SGML-Einführung

Oliver Corff Juni 1998

SGML steht für **S**tandard **G**eneralized **M**arkup **L**anguage und bezeichnet ein vereinheitlichtes Verfahren zur *inhaltlichen* Auszeichnung von elektronisch erstellten Texten, deren konkretes Erscheinungsbild vom jeweiligen Ausgabemedium bestimmt wird. In SGML geschriebene Texte lassen sich leicht pflegen; SGML ist daher besonders für Texte geeignet, die sich oft ändern oder an denen mehrere Autoren gleichzeitig arbeiten. Software zur Verarbeitung von SGML-Texten steht für eine große Anzahl von Betriebssystemen bereit; in den Anwendungsbeispielen dieses Artikels wird das kostenlose Softwarepaket `sgml-tools` von Greg Hankins, Tom Gordon, Alexander Horz und Cees de Groot vorgeführt.

Inhaltsverzeichnis

1 Einführung: Was ist SGML?

SGML bezeichnet ein standardisiertes Verfahren zur inhaltlichen Markierung elektronisch erstellter Texte. Ein mit solchen Markierungen versehener Text wird von einem SGML-Parser gelesen und über automatisierte Zwischenstufen in einer auf das gewünschte Ausgabemedium angepaßten Form wiedergegeben. Im Idealfall hat der Autor eines SGML-Textes nur geringe Einflußmöglichkeiten auf die Art, wie die Markierungen auf der Zielplattform umgesetzt werden; geringfügig mehr Definitionsmöglichkeiten verbleiben dem Autor meistens jedoch (auch das ist medienabhängig) bei der Wahl bestimmer Schriftattribute. Im Gegensatz zu diesen eingeschränkten Möglichkeiten der Gestaltung des Erscheinungsbildes hat der Benutzer eines SGML-Systems jedoch umfassende Möglichkeiten, seinen Text zu gliedern und zu strukturieren; in jedem Falle wird die Gegebenheiten des Ausgabemediums optimal zur Wiedergabe dieser Strukturierung genutzt, wie man am Vergleich von gedruckter Ausgabe und HTML-Ausgabe des semantisch gleichen Dokuments sehen kann.

Jedes SGML-System besteht aus mindestens zwei Teilen, der Dokumentendefinition (abgkürzt DTD, *Document Type Definition*), und einem Parser.

1. Die Dokumentendefinition enthält die Beschreibung aller zulässigen Textauszeichnungen, und
2. der Parser erzeugt aus dem Quelltext unter Verwendung der Dokumentendefinition den Text für ein bestimmtes Zielmedium.

Es muß hervorgehoben werden, daß SGML, bei aller oberflächlichen Analogie zu bekannten auszeichnungsbasierenden Formatiersystemen wie `roff`, LaTeX oder HTML, *kein* weiteres *Formatiersystem* ist. Kein SGML-Parser erzeugt aus der Quelldatei jemals ein fertiges Dokument; SGML ist eine Sprache zur *Beschreibung* von Auszeichnungssystemen. SGML-Systeme erzeugen daher lediglich den LaTeX-, HTML- etc. Quelltext, aus dem dann das fertige Dokument des jeweiligen Zielformats entsteht.

Mit SGML-Systemen lassen sich daher aus einem Quelltext Texte für die unterschiedlichsten Zwecke erstellen. Das System `sgml-tools`, dessen Bedienung hier beispielhaft vorgestellt wird, kennt folgende Ausgabeformate:

HTML

Das Kommando `sgml2html` wandelt eine SGML-Quelle in ein HTML-Dokument um, das nötigenfalls über mehrere Dateien aufgeteilt ist, die von einem gemeinsamen Inhaltsverzeichnis zusammengehalten werden.

info

Das Kommando `sgml2info` erzeugt Texte im GNU-spezifischen Hypertextformat `info`, im dem ein Großteil der GNU-Software dokumentiert ist. Ein `info`-Modus ist Bestandteil des Editors `emacs`, ohne `emacs` können die Texte entweder mit `info` oder `tkinfo` gelesen werden.

