NewsML for dummies

This version: draft2
Last revision date : September 30 2001
Author: L. Le Meur
Direction Technique - Direction Développement Multimédia AFP

This document is an introduction to the NewsML [1] structure, and its use by AFP ; it is intended to help XML developers understand the logic of NewsML, under its apparent complexity. It doesn't preclude the study of the NewsML functional specifications (www.newsml.org).

AFP now distributes its multimedia services in the NewsML format. This document describes the AFP implementation of NewsML, and constitutes a guideline for developers of news systems receiving those AFP services.

Presentation

NewsML is a **media independent** standard for describing news in an electronic service environment. NewsML defines an **XML** based language for expressing the **structure** of news , associated **metadata**, and **relationships** between news, throughout their lifecycle.

NewsML introduces several types of *objects*, each of them representing a level of news information ; from the lower level to the upper, they are :

- content item (**ContentItem**) used to wrap information content,
- news component (**NewsComponent**) used to structure and describe news information,
- news item (**NewsItem**) used to identify news information.
- news document (**NewsML**) used to transport news information,
- news package (e.g. **MIME** multipart-related envelope) used to package news data.

NewsItems, NewsComponents and ContentItems are generically referred as *news objects*.

The granularity of those objects has been chosen to get clear and non overlapping functions in the architecture.

The current NewsML version is v1.0, ratified in october 2000 by IPTC members.

---

[1] NewsML is developed by the IPTC, the international consortium of news publishers and vendors.

1 NewsML description

In this description, we use a bottom-up logic (from the physical content up to the abstract news management concept), inverse from the functional specifications logic, but perhaps easier to understand for NewsML crookies.

1.1 NewsML describes the structure of news items

(extract of IPTC members memos)

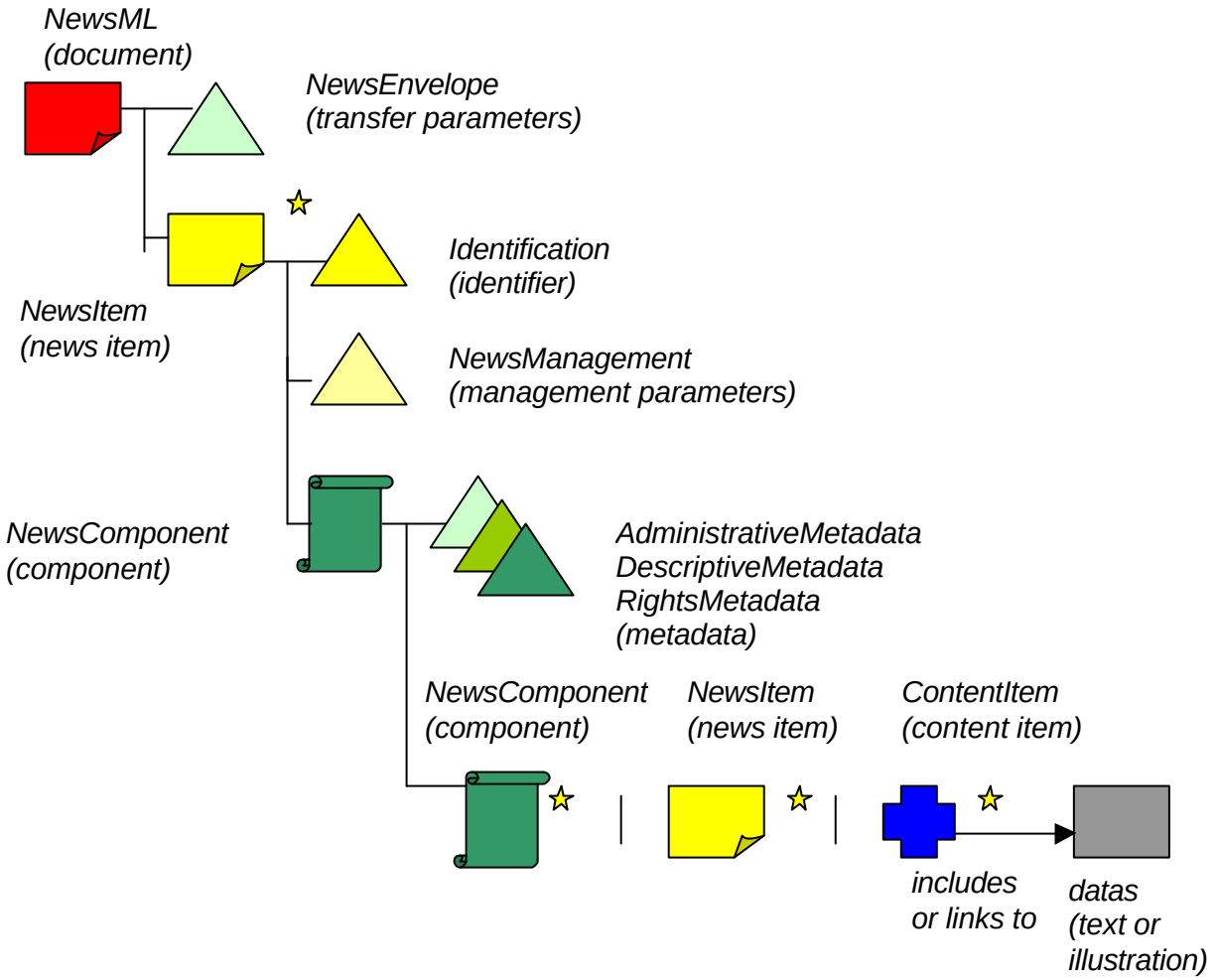
"NewsML focus is on structure: it does not represent layout or other presentational features. Such features can be represented using stylesheets, or other presentation languages based on XML technology such as XHTML or SMIL.

