SGML – technologia opracowania dokumentów elektronicznych

Włodzimierz Wypych

Tekst + struktura = dokument

Tekst jest przekazem sformułowanym w pewnym języku i utrwalonym za pomocą pewnego alfabetu. Alfabet, którym się posługujemy, zawiera, poza literami, cyframi i różnymi symbolami, także znaki interpunkcji, separatory (spacja) i wyróżniki (cudzysłów), które pozwalają ujawnić strukturę tekstu. Bez tych znaków nasze teksty stanowiłyby *scriptio continua* i wyglądałaby tak, jak rękopisy z epoki Karolingów, które trudno odczytać i trudno nazwać tekstami.

Nasze wyobrażenie o tekście ukształtowała tradycja Gutenberga. Kilku stuleci nauczyło nas, jak forma ma służyć treści, jak układ typograficzny ujawniać strukturę tekstu, by ułatwiać jego odczytanie i zrozumienie. Dlatego wyróżniamy tytuły, tekst dzielimy na akapity, a odmiennym krojem pisma wyróżniamy te elementy, które uznajemy za godne podkreślenia.

Niegdyś, w czasach składu ręcznego, o takich wyróżnieniach decydował redaktor techniczny. Uzbrojony w kolorowy ołówek, wprowadzał do tekstu instrukcje dla zecera składającego tekst, wskazując, jakim stopniem złożyć tytuły, gdzie wprowadzić kursywę itd. W Polsce nazywano to adiustacją techniczną, w Ameryce krótko: → *markup*.

Technika komputerowa istotnie ułatwiła proces przygotowania tekstu do druku, ale zasada pozostała ta sama. Programy do składu tekstów (wliczając w to edytory wykorzystywane w biurach) wprowadzają do tekstu specjalne znaczniki, będące jak gdyby rozwiniętą formą interpunkcji. Znaczniki te są odpowiednio interpretowane przez oprogramowanie, gdy tekst odtwarzany jest na „urządzeniach wyjściowych", takich jak ekran monitora, drukarka lub naświetlarka produkująca klisze dla drukarni.

Teksty zawierające informacje o swojej strukturze nazywamy tekstami sformatowanymi albo → dokumentami, dodając nazwę programu, za pomocą którego zostały złożone. Mówimy więc o dokumentach MS Worda lub dokumentach PageMakera, podkreślając w ten sposób, że informacje o ich strukturze zapisane zostały w specyficznym formacie każdego z tych programów.

Od znaków do tagów

Technika komputerowego składu tekstów przeszła od lat sześćdziesiątych, kiedy podjęto pierwsze próby jej zastosowania, kilka etapów rozwoju, których świadectwem są także zmiany sposobu znakowania tekstów.

W pierwszych systemach wykorzystywano tzw. znakowanie proceduralne. Do tekstów wprowadzano znaczniki, które stanowiły sekwencje kodów sterujących bezpośrednio urządzeniami wyjściowymi. Sekwencja zapisana w postaci:

```
^v50^h30^f002^fs18
```

mogła oznaczać: *przesuń głowicę o 50 punktów w dół i 30 punktów w prawo, zastosuj krój pisma (font) numer 2, zmień stopień pisma na 30 punktów*. Tak specyficznie oznakowany tekst mógł być odtworzony jedynie na urządzeniu określonego typu czy nawet tylko na pewnym jego modelu.

Znakowanie proceduralne stało się wkrótce anachroniczne. W latach osiemdziesiątych pojawiły się bowiem bardziej elastyczne rozwiązania, które dały początek oprogramowaniu typu Desktop Publishing (DTP).

Zamiast kodów sterujących bezpośrednio urządzeniami wprowadzone zostały znaczniki opisujące strukturę tekstu, podobne do tych, jakich używali redaktorzy techniczni. Były zapisywane np. jako `^par` – (od ang. paragraph – początek akapitu) lub `^b` (bold – czcionka pogrubiona), a więc w sposób czytelny już nie tylko dla maszyny, ale i dla ludzi. Znaczniki te były oczywiście nadal przekodowywane w urządzeniu wyjściowym na operacje typu *przesuń głowicę..., zmień krój pisma...*, ale operator składający publikację nie musiał już znać szczegółów realizacji tych instrukcji. Wystarczyło, że widział na ekranie efekty wprowadzonego do tekstu oznakowania.

