デジタル・ドキュメント 14－2
（1998. 9. 25）

SGMLデータカートリッジによる文書管理システムの構築

矢島正樹　藤津真一　大野邦夫
yajima@inse.co.jp　fuji@inse.co.jp　ohno@inse.co.jp

ＩＮＳエンジニアリング株式会社

SGMLデータカートリッジを用いた、技術文書の配布・管理を効率的に実施するシステムを構築した。本システムでは、汎用の RDB でデータ管理を行いつつ、多様な出力メディアに対応可能なシステムを実現しているが、それらはデータカートリッジの機能を有効に使うことにより実現された。従来、SGML を用いたシステムの場合、煩雑で使い難いという先入観に支配されがちであったが、ネックになる部分をカスタマイズし、既存ファイルを極力外部エンティティとして扱う手法を用いることにより従来の課題を大幅に解決することができた。

Development of Document Management and Distribution System Basd on SGML Data Cartridge

Masaki Yajima　Shinichi Fujitsu　Kunio Ohno
yajima@inse.co.jp　fuji@inse.co.jp　ohno@inse.co.jp

INS Engineering Corporation

Digital Document Management and Distribution System based on SGML Data Cartridge has been developed. Conventional RDB is used for the data storage of the Data Cartridge which enables to interchange the SGML data to multiple output media as Web browser, SGML viewer, CD-ROMs, and printed papers. SGML systems have been thought complicated and difficult to use. But, the difficulty has been overcome through the customizing effort to modify the bottlenecks and make use of the conventional files as external entities.

まえがき

インターネットの普及に伴い、従来の紙ベースの文書管理システムから電子化されたデジタルドキュメントへの移行が進展している。しかしながら、デジタルドキュメント化は、新規システムでない限り全面的に移行することは困難であり、一般には既存システムとの融合が望まれている。我々は、一昨年以来、SGML を用いるマルチメディア情報を含む文書管理システムの枠組みを検討してきた[1][2][3]。

マルチメディア情報を SGML で管理することの基本的な考え方は、情報実体をその性質により分類し、その分類毎に異なったコントロールを可能とする枠組みを設定することにある。具体的には、その枠組みをタグによる属性と内容実体のペアとして管理し、タグ属性の定義と相互関係を DTD で定義し、情報実体を、DTD の枠組みの要素群に分解し、ツリーとして関係付けることにある。このアプローチは、各種情報を分類定義する上で普遍的であり、既存メディアからマルチメディア、分散環境上のハイパーメディアをも包含するものである（図１）。

これらの分類された情報への操作（オペレーション）は情報のタイプ（型）毎に決まるものであり、これらの操作は、対応する情報タイプへの API 群を構成する。我々の開発したデータカートリッジは、SGML という情報の枠組みを管理する基本レベルの API を提供している。SGML は基本的な枠組みに過ぎないので、具体的な応用分野毎に DTD と対応する API が定義されることになる。そのような意味で、SGML データカートリッジは、XML と同様に拡張可能な枠組みである。

ここではそのような SGML データカートリッジ[2]の特徴を生かして開発された、大手家電メーカの技術文書配布管理システムを紹介する。日本の大手家電メーカは、その製造技術と品質管理技術により世界規模のビジネスを展開し、激しい競争環境の下で新機種を半年毎に出荷している。当然、関連するドキュメント類も必用とされ、それらを作成、製本して国内および海外に点在する数多くのサービス店、販売店に配布するためには膨大なコストと時間を費やしている。

今後、一般消費者のニーズの多様化や新技術を用いた製品開発のスピードアップに伴い、製品の種類はさらに増加し、リリース期間は短縮されると予想されている。その場合、従来の紙のみによるドキュメント類の配布方式では、迅速な対応は著しく困難になると考えられる。以上の背景の下に、今回のシステム導入が検討された。