The needs to express complex structures are a requirement of NewsML, but NewsML representation is kept as simple as possible for news content commonly used.

Developed in news community, NewsML must not become a new all-purpose schema language, or abstract description framework like RDF or XML-Data. Thus NewsML definitions are directly understandable by people from the news environment.

NewsML provides support for all media-types in use today. Knowing the rapid pace of development in IT fields and news environment, NewsML must be easily extensible. This extensibility is achieved by the definition of generic tags, and by the standard means developed by the XML community, especially the namespace mechanism."

1.1.1 NewsML illustrated



1.1.2 NewsML wraps content

1.1.2.1 Concept of ContentItem

A **ContentItem** represents the unit of content managed in a news environment ; it's the representation of a textual or binary *resource* (data block). Examples of ContentItems are a text body, a digital picture, a vector graphic, an audio or video sample, or an animation. ContentItems possess some physical metadata, called *characteristics*. ContentItems are never found as stand-alone pieces of information in a news environment ; they are instead included in *news objects*, where they are locally identified .

1.1.2.2 Content media types

ContentItems fall in different categories (medias). Possible "**MediaTypes**" of ContentItems defined in NewsML 1.0 are:

- '**Text'** : represents a text body, e.g. a wire story, an article content, or a report. Such a string can be a plain text, or an XML text, with internal markup.
- '**Graphic'** represents a still graphic, in a bitmap or vector form.
- '**Photo'** represents a digital picture, a snapshot of a real world situation.
- '**Audio'** represents a digital audio sequence.
- '**Video'** represents a digital video sequence ; the sequence may include a soundtrack
- '**Animation'** represents a graphic animation, in a bitmap or vector form. This dynamic graphic can be a 2D or 3D work, and can be interactive, or a pure animation.

Those media types are found in a "topicset" (see below) called 'topicset.iptc-mediatype.xml'. Other named medias could be added in future NewsML revisions ; providers can also create their own types.

1.1.2.3 Content format and characteristics

ContentItems also get a **Format**, a **MimeType** and a **Notation**. The current IPTC *Format* topicset (topicset.iptc-format.xml) is a first shot, and AFP uses its own. The *MimeType* metadata is somewhat redundant with Format, follows a fuzzy registration mechanism, and mixes file format and information about the application needed for content edition. The *Notation* metadata is a third alternative way of giving a content format, coming from the XML world.

ContentItems accept different **Characteristics**. They hold information about the composition of a digital resource ; they are inherent attributes of a ContentItem and could be automatically retrieved from a digital content, but are provided in NewsML for easy search, retrieval, or representation purpose.

A **Size** (in bytes) is the only currently defined physical metadata ; IPTC will introduce other *characteristics* in the future (there's a current study on audio/video/photo standard characteristics), but such metadata can already be freely added by providers using a generic **Property** element.