LaTeX

Das Kommando `sgml2latex` erzeugt Texte, deren Erscheinungsbild für Printmedien (Bücher, Zeitschriften) optimiert ist. Wahlweise läßt sich der Ausgabetext als PostScript-, als DVI- oder als TeX-Datei abspeichern.

LyX

Das Kommando `smgl2lyx` erzeugt Texte, die mit der auf LaTeX basierenden WYSIWYG-Oberfläche LyX weiterverarbeitet werden können. Im Gegensatz zur LaTeX-Variante verfügt diese Ausgabeart über keinen Befehl zur Erstellung der Druckvorlagen, da diese mit der Ausgabe des LaTeX-Systems identisch wären.

RTF

Das Kommando `sgml2rtf` erzeugt Dateien im Microsoft-spezifischen Rich Text Format, einer ASCII-Beschreibung der Textdatei, die in vielen Punkten (Kommandosyntax, Befehlsorthographie) stark an ein korrumpiertes TeX erinnert. Die mit `sgml2rtf` erstellten Dateien sind so aufgebaut, daß sie sich problemlos vom Textcompiler `hc31.exe` lesen lassen können, so daß sich Texte für das Hilfesystem von Microsoft Windows damit erstellen lassen.

txt

Das Kommando `sgml2txt` erzeugt aus einer SGML-Vorlage einen einfachen ASCII-Text oder wahlweise auch einen Text im Zeichensatz ISO-8859-1, so daß auch Umlaute und Sonderzeichen korrekt dargestellt werden können. Neben der Ausgabe einfacher Textdateien ist `sgml2txt` auch zur Erzeugung von Quelldateien auf `groff`-Basis fähig, so daß sich leicht die bekannten `manpages` (Bedienungsanleitungen der Unix-Systeme) erzeugen lassen.

1.1 Anwendungsbeispiele von SGML

1.1.1 Dokumentation von Computersoftware

Eines der umfassendsten frei zugänglichen Beispiele für die gelungene Verwendung von SGML ist die Serie der Linux-HOWTO-Anleitungen. Diese Texte geben Informationen zu den verschiedensten Fragestellungen des Betriebssystems Linux; geschrieben werden diese Texte von zahlreichen Autoren, aber gesammelt und gepflegt werden sie an zentraler Stelle. Die Koordinatoren des Linuxdoc-Projektes entwickelten eigens zu diesem Zweck eine zunächst relativ spezialisierte Variante eines SGML-Systems, das auf die spezifischen Bedürfnisse der Software-Dokumentation zugeschnitten war. So wurde es möglich, die Texte nicht nur im Internet und als Onlinedokumentation im HTML-Format anzubieten, sondern auch die Erstellung gedruckter Handbücher (etwa der Linux-Bibel etc.) erwies sich als sehr einfach, wobei von Anfang an ein einheitliches Erscheinungsbild und damit eine einfache Orientierung des Lesers sichergestellt ist.

1.1.2 Bedienungsanleitungen umfangreicher technischer Systeme

Ähnlich wie die Dokumentation komplexer Computersysteme umfaßt auch die technische Dokumentation größerer Industrieanlagen oft Tausende von Seiten. Obwohl die einzelnen Komponenten einer Industrieanlage standardisiert sind, ist die einzelne Anlage fast immer ein auf den Kundenwunsch hin zugeschnittenes Einzelobjekt. Als Beispiel möge ein Umspannwerk dienen, das die Hochspannung eines städtischen Verteilungsnetzes in die für die Einspeisung ins Haushaltsnetz nötige Niederspannung umwandelt. Ein Umspannwerk enthält wenigstens einen Transformator, darüberhinaus aber umfangreiche Schaltanlagen, mit denen einerseits der Transformator vom Netz getrennt werden kann und andererseits der Kurzschlußschutz bei Unfällen realisiert wird. Darüberhinaus gibt es umfangreiche Regel- und Kontrollsysteme, die zur Aufrechterhaltung auch des normalen Betriebs notwendig sind.