システムの導入にあたり、以下のような効果を期待した。

(1) 迅速な配布／閲覧
(2) 印刷／配送コストの削減
(3) ドキュメントデータの一元管理
(4) 多様な媒体（クロスメディア）への出力

以上のシステムは、電子的な管理と配布を行うことから電子化ドキュメントとして実現されるが、既存の利用者の全てを電子化ドキュメントのユーザに移行させ

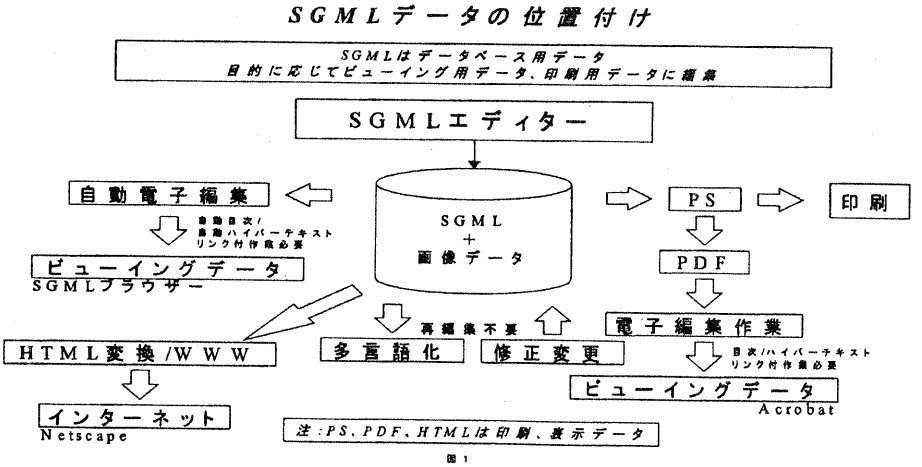


図１

るのは無理なので、現行の紙による方式を包含し得る方式で実現する必用がある。以上を実現するための、具体的システムの枠組みとして、既存方式を包含するSGML の適用が検討され、さらに以下のような要求が出された。

(1) イントラネット／インターネットを利用したドキュメントデータの登録／更新／検索
(2) 登録／更新手続きの簡略化
(3) 検索時における、SGML ドキュメントの部分木参照

以上の要求を満たすためには、データの一元的な管理とデータ保護の観点からデータベースの適用が有効である。そこで SGML による文書管理を汎用の RDB で行える SGML データカートリッジの適用が試みられた。

2. システムの構成

2.1. システム概要

システム全体は、SGML 文書の作成・編集系、SGML 文書管理系、および複数メディアへの出力・印刷系の 3 つの部分で構成され、基本的には作成・編集／蓄積・管理／出力・配布という文書のワークフローに対応している。ここでは、SGML データカートリッジに関係が深い、文書管理系（登録、更新、検索機能）システムを中心に述べることとする。

2.2. SGML 文書管理システム

2.2.1. 文書管理サーバー

文書を SGML データベースで管理するとは言っても、管理する文書の全てを SGML 化する必要はない。SGML 化せなばならないのは、文書の構成要素をエレメントとして管理する必要があるものだけであり、他は外部エンティティとしてファイルで管理すれば良い。本システムで SGML 化されているのは一部の技術文書（主に製品毎のサービスマニュアル、技術関連情報など）である。文書管理サーバーで使用する RDB によって管理するのは SGML 化されたドキュメントのみで、その他は文書管理サーバーのファイル・システムにファイルのまま管理される。文書管理サーバーではWWWサーバーとして、全てのドキュメントの閲覧を可能にするよう構成されている。また、WWWサーバーとしての機能を十分に確保するため、クライアントからの検索／参照要求の多い日中の時間帯ではスプール領域（参照・出力用のバッファ領域）に置かれた SGML ドキュメントの登録／更新作業は行わないようにしている。

2.2.2. SGML ドキュメントの格納

各部署で作成された SGML ドキュメントは文書管理サーバー上のスプール領域へファイル形式で転送される。転送されるデータは SGML ファイル、DTD、外部エンティティファイル（PDF で作成された回路図や TIFF、CGM 形式で作成された解説図等）、それらの構成を記述したカタログファイル、および登録／更新タグファイルがある。 なお、登録／更新タグファイルは、文書管理サーバーで自動登録、更新処理を行うために必要な情報を記述したものである。