1.1.2.4 Inclusion or reference of content

The data of a ContentItem can be included explicitly in its **DataContent** sub-element, or referenced by a URL (Uniform Resource Locator) via its **Href** attribute. If included in-line, binary parts are encoded as text data (e.g. base64), wrapped in the **Encoding** sub-element of DataContent. Be aware that this recursive feature allows for the representation of multiple encoding.

Note : a URL *isn't* a global id; a resource can be freely duplicated, so identical resources can be referenced by different URL's. Even if not duplicated, a single resource can be pointed at by different URL's, using alternative syntaxes (e.g. IP numbers vs host names). Users must also be warned that the referenced resource can be shared between several news items without any warning in NewsML.

1.1.3 NewsML structures news information

1.1.3.1 Concept of news object

A **news object** represents a piece of news : it adds *publishing* information to content, and so offers a point of view that relates to a specific point in time. News objects possess a **local or global identification**, and

a set of **news metadata** that relate them to a time and to a source (person or organisation) whose point of view they represent.

1.1.3.2 Concept of NewsComponent

A NewsComponent acts as

- a *container* of news objects,
- a *news metadata* handler,
- a anchor for news object *roles*,
- a *news lines* handler.

NewsComponent as a container of news objects

A **NewsComponent** is a *container object* that groups together other news objects (NewsItems, NewsItemRefs or NewsComponents) or ContentItems *in the context* of a NewsItem (see below).

A NewsComponent adds a hierarchical structure to the documents, a structure that doesn't represent layout, but news information. For example a NewsComponent can represent a picture with its caption, or several text parts in different languages.

A NewsComponent collection cannot include a mix of NewsItems/NewsItemRefs, NewsComponents and ContentItems : NewsML considers those as different beasts.

During IPTC discussions, some participants thought of a NewsComponent as a replicate of a RDF container , that could act as a 'Bag' (unordered list), a 'Sequence' (ordered list), or an 'Alternative' (a list of alternative resources). This idea evolved and the representation of the order of the news objects was put in stand-by mode.

The **EquivalentsList** attribute is used to stand that the elements of a NewsComponent collection are alternative objects : for example, several text parts in different languages are indeed alternative pieces of information, as are several pictures from the same scene in different formats, or alternative articles - one for the Web, one for the WAP - related to the same story.

Such *equivalent* collections get a **BasisForChoice** element, which gives the name of a property (the criteria) that significantly differs between those objects. BasisForChoice uses an **XPath** syntax to point at the chosen property (language, format ...).

Note : *XPath isn't so difficult to use*.
"Role/@FormalName" uses the Role as a basis for choice (points at the FormalName attribute of the Role element child of the current node).
"DescriptiveMetadata/Language/@FormalName" uses the Language as a basis for choice.
"ContentItem/Format/@FormalName" uses the content format as a basis for choice.

NewsComponent as a news metadata handler

A **NewsComponent** holds administrative, descriptive, rights metadata and NewsLines in the NewsML environment, but has no *globally unique identifier*.

NewsLines and news metadata are described below.

Local identification of NewsComponents

A local identifier is usually associated with a NewsComponent, in the context of a news document. So NewsComponents can be pointed at from another part of the document.

As NewsComponent don't get a globally unique identifier, they're not *reusable* in a NewsML environment : there's no way of knowing that two NewsComponents are identical, and no way to directly refer to an external NewsComponent.

1.1.3.3 Concept of Role

Each NewsComponent in a news document can get a **Role** sub-element. A Role is "the distinguishing characteristic of a NewsComponent, or its relationship to the others with which it is associated within the same containing NewsComponent" (IPTC).

A Role indicates why a news object is present inside its container. Roles can represent dependencies or logical links between objects, or can show the possible use of an object for a client application.

For example, in a news document that holds two NewsComponents representing photos, the Role of the first image can be 'FullRes' (full resolution), and the Role of the second image can be 'Thumbnail'. A NewsComponent representing an article can have the 'Main' Role, and another have the 'Sidebar' Role.

1.1.3.4 Concept of NewsLines

News lines are characteristic properties of *news objects* such as **slug, headline, byline**, **dateline, copyright information**.

They represent information in a human readable form and are characteristically displayed alongside the content of an object.

Using NewsLines, providers state some presentation rules for the information they deliver.