Znakowanie tego typu, nazywane *znakowaniem opisowym*, pozwalało korzystać z urządzeń i oprogramowania różnych producentów. Na znakowaniu opisowym opiera się współczesne DTP, programy takie jak TEX, PageMaker, Ventura, QuarkXpress. Powstały też formaty „uniwersalne" umożliwiające przenoszenie dokumentów między różnymi systemami i platformami sprzętowymi, takie jak RTF (Rich Text Format) i PostScript.

Sama już jednak różnorodność tych formatów stwarza poważne kłopoty a ich konwersja nie zawsze jest zadowalająca. Elementy wyróżnianie w jednym systemie nie muszą mieć bowiem odpowiednika w innym, a więc pewne informacje o strukturze tekstu mogą być gubione. Często zdarza się, że teksty złożone w jednym

systemie muszą zostać po prostu zredukowane do → *czystego tekstu* i w innym systemie składane od nowa.

Istnieje jeszcze jeden, nawet ważniejszy, powód dla którego żadnego z tych formatów nie można uznać za uniwersalny. Wynika to z samej ich genezy i przeznaczenia – powstały w celu przygotowania tekstów do druku i nie mogą sprostać nowym potrzebom, które stworzył postęp techniki. Dziś coraz więcej dokumentów publikuje się w postaci elektronicznej, na CD-ROM i w sieciach komputerowych, przede wszystkim w Internecie. Coraz większą rolę odgrywają elektroniczne repozytoria, bazy danych, w których można szybko odszukać potrzebne informacje. W końcu też (bez druku nie możemy się bowiem obyć) coraz częściej powstaje potrzeba zamiany dokumentu drukowanego na elektroniczny lub odwrotnie.

Zaspokojenie tych wszystkich potrzeb wykracza poza możliwości systemów DTP. Służą do tego nowe techonologie. Okazuje się, że Desktop Publishing z całym swoim instrumentarium i doświadczeniem stanowi dziś tylko część znacznie szerszej dziedziny techniki, którą nazywamy elektronicznym edytorstwem *(electronic publishing)*. Jej fundamentem jest SGML.

Geneza SGML

Zadanie stworzenia uniwersalnego systemu oznakowania tekstów zapewniającego ich przenaszalność pomiędzy różnymi systemami składu podjął w końcu lat sześćdziesiątych zespół ośrodka badawczego IBM, którym kierował Charles Goldfarb.

Rezulatem podjętych tu prac stało się nowe ujęcie problemu znakowania tekstów i nowa technologia, która po latach prób i udoskonaleń, już jako w pełni dojrzała, uzyskała w 1986 roku status międzynarodowego standardu pod niezbyt oczywistą nazwą: ***International Standard ISO 8879 Information Processing – Text and Office Systems – Standardized Generalized Markup Language (SGML).***

Intencją Goldfarba było uogólnienie zasad znakowania opisowego. Świadczy o tym sama nazwa nadana tej technologii: generalized markup.

„Znakowanie powinno opisywać strukturę dokumentu (...) a nie określać sposób przetwarzania dokumentu" – stwierdzał Goldfarb. Aby jednak dojść do całkiem ogólnego opisu, należało przede wszystkim wyraźnie rozróżnić to, co w dotychczasowych systemach składu tekstów nie było (i nie mogło być) rozróżniane, mianowicie elementy struktury tekstów oraz formy ich prezentacji (typografii).

Prawdziwie ogólne znakowanie nie może ograniczać się jedynie do elementów typograficznych, ale powinno obejmować wszystkie elementy, które z jakichkolwiek powodów uznane zostaną za ważne. Również więc elementy semantyczne, dowolne wyrażenia, niezależnie od tego, czy miały one zostać wyróżnione typograficznie, czy też nie.

