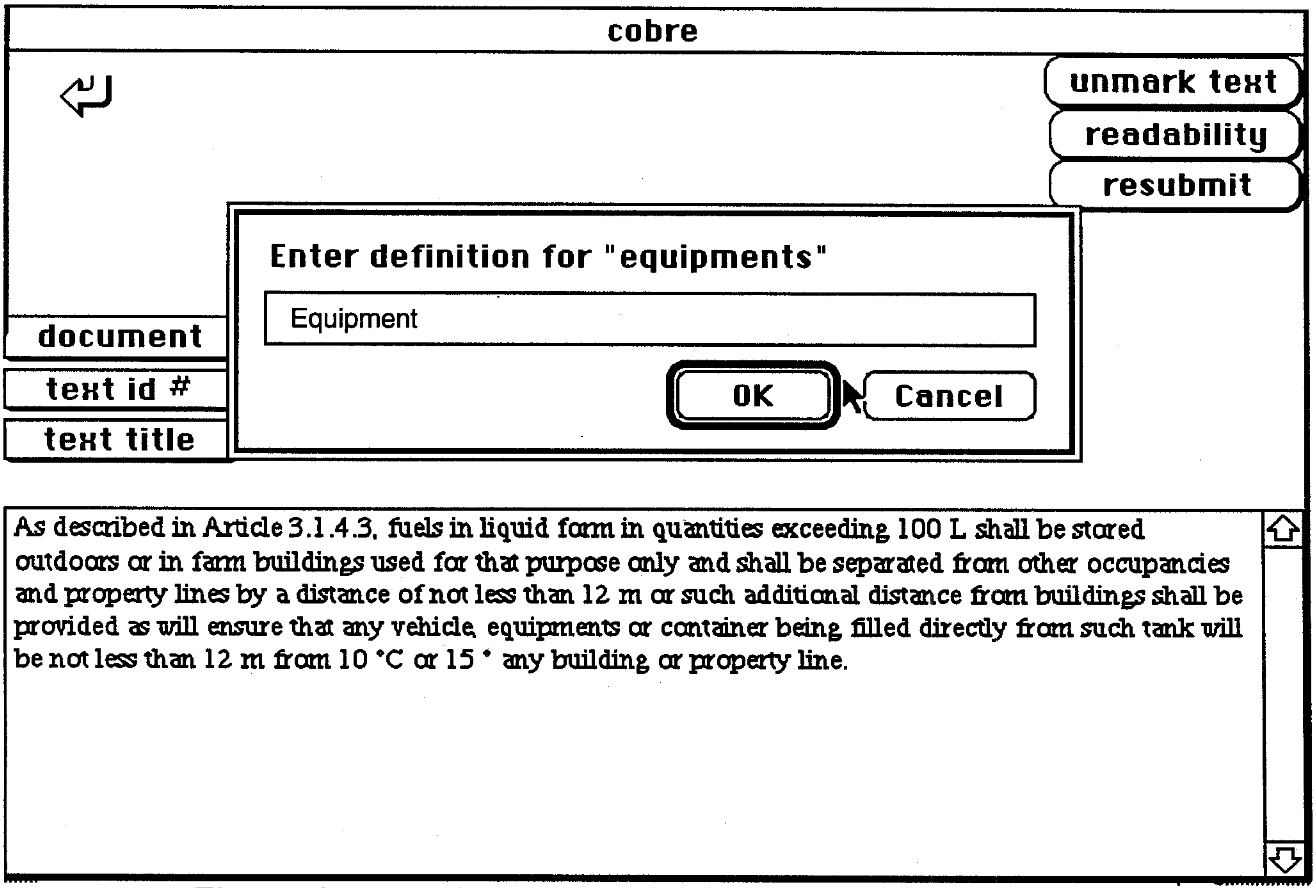
**Figure 2.** Article Text entered into the Form Slots in the amanuensis system.



Would you like to define "equipments"?

As described in Article 3.1.4.3, fuels in liquid form in quantities exceeding 100 L shall be stored outdoors or in farm buildings used for that purpose only and shall be separated from other occupancies and property lines by a distance of not less than 12 m or such additional distance from buildings shall be provided as will ensure that any vehicle, equipments or container being filled directly from such tank will be not less than 12 m from 10 °C or 15 ° any building or property line.