When several news objects grouped inside a NewsComponent share the same NewsLines, those are usually attached to the container NewsComponent, to avoid duplication.

News lines are inherently textual and may be present in multiple languages. (The "xml:lang" attribute may be used to specify the language used in a News lines element).

1.1.4 NewsML identifies and manages news information

1.1.4.1 Concept of NewsItem

A **NewsItem** is a *news object* that gets *identification* and *management* metadata. This is the unit of interchange in a news environment, an entry point in a web of included news objects, and thus it can be seen as a *publishable* object.

Be aware that NewsItems, being uniquely identified, are the only *reusable* news objects in a NewsML environment.

A NewsItem can represent a mono-media resource (Text, Photo, Graphic, Audio, Video, Animation), a multimedia resource, or a collection of resources.

A Multimedia NewsItem is a package of different NewsComponents, from mixed medias, usually created by a multimedia service. For example, an editorial column composed of a text and some photos; a thumbnail picture can be transmitted with this kind of document.

A Collection NewsItem is a group of related news objects ; included ContentItems usually share the same media (for example the five best pictures of the day). A Collection usually contains only links to other NewsItems ; the **NewsItemRef** element then provides a *pointer* to a NewsItem that is deemed to replace the NewsItemRef element.

1.1.4.2 Identification of news

The **Identification** element of a NewsItem handles a *globally unique news identifier*, a *name* and a *date*.

The globally unique news identifier (more or less equivalent to the IIM UNO (Unique Name of Object)) is constituted of four elements :

Provider Id	The domain name of the provider (“afp.com”).
Date Id	In practice, the creation date of the news item, in iso compact format (YYYYMMDD).
NewsItem Id	A news item identifier, unique in the publisher’s namespace for the given Date Id.
Revision Id	A revision number, that refers to an editorial decision to change the content The revision number is used to trace the evolution of a document or content through its life cycle. The default is ‘1’, and the value is incremented when revisions occur.

The news identifier offers two alternative way of identification :

- The **PublicIdentifier** is a NewsML URN (Uniform Resource Name), a kind of URI that identifies an object, but doesn't locate it explicitly. A NewsML URN is of the form “**urn:NewsML:ProviderId:DateId:NewsItemId:RevisionId**”**.**
- The set made of 4 elements {**ProviderId**, **DateId**, **NewsItemId**, **RevisionId**}

The optional **NameLabel** is a human readable label used to select a news item ; it’s not intended to be globally unique. Note that this label can’t be given is several languages, this is one of the issues of NewsMLv1.0.

The **DateLabel** is a free formatted label that indicates the publishing date of the news, as found in the *dateline* classical information.

1.1.4.3 Management of news

The **NewsManagement** element of NewsItem handles at least the *type* of the information, the *creation date* of the document, its *date of last modification*, and the *status* of the NewsItem. Other optional metadata can be set, as the *importance* of the information, *inheritance* and *associations* between NewsItems, or *special instructions* given to the recipient of the information.

NewsItems fall in different categories. Main **NewsItemTypes** defined in the current IPTC topicset called “topicset.iptc-newsitemtype.xml” are:

- **‘News’** : Basic coverage of a news event, such as articles, photos, video or audio reports.
- **‘Data’** : Non-narrative information, usually not eligible for journalistic intervention or modification, or information routed by the provider from a third party to the user. Examples are sports results and stock prices.
- **‘Advisory’** : A non-publishable communication providing information about the existing or planned coverage of news events.

The **FirstCreated** date is a system date associated with the NewsItem creation (the date the information was entered in a NewsML compliant system) : be aware that it’s not necessarily the date of origin of the information (the IIM date of intellectual creation). The **ThisRevisionCreated** date is associated with the current revision of the NewsItem.

NewsML provides mechanisms for the efficient handling of changes to NewsItems over time.
The **Status** element gives information about the NewsItem usability. Possible Status defined in the current IPTC topicset called “topicset.iptc-status.xml” are:

- **‘Usable’** : The NewsItem and its content may be published without restriction.
- **‘Embargoed’** : Neither the NewsItem nor its content may be published until released for publication by the provider.
- **‘Withheld’** : Neither the NewsItem nor its content may be published until further notice.
- **‘Canceled’** : Neither the NewsItem nor its content may be used under any circumstances. If the NewsItem or its content has been published, the publisher must take immediate action to withdraw or retract it, as may be legally necessary.