Znakowanie nie może przesądzać o formie, w jakiej dokument będzie prezentowany odbiorcy. Jeśli w tekście zostały wyróżnione np. różnego typu tytuły, to nie powino to decydować o tym, jaki krój i stopień pisma zostanie użyty do ich prezentacji. Wszystkie wyróżnienia wprowadzone do tekstu powinny zachować trwały walor ogólności, natomiast o sposobie ich prezentacji powinno się rozstrzygać, przygotowując dokument do druku. Przygotowanie to sprowadza się do określenia, które z wyróżnionych elementów mają uzyskać swój wyraz w typografii i jak mają być odtworzone. Wystarczy wskazać, jakim krojem pisma ma być złożony każdy z wyróżnionych rodzajów tytułów i za pomocą odpowiedniego konwertera stworzyć dokument drukowalny, w którym oznakowanie ogólne zostanie zastąpione oznakowaniem przyjętym (możliwym do zrealizowania) w określonych urządzeniach wyjściowych. Sam dokument źródłowy pozostaje jednak nie zmieniony i może stanowić podstawę do przygotowania w podobny sposób innych (odmienych typograficznie) jego wydań.

Ten prosty pomysł okazał się rewolucyjny w konsekwencjach. Uwalniając znakowanie od służby na rzecz składu tekstów, Goldfarb stworzył fundament elektronicznego edytorstwa.

SGML od środka

SGML nie jest jeszcze jednym językiem opisu dokumentów tekstowych, nie oferuje żadnego zestawu znaczników do formatowania tekstu. Wbrew temu, co sama nazwa sugeruje, SGML nie jest językiem, w każdym razie nie w tym sensie, który pozwalałby postawić go obok takich języków, jak RTF czy PostScript. SGML operuje o piętro wyżej. Ściśle biorąc, SGML jest *metajęzykiem* – sformalizowanym systemem zasad, które pozwalają zdefiniować języki opisu (zasady znakowania) różnego typu dokumentów. Tak właśnie należy rozumieć określenie *generalized markup language*.

SGML ma charakter uniwersalny, nie jest związany z żadną platformą sprzętową, systemem operacyjnym czy oprogramowaniem. Nie przesądza też o medium, w jakim dokumenty zostaną udostępnione, a w każdym razie jego zastosowania nie sprowadzają się jedynie do przygotowaniu tekstów do druku.

Podstawowym pojęciem SGML jest pojęcie typu dokumentu. O typie dokumentu decyduje jego struktura. Sformalizowana definicja, tzw. DTD (Document Type Definition), która określa sposób wyróżnienia elementów strukturalnych dla pewnej klasy dokumentów,

definiuje znaczniki (tzw. tagi) oraz reguły ich użycia. Można powiedzieć, że DTD stanowi gramatykę (słownik i reguły syntaktyczne) formalnego języka, który pozwala opisywać dany typ dokumentów.

Użycie znaczników w dokumencie SGML przypomina użycie nawiasów do zaznaczenia struktury wyrażeń matematycznych. Znaczniki są zwyczajowo zapisywane w postaci: <nazwa> dla „nawiasu" otwierającego i </nazwa> dla zamykającego, gdzie nazwa jest indentyfikatorem tego znacznika. Nazwa może być w zasadzie dowolna, choć rozsądek nakazuje, aby znacznik sugerował, co jest nim oznakowane. Jeśli trzeba np. wyróżnić występujące w tekstach nazwy firm, to najlepiej wprowadzić znacznik o nazwie <firma>, wówczas oznakowanie

```
<firma>Supergraph</firma>
```

będzie całkowicie czytelne.

Znaczniki służą do wyróżnienia dowolnych elementów struktury tekstów, które mają znaczenie semantyczne (jak w wyżej podanym przykładzie) lub dotyczą formy typograficznej. Np. za pomocą znaczników <tytul1>, <tytul2>, <tytul3> itd. można ustalić hierarchię tytułów.

