

新聞業界向け XML コンバータ

"XML Converter" for the Newspaper Industry

小柳恵介 *
Keisuke Oyanagi

鈴木義昭 *
Yoshiaki Suzuki

梅津良昭 *
Yoshiaki Umetsu

山田浩司 **
Hiroshi Yamada

城竈 進 **
Susumu Joho

草開 康弘 **
Yasuhiro Kusabiraki

* ソリューションビジネス本部 システム事業部 第一システム統括部 第一システム部

** PFU アクティブラボ株式会社 第二システム部

新聞業界では、ニュースコンテンツの一元管理、多メディアへの展開を急速に進めようとしている。

PFU と PFU アクティブラボ (株) は、新聞業界向けに多様な電文フォーマットの XML 変換に対応し、変換性能に優れていて、定義ファイルの変更により変換方式の変更が柔軟に行える『XML コンバータ』を開発した。

本稿では、XML コンバータの開発の背景とねらい、製品の概要、特長、今後の展開について述べる。

The newspaper industry is rapidly moving to promote the concentrated management of their news content and to expand their activity to other media. PFU and PFU Active Labs Limited has developed "XML Converter" for the newspaper industry. This is a tool that helps to convert messages of various digital formats into XML. It is capable of realizing high level of conversion and making such conversion highly flexible by altering the user-defined files.

This paper explains the background and purpose of development of our "XML Converter", its product outline, product features and future roadmap.

1 まえがき

新聞業界では、XML (Extensible Markup Language) ベースで、ニュース配信の標準フォーマットである NewsML を活用して、ニュースコンテンツの一元管理、多メディアへの展開を進めようとしている。まず記事を配信する通信社、それから新聞社へと急速に NewsML が普及しようとしている。

PFU と PFU アクティブラボは、XML の黎明期から実用化に取り組み、啓発・普及活動として XML コンソーシアムの活動にも参加し、XML に関する情報をいち早くキャッチして、XML 変換/編集ツールの開発に取り組んできた。その取り組みの中から、新聞業界向け XML 変換および編集ツールとして『XML コンバータ』を開発した。

2 開発の背景とねらい

2.1 開発の背景

(1) 新聞業界の動向

1998 年、英国ロイター通信社は、インターネットなどを通じてテキスト、画像、動画などの異なる性質を持つニュース素材を組み合わせて配信する方法として NewsML を発案し、1999 年に通信社の国際的な協議機関である国際新聞電気通信評議会 (IPTC) に NewsML を「次世代のニュース配信フォーマット」として標準化することを提案した。

NewsML のソース内には、IPTC 規定のニュースジャンル、通信社・新聞社などの情報 (取材日、場所、著作権、見出し等) を挿入することができ、国・メディアを問わず、同じテーマでニュースを横断して検索するこ

とが可能である。また、テキストだけでなく画像など様々なニュース素材を組み合わせ、インターネット・携帯電話・テレビなどの紙面以外のメディアに対する配信も可能である。

IPTC は NewsML の標準化を検討し、2000 年 10 月に正式に承認し、データ構造を定義した文書型定義 (DTD) である「NewsML ver1.0」がリリースされた。

国内では、社団法人日本新聞協会 (NSK) が 2001 年 8 月に NewsML を日本で使用するための使用ガイドライン¹⁾を公表した。2003 年 12 月には、通信社最大手の社団法人共同通信社が NewsML による電文配信を開始し、2008 年に従来形式の配信を終了することを表明している。そのため、大手新聞社、地方新聞社においても NewsML への対応が急務となっている。