In the case of an embargo, the embargo limit can be set using the **StatusWillChange** element, which **DateAndTime** sub-element gives the end time, and **FutureStatus** is set to 'Usable'.

The importance of a NewsItem for editorial examination (from the provider's point of view) can be set via the **Urgency** element (equivalent to the IIM definition) using a numeric value from 1 to 8, 1 being most important. Be aware that this importance is in reality tied to the news item targeted audience. Many people in the news world think that a news agency can't set a global importance level to an information: an Importance attribute of the *OfInterestTo* descriptive metadata is here for this reason. Also don't mix importance and transmission priority: the latter in given by a NewsML NewsEnvelope sub-element called Priority.

If a NewsItem has an interesting inheritance relationship with another NewsItem, this can be described using the **DerivedFrom** element, which *NewsItem* attribute is set to the parent NewsItem NewsML URN.

If a NewsItem is clearly associated with other NewsItems, this can also be described using the **AssociatedWith** element, which *NewsItem* attribute is set to the associated NewsItem NewsML URN. In NewsMLv1.0, the semantic of this association can only be given via the *Comment* sub-element. AFP will study a controlled vocabulary for those association semantics.

The **Instruction** element is deemed to contain a controlled instruction from the news provider to the recipient of a NewsItem ; it can also give the status of previous revisions of a revised document. AFP doesn't use this field at this stage ; any new revision of a NewsItem should replace the previous one.

1.2 NewsML describes news information

Metadata of some kind are associated with every NewsML objects; metadata are mainly intended to assist production tasks, routing, search and retrieval.

NewsML recognises a number of different categories of *metadata*, and defines specific *metadata* properties within each category. It also provides a mechanism whereby the properties recognised within each category can be extended over time through the use of *controlled vocabularies*.

Metadata provide information about objects, from several angles : **physical characteristics** (what do they look like), **administration** (who made them, when, where), **description** (what's inside), **rights** (how they can be used). Physical characteristics have already been described, let's see the other classes.

1.2.1 Administrative metadata

Administrative metadata are intended to provide **who**, **when** and **where** information about a news object. As it happens, only the 'who' part has been developed in NewsMLv1.0. 'when' information like creation date is found in the news management set, and 'where' information is left to providers as extensions of NewsMLv1.0.

Current administrative metadata are : **FileName** (a proposed document file name), **SystemIdentifier** (a document URL), **Source** (the organization at the origin of the information), **Provider** (the organization that publishes the information), **Creator** (the name of the people that created the news object), **Contributor** (the name of a contributor to the news object ; there's no way to state what the contribution was).

Usually, two NewsComponents of a same news document will share the same administrative metadata (source, provider, creator, contributor) ; but a collaboratively created document could have different creators for different news objects.

Be aware that a FileName and SystemIdentifier are NewsItem level metadata, and are not logically associated with the NewsComponent.

1.2.2 Descriptive metadata

Descriptive metadata provide **what** information about a NewsComponent. They are more subjective that administrative metadata, and can't be set automatically. They provide a "point de vue", a "mise en regard de l'actualité", and thus make an object a *new object.*

Current descriptive metadata are : **Language** (that appear in the NewsComponent), **Genre** (the editorial class of the information), **SubjectCode** (the IPTC subject codes, with their three levels), **OfInterestTo** (the targeted audience of the NewsComponent), **TopicOccurence** (signals what topics appear in the NewsComponent).

A NewsItem can support an xml:lang attribute, using the iso 2 letters syntax. But AFP decided to use the Language metadata (multi-valued) instead, still using the iso 2 letters syntax.

1.2.3 Rights metadata

Rights give information about **ownership**, **copyright**, and **usage rights**. Copyrights have a name and a year parameter. Usage rights have a type, geography limits, restrictions a start date, an end date.

AFP didn't fully study this part of the NewsML specifications yet.

1.2.4 Metadata extensibility

The NewsML DTD and Schema offers a complete set of metadata elements in each category; extensibility is provided by the use of a "generic" **Property** metadata element, with a **FormalName** attribute that give a class to the property, and a **Value** attribute that gives its value. Property can be used recursively, allowing the creation of complex extended metadata.

