

## NewsML プロセッサの実装

森下亜里

日本アイ・ビー・エム(株) ソフトウェア開発研究所

XML ベースのニュース用の規格である NewsML 文書をモデル化し、操作するためのツール、NewsML プロセッサの実装を行った。XML 文書の汎用的なデータモデルとアクセスのための API としては DOM があり、これを利用して NewsML 文書を扱うこともできる。しかし、これは汎用的であるため、アプリケーションからは使用しにくい。NewsML プロセッサは、NewsML の各構成エレメントをひとつのクラスで表現するようにし、アプリケーションからの操作性を向上した。また、NewsML DTD に準拠しているかを検証するバリデーターの提供、およびアプリケーションの利便性を向上させるメソッドの提供も行った。本文では、NewsML プロセッサ実装の経験事例を報告する。

### はじめに

NewsML とは、国際新聞電気通信評議会 (IPTC, International Press Telecommunications Council) によって標準化された XML ベースのニュース用の規格である。ニュースの発生から修正、削除、送信というサイクルを通じて、ニュースを管理、表現するための標準である NewsML は、多くの新聞社、通信社で採用を検討されている。この NewsML V1.0 に準拠した NewsML 文書をモデル化し操作することを目的にしたツールが、NewsML プロセッサである。XML 文書の汎用的なデータモデルとアクセスのための API としては、DOM (Document Object Model) があり、これを利用して、NewsML 文書を扱うこともできる。しかし、DOM で扱うオブジェクトや、それを操作するメソッドは、Element、getChildNodes などの抽象的な名前、汎用的な操作方法になっているため、NewsML を扱うアプリケーションからは負担が大きい。そこで、NewsML に特化したモデルの作成、および操作をする NewsML プロセッサが必要とされた。筆者らは、NewsML 文書を扱うアプリケーション開発のワークロード軽減を目指して、NewsML プロセッサの開発を行った。本文では、NewsML プロセッサ実装の経験報告を行う。

NewsML 文書の構成は、図 1 のように、NewsML を文書ルートとしたツリー構造である。各エレメントの下には、複数の階層があり、エレメントの総数は 100 を超える。主要なエレメントの内容は、次のとおりである。

#### NewsEnvelope

配送 ID、送信元、受信先、送信日時、優先度、あて先などの情報を含む配信情報部である。

#### NewsItem

発行可能なニュース単位である。NewsML の中に複数の NewsItem エレメントを含むことができる。

#### Identification

NewsItem を識別する識別子である。

#### NewsManagement

NewsItem の管理情報である。ニュースのタイプ、作成日付、修正日付、緊急度などの情報を含んでいる。

#### NewsComponent

記事や写真、グラフィックスなどのオブジェクトの入れ物である。

#### ContentItem

ブレンテキストなど生データをそのまま格納できる。または、Href を使用して外部ファイルを参照することもできる。

### NewsML の構成

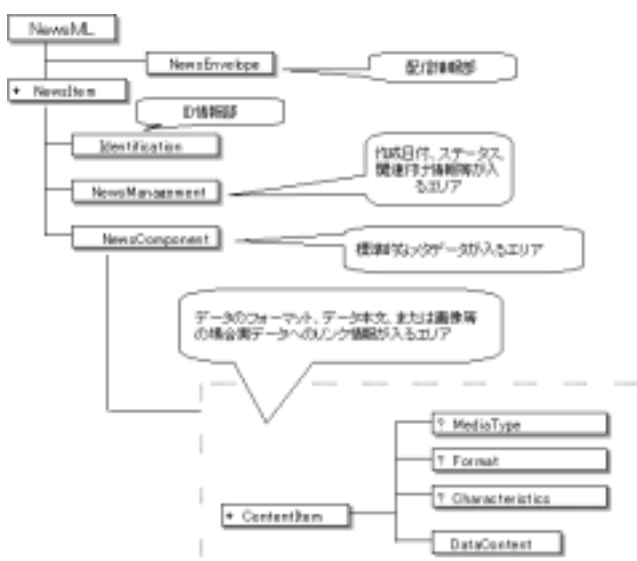


図 1. ニュース ML の構成

### NewsML プロセッサの概要

NewsML プロセッサの機能は、以下の 4 つに分けられる。

1. NewsMLBuilder
2. NewsMLWriter
3. NewsMLValidator
4. モデルの操作

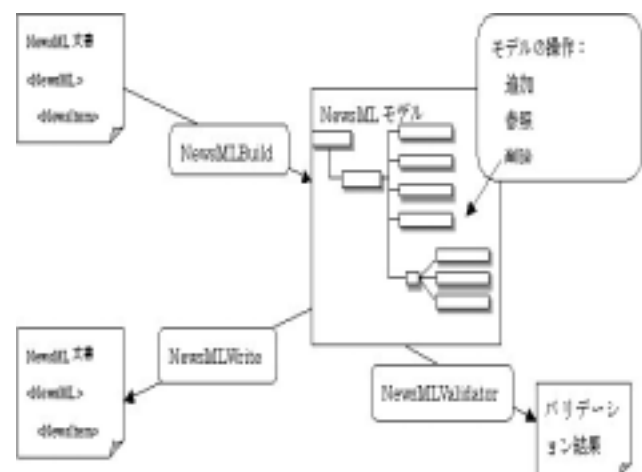


図 2. NewsML プロセッサの概要図

### 1.) NewsMLBuilder

NewsMLBuilder は Reader または InputStream から、XML 文書を読み込み、データモデルを作成する。内部的には、標準の SAX パーサーを使用している。