(2) NewsML の特長

NewsML の特長として次のものが挙げられる。

- 1) XML やその他標準、仕様を基に、ニュース素材に対してコンパクトで、しかも拡張性が高い、柔軟な構造化の枠組みを提供する。
- 2) 電子的なニュースアイテム (ニュースの最小単位: 1 本の記事)、ニュースアイテムの集合、それらの間の関係、および関連のデータの意味を記述したデータであるメタデータの表現をシステムにもたらす。
- 3) NewsML は同じ情報の複数表現の規定を許し、任意のメディアタイプ、フォーマット、言語、符号化を混在して使用できる。
- 4) ニュースのライフサイクルのあらゆる場をサポートし、ニュースアイテムの繰返しの改版を許す。

2.2 開発のねらい

NewsML は上記のように複雑に構造化された XML ベースのフォーマットである。また、新聞社としてはすべての電文フォーマットが NewsML に切り替えられるまでは従来の電文にも対応しなければならない。

PFU では、既に提供しているコンフィグレーションパッケージ『コンフィグ Pro』に同梱している、CSV から XML に高速変換可能なツールをベースに、新聞業務に対応可能な変換ツールを『XML コンバータ』として提案し、開発することにした。開発に際しては以下の課題を解決することを目標とした。

(1) 多様な電文フォーマットへの対応

複雑な構造を持つ NewsML と既存電文フォーマッ

トへ対応できること。

(2) 高速変換

大量の記事情報を所定時間内に NewsML 又は XML に変換できること。

(3) 変換方式の柔軟な変更

次々に通信社から開示される新たな電文フォーマットに対して、プログラムを変更することなく、定義ファイルの変更・追加だけで変換方式を変更できること (対応費用・時間を最小限に抑えることが可能)。

3 本製品の概要

3.1 本製品に求められる要件

(1) 多様な電文フォーマットへの対応

NewsML の仕様としてコンテンツのデータ形式までは定義しておらず、実際に配信される NewsML の中には、CSV データや特定の文字を区切りとする可変長データが埋め込まれている場合がある。

また、新聞業界の流通データ形式は NewsML へ移行する流れとなっているが、依然既存システムを使用している通信社もある。配信データは固定長形式などの非 XML データであり、新聞社側のシステムが XML で情報流通を行う場合、これらの非 XML 形式の既存データを XML 形式に変換する必要がある。

新聞社でのデータ流通には多様なデータの変換が必要となるため、データ変換エンジンにはそれに追従できるだけの機能および拡張性が必要である。

(2) 高速変換

新聞社では、通信社からの記事データ受信から校正印刷開始までの所要時間として数秒というレベルを求められる。新聞システムでは、データ受信から数々の各種データ処理を行った後に校正印刷するが、処理経路が長くそれぞれの処理時間を要する。そのため、受信した NewsML のデータ変換処理についても相応の性能を求められる。また、選挙、オリンピック等のイベントがあるときは、通信社より大量の記事が送信されてくるが、その場合ピーク時で 1 秒あたり 10 件程度の NewsML を受信するため、それを処理できるだけの変換性能が必要である。

(3) 変換方式の柔軟な変更

(1)項でも述べたように、NewsML 内のコンテンツは様々な形式となるが、さらに通信社から新しい電文フォーマットの開示も予定されており、今後はコンテンツの種類が増えていくことが見込まれる。

NewsML データを受信する新聞社側でも、増加するコンテンツ形式に対応していく必要があり、システムのメンテナンス面およびコスト面から、変換定義の追加・変更だけで容易に、しかも柔軟に変換方式を変更できる仕組みが必要である。

3.2 構成

本製品は、XML コンバータ本体と定義ファイル保守ツールから構成される(図-1参照)。

(1) XML コンバータ本体

Java^{注1)}仮想マシン上で動作する変換エンジン部分である。変換エンジンは上位アプリケーションからAPIにより呼び出され、指定された定義ファイルの変換定義に基づいて入力データを変換し、変換後の出力データを上位アプリケーションに返す。