2.2.3. 自動処理

文書管理サーバーで自動登録／更新処理を行うためにはスプール領域におかれた SGML ドキュメントデータに必要な情報が網羅されていることが必要である。本システムで用いられている DTD には、製品のモデル番号、名称、および文書種別（例えば、サービスマニュアルといった種別）を記述するタグが定義されており、それらを参照することにより自動的に取得可能になる。ただしそれ以外に必要な情報に関しては、登録／更新タグファイルを作成して対応することにした。登録／更新タグファイルは SGML ドキュメントの作成元である各部署で作成する。このファイルには作成部署名、依頼する処理の種類（登録／更新）、他地域にある文書管理サーバーへの配送先指定、コメント文、更新処理時の情報（SGML 文書の全体あるいは部分的な更新、または外部エンティティの更新なのかが明記してある）等が含まれる。これらスプール領域に置かれた SGML ファイルは文書管理サーバーに常駐しているスプール管理 AP により一括管理される。

2.2.4. スプール管理 AP

スプール管理 AP はスプール領域に格納された SGML ドキュメントの状態管理の他に、データベース（以下 DB）アクセスを含む SGML ドキュメント処理への待ち行列管理（キューイング）も行う。DB への登録／更新処理は登録／更新 AP がスプール管理 AP に処理待ちになっているドキュメントがあるかどうかを問い合わせるようにしている。処理待ちのドキュメントが存

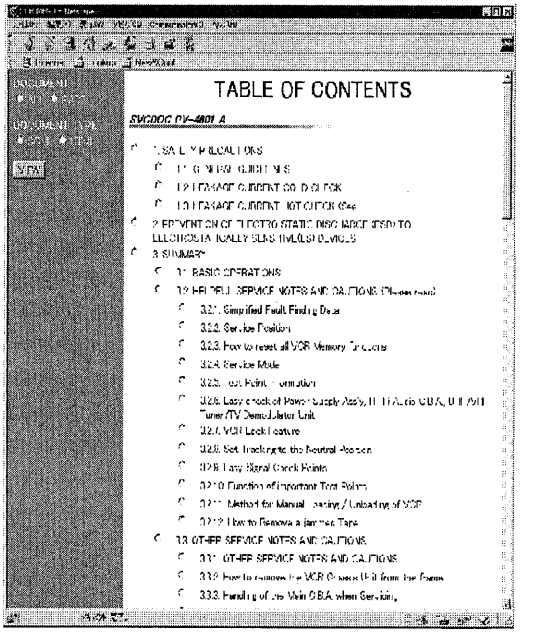
在する場合、スプール管理APは登録／更新APにそれらの情報を通知する。登録／更新APはDBへの処理後、結果をスプール管理APに通知する。これによりスプール管理APは各SGMLドキュメントの状態（登録／更新待ち、登録／更新処理中、処理済み）を把握することが可能となる。これらの情報はWebブラウザを使用して管理、操作することも可能である。

2.2.5. 登録AP

登録APは比較的検索処理の少ない夜間に自動的に起動される。スプール管理APに対して登録待ちSGMLドキュメントの問い合わせ要求を通知する。登録待ちSGMLドキュメントがある場合、登録APはスプール管理APからの応答情報に基づいてDBへの登録処理を行う。本システムのSGMLドキュメントはパーツ表やイメージファイル、また印刷用の属性情報が多く、パーサーを通してエレメントツリーに展開する場合、一文書当たりのデータ量が膨大になってしまう。検索処理時にユーザーが指定する検索条件にパーツ表内のデータや印刷属性の項目は使用しない事、また、エレメントに展開されたデータからSGMLドキュメント（インスタンス）への再構成（全体、またはユーザーが指定した部分木）にかかる時間を短縮するという考えから、パーツリスト（tableタグで定義されている）は外部エンティティ（外部参照ファイル）として分離し、そのためのエンティティ宣言を追加するようにした。変更されたSGMLドキュメントはSGMLデータカートリッジを経由して格納される。