**Figure 3.** Identifying an undefined term.



Enter definition for "equipments"

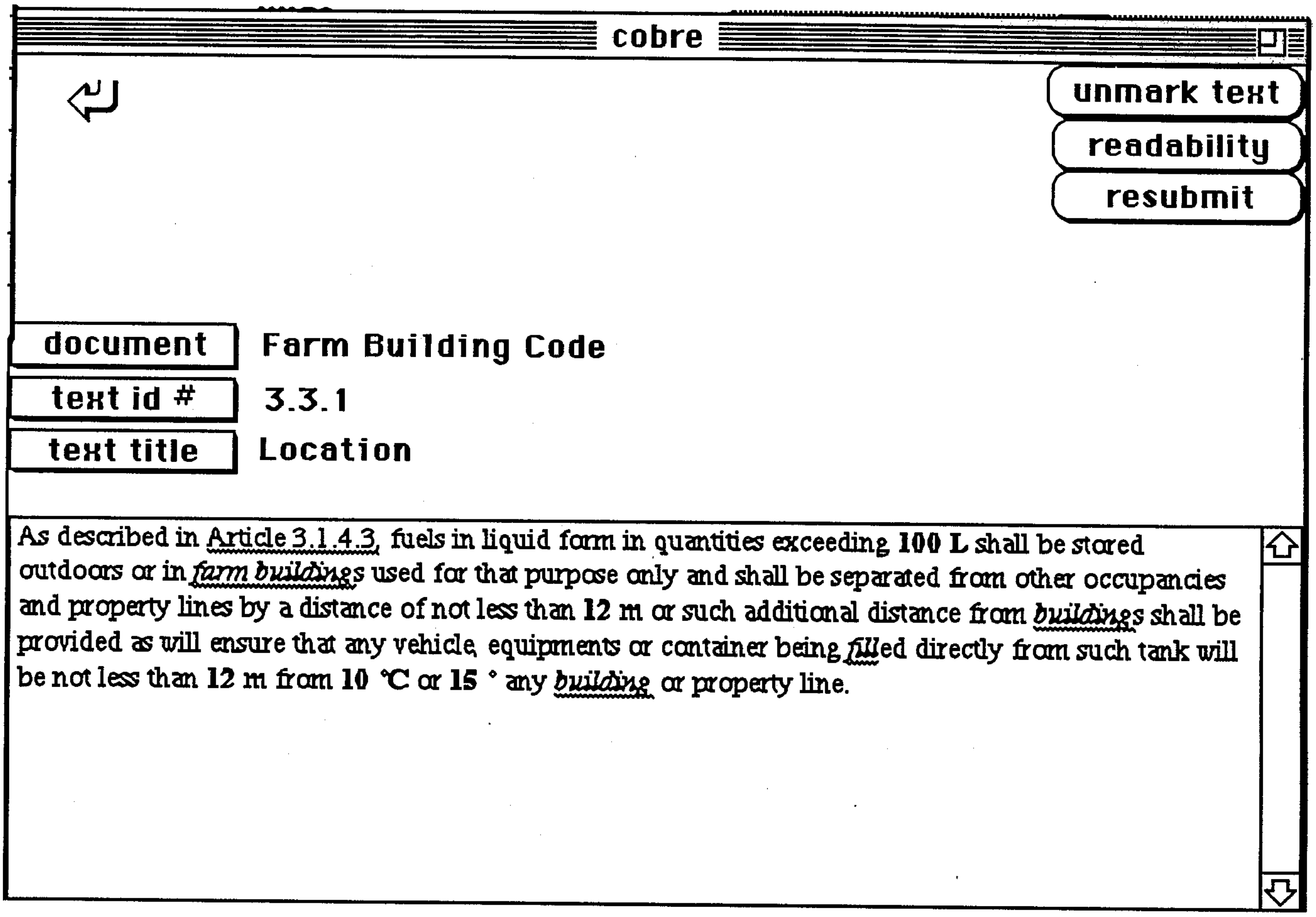
Equipment

As described in Article 3.1.4.3, fuels in liquid form in quantities exceeding 100 L shall be stored outdoors or in farm buildings used for that purpose only and shall be separated from other occupancies and property lines by a distance of not less than 12 m or such additional distance from buildings shall be provided as will ensure that any vehicle, equipments or container being filled directly from such tank will be not less than 12 m from 10 °C or 15 ° any building or property line.

**Figure 4.** Entering a definition for a new defined term.

Once the required slots had been completed the system would then proceed to parse the text of the article to identify the existence of a number of key components. These included defined terms; units of measurement; references to known Standards; internal references within the document and; if terms were discovered that were not in the lexicon then the user would be asked if they wished to have the new term added as a defined term within the document(See Figures 3 and 4.)