1) 変換機能

本製品では、入力データ形式や変換方法の違いに応じて表-1に示す六つの変換機能を提供している。

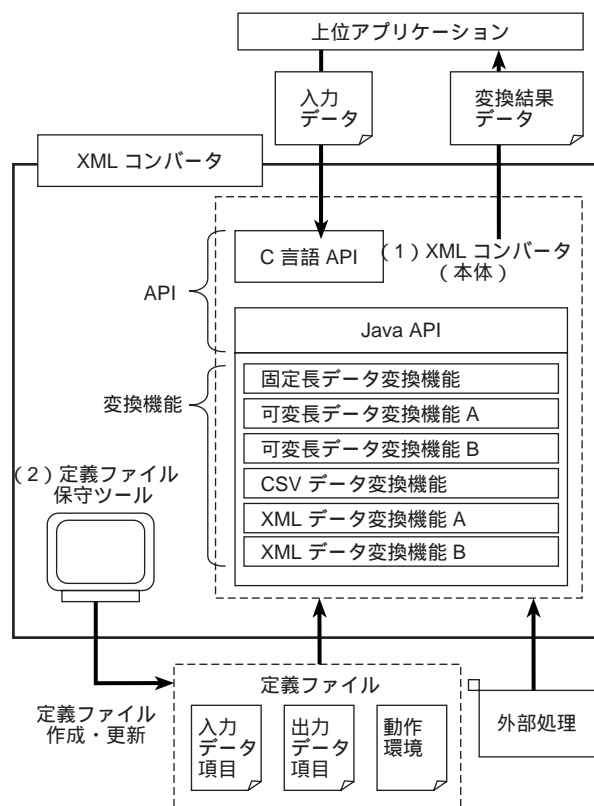


図 1 XML コンバータの構成
(Fig.1-Configuration of XML Converter)

注 1) Java およびすべての Java 関連の商標およびロゴは、米国およびその他の国における米国 Sun Microsystems, Inc. の商標または登録商標である。

2) API

Java API と C 言語 API を用意している。表-1の六つの変換機能はいずれの API から同等に利用できる。また、Java API, C 言語 API は Windows^{注2)}, Solaris^{注3)}, Linux 上で使用できる。

3) 定義ファイル

定義ファイルは独自形式であり、一つの変換処理定義に対して表-2に示す三つの定義ファイルを用意

表 1 機能概要

分類	機能
固定長データ変換機能	バイト単位で区切られた固定長データ(16進データを含む)をデータ変換する。
可変長データ変換機能 A (終了文字区切り)	任意の文字で区切られた項目を識別して、データ変換する。
可変長データ変換機能 B (開始文字区切り)	項目先頭の開始文字を識別して、データ変換する。
CSV データ変換機能	CSV データの各項目を識別して、データ変換する。
XML データ変換機能 A	XML データの要素・属性を識別してデータ変換する。入力 XML を先頭から読み込み、要素・属性の識別イベントに応じてデータ出力を行う。
XML データ変換機能 B	XML データの要素・属性を識別してデータ変換する。XML データ内の対象項目を XPath ^{注1)} ライクな指定方法で抽出し、出力データ形式定義ファイルの所定の箇所に文字列を埋め込む。

注 1) XML Path Language の略。XML ドキュメントの一部をアドレスリングするための言語である。

表 2 定義ファイル概要

分類	定義ファイル概要
入力データ項目定義ファイル	入力データのフィールド値のフォーマットを定義する。
出力データ項目定義ファイル	出力データの雛型を記述するテキストファイルである。
動作環境ファイル	変換に使用する入力・出力データ項目定義ファイルや、データのエンコーディング情報など、XML コンバータの動作に関する各種パラメータを指定する。

注 2) Windows は、米国 Microsoft Corporation の米国およびその他の国における登録商標である。

注 3) Solaris およびすべての Solaris に関連する商標およびロゴは、米国およびその他の国における米国 Sun Microsystems, Inc. の商標または登録商標である。

