Structuring Documents: the Key to Increasing Access to Information for the Print Disabled

Bart Bauwens, Jan Engelen, Filip Evenepoel

Katholieke Universiteit, Leuven
Kardinaal Mercierlaan 94, B-3001 Leuven, Belgium
Email: engelen@esat.kuleuven.ac.be

Chris Tobin, Tom Wesley

University of Bradford
Department of Computing, Bradford, BD7 1DP, United Kingdom
Email: t.a.b.wesley@bradford.ac.uk

Abstract. There is a growing conviction that the Standard Generalized Markup Language, SGML, can play an important role as an enabling technology to increase access to information for blind and partially sighted people. This paper reports on mechanisms that have been devised to build in accessibility into SGML encoded electronic documents, concentrating on the work done in the CAPS Consortium—Communication and Access to Information for People with Special Needs, a European Union funded project in the Technology Initiative for Disabled and Elderly People (TIDE) Programme—and by ICADD, the International Committee on Accessible Document Design. The CAPS follow on project, HARMONY is briefly described.

1 Introduction

The main work of the CAPS Consortium (Communication and Access to Information for People with Special Needs, a European Union funded project in the Technology Initiative for Disabled and Elderly People (TIDE) Programme) is directed to developing methods to increase the access to information for the print disabled [1]. People with print disabilities include the blind, the deaf-blind, the partially sighted, the dyslexic and those with motor impairments which make it difficult to physically control paper documents.

A previous paper [2] has indicated that one of the significant limiting factors for the print disabled is the difficulty they face in accessing the predominant form of information provision, which is almost entirely oriented to printed and other visual forms. The proportion of information easily accessible to the print disabled is very small. This is so for two main reasons:

- provision of information in forms suitable for the print disabled is regarded as a peripheral activity in comparison to provision for normally sighted persons;
- the means to produce information in forms suitable for the print disabled, such as braille, large print or synthetic speech, are slow, manually intensive and divorced from the initial information creation and distribution processes.

CAPS believes that a vital factor in creating the environment for such improvements

is to develop methods in which the provision of information for the print disabled is, as far as possible, an automatic supplementary process related to the normal information creation processes. Technologically, this can be achieved through application of the developments of *standardised structured electronic documents.*

Electronic documents are the key to linking into the commercial information production processes, as increasingly these processes are electronically based; and to the transformations required to make the information accessible to the print disabled.

The importance for the print disabled of *structure* in electronic documents can be realised when it is considered how the normally sighted reader obtains a significant amount of information from the layout of a document—titles in bold, bulleted indents, emphasised sections in italics. These are crucial when browsing through a large document. To make this information available to aid the print disabled user to browse—or navigate—within a document, the structure needs to be defined explicitly within the electronic document.

However, the transformations into forms accessible for the print disabled are, in many cases, non–trivial and this points to the need for *standardised* structured electronic documents. Having once made the transformations for *standardised structured* documents, significant amounts of information can rapidly be made available to the print disabled as the use of such standards grows in the commercial world.

An important document system standardised by ISO, the International Standards Organisation is the Standard Generalized Markup Language (SGML) [3]. The CAPS Consortium has recognised the potential of this standard for increasing the access to information for the print disabled, and Engelen and Wesley [4] have given a general account of this potential. The use of SGML as an internal format in the publishing industry is growing rapidly as publishers recognise the value and power of using a truly international standard for encoding their electronic documents.

A major part of the CAPS Project is devoted to developing methods whereby SGML is used at the heart of a generic model for dramatically improving the access to information for the print disabled. Within the current phase of the Project, which will complete at the end of 1994, a Pilot Electronic Library is being set up. Access will be provided interactively using synthetic speech on both an adapted work station and also through the home telephone using a voice response system with high quality real time text to speech. In addition, the provision of information through the off line production of braille and large print versions is being investigated.

This paper describes the techniques that the CAPS Consortium has devised to incorporate accessibility into SGML encoded documents.