SGML umożliwia bardzo precyzyjne zdefiniowanie oznakowania. Pozwala określić reguły zagnieżdżania znaczników, czyli wskazać te, które mogą wystąpić jedynie we fragmencie tekstu wyróżnionym innymi znacznikami. Strukturę antologii złożonej z nowel kilku autorów, można by oznakować następująco:

```
<antologia>
...
<nowela>
<tytuł>...</tytuł>
< autor>...</autor>
<tekst>...</tekst>
</nowela>
...
<nowela>
<tytuł>...</tytuł>
< autor>...</autor>
<tekst>...</tekst>
</nowela>
...
</antologia>
```

Można określić, które znaczniki muszą zostać domknięte, a które nie. Jeśli tekst zawiera wyliczenie lub, jak w podanym wyżej przykładzie, jest sekwencją podobnych elementów, to koniec jednego elementu wyznaczony jest przez znacznik wyróżniający element następny, oczywiście poza ostatnim, który kończy znacznik zamykający całą sekwencję. W powyższym przykładzie można by więc zrezygnować ze znacznika zamykającego </nowela>.
SGML pozwala określić, które elementy muszą w danym typie dokumentu wystąpić, a które są dopuszczalne, i wskazać, czy chodzi o występienie jedno- czy wielokrotne.

Znaczniki mogą zostać sparametryzowane. Można związać z nimi atrybuty posiadające wartości, np.
<firma typ=producent>
lub
<firma typ=dystrybutor>.
Atrybutem może też być referencja (odsyłacz) do innych dokumentów, np:
<zobacz="inny.dokument">.

W definicji dokumentów można wprowadzić pewne stałe, zdefiniowane „jednostki tekstowe", tzw. encje *(entities)*. Encją może być wyrażenie, a także pojedynczy znak. Np. literę „ą" można zdefiniować jako „ą" i zapewnić przez to jej jednoznaczność i przenaszalność niezależną do systemów kodowania znaków narodowych w różnych systemach komputerowych.

Oczywiście każdy tekst może być oznakowany na tyle sposobów, na ile może być odczytywany. Przyjęte DTD jest zawsze projekcją pewnego odczytania tekstu i jego przewidywanego wykorzystania. SGML dopuszcza użycie w ramach jednego dokumentu znaczników należących do różnych DTD, które pozostają dla siebie wzajemnie niewidoczne. Można dzięki temu oznakować np. *Volumina legum* z punktu widzenia historyka prawa, językoznawcy lub wydawcy, który by chciał dokonać reedycji tego dzieła.

Można stworzyć DTD zarówno do opisu już istniejących dokumentów (publikacji archiwalnych), jak też dokumentów, które mają dopiero powstać. Takie „aprioryczne" DTD może być wykorzystane przez wyspecjalizowany edytor, który pozwala, już podczas wprowadzania tekstu, wyróżniać w nim wcześniej zdefiniowane elementy.

Dokumenty SGML są tekstami ASCII, które można tworzyć i odczytywać za pomocą każdego nieformatującego edytora (choć wyspecjalizowany edytor jest bardziej przydatny). Wprowadzane do tekstu oznakowanie może być sprawdzone automatycznie, przez specjalne oprogramowanie (tzw. parser), który konfrontuje definicję języka zapisaną w DTD z jego użyciem w danym dokumencie.

Zastosowania SGML

- **Projekty wydawnicze**

Stworzenie DTD dla złożonych, „gęstych" w swej warstwie znaczeniowej, dokumentów (takich jak różnego

rodzaju publikacje leksykograficzne) jest zajęciem trudnym, wymagającym połączenia kompetencji informatycznych i merytorycznych. Może więc być kosztowne.
Zdecydowana większość producji wydawniczej to jednak publikacje dotyczące dość wąskiej tematyki o stosunkowo prostej formie graficznej (jak np. artykuły w czasopismach fachowych). Zdefiniowanie języka opisu takich dokumentów nie jest skomplikowane, a ich oznakowanie zbyt pracochłonne.