する。

上位アプリケーションからは、動作環境ファイル名を指定して変換処理を実行する。

本製品では独自の変換定義方法を採用している。他社製の XML データ変換エンジンでは XSLT^{注4)}を定義ファイルとしているものが多いが、新聞業界向けでは XML 形式以外のデータ変換に対する要件が多いため、XSLT に対応することは困難である。

(2) 定義ファイル保守ツール

XML コンバータの利用に当たり、あらかじめ定義ファイルを作成する必要がある。本製品では定義ファイルを容易に作成するため、定義ファイル保守ツールを用意している(図-2参照)。

当保守ツールは .NET Framework をベースとした Windows アプリケーションであり、定義ファイルを視覚的にわかりやすく扱うことができる。また、XML コンバータ本体は通常サーバで運用するが、サーバ上に配置された定義ファイルの更新機能や、定義作成ウィザード機能などにより利便性を高めている。

3.3 変換機能詳細

(1) 基本変換機能

XML コンバータは 6 種類の変換機能を持つが、こ

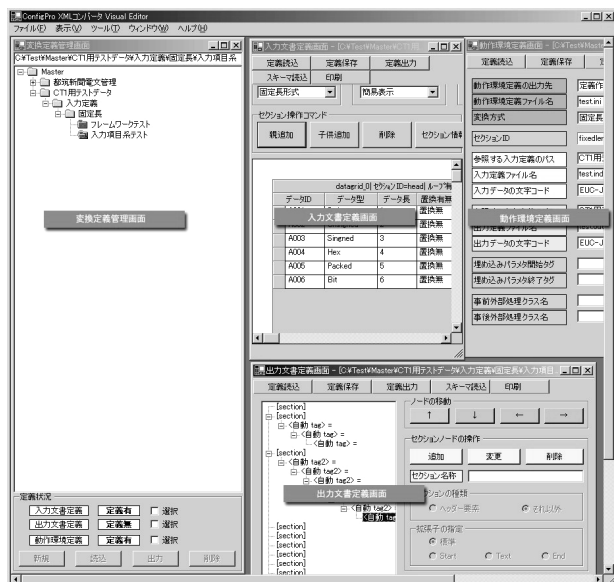


図 2 定義ファイル保守ツール
(Fig.2-Maintenance tool for user-defined files)

注4) XSL Transformations の略。XML で記述された文書を他の XML 文書に変換するための言語である。

なお、XSL は Extensible Stylesheet Language の略。XML 文書の見栄えを記述するための言語である。

こでは利用度数の高い 2 種類の変換機能について概要を説明する。

1) 可変長データ変換機能 A

任意の区切り文字で区切られた項目を識別して、データ変換する機能である。

変換処理のイメージを図-3に示す。

区切り文字を用いた可変長データは、従来の新聞システムの標準的なフォーマットである。システムの XML 化が進むなかで可変長データを XML へ変換する仕組みが必要となり、当変換機能の開発に至った。

当変換機能では、入力項目の識別時にイベント駆動的に出力データを生成するため、高速で、しかもメモリ消費が少なく、大きなデータの変換に適している。

2) XML データ変換機能 B

XML データ内の対象項目を XPath ライクな指定で抽出して一度 XML コンバータのメモリ上にツリー状のデータとして保持しておき、出力データ形式定義ファイルの所定の箇所に文字列を埋め込んでいる。

変換処理のイメージを図-4に示す。

XML データ変換機能 B では出力に必要な情報を一度メモリにすべて蓄えるため、他の変換機能に比べてメモリ使用量が大いだが、抽出した文字列を任意の箇所に出力できるため、柔軟性の高い変換処理が可能である。

(2) 変換機能の拡張

1) 変換機能の再帰的呼び出し

先に説明した各種変換機能について、別種の変換を行う変換機能の再帰的な呼び出しを可能としている。

例えば、NewsML のソース中に CSV データが含まれることがあるが、変換機能の再帰的な呼び出しにより、1 回の変換処理で純粋な XML データに変換可能である(図-5参照)。