2 An Overview of SGML

The complexity of real world documents is often not fully understood. SGML, in creating a standard for the description of such documents, is therefore, itself complex. A useful practical technical description is by van Herwijnen [5], while perhaps the easiest general guide is in the Text Encoding Initiative (TEI) Guidelines [6]. This overview draws particularly from the latter.

SGML is an international standard for the description of marked up electronic text. More exactly, SGML is a *metalanguage*, that is, a means of formally describing a language, in this case, a *markup language*. Historically, the word markup has been

used to describe annotation or other marks within a text intended to instruct a compositor or typist how a particular passage should be printed or laid out. As the formatting and printing of texts was automated, the term was extended to cover all sorts of special markup codes inserted into electronic texts to govern formatting, printing, or other processing. Generalising from that sense, *markup* is defined as any means of making explicit an interpretation of a text.

By *markup language* is meant a set of markup conventions used together for encoding texts. A markup language must specify what markup is allowed, what markup is required, how markup is to be distinguished from text, and what the markup means. SGML provides the means for doing the first three; application programs which process the markup are written with an understanding of what the markup means.

2.1 Characteristics of SGML

There are two characteristics of SGML which distinguish it from other markup languages: its emphasis on descriptive rather than procedural markup and its document type concept.

Descriptive Markup. A descriptive markup system uses markup codes to provide names to categorise parts of a document. Markup codes such as `<para>` simply identify a part of a document as a 'para'. By contrast, a procedural markup system defines what processing is to be carried out at particular points in a document. In SGML, the instructions needed to process a document for some particular purpose (for example, to format it) are sharply distinguished from the descriptive markup which occurs within the document. Usually, they are collected outside the document in separate procedures or programs.

With descriptive instead of procedural markup the same document can readily be processed by many different pieces of software, each of which can apply different processing instructions to those parts of it which are considered relevant. For example, a braille conversion program can apply appropriate rules for the correct formatting of paragraphs in the various national literary braille codes.

Types of Document. Secondly, SGML introduces the notion of a *document type*, and hence a *document type definition* (DTD). Documents are regarded as having types, just as other objects processed by computers. The type of a document is formally defined by its constituent parts and their structure. The definition of a report, for example, might be that it consisted of a title and possibly an author, followed by an abstract and a sequence of one or more paragraphs. Anything lacking a title, according to this formal definition, would not formally be a report.

If documents are of known types, different documents of the same type can be processed in a uniform way. This separation of the definition of the document *type* from the document *instance* is a key feature for our purposes. If a particular DTD can be made accessible, then all document instances are made accessible.

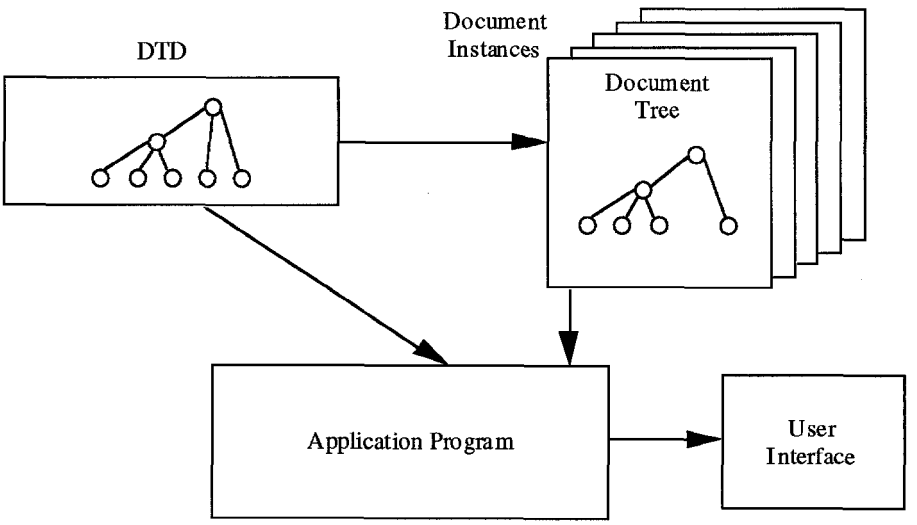
2.2 Defining SGML Document Structures: the DTD