Koszty związane z wprowadzeniem technologii SGML szybko mogą się zwrócić, a praca włożona w oznakowanie dokumentów SGML nie zostanie zmarnowana, gdy trzeba będzie przygotować nowe wydanie, zmienić je i uzupełnić. Jeśli natomiast podstawą reedycji miałby być dokument oznakowany typograficznie, to przygotowanie nowego wydania może wymagać znacznie większego nakładu pracy, a nawet opracowania całej publikacji od nowa.

Ten wzgląd właśnie sprawił, że SGML budził od początku zainteresowanie amerykańskich wydawców, a ich stowarzyszenie (Associaton of American Publishers) odegrało ważną rolę w promowaniu SGML jako efektywnej i ekonomicznej technologii wydawniczej.

Nie tylko to jednak stanowi o przewadze SGML na znakowaniem stosowanym w DTP. Jego potencjał ujawnia się najpełniej wówczas, gdy potrzebne staje się odszukanie informacji zawartej w dokumentach tekstowych.

• Bazy danych – elektroniczna dokumentacja

Dokumenty SGML, dzięki zawartym w nich znacznikom, można łatwo indeksować i tworzyć na tej podstawie różnego rodzaju skorowidze i zestawienia. Można je gromadzić w bazach danych, które, zaopatrzone w odpowiednie mechanizmy, pozwalają szybko i precyzyjnie zlokalizować poszukiwaną informację.

Indeksacja tekstów już wcześniej oznakowanych zapewnia znacznie większą trafność wyszukiwania niż przy wyszukiwaniu pełnotekstowym (opartym na indeksowaniu wszystkich wyrażeń w zgromadzonym zbiorze tekstów). W tym bowiem przypadku odszukane zostaną wszystkie dokumenty, które zawierają podane wyrażenie, a nie tylko te, w których zostało ono wyróżnione.

Z tego też powodu SGML był od początku wykorzystywany do opracowania elektronicznej dokumentacji.

Jedno z pierwszych zastosowań SGML w tej dziedzinie, miało charakter ściśle militarny. W 1988 roku, w ramach inicjatywy Departamentu Obrony o nazwie CALS (Computer-aided Acquisition and Logistic Support) SGML przyjęty został jako standard elektronicznej dokumentacji i „narzucony" dostawcom sprzętu dla armii amerykańskiej. Tą drogą SGML trafił do przemysłu i jako „industrial standard" wykorzystywany jest od tego czasu do tworzenia elektronicznej dokumentacji w wielu korporacjach.

Innym bardzo ważnym, i tym razem ściśle pokojowym, zastosowaniem SGML była inicjatywa znana pod nazwą TEI (Text Encoding Initiative), którą zrodziły potrzeby środowisk akademickich i naukowych. W ramach TEI zespoły fachowców opracowały prototypy języków (tzw. guidelines) służących do opisu tekstów, często bardzo specyficznych, z jakimi mają do czynienia przedstawiciele różnych dyscyplin naukowych: lingwiści, leksykografowie, historycy, matematycy itd.

• Nowe media

W czasach, gdy SGML powstawał, nie istniała jeszcze technika zapisu na dyskach optycznych (CD-ROM), nie istniał też World Wide Web. Te nowe media rozwinęły się dopiero w ostatnich latach i otworzyły przed SGML całkiem nowe zastosowania.

SGML stał się podstawą wyspecjalizowanych języków służących do przygotowania publikacji multimedialnych, zawierających sekwencje audio i video. Najbardziej znanym z nich jest HyTime (Hypermedia/Time-Base Structuring Language).

Jeszcze większe znaczenie ma SGML dla publikacji w Internecie. HTML – język opisu dokumentów hipertekstowych dostępnych w ramach World Wide Web, został zdefiniowany w SGML. W ten sposób twórca tego języka, Tim Berners-Lee, zapewnił mu solidny fundament i szanse rozwoju.