2) 外部処理

新聞社でのデータ変換では、標準的なデータ変換機能ではサポートされないようなデータ変換要件が多い。それに対して本製品では、外部処理呼び出しという機構を設けることにより、様々な場面で臨機応変に機能拡張できる柔軟性を持たせている。

外部処理自体は Java プログラムとして作成する必要があるが、Java の入力・出力データ項目定義ファ

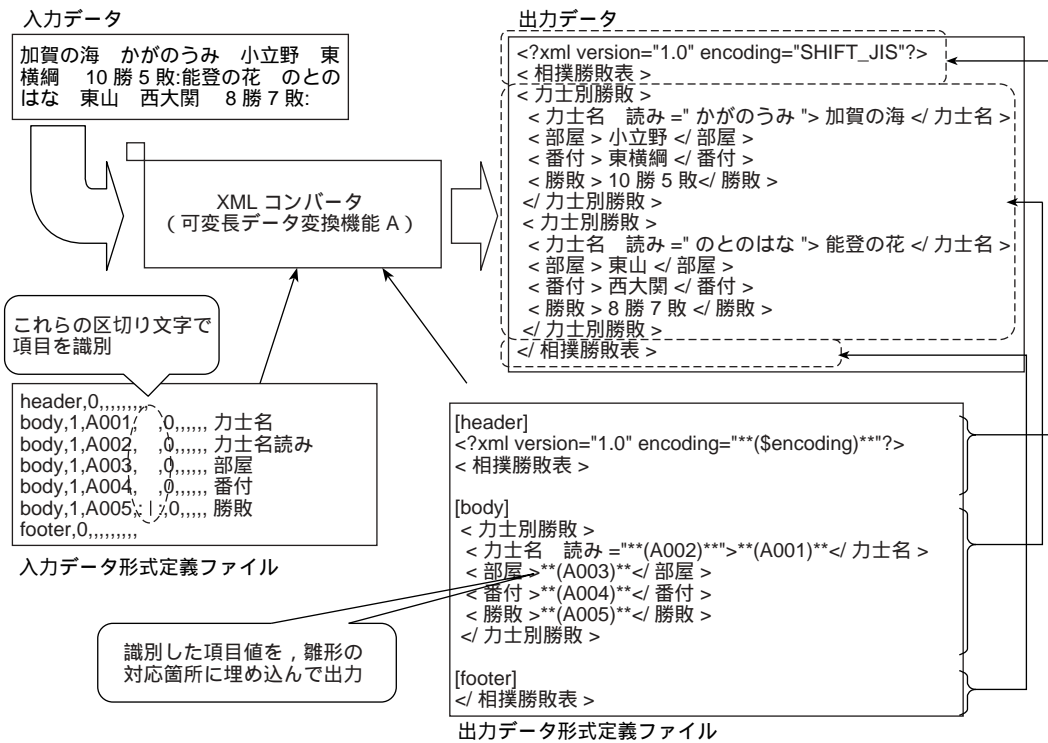


図 3 可変長データ変換機能 A の処理概要
(Fig.3-Overview of how "Flexible length data conversion function A" processes data)