A DTD is expressed in SGML as a set of declarative statements, using the syntax defined in the standard. The DTD defines what markup is allowed, the major unit of

markup being an *element.* A fragment of a DTD for a publication of bibliographic information might be:

```
<!ELEMENT publ  - -    (date, title, sec1+) >
<!--            Publications have a date to identify this occurrence
                along with the title; the title is followed by at
                least 1 but typically many sections. -->
<!ELEMENT sec1  - -    (heading, ((data, sec2*) | sec2+ )) >
<!ELEMENT sec2  - -    (heading, ((data, sec3*) | sec3+ )) >
<!ELEMENT sec3  - -    (heading, data) >
<!--            Sections may be nested to three levels. Each section
                has a heading followed by a mixture of subsections
                and data. -->
```

As can be seen from the comments embedded within the fragment, the DTD defines the properties of a set of document instances. Thus, the DTD is a formal, parsable specification of the structure of a class of documents. However, SGML imparts no semantic significance to the elements defining the structure, such as 'heading' or 'title'. Of course, it is good practice to name elements so that they can generally be understood; however, the semantic significance is generally incorporated in the application programs which process SGML documents. This process can be illustrated diagrammatically:

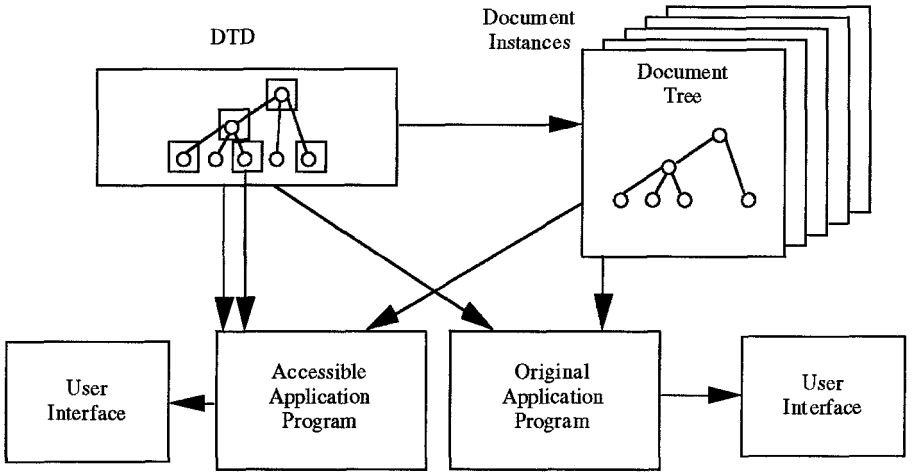


in which an application program, guided by the DTD, delivers a set of document instances to a given user interface. For our purposes, it is important to observe that the application program is likely to provide the information without consideration of accessibility for people with print disabilities.

3 SGML Associated Specifications

3.1 General Principles

There is a need to add extra information to DTDs to enable access for people with print disabilities to documents conforming to those DTDs. CAPS has called this extra information, *Associated Specifications*. An accessible application model is shown diagrammatically:



This model shows a number of fundamental features. Firstly, Associated Specifications, in the form of SGML conformant syntax are added to the DTD, *not* to the document instances. Thus, a whole class of documents is made accessible, without the major resources required to edit individual documents. Secondly, the Associated Specifications are added to the DTD—nothing is removed; thus, the original application program can be used without modification. This conforms to the principle that accessibility should be a by-product of the normal production processes. Finally, it should be noted that no new document formatting standards are needed—the Associated Specifications can be fully defined in SGML.