Die Dokumentation ein und desselben Objekts wendet sich oft gleichzeitig an ganz unterschiedliche Interessenten: Der Auftraggeber und Käufer einer solchen Anlage möchte etwas anderes lesen als der Sicherheitsfachmann, der die Anlage genehmigt; der Wartungsingenieur wiederum hat andere Wünsche an die Dokumentation als der Chef der Verkaufsabteilung. Allen aber ist gemeinsam, daß sie sich darauf verlassen können müssen, daß die Dokumentation exakt das Produkt beschreibt, mit dem sie es zu tun haben, auch wenn vielleicht nur der Aspekt ausführlicher dargestellt wird, an dem die jeweilige Zielgruppe interessiert ist.

1.1.3 Innerbetriebliche Dokumentation (Firmenhandbücher)

Der Name "Handbuch" für die Dokumentation der innerbetrieblichen Regeln und Abläufe einer Firma, einer Behörde oder nahezu jeder größeren Organisation ist irreführend, da es sich meist um Aktenordner voll des unterschiedlichsten Materials handelt. Typische Bestandteile einer innerbetrieblichen Dokumentation umfassen neben einer Beschreibung der Arbeits- und Produktionsabläufe ausführlich die Pflichten und Rechte aller Mitglieder und Organisationseinheiten der Einrichtung; für Aktiengesellschaften sind Regeln wichtig, die Form, Art und Umfang der Öffentlichkeitspflicht festlegen (Aktionärsversammlung, Geschäftsbericht, etc.; letzten Endes ist auch der Geschäftsbericht ein Teil der innerbetrieblichen Dokumentation); das Verhalten im Ausnahmefall muß festgelegt werden; Musterverträge für sämtliche zu erwartenden Transaktionen werden hinterlegt.

1.1.4 Verträge und Satzungen

Verträge werden zwischen mehreren Parteien zwecks Festlegung von Handlungen, Rechten und Pflichten innerhalb eines gemeinsam definierten Kontextes geschlossen. Ein relativ einfacher Vertrag ist ein Vertrag über eine Wohnungsvermietung mit Klauseln für Warmwasserabrechnung und Kleintierhaltung; ein umfangreicher Vertrag kann viele Hundert Seiten umfassen und regelt beispielsweise sämtliche Handlungsschritte und Kapitalfragen zweier Unternehmen, die gemeinsam ein Joint Venture bilden. Solche Verträge enthalten neben dem eigentlichen Vertragstext oft noch umfangreiche Anlagen und Zusatzabkommen mit Konstruktionsplänen oder Vertriebsvereinbarungen.

Satzungen von Vereinen und Körperschaften stellen eine Sonderform des Vertrages dar, der alle Mitglieder einer Körperschaft bindet; auch hier kann es erforderlich sein, das Erscheinungsbild des fertigen Textes den übrigen Dokumenten der Körperschaft anzugleichen.

1.1.5 Informationsangebote im WWW

Wenn es sich nicht um die private Homepage einer Privatperson mit Informationen über Hobbies, Haustiere und Lebensgefährten handelt, wachsen Angebote, die oft aktualisiert werden und über komplexe Sachverhalte informieren, oft mit großer Geschwindigkeit. Aufgrund der Eigenheiten der Übertragungswege und Wiedergabemedien ist hier eine rasche und einheitliche Benutzerführung gestützt auf ein übersichtliches Inhaltsverzeichnis von besonderer Wichtigkeit.

Alle genannten Beispiele zeichnen sich durch mehrere große Gemeinsamkeiten aus:

- Fast immer sind viele und wechselnde Autoren an der Entstehung der Dokumentation beteiligt;
- die Autoren benutzen häufig unterschiedliche Computersysteme: Der Entwicklungsingenieur sitzt vor einer Unix-Workstation, der Kaufmann sitzt an einem Terminal eines Großrechners mit der kompletten Firmendatenbank, aber das Chefsekretariat eines Unternehmens oder einer Organisation benutzt PCs mit Windows und Word als Software.
- Trotz der unterschiedlichen geistigen und technischen Herkunft der Texte muß ein einheitliches Erscheinungsbild des Gesamtwerks zwingend gewährleistet sein.
- Oft besteht ein ausgeprägter Wunsch, die gleiche Information über verschiedene Medien zu verbreiten, etwa als HTML-Datei ebenso wie als gedrucktes Buch oder als erweiterbare Loseblatt-Sammlung.