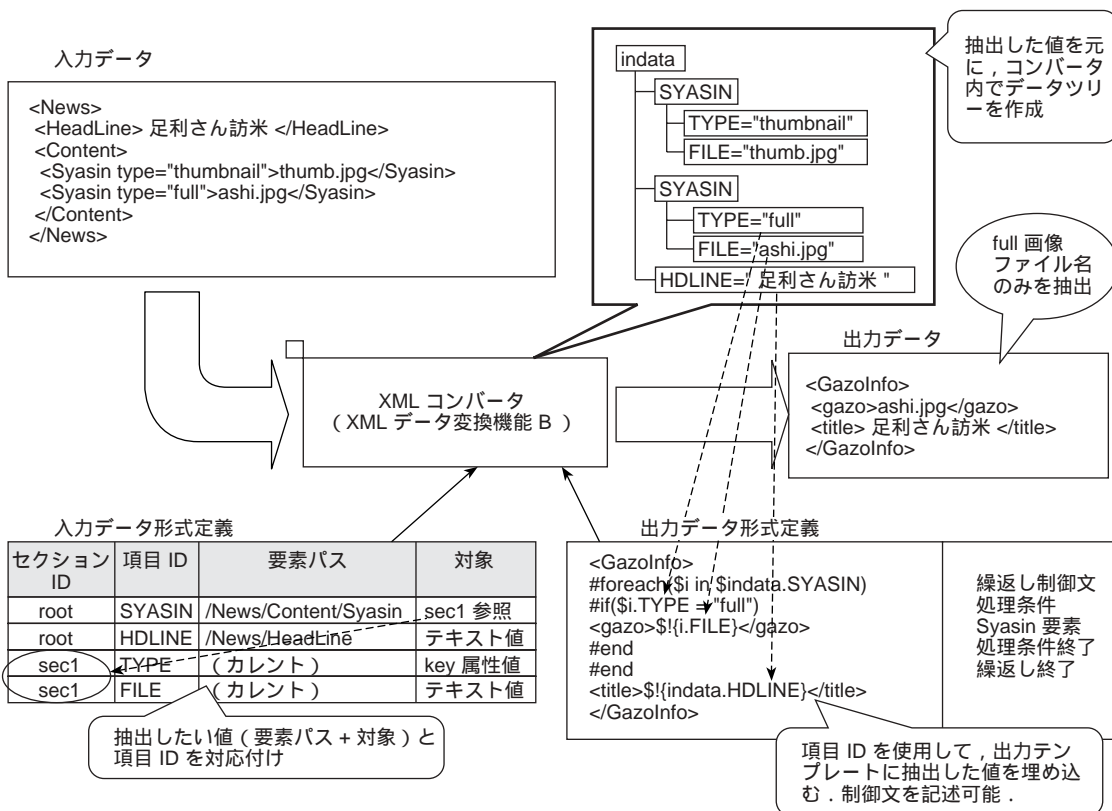


図 4 XML データ変換機能 B の処理概要
(Fig.4-Overview of how "XML data conversion function B" processes data)

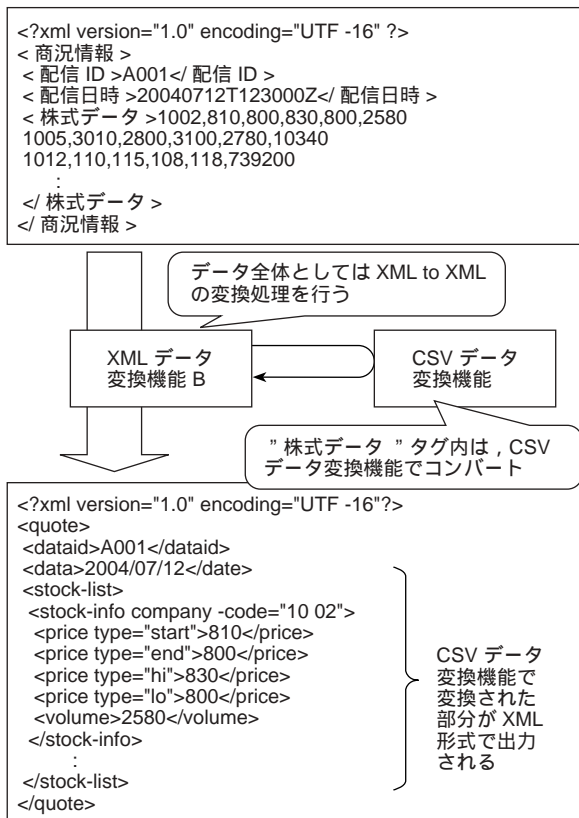


図 5 変換機能の再帰的呼び出し例 (Fig.5-Example of a recursive call of "Conversion function")