CAPS recognised the following main design principles for Associated Specifications which should allow:

- the transformation of electronic documents to a variety of accessible formats, such as braille, Moon and large print;
- the possibility of associating a text string with each element class that would ease the understanding of the meaning of the element class for an end user during interactive use;
- the addition of extra text that would normally be generated in the ink print version so that it may also occur in accessible versions;
- elements with the same generic identifier to have different treatment depending upon their context or relative position in the document;
- interactive operations that use the structure, such as overviews using headings and navigation through a document from heading to heading;
- multiple languages—any product that is to be distributed across Europe or wider

must be able to cater for an arbitrary set of languages.

In developing the concept of Associated Specifications, CAPS has worked closely with ICADD, the International Committee on Accessible Document Design [7]. This Committee, a non–profit organisation, incorporated in the State of New Hampshire, has the aim of developing techniques and raising awareness to enable documents to be made available to persons with print disabilities at the same time and at no greater cost as the print enabled community enjoys. Not surprisingly, given the common nature of both the needs and the potential technological solutions, the approaches taken by CAPS and ICADD have converged. CAPS in addition is gaining advantage from the legislative push being provided by ADA, the Americans with Disabilities Act.

In a remarkable development, ICADD has managed to have its mechanisms for accessibility incorporated into a new ISO Standard DTD for Electronic Manuscript Preparation and Markup [8]. This is, as far as is known, the first time that disability issues have been directly incorporated into a standard for *commercial* use. If, as seems likely, publishers start using the standard, accessibility will be automatically built in to any document instances that are produced.

3.2 The Basic Mechanism

The Associated Specifications developed by CAPS are fully documented in [9] and described in [10]. They are a superset of the ICADD mechanism detailed in [8], in which a number of fixed SDA (SGML Document Access) attributes are used to map an arbitrary DTD into a simple fixed ICADD DTD. In this paper it is only possible to sketch briefly the technical details.

The Base Tag Set. The complexity of real world documents means that DTDs used commercially are complex, and define many elements. The Book DTD defined in ISO 12083 has, for example, about 150 elements. Given the relative simplicity of the formatting available in braille, ICADD defines a simple set of 22 elements which can guide the production of accessible forms. A group of elements (`h1` through `h6`) is used to define a hierarchy of headings; another group contains 'inline' elements such as `b` (bold), `it` (italics), `lang` (language). These elements may occur within, for example, the text of a paragraph and indicate another type of processing, for example, a `lang` element will switch the language.

Simple Mappings. One of the four fixed SDA (SGML Document Access) attributes is named `SDAFORM`:

```
<!ATTLIST  sectitle  SDAFORM    NAME   #FIXED  "h1">
```

In this example, the attribute `SDAFORM` indicates that wherever a `<sectitle>` element occurs, it should be mapped to an `h1` element.

One can declare simple context–sensitive mappings as follows:

```
<!ATTLIST  section   SDARULE    NAMES  #FIXED  "title h1">
<!ATTLIST  chapter   SDARULE    NAMES  #FIXED  "title h2">
```

This example defines two rules: the first is intended to be used within section elements, the second within chapter elements. Within sections the title will be mapped to `h1` headings, while within chapters the title will map to `h2` headings. Complex

mappings can be defined by more advanced techniques: if `title` appears in a `chapter` within a `part`, map to an `h2`; if the `chapter` is not in a `part`, map to an `h1`.

The third and fourth SDA attributes, named `SDAPREF` and `SDASUFF`, enable the replacement of start–tags (`SDAPREF`) or end–tags (`SDASUFF`) by specified text, for example:

```
<!ATTLIST  author    SDAPREF    CDATA  #FIXED  "Author name:">
```

3.3 Interactive Applications

Transforming the Original DTD. CAPS accepts that for non–interactive processes such as braille or large print on paper the transformation of a document so that it conforms to the simpler ICADD DTD seems acceptable. However, for interactive processes this loss of structure is *not* acceptable. For interactive applications, users may require much of the original document structure, and need to interpret or view it by means of its Associated Specifications.