2.2.6. SGMLドキュメントの検索

利用者は、Webブラウザを使用して文書管理サーバーにアクセスし、閲覧対象とするドキュメントを指定する。利用者がある特定の技術文書を選択した場合、文書管理サーバーからそのドキュメントに対する履歴情報（バージョン情報）が表示され、閲覧したいバージョンを選択する。文書管理サーバーでは選択されたバージョンでのインデックス情報（3～4階層までの見出しタグの情報であり、文書管理サーバーへのSGMLドキュメント登録時に作成される）を表示する。ユーザーが参照したいインデックスを選択すると、検索APは、展開されたエレメントテーブルから、ユーザが指定した部分木を抽出して再構成する。DBから抽出されたSGMLドキュメントデータは、テンポラリー・ディレクトリ上に置かれる。テンポラリー・ディレクトリは、文書管理サーバーの仮想ディレクトリでもあり、クライアントへはその参照先URLを記述したSGMLドキュメントを渡す。クライアントのWebブラウザでは文書管理サーバーから受け取ったSGMLドキュメントを起動したSGMLビューワーに渡す。このSGMLドキュメントはSGMLビューワー自体が文書管理サーバーにアクセスし、必要なSGMLドキュメントデータを取得する為に使用するためのものであり、表示はされない。

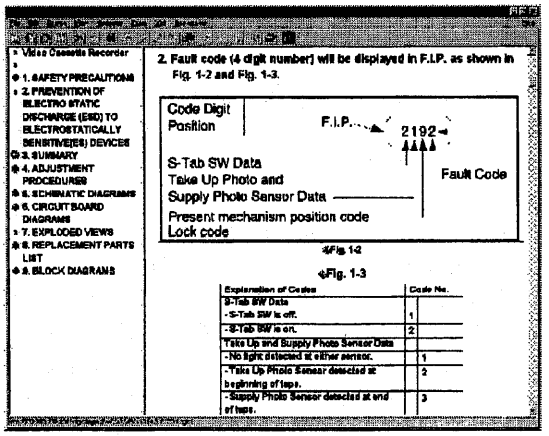


2.2.7. SGMLビューワー

市販の各種SGMLドキュメント用ビューワーの機能、操作性を評価した結果、適当なものが見つからず、専用のカスタマイズ製品が必要となった。そこでSYNEX社のSGML/HyTime Browser EngineであるViewPortを独自にカスタマイズすることとした。このSGMLビューワーはCD-ROMで配布されたSGMLドキュメントデータとWebブラウザからのヘルパーアプリケーションとしてネットワーク経由での参照を可能にしている。ネットワーク経由での参照の場合、SGMLビューワーは最初に取得するSGMLドキュメントが置かれているURL情報が記述されたSGMLファイルを読み込む。続いてそのURL情報を基に文書管理サーバーにアクセスする。SGMLビューワーが最初に取得するファイルは、SGMLドキュメントを構成するファイル情報が記述しているカタログファイルである。このカタログファイ

ル内の情報により表示する SGML ドキュメントデータを順次取得する。

利用者の多くは、各地に点在するサービス店や販売店であり、そこに設置されたパソコンがクライアントマシンとなる。それらの接続形態も比較的転送スピードが確保されている場合からそうでない場合まで様々であり、また SGML ドキュメント全体を表示する場合、そのデータ量は平均十数メガバイト以上にもなるので、それらを利用者にストレスなく表示させる方法が課題になった。SGML ビューワーは SGML ファイルを取得後、記述されているエンティティ宣言から外部参照ファイルを順次取り込む仕様になっている。そこで SGML ビューワーで直接表示される TIFF 等のイメージファイルの宣言を極力冒頭に記述し、そうでない PDF 等の宣言はその後に記述するようにした。SGML ビューワーはカタログファイル、SGML ファイルを取得した後、新たに Web サーバーからのファイル取得専用のアプリケーションをバックグラウンドで起動する。このファイル取得用アプリケーションは、SGML ファイルで宣言されている外部エンティティ・ファイルを順次読み込むためのものである。また SGML ビューワーはウィンドウ表示領域の近くに宣言されているイメージファイルを自動的に取り込む機能も持っており、対象のファイルがローカル・ディスクに存在しない場合には、自ら Web サーバーから取得するようになっている。既に等外ファイルがローカルディスクに取り込まれている場合には、そのデータを使用する。以上によって初期画面表示までの時間が短縮され、SGML ドキュメントデータを全て取り込むまでクライアントの画面には何も表示されないという事態が解消された。