1.2 Beziehung von SGML und HTML

HTML, die Hypertext Markup Language, ist das Textauszeichnungssystem, mit dem die Texte des WWW erstellt werden. Gemeinsam mit SGML hat HTML, daß das Erscheinungsbild bestimmter Auszeichnungen erst vom Textbetrachter festgelegt wird und der Autor eigentlich keinen Einfluß auf z.B. die Wahl der Schriftgröße bei Überschriften haben sollte. Dennoch sind SGML und HTML grundlegend verschieden. Mit SGML werden Auszeichnungssysteme beschrieben, während HTML ein Auszeichnungssystem ist. Anders gesagt: SGML ist ein abstraktes, HTML ist ein konkretes System. Im grammatikalischen Vergleich beider Systeme wird deutlich, daß HTML nur eine Untermenge der Möglichkeiten von SGML bietet.

So gibt es als Beispiel einer Übereinstimmung in HTML wie in SGML Befehle zur Erzeugung Listen, aber während SGML unterschiedlich formatierte Überschriften als Resultat einer Auswertung von Gliederungsebenen erzeugt, "versteht" HTML nichts von hierarchischer Dokumentengliederung und muß konkret mitgeteilt bekommen, aus welcher Ebene eine Überschrift ausgewählt werden soll.

Aufgrund dieser fundamentalen Einschränkungen ist es daher nur unvollständig möglich, einen HTML-Text auf einem anderem Medium als dem Textbetrachter ideal auszugeben; das Resultat läßt ästhetisch zu wünschen übrig und wirkt immer wie eine Nachbildung der bildschirmtypischen HTML-Darstellung mit Textprozessormitteln.

2 Grundlegende SGML-Befehle

SGML kennt zwei grundlegende Arten von Befehlen zur Erzeugung bestimmter Effekte. Kommandos, die Text in irgendeiner Form verändern, etwa Schriftenauszeichnungen, stehen in spitzen Klammern, Die in spitzen Klammern stehenden Kommandowörter umschließen oft paarweise einen Bereich, wobei im zweiten Klammerpaar durch einen Schrägstrich vor dem Kommandowort angezeigt wird, daß hier die Wirkung des soeben gesetzten Kommandos aufgehoben wird, wie hier an der `abstract`-Markierung gezeigt wird:

```
<abstract>
        Der Text beschreibt ein Verfahren zur Gewinnung
        von Blättern aus Blattläusen, die mit 3,4-Isofoliol
        behandelt wurden.
</abstract>
```

Kommandos, die von sich eine begrenzte, vorhersehbare Menge Text erzeugen, etwa Befehle zur Erzeugung von Inhaltsverzeichnissen, (<`toc`>) müssen nicht in der Wirkung aufgehoben werden. Wenn das Textargument des Befehls keinen Schrägstrich enthält, kann der Befehl auch so abgekürzt werden (hier am Beispiel des Fettdrucks): <`bf/Fettdruck/`.

Kommandos zur Erzeugung einzelner Sonderzeichen (*entities* im SGML-Jargon) werden mit dem Kaufmanns-Und `&` eingeleitet und durch ein Semikolon abgeschlossen. Die `sgml-tools` kennen eine ganze Reihe solcher Sonderzeichen, die teilweise auch in `verbatim`-Umgebungen benutzt werden müssen, da auch hier bestimmte Zeichen nicht wörtlich, sondern immer als Anweisung interpretiert werden. Das augenfälligste Sonderzeichen ist natürlich die spitze Klammer, sie wird als `<` (öffnend) bzw. `>` (schließend) wiedergegeben. Ebenso hat die Kombination aus öffnender spitzer Klammer und Schrägstrich (</) eine eigene Bezeichnung: `&etago;`. Hier folgt eine vollständige Liste:

&half ($\frac{1}{2}$)

Vertikaler 1/2-Bruch

&frac12 (1/2)

Typographischer 1/2-Bruch

&frac14 (1/4)

Typographischer 1/4-Bruch

&frac34 (3/4)