イルの中に使用する外部処理名 (Java クラス名・メソッド名) を記述することで外部処理の機能呼び出すことができ、XML コンバータ本体をカスタマイズすることなく、機能拡張が可能である。

図 - 6 に、入力側の数値 (1, 2, .. など) を丸つき数字 (①, ②) に変換して出力する例を示す。

3.4 変換性能

本製品では、高速な XML パース処理を可能とする富士通 XMLSDK V5.1 SAX^{注5)}パーサの採用や、各種の高速化手法を組み合わせることで性能向上を図っている。

その結果、可変長データ変換機能 A では、標準的なサイズの電文変換 (入力データ: 0.8 K バイト, 出力データ: 3 K バイト) で 1 電文あたり約 5 ミリ秒で

注5) Simple API for XML の略。XML 文書をアクセスするための API で、XML 文書を読み込みながら処理を行うイベント駆動型のインターフェースである。SAX パーサは SAX の構文解析プログラムである。

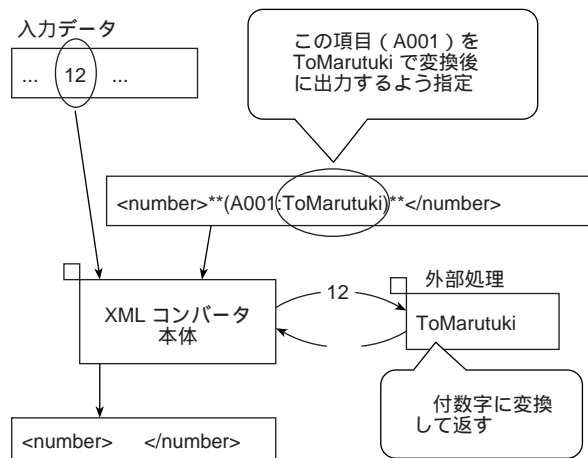


図 6 外部処理の使用例 (Fig.6-Example of external processing)

データ変換が行える。また、XML データ変換機能 B では、NewsML としては最大サイズとなる電文 (約 500K バイト) を変換した場合でも、1 電文あたり約 100 ミリ秒でデータ変換できており、新聞社での性能要件 (ピーク時で 1 秒あたり 10 件程度) をほぼ満足する性能を実現している。

XML データ変換性能について、他社製 XSLT エンジンと比較を行った。XML コンバータは他社 XSLT に比べ良好な変換性能をそなえている (図 - 7 参照)。

4 適用事例

新聞システムで XML コンバータが適用されている部分を図 - 8 の網がけ部分に示す。

(1) 外部通信社からの入稿

通信社から送られてくるニュース素材は NewsML 化されている。NewsML であるため構造の変換は不要であるが、これを素材データベースに格納する際、システム上の管理情報を付加した内部形式への変換に XML コンバータを用いる。

また、すべての通信社のニュース素材が NewsML 化されているわけではなく、様々な電文フォーマットで送られてくる。これらも XML コンバータを用い内部形式に変換する。

(2) 提携新聞社、Web 公開システムへの配信

提携新聞社あるいは Web サイトでニュースを公開するシステムへのニュース素材配信の際、XML コンバータを用いて内部形式から NewsML へ変換する。

5 今後の展開

これまで新聞業界向けに XML コンバータを開発し販売してきたが、それ以外の分野においても、データの XML 化が進むにつれ、高度で、しかも複雑なデータ変換が必要となることが予想される。XML コンバータは、多様なフォーマットに対応し、変換性能に優れていて、変換方式の変更が柔軟に行えることから、今後も様々な適用場面が出てくるものと考えられる。以下のようにプロモーション活動を展開し、適用分野の拡大を図る。