2.2.8. 更新 AP

SGML ドキュメントを更新する場合、SGML ドキュメントデータを文書管理サーバーからクライアント端末にダウンロードする必用がある。予め編集する範囲が決まっている場合（例えば SGML ファイルのみやパーツリストのみの場合）には必要最小限のデータのみをダウンロードしている。当初、SGML ファイルの編集の際に、編集個所が限定される場合には編集取り出し範囲を選択可能とし SGML ファイルの部分木のみをダウンロードする方法を考慮していた。しかし、編集側ではダウンロードした SGML ファイルを直接編集するのではなく、コンバータを通して他のエディタ形式（RTF 等）に変換されたドキュメントを編集するので、このためにパーサーで使用する数種類の部分木に対応した編集用 DTD を用意しなければならない。これは DTD の管理面で問題があるので取りやめ、SGML ファイルの更新時は必ずファイル全体をダウンロードさせるようにした。SGML ドキュメントは一文書当たり数十メガバイトであるが、その大部分は TIFF、CGM 及び PDF ファイルであり、SGML ファイル本体は数百キロバイトにすぎない。また、SGML ファイルはテキストファイルなので SGML ファイル全体をダウンロードする場合でも ZIP や LHA 形式で圧縮すれば転送効率も良くなるということも理由になった。クライアントで編集したドキュメントは再度コンバータを経由して SGML ドキュメントを作成し、登録／更新タグファイルと一緒に文書管理サーバーのスプール領域へ転送する。文書管理サーバーでは登録処理と同様にスプール管理 AP から更新 AP にデータが渡され、SGML データカートリッジを経由して DB に格納される。

3. データカートリッジ

3.1. SGML データカートリッジ

本システムでは弊社が開発した SGML データカートリッジを使用した。SGML データカートリッジは AP I で提供しており、ソース API とエレメント API に大別される。ソース API で SGML ファイル、DTD、SGML 宣言、外部エンティティを DB 内に格納する。SGML データカートリッジは以下のようなテーブル群から構成されている。

・ディレクトリテーブル（DB 内でファイルシステムのように扱うために使用）

・ファイルバージョンテーブル（ファイル毎の履歴を管理）
・エレメントテーブル（SGML ファイルを展開した情報を管理）
・ファイル実体テーブル（DB 内でファイルの実体そのものを管理）
・エンティティ情報テーブル（SGML ファイルに記述されているエンティティ情報を管理）
・文書バージョンテーブル（SGML ファイルの履歴管理）
・文書バージョン構成情報テーブル（SGML ドキュメントを構成する DTD、外部エンティティ等の履歴管理）

SGML ファイルは SGML データカートリッジに内臓されたパーサーを通してエレメントツリーに展開され、エレメントテーブルやその他のテーブルに格納する。エレメントツリーに展開されたデータは個々にオブジェクト識別子を付与する。これにより履歴情報の管理や指定したバージョン、部分木での取り出しや更新が可能となる。

3.2. SGML データカートリッジの API

SGML データカートリッジは、以下のような API 群を提供している。

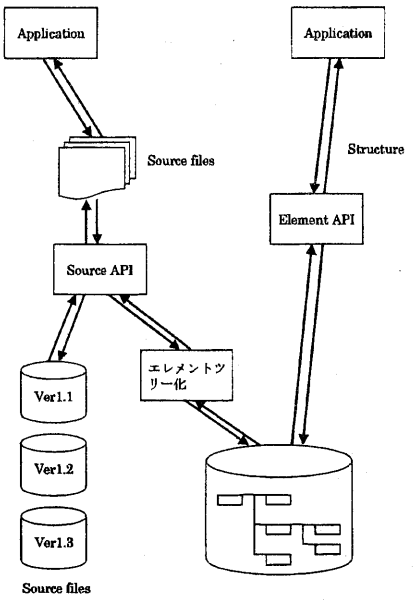
・セッション API
サーバへの接続とその解除を行う。ユーザの接続ごとにデータベース単位でのユニークな数値を作り出し、エレメントに付与されるオブジェクト識別子の一部として用いる。セッション内でのデータベース内容の変更の確定、取り消しはセッション・コミット、セッション・ロールバックの API を使用する。

・アクセス管理 API
アクセス権を管理するためのサーバ上のデータベースの内容を保守するための機能を持ち、これらに使用するデータを、データベースのユーザテーブル、グループテーブル、ユーザ/グループ対応テーブルに格納する。