Typographischer 3/4-Bruch

&frac18 (1/8)

Typographischer 1/8-Bruch

&frac38 (3/8)

Typographischer 3/8-Bruch

&frac58 (5/8)

Typographischer 5/8-Bruch

&frac78 (7/8)

Typographischer 7/8-Bruch

&sup1 (^{1})

Hochgestellte 1

&sup2 (^{2})

Hochgestellte 2

&sup3 (^{3})

Hochgestellte 3

&plus (+)

Pluszeichen

&plusmn (±)

Plusminus-Zeichen

&lt (<)

Spitze Klammer, öffnend

&equals (=)

Gleichheitszeichen

&gt (>)

Spitze Klammer, schließend

&divide (÷)

Divisionszeichen

&times (×)

Multiplikationszeichen

&curren ({curren})

Währungssymbol

&pound (£)

Zeichen für “Pfund”

&dollar ($)

Dollarzeichen

&cent ()

Cent-Zeichen

&yen ({yen})

Yen-Zeichen

&num (#)

Nummern- “Hash”-Zeichen (“Gartenzaun”)

&percnt (%)

Prozent-Zeichen

&amp (&)

Kaufmanns-Und

&ast (*)

Stern

&commat (@)

“Klammeraffe”

&lsqb ([)

Linke eckige Klammer

&bsol (\)

Rückwärtsschrägstrich

&rsqb ( )]

Rechte eckige Klammer

&lcub ({)

Linke geschweifte Klammer

&horbar (–)

Horizontalstrich

&verbar (|)

Vertikalstrich

&rcub (})

Rechte geschweifte Klammer

&micro (μ)

Griechisches My (Mikro-Präfix)

&ohm (Ω)

Großes griechisches Omega (Ohm-Symbol)

&deg (°)

Kleiner hochgestellter Kreis (Grad-Zeichen)

&ordm ({ordm})

Männliche Ordinalzahl

&ordf ({ordf})

Weibliche Ordinalzahl

&sect (§)

Abschnittszeichen

&para (¶)

Absatzzeichen

&middot (·)

Punkt auf halber Höhe

&larr (←)

Pfeil nach links

&rarr (→)

Pfeil nach rechts

&uarr (↑)

Pfeil nach oben

&darr (↓)

Pfeil nach unten

&copy (©)

Copyright-Zeichen

&reg ()

R im Kreis ("Registered")

&trade ((TM))

Warenzeichen

&brvbar ({brvbar})

Durchbrochener Vertikalstrich

&not (¬)

Logische Verneinung

&sung ({sung})

sung-note sign

&excl (!)

Ausrufezeichen

&iexcl (¡)

Umgekehrtes Ausrufezeichen

&quot (")

Doppeltes Anführungszeichen

&apos (')

Apostroph (einfaches Anführungszeichen)

&lpar (()

Linke Klammer

&rpar ())

Rechte Klammer

&comma (,)

Komma

&lowbar (_)

Unterstrich

&hyphen (-)

Bindestrich

&period (.)

Punkt

&sol (/)

solidus

&colon (:)

Doppelpunkt

&semi (;)

Semikolon

&quest (?)

Fragezeichen

&iquest (¿)

interrobang

&laquo («)

Linke doppelte spitze Klammer

&raquo (»)

Rechte doppelte spitze Klammer

&lsquo (‘)

Linkes einfaches Anführungszeichen

&rsquo (’)

Rechtes einfaches Anführungszeichen

&ldquo (“)

Linkes doppeltes Anführungszeichen

&rdquo (”)

Rechtes doppeltes Anführungszeichen

** ( )**

Hartes Leerzeichen

&shy ()

Weicher Bindestrich

Je nach Ausgabemedium ist es denkbar, daß nicht alle Zeichen korrekt wiedergegeben werden können. Dann steht in der Wiedergabe für das Beispiel nur ein leeres Klammerpaar!