### 2.) NewsMLWriter

データモデルから XML 文書の変換を行う。出力先として、Writer または OutputStream が指定できる。

### 3.) NewsMLValidator

データモデルが NewsML の DTD に準拠しているかの整合性を検証し、エラーを出力する。

### 4.) モデルの操作

データモデル上の、アイテムを追加、削除、取得することができる。

## DOM との比較

### 1.) 操作性

NewsML プロセッサでは、各エレメントを独自クラスで表現している。つまり、NewsML、NewsEnvelope といったクラスを提供する。PublicIdentifier エレメントのデータを取得する場合のコード例は以下のとおりである。NewsML プロセッサを使用すると、簡単でわかりやすいコードで記述できることがわかる。

< DOM 使用のコード例 >

```
Element root=doc.getDocumentElement();
Element newsItem= getFirstElement(root,"NewsItem");
Element identification=
    getFirstElement(newsItem,"Identification");
Element newsIdentifier=
    getFirstElement(identification,"NewsIdentifier");
Element publicIdentifier=
    getFirstElement(newsIdentifier,"PublicIdentifier");
id=publicIdentifier.getFirstChild().getNodeValue();
```

< NewsML プロセッサ使用のコード例 >

```
NewsItem newsItem=newsml.getNewsItem(0);
id=newsItem.getIdentification().
    getNewsIdentifier().getPublicIdentifier();
```

### 2.) パフォーマンス

同一モデルをビルドするのにかかる時間は、DOM とほぼ同じである。モデル内のエレメント・データ取得にかかる時間は、NewsML プロセッサの方が早い。これは DOM の場合は目的のデータをエレメント名や属性名で検索しなければならないからである。

### 3.) 検証機能

NewsML プロセッサでは、通常の DOM パーサーで可能な、DTD による検証、すなわち必須エレメント、必須属性が存在するか、また予期しないエレメント、属性がないかなどの検証に加え、さらに高度な検証機能を行う。例えば、日付時刻フィールドに関しては、ISO8601 で規定された形式に沿っているかの検証も行う。また、RevisionId エレメントのエレメント・データは 1 以上の正の整数値で、PreviousRevision 属性値より大きくなければならないが、NewsML プロセッサではこれを検査することが可能である。

### 4.) 追加機能

さらに、NewsML プロセッサでは、DOM にはない以下の機能を追加した。

#### ・エレメントの内容を考慮した処理

エレメント同士の関係が重要な意味を持つ場合は、これを考慮にいたした処理を行なっている。たとえば、HeadLine と SubHeadLine は、ともに NewsLines エレメントの子エレメントであるが、これは組になって意味をもつので、その関連付けをモデルでも保持できるようにした。

#### ・日付データの処理

NewsML 上ではストリングである日付データを Java の Date オブジェクトとして取り扱えるようにしている。アプリケーションからは値を Date として読んだり、Date をセットすることが可能である。これによりアプリケーションの日付、時刻の処理が飛躍的に簡単になった。

#### ・クローン機能

NewsML プロセッサの任意のオブジェクト、例えば NewsItem 単位でのクローン機能を実装した。これにより新しいリビジョンの記事を作成したり、他の記事から任意の部分をコピーすることが可能になった。

#### ・ユーティリティ機能

NewsML を使用するアプリケーションでよく使われるであろう一般的な処理に関して、簡単にできるようなユーティリティ機能を提供した。例えば、リビジョンアップ機能などである。

## ユーザー拡張モデルへの対応

NewsML には、ユーザーが拡張して使用できる機能がある。この機能を利用して、例えば、移行時期において従来のシステムで使用しているデータを保存したり、また他と差別化するために独自のメタデータを定義し提供することができる。

### 1.) ユーザー拡張モデルのしくみ

NewsML の、汎用的なエレメントである Property および Metadata に独自の定義を行なうことで拡張する。

### 2.) 拡張モデルの作成支援

しかし、操作性を向上させるため、拡張されたモデルを NewsML プロセッサに組み込むための機能を提供した。

### 3.) 検証機能

拡張モデルを定義する定義ファイルに従い、ユーザー拡張部分についても検証が行なえるようにした。

定義ファイルは XML で簡単に記述可能であり、モデルの変更、追加などにも柔軟に対応ができる。また、この検証についてはデータタイプ、最小出現回数、最大出現回数、といったより高度な機能を実装している。

## まとめ

メディア業界はもちろんのこと、多くの業界で使用されていくことは間違いのない NewsML であるが、XML の標準の API ( DOM, SAX ) だけで取り扱うのは決して簡単なことではない。この NewsML プロセッサを開発したことで、アプリケーションの処理が簡単になり、生産性が飛躍的に向上することが十分に実証された。