HTML jest dziś bez wątpienia najszerzej znaną aplikacją SGML, której używa każdy, kto publikuje cokolwiek w Internecie. Zapewne nie wszyscy jednak zdają sobie sprawę, czym się posługują. Nie zawsze wyjaśniają to popularne podręczniki do HTML i prasa komputerowa. W poważnym skądinąd tygodniku (*Computer-World* 34/97, str. 38) można np. przeczytać, że HTML „stanowi podzbiór" SGML. Nie wyjaśniono jednak, czego zbiorem miałby być sam SGML.
HTML jest dość prostym językiem, zorientowanym głównie na prezentację dokumentu w oknie internetowej przeglądarki. Jego możliwości są w tym względzie znacznie skromniejsze niż te, jakimi dysponują systemy DTP. Jeszcze mniej oferuje HTML w zakresie wyróżnienia elementów semantycznych. Rozwinięcie tych możliwości wymagałoby jednak nie tylko rozszerzenia definicji tego języka, ale też stworzenia nowego oprogramowania (nowego typu przeglądarek), które pozwoliłoby z tych możliwości korzystać.

Jest na to chyba już za poźno. Beniaminkiem głównych producentów narzędzi internetowych (przede wszystkim Microsoftu) stał się XML (eXtensible Markup Language), który można uznać za próbę wykroczenia poza ograniczoność HTML w zakresie elementów semantycznych. XML nie zakłada, jak HTML, skończonego zestawu znaczników, ale pozwala użytkownikom stworzyć własne, odpowiadające strukturze opracowywanych dokumentów. Umożliwia zaprezentowanie tych dokumentów w różnych perspektywach (wybiórczo), zgodnie z potrzebami różnych odbiorców. Daje duże możliwości definicyjne, a jednocześnie praktyczne jego wykorzystanie nie wymaga wprowadzenia całego tego aparatu formalnego, jaki wiąże się z zastosowaniem SGML.

Za prostotę XML trzeba wprawdzie zapłacić uniwersalnością zastosowań, ale XML, jako „uproszczony SGML", może być z pożytkiem wykorzystywany do opracowania typowych publikacji, jakie powstają w większości wydawnictw. W końcu nie każdy jest wydawcą powszechnej encyklopedii lub tworzy elektroniczną dokumentację dla platformy wiertniczej. Wielu wydawców opierało się dotąd przed wdrożeniem SGML uznając go, poniekąd słusznie, za „armatę na wróble". Skrót SGML dowcipnie rozwinięto nawet jako *Sounds Good, Maybe Later* – „brzmi dobrze, ale może poźniej". XML te opory powinien przełamać.

XML może zostać wykorzystany do przygotowania publikacji w Internecie (choć z pewnością nie zastąpi HTML). Ma też szanse wejść do wydawnictw i stać się „standardem wydawniczym". Doskonale nadaje się do tego, aby być podstawowym formatem wszystkich tekstów opracowywanych w wydawnictwie, nie przesądzając o medium, w jakim zostaną one opublikowane (druk, CD-ROM, Internet). Wskazuje na to m.in. to, co powiedziano i pokazano na ostatniej imprezie Seybolda w San Prancisco i co relacjonujemy w bieżącym numerze Nowe Mediów.

[Dziękuję Mariuszowi Olko za cenne uwagi i sugestie, które pozwoliły uczynić niniejsze wprowadzenie w problematykę SGML bardziej klarownym – autor.]

Źrodła

- Charles Goldfarb, The SGML Handbook, Clarendon Press 1995, autorytatywny wykład twórcy SGML.

Nie trzeba jednak sięgać od razu do samej biblii, można zacząć od czegoś lżejszego.

- Eric van Herwijnen, Practical SGML, wyd 2, Kluwer 1994

Najbardziej wydajnym źródłem informacji dotyczącej SGML jest dziś World Wide Web.