・ファイル API
データベース上に OS のファイルシステムと同様なディレクトリ/ファイルのツリー構造を持たせ、SGML インスタンス、DTD、SGML 宣言、外部エンティティを各々ファイルとして管理する。ファイル関連 API は、外部エンティティ・ファイルなどを格納、取得を行う。また、文書バージョンの管理を行う。

・ソース API
SGML ファイルを内部のパーサによりエレメントツリー化してデータベースに格納する。文書の更新に関しては、バージョン毎にソースファイル、エレメントツリーを格納する。

・エレメント API
ソース API の機能を利用して格納してエレメントテーブルの内容をメモリ上にツリー構造で展開して、エレメント単位の部分木取り出しを行う。



4. DTD (Document Type Definition)

4.1. DTD の構成

今回のシステムで使用する DTD を作成するため、先行して各種資料におけるドキュメント構造の調査が行なわれた。製品により様々なレイアウト・スタイルが存在する中から、骨子となる共通構造を抽出し基盤 DTD を作成し、次に各ドキュメントイプ別の差異を定義したタイプ別 DTD を作成した。これには SGML ドキュメントをクロスメディアとして出力する機能、例えば、

印刷システムで処理するために必要な印刷属性などを定義している。

同時に SGML ビューワーで参照する際に使用するナビゲーションのスタイルシートに対応する DTD の作成も行った。クライアントが CD-ROM、Web ブラウザまたは印刷物等の様々なメディアで常に同様にドキュメントの参照を可能とすることが考慮された。以上により現行の印刷物のレイアウト、品質を確保すると同時に、クロスメディアとしての自由度と統一性を持たせることが可能となった。

4.2. システムによる DTD のカスタマイズ

先述した通り、データカートリッジを使用した場合、SGML ドキュメントは RDB 内においてエレメントツリーに分解して格納されている。各エレメントにはデータカートリッジによってユニークなオブジェクト識別子（OID）が付与され、それにより全てのエレメントが管理される。そのために、OID の付与および取り出しを可能とするようにタグの属性が追加され、それに伴い DTD のカスタマイズが行われている。その結果、クライアント・プログラムが SGML ドキュメントを DB から取り出す際には、全てのタグの属性にオブジェクト識別子が付与されるようになり、その値を用いてエレメント単位での履歴管理を実現している。

5. 考察

5.1 データカートリッジの有効性

以上、大手家電メーカの技術文書のライフサイクルを管理しつつ、各種メディアに出力し、様々な手段で配布可能なシステムについて紹介した。このシステムの枠組みは、既存の紙による配布、CD-ROM による配布、インターネット上のウエブブラウザによる参照、同ウエブブラウザに SGML ビューワをプラグインした参照機能と、幅広い配布、参照機能を提供するので、かなり普遍的な参照・配布システムと考えることが可能である。

本システムの実現にあたり、参照に要する待ち時間の短縮が繰り返し検討された。そのために、プロセスの処理順序や処理上の優先順位の管理が工夫された。それに伴い、DTD のカスタマイズも行われた。専用の SGML ビューワが開発されたのも、参照時の待ち時間、表示速度に対する要求条件が関係している。以上の対応を試みる毎に、参照性能は確実に向上し、実用に耐えるシステムが構築された。

一般に、SGML を用いたシステムの場合、DTD が必要でパーサでチェックせねばならない等、煩雑で使い難いという先入観に支配されがちであるが、本システムのように、ネックになる部分をカスタマイズしたり、既存のファイルを極力外部エンティティとしてそのまま扱うような手法を用いると、従来言われていた欠陥を殆ど意識せずに適用することが可能となる。DTD による枠組みの整理と、部分木の取り出し、個別のタグの性質を用いた API の定義等の考え方は、オブジェクト指向設計の考え方に近いものがある。XML の DOM[4]などは、この考え方をさらに推し進めて、API を OMG の IDL[5]化したものである。

5.2 IETM としての見方

本システムで実現した、Web を使用したドキュメントの登録、更新、検索を行うという機能は、クライアントのビューワと文書管理サーバー AP との会話型電子化マニュアルとして捉えることができる。会話型電子化マニュアルとしては IETM (Interactive Electronic Technical Manual) と称されるコンセプトが広く流通しており、今後の電子化された技術文書のモデルとなっている。IETM について、DoD は、次のような５階層にクラス分けしている[6]。