2.1 Die einfachste SGML-Datei

Die einfachste SGML-Datei besteht mindestens aus drei Zeilen. Die erste Zeile informiert den SGML-Parser über die verwendete Dokumentendefinition, die beiden folgenden Zeilen des folgenden Listings eröffnen und schließen den Dokumententext.

```
<!doctype linuxdoc system>
<article>
</article>
```

2.2 Schriftauszeichnung

Für die Schriftauszeichnung stehen eine Reihe von Befehlen bereit, mit denen sich Schriften **fett** (`bf`), *kursiv* (`it`), als `Schreibmaschinentype` (`tt`) sowie *hervorgehoben* (`em`) darstellen lassen.

2.3 Listen

Listen können numeriert oder nicht numeriert sein. Im ersten Fall wird die Umgebung mit `enum` eingeleitet, im zweiten Fall mit `itemize`. Jedes Listenelement wird mit `item` bezeichnet.

2.4 URL-Angaben

Verweise auf Internet-Adressen sind so möglich, daß die Adresse durch die Titelangabe verdeckt ist, im Falle eines HTML-Dokuments aber anklickbar wird. Die entsprechende Markierung heißt `htmlurl`.

```
<htmlurl
url="http://userpage.fu-berlin.de/~corff/SGML/"
        name="Einführung in SGML">
```

Im Text erscheint dieser Quellverweis dann so: `Einf\"uhrunginSGML`

3 Verfügbare Software für SGML

Es gibt eine ganze Reihe teilweise sehr teurer SGML-Systeme für den kommerziellen Bedarf für unterschiedliche Plattformen. Eine Übersicht findet sich auf mehreren der später genannten WWW-Seiten.

Neben der kommerziellen gibt es auch freie Software für SGML-Systeme, hier soll exemplarisch das Paket `sgml-tools` vorgestellt werden, mit dem auch die Linux-Dokumentation erstellt wurde. Das SGML-Paket `sgml-tools` für Unix-Systeme ist verfügbar bei `http://pobox.com/\~{}cg/sgmltools`. Bei diesem Paket finden sich eine Beispieldatei und auch eine (englische) Einführung in die wichtigsten Befehle. Es muß allerdings hervorgehoben werden, daß die verwendete DTD (Document Type Definition — Dokumentendefinition) erheblich mächtiger ist, als die Beispieldatei es erahnen läßt. So lassen sich beispielsweise Fußnoten und Querverweise in die Texte einbauen, ebenso lassen sich Indices erstellen. Die wichtigsten Fähigkeiten von LaTeX werden weitgehend unterstützt. Die nötigen SGML-Befehle muß man, bis eine vollständige Dokumentation der verwendeten DTD vorliegt, allerdings leider noch durch Suchen in den Systemdateien von `sgml-tools` ermitteln. Dies gilt jedoch nur für sehr spezialisierte Befehle, alle grundlegenden Dinge (und damit 99% der täglichen Arbeit) sind in der beiliegenden Dokumentation erklärt.

4 Ein konkretes Anwendungsbeispiel der `sgml-tools`

Als praktisches Beispiel für die Anwendung von SGML möge dieser Text selbst dienen: er ist mit einem einfachen ASCII-Editor (in diesem Fall dem `vi`) geschrieben; alle Einrückungen des Quelltextes wurden nur zur besseren Übersicht eingeführt und wirken sich *nicht* auf das Erscheinungsbild des fertigen Textes aus.

```
$ vi SGML-Einfuehrung.sgml                          # editieren
$ sgml2html SGML-Einfuehrung                        # HTML erzeugen
$ netscape SGML-Einfuehrung.html                    # HTML betrachten
$ sgml2latex -p a4 --language=de -c latin -o dvi\\
        SGML-Einfuehrung                            # LaTeX erzeugen
$ xdvi SGML-Einfuhrung                              # LaTeX betrachten
$ ...
```

5 Weiterführende Informationen im WWW

Im WWW findet sich mittlerweile eine umfangreiche Sammlung von Texten über SGML ebenso wie Software.

- `SGMLdocumentintroducingyoutoSGML`
- `SGML:OverviewandGeneral`
- `TheSGML/XMLWebPage-Home`
- `SGML-Toolshomepage`
- `SGML:AnswerstoBasicQuestions`
- `SGMLPRIMER:Introduction`
- `PublicSGML/XMLSoftware`