Interactive Explanations. Users of interactive applications may need contextual information about where they are in a document. By keeping the original structure, this information could be provided by means of the element names of the SGML document. However, this is not appropriate—users should not be confronted with abbreviations for elements as declared in the DTD, which are designed for automatic processing by SGML aware applications. CAPS has therefore defined, for interactive applications the attribute, `SDAEXPL`, to explain element names. The following example associates with an element called `npinfo` the text `"Newspaper Information"`:

```
<!ATTLIST  npinfo    %SDAEXPL;   "Newspaper Information" >
```

4 HARMONY

As the current CAPS Project draws to a close, the work will be continued in a new EU TIDE funded project, HARMONY, Horizontal Action for the Harmonisation of Accessible Structured Documents [11].

One of the main CAPS results is the development of the CAPSNEWS DTD as an interchange format for electronic newspapers for the print disabled [12]. However, despite the technological progress, CAPS has recognised that the effort so far allocated to ensuring the wider acceptance—particularly by major publishing companies—of such developments and standards has been inadequate.

HARMONY will address the non technological barriers to the growth of electronic newspapers for the print disabled. Its main objective is "to increase the *quantity and quality* of information accessible to print disabled people—especially in daily newspapers—by stimulating the publishing community via a process of involvement, lobbying and standardisation, and by encouraging them to adapt their existing electronic production systems to make use of appropriate new document structuring concepts." It will also maintain a technical watch on developments in the SGML area so that they may be used as soon as possible for the print disabled.

Acknowledgement

This work has been partly funded by DGXIII of the Commission of the European Union, under its Technology Initiative for Disabled and Elderly People (TIDE) Programme. The other CAPS partners are the Royal National Institute for the Blind (UK), Sensotec BV (BE), Infovox (SE) and the Handicap Institute (SE).

References

1. Full details of the CAPS Consortium can be obtained from the Coordinator, Professor Jan Engelen at the Katholieke Universiteit, Leuven, Belgium. The Consortium maintains an ftp site, gate.esat.kuleuven.ac.be in the directory /pub/CAPS and its sub directories, which provides access to its latest public documents.
2. J. Engelen, J. Baldewijns: Digital Information Distribution for the Reading Impaired: from Daily Newspapers to Whole Libraries. In: The 3rd International Conference on Computers for Handicapped Persons. Vienna, 1992, pp. 144–149
3. ISO 8879 : 1986 Information processing—Text and Office systems—Standard Generalized Markup Language (SGML). International Organisation for Standardisation
4. J. Engelen, T. Wesley: SGML—A Major Opportunity for Access to Information. In: The Seventh International Conference: Technology and Persons with Disabilities. CSUN Los Angeles, 1992, pp. 593-598
5. E. van Herwijnen: Practical SGML, Second Edition. Kluwer, 1994
6. A Gentle Introduction to SGML. In: C. Sperberg–McQueen, L. Burnard (eds.): Guidelines for Electronic Text Encoding and Interchange. 1994. Up to date information can be obtained from L. Burnard, Oxford University Computing Services, 13 Banbury Road, Oxford, OX2 6NN.
7. The latest information about the International Committee on Accessible Document Design (ICADD) can be obtained from the President, Michael G. Paciello, 110 Spit Brook Road, Nashua, NH. USA 03062, phone: +1 603 881 1831, Email: Paciello@Shane.Enet.Dec.Com.
8. ISO 12083 : 1993 Information processing—Text and Office systems—Electronic Manuscript Preparation and Markup. International Organisation for Standardisation
9. CAPS Deliverable D1, Development of SGML Associated Specifications, August 1993 and Addendum to Deliverable D1, November 1993. Obtainable from the Coordinator [1]
10. B. Bauwens, J. Engelen, F. Evenepoel, C. Tobin, T. Wesley: SGML—An Enabling Technology for the Reading Impaired. In: SGML Europe '94. Montreux: Database Publishing Ltd, Swindon, UK 1994
11. Full details of the HARMONY Consortium can be obtained from the Coordinator, Professor Jan Engelen at the Katholieke Universiteit, Leuven, Belgium
12. European Interchange Format, CAPSNEWS DTD, Version 2.0, November 1993. Obtainable from the Coordinator [1]