- Najważniejszy adres: http://www.sil.org/sgml/sgml.html – „oficjalna strona" SGML – wszechstronny i starannie opracowany przez Robina Covera katalog obejmujący wszystko, co tylko z SGML się wiąże. Zawiera m.in. szczegółową bibliografię, kalendarium i wiele przekrojowych zestawień.

- Dokumentację i narzędzia SGML (parsery) można uzyskać z dwu (częściowo pokrywających się) archiwów, norweskiego: ftp://ftp.ifi.uio.no/pub/SGML/TEI/

oraz brytyjskiego:

ftp://info.ex.ac.uk/pub/SGML/.

- http://www.falch.no/~pepper/sgmltool/ – Whirlwind Guide to SGML – przegląd dostępnych narzędzi SGML i informacje o ich producentach.

- Najbardziej autorytatywnym źródłem informacji na temat HTML i XML są publikacje W3 Consortium: http://www.w3.org/hypertext/WWW/MarkUp/.

Poczynania związane z SGML w Polsce, w tym (nieliczne) polskie publikacje na ten temat, monitoruje Rafał Księżyk na stronie SGML in Poland pod adresem http://www.fuw.edu.pl/~ksiezyk/sgml.html.

Glossarium

czysty tekst – (plain text) jest w zasadzie fikcją, każdy tekst zawiera pewne informacje strukturalne, choćby znaczniki końca wiersza czy akapitu (minimum formatu), których usunięcie zamieniło by go w nieczytelne scriptio continua.

dokument – coś więcej niż tylko → *czysty tekst*, tekst sformatowany, wzbogacony o informacje dotyczące jego stuktury (rich text).

dokument SGML – tekst, w którym za pomocą odpowiednich znaczników zdefiniowanych w postaci → DTD wyróżnione zostały pewne elementy strukturalne.

DTD (Document Definiton Type) – formalna definicja oznakowania pewnego typu dokumentów, podająca reguły wyróżniania elementów strukturalnych, zestaw znaczników (tagów) i → *encji*.

element – każdy fragment struktury tekstu, który może zostać wyróżniony za pomocą odpowiednich znaczników. Rozróżniamy elementy semantyczne (coś znaczącego) i typograficzne (jakoś wyglądającego).
encja (ang. *entity* – od łac. ens – *istność*) – jednostka leksykalna, której znaczenie zostało wcześniej zdefiniowane – wyrażenie, które w dokumencie SGML zastępuje inne.

HTML (HyperText Markup Language) – język zdefiniowany w SGML służący do opisu hipertekstowych dokumentów World Wide Web („stron WWW"). Pierwsza wersja tego języka została opublikowana w 1992 r., obecnie obowiązująca oznaczona 3.2, wkrótce zostanie zastąpiona kolejną wersją z numerem 4. Rozwojem HTML zajmuje się W3 Consortium (W3C), grupujące producentów narzędzi internetowych (m.in. firmy Microsoft, Netscape, SoftQuad).

markup – jest terminem przyjętym w anglosaskim edytorstwie i oznacza zarówno czynność, jak i rezultat adiustacji technicznej (w odróżnieniu od adiustacji merytorycznej). *Chicago Manual of Style*, biblia amerykańskich wydawców, definiuje *markup* jako *the process of marking manuscript copy for typesetting with directions for use of type fonts and sizes, spacing, indentation, etc.* Zgodnie z utartą już w języku polskim konwencją, *markup* w sensie czynności tłumaczymy jako *znakowanie* (lub żargonowo: *tagowanie*, ale nigdy: *oznaczanie czy markowanie*), natomiast *markup* w sensie rzeczowym, jako *oznakowanie*. Oznakowanie jest rezultatem znakowania czyli wstawienia do tekstu ustawienia odpowiednich znaczników. Znakowanie oparte jest na pewnej formalnie zdefiniowanej normie. Dokumenty SGML oznakowane są na podstawie pewnego → *DTD*.

parser – (od ang. parse – analizować, dokonywać rozbioru gramatycznego) program, który sprawdza typ dokumentu, tzn. określa, czy struktura dokumentu odpowiada definicji sformułowanej w postaci DTD.