- (1) PFU アクティブラボ公開 Web サイトの『商品紹介』(<http://www.pfu.fujitsu.com/pal/>) に XML

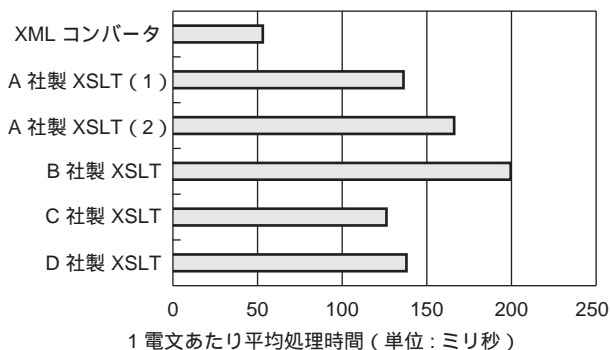


図 7 XML コンバータと他社製 XSLT の性能比較
(Fig.7-Comparison between XML Converter and other vendors' XSLTs)

コンバータの商品情報、適用事例を掲載する。

- (2) 富士通および富士通パートナーの協力を得てそれぞれが設けている Web サイトの XML ソリューションページとの相互リンクを図り、広範な業種での適用事例の閲覧を可能とする。
- (3) 公開 Web サイトを通じて XML コンバータの体験版を配布し、XML コンバータの高速変換をはじめとする特長を実感していただけるようにする。

6 むすび

今回、NewsML やその他電文フォーマットを効率的に XML データに変換するために、新聞業界向けの XML 変換および編集ツールとして XML コンバータを提供した。

今後、新聞業界は勿論、その他メディアの分野でも XML コンバータの採用が期待される。本製品の特長である、XML データ変換の多様なフォーマットへの対応、高速変換、変換方式の柔軟で容易な変更をアピールし、XML コンバータ・パッケージ採用の拡大、他業種への適用拡大を図っていきたい。

参考文献

- 1) 日本新聞協会 NewsML レベル 1 解説書 (第 1.0.3 版)
<http://www.newsml.jp/>

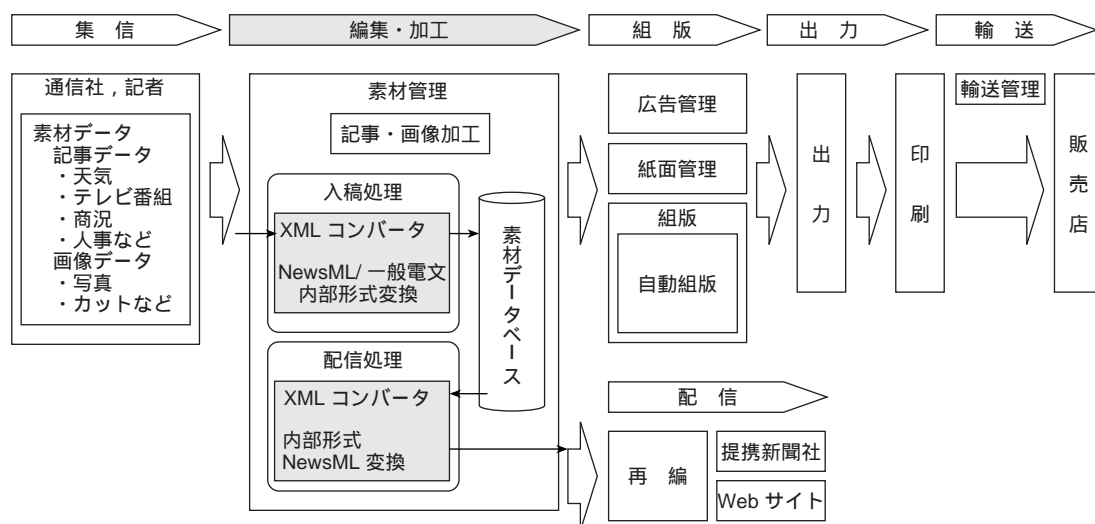


図 8 適用事例
(Fig.8-Case of application of XML Converter)