- クラス１：改ページ形式の表示
- クラス２：ページ指向のハイパーテキスト
- クラス３：コンテキスト管理されたフレーム指向のハイパーテキスト
- クラス４：クラス３を階層構造とし、表示手順の管理を可能としたもの
- クラス５：インテリジェント・エージェント形式の要素（ルールシステム、ニューロネットワーク等）統合したシステム

本システムでは SGML ドキュメントの文書構造に沿ったエレメントツリーでの文書管理を行い、Web ブラウザを使用した登録・更新・検索機能を実現している。これは上記のクラス４に相当するものと考えている。クラス 5 は、エージェント機能をベースにしたもので、FIPA[7]などによる標準化の成果を待たねば実現は不可能と考えられるので、現状の IETM はレベル４で実現されるのが妥当であろう。その面では、本システムの枠組みが、効果的であると考えられる。

5.3 データカートリッジの改良

今回用いたデータカートリッジは、DocTor/SGML V2.1と呼ばれる製品である。今回の製品でも基本的には十分な機能と性能を達成しているが、今後想定されるより大規模なシステムの構築に対処するために以下のような改良を考慮している。

・エレメントツリー
特定の単一ドキュメントに関する全ての履歴情報を単一のエレメントツリーで管理することを検討している。そうすることにより、変更された個所のみを差分情報として管理し、履歴に関しては随時差分情報を追加する事によりデータ領域を効率的に管理する。

・属性データのエレメントツリーからの分離
属性データをエレメントツリーとは別に管理する事によりインスタンスの検索スピードを上げる。

6. あとがき

以上、SGML データカートリッジを用いた、技術文書の配布・管理を効率的に実施するシステムについて紹介した。このシステムは、ほぼ順調に稼動しており、開発依頼元の顧客によりその概要が紹介されている[8]。

従来、SGML を用いたシステムの場合、煩雑で使い難いという先入観に支配されがちであったが、本システムのように、ネックになる部分をカスタマイズしたり、既存のファイルを極力外部エンティティとしてそのまま扱うような手法を用いると、従来言われていた欠陥を殆ど意識せずに適用することが可能となっている。今後、IETM のように、技術文書を電子化しインターネットにより配布するシステムのニーズが幅広い分野で高まると予想されるが、本システムはそのための具体的なモデルとなり得るものと考える。

最後に、本システムの使用者であり開発依頼元である、松下電器産業株式会社の北川勝巳氏に深く感謝します。また、システム構築に多大な貢献をされた株式会社アプローチの Moren Beyer 氏、INS エンジニアリング(株)の大塚正二氏、宮地文樹氏に感謝します。

文献

[1] 大野、佐藤；"ORDB によるマルチメディア・ドキュメントの管理", 情報処理学会デジタルドキュメント研究会研究報告（デジタルドキュメント, 7-5）, (1997.5.23)

[2] 大野；"ミドルウエアによる SGML/XML 文書管理の枠組みの検討",情報処理学会デジタルドキュメントシンポジウム講演論文（S1-5）, (1998.1.31)

[3] K.Ohno, M.Beyer ; "Dvelopment of SGML/XML Middleware Component", Proc. on SGML/XML Europe'98, (1998.5)

[4] Lauren Wood; "The Web Document API," Conference Procs. on SGML/XML'97, pp.445-448 (1997)

[5] OMG ; "共通オブジェクト・リクエスト・ブローカ - 構造と仕様 - CORBA 1.1", 創研プランニング, (1992)

[6] M. Dugand-Saenz; "Creating IETMs with WEB technology," Conference Procs. on SGML/XML'97, pp.493-499 (1997)

[7] Christophe Vermeulen ; "Software Agents Using XML for Telecom Service Modelling: A Practical Experienc", Prc. SGML/XML Europe'98, (1998.5)

[8] 北川勝巳；"サービス技術資料の電子化：SGML によるクロスメディアからネットワーク化への対応", SGML/XML ソリューション・ワールド'98 TOKYO 講演資料, (1998.7)