Once all the components of the new article have been identified by the system, often through the process of a dialogue with the author, the completed text is presented in a visual form so that the author can see that the different components of the text have been captured by the system (see figure 5.)



document Farm Building Code

text id # 3.3.1

text title Location

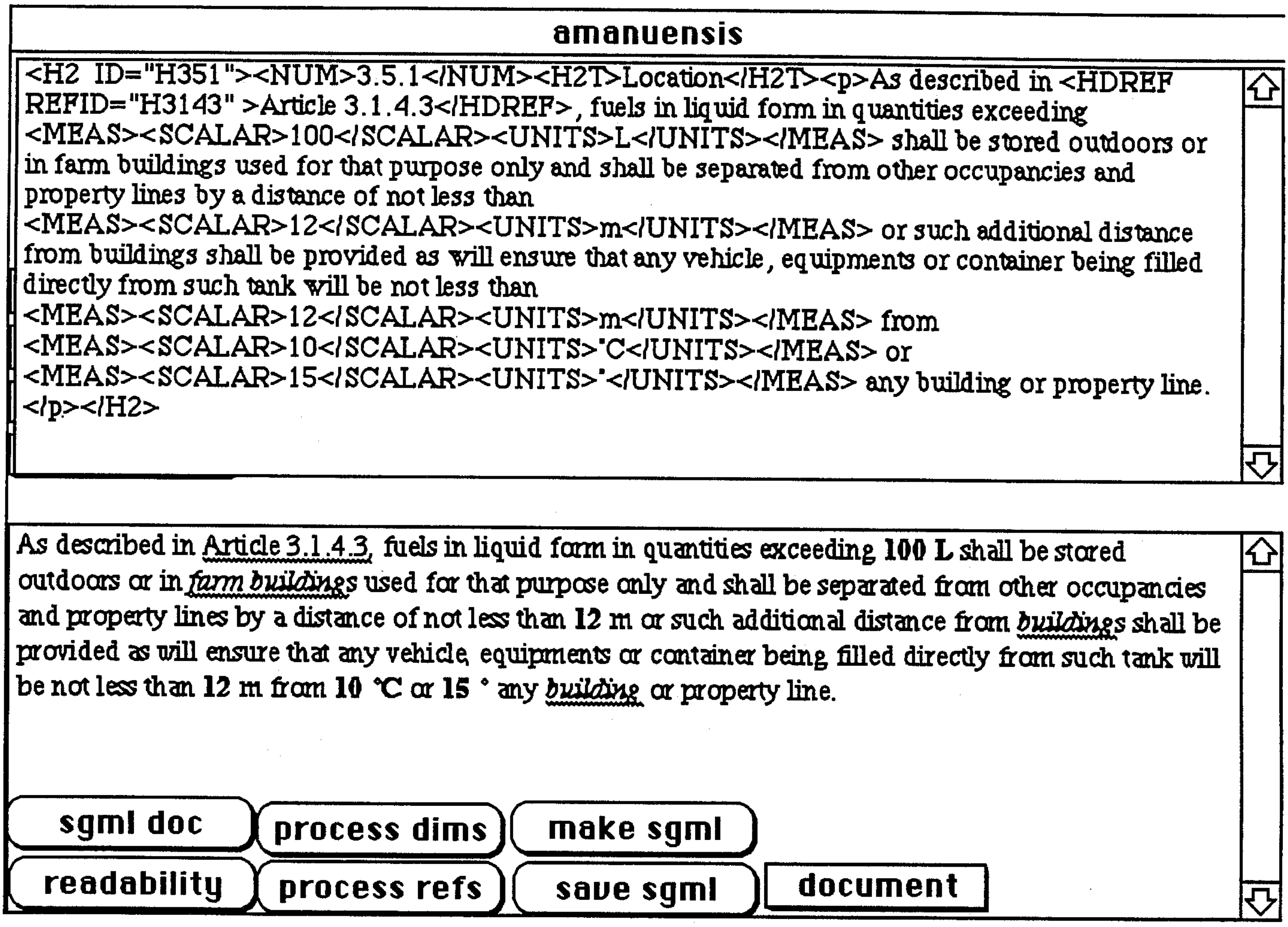
As described in Article 3.1.4.3, fuels in liquid form in quantities exceeding **100 L** shall be stored outdoors or in *farm buildings* used for that purpose only and shall be separated from other occupancies and property lines by a distance of not less than **12 m** or such additional distance from *buildings* shall be provided as will ensure that any vehicle, equipments or container being *fill*ed directly from such tank will be not less than **12 m** from **10 °C** or **15 °** any *building* or property line.

**Figure 5.** Article text using visual features (bold, italics, etc.) to indicate the different components of the text.

In addition to the visual features of the text the author can also view the SGML tagged version of the article that has been generated in the background within the system (see figure 6.). When the system saves the authors work it is the SGML marked up version of the article that is saved and when this is accessed by the author on a subsequent occasion the system will interpret the SGML markup and display only the visual cues that identify the components of the text (Figure 5).

Work is still continuing on developing these authoring tools as it is clear that authors do not wish to have to work directly with the SGML marked-up text. It is also clear that these tools also provide valuable help in ensuring that the documents are internally consistent (cross reference to the correct articles, etc.) as well as ensuring that they reference the current versions of external documents (Standards, etc.).

One final aspect of the authoring process that we are also attempting to capture are the rational underlying the articles that are part of the code documents. This follows on from the work of Stone and Tweed [Stone91] who indicated that capturing the argumentation that occurred during the process of developing a new code article could be useful to code enforcement personnel at a later date when seeking to interpret the intent of the article. We are currently investigating making similar texts, which outlines the objective of the article, available in electronic versions of our documents.



```
<H2 ID="H351"><NUM>3.5.1</NUM><H2T>Location</H2T><p>As described in <HDREF
REFID="H3143" >Article 3.1.4.3</HDREF>, fuels in liquid form in quantities exceeding
<MEAS><SCALAR>100</SCALAR><UNITS>L</UNITS></MEAS> shall be stored outdoors or
in farm buildings used for that purpose only and shall be separated from other occupancies and
property lines by a distance of not less than
<MEAS><SCALAR>12</SCALAR><UNITS>m</UNITS></MEAS> or such additional distance
from buildings shall be provided as will ensure that any vehicle, equipments or container being filled
directly from such tank will be not less than
<MEAS><SCALAR>12</SCALAR><UNITS>m</UNITS></MEAS> from
<MEAS><SCALAR>10</SCALAR><UNITS>°C</UNITS></MEAS> or
<MEAS><SCALAR>15</SCALAR><UNITS>°</UNITS></MEAS> any building or property line.
</p></H2>
```

As described in Article 3.1.4.3, fuels in liquid form in quantities exceeding **100 L** shall be stored outdoors or in *farm buildings* used for that purpose only and shall be separated from other occupancies and property lines by a distance of not less than **12 m** or such additional distance from *buildings* shall be provided as will ensure that any vehicle, equipments or container being filled directly from such tank will be not less than **12 m** from **10 °C** or **15 °** any *building* or property line.

**Figure 6.** A display of both the SGML Coded version of the article text as well as the text using visual cues (Bold, italics, etc.).

## Conclusions

The move to using SGML as a method of marking up our documents has proven to be valuable. Not only have we been able establish a method of publishing in both paper and electronic format from single source documents but we have also managed to enrich our documents with additional information which can be used for other applications. We have begun to produce knowledge based applications which are able to make use of this information (e.g. MiniCode system) and we expect to see further developments of such tools in the future. Many of these new tools are likely to be hybrids and to represent a marriage of expert systems and hypertext systems (ExperText) where the intent is to provide intelligent dynamic documents which are configurable to the users needs.

As the codes development process becomes more complex, it is clear that we need to make the best use of our computing resources to help manage the process. The existing ODMS and authoring tools are only a start in this direction. They will enable us not only to produce better code documents but also enable us to provide the users of the documents with a clear indication of the objectives which each article is attempting to achieve.

## Acknowledgments

We would like to acknowledge our thanks to Daniel Jaliff who was responsible for programming the Amanuensis and Cobre authoring tools.

## References

[Bourdeau91] Bourdeau, Marc. The CD-REEF: The French building technical rules on CD-ROM. In, Computers and Building Regulations, VTT Symposium Series, No. 125, Espoo, Finland, 27-29 May, pp. 1117-133, 1991.

[Cornick90] Cornick, S.M., Leishman, D.A., Thomas, J.R., "Incorporating Building Regulations into Design Systems: An Object Oriented Approach". ASHRAE Transactions, Vol. 96(2), pp. 542-549, 1990.

[ISO 8879] ISO 8879:1986 Information Processing -- Text and Office Systems -- Standard Generalized Markup Language (SGML). International Organization for Standardization. Ref. No. ISO 8879:1986 (E). Geneva/New York, 1986.

[Stone,91] Stone, D. & Tweed, C., "Representation issues in the design of an intelligent information system for building standards". In, Computers and Building Regulations, VTT Symposium Series, No.125, Espoo, Finland, 27-29 May, pp. 237-249, 1991.

[Thomas92] Thomas, R., Cornick, S.M., Leishman, D.A., & Vanier, D.J. (1992) "Research at The Advanced Construction Technology Laboratory." In, D.E. Grierson, G. Rzevski & R.A. Adey (Eds.) Applications of Artificial Intelligence in Engineering VII. pp. 35 - 50, Co-published by Computational Mechanics Publications: Southampton U.K. & Elsavier Applied Science: London U.K, 1992.

[Vanier89] Vanier, Dana J., "Computerization of Building Regulations", In Proceedings of the International Conference on Municipal Code Administration, Building Safety, and the Computer, Winnipeg, Manitoba, 24-28 Sep., pp. 43-62, 1989.