1.2.5 Controlled vocabularies

NewsML doesn't provide a fixed set of values for any particular metadata attribute *value*, but instead uses a *controlled vocabulary* mechanism. Enumerated values for metadata attributes are not specified in NewsML DTD and schema, but in separate collections of names named *topicsets*.

Default vocabularies are defined by IPTC ; extensibility can be provided by the use of alternate vocabulary schemes, published by named authorities.
A *Catalog* of the topicsets used by a provider is represented at the top of any NewsML document as a file pointed at via an URL, and "http://www.iptc.org/NewsML/catalog/catalog.IptcMasterCatalog.xml" is the current URL of the IPTC catalog.

1.3 NewsML transports news information

1.3.1.1 Concept of NewsEnvelope

NewsML is used to transmit NewsItems between editorial systems. The **NewsEnvelope** provides transmission information data such as routing, and some basic workflow information.

Transmission information is thus separated from the publishable content of a NewsItem. This separation allows for the inclusion of a NewsItem within a NewsItem, without recursive inclusion of production information.

A NewsEnvelope gets a transmission identifier, **TransmissionId**. In case of transmission retries, this identifier stays the same but a **Repeat** numeric attribute is incremented (1 for the first retrie).

SentFrom and **SentTo** both give information about the origin and destination parties. The **Party** sub-element uses a controlled vocabulary specific to the provider.

DateAndTime is set to the transmission time

NewsService and **NewsProduct** use a controlled vocabulary specific to the provider. Products are subsets of services, using the IPTC IIM definition.

The **Priority** element sets the transmission priority, using a numeric value from 1 to 8, 1 being most important.

1.3.1.2 Concept of news package

Currently, the NewsML document and associated resources are transmitted side by side in different files stores in the same folder.

As a more compact future alternative, the included NewsItems and all related *resources* (the raw content) can be aggregated in a physical transport package, especially a **MIME envelope** (a better mechanism for binary transport is still to be found in the XML world). A naming system is used to identify the parts inside the package.

2 The AFP implementation

2.1 Simple text document

The simplest document is a flash story : it contains only mandatory elements (identification, creation date, status), plus a headline, the provider, and a copyright.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE NewsML PUBLIC "urn:newsml:iptc.org:20001006:NewsMLv1.0:1" "http://www.afp.com/dtd/NewsMLv1.0.dtd" [
    <!ENTITY % nitf SYSTEM "http://www.afp.com/dtd/nitf-2-5.dtd">
    %nitf;
]>
<NewsML>
    <Catalog Href="http://www.afp.com/dtd/AFPCatalog.xml"/>
    <NewsEnvelope>
        <DateAndTime>20000811T0818Z</DateAndTime>
    </NewsEnvelope>
    <NewsItem>
        <Identification>
            <NewsIdentifier>
                <ProviderId>afp.com</ProviderId>
                <DateId>20000811</DateId>
                <NewsItemId>010607144425.x6pxrl6k</NewsItemId>
                <RevisionId PreviousRevision="0" Update="N">1</RevisionId>
                <PublicIdentifier>urn:NewsML:afp.com:20000811:010607144425.x6pxrl6k:1</PublicIdentifier>
            </NewsIdentifier>
        </Identification>
        <NewsManagement>
            <NewsItemType FormalName="News"/>
            <FirstCreated>20000811T0818Z</FirstCreated>
            <ThisRevisionCreated>20000811T0818Z</ThisRevisionCreated>
            <Status FormalName="Usable"/>
        </NewsManagement>
        <NewsComponent>
            <NewsLines>
                <HeadLine>George W. Bush president (television)</HeadLine>
                <CopyrightLine>© 2001 AFP</CopyrightLine>
            </NewsLines>
            <AdministrativeMetadata>
                <Provider>
                    <Party FormalName="AFP"/>
                </Provider>
            </AdministrativeMetadata>
        </NewsComponent>
    </NewsItem>
</NewsML>
```

You feel it’s big ? perhaps, but the information you’ve got there allows a perfect handling of the story.

Let’s check the information here :

Encoding

The xml encoding is currently iso-8859-x : most of the current system support this type of charset, but AFP will move to UTF8 when the majority of our customer systems support it.

Document validation

NewsML and NITF dtd references are set in the AFP documents using the following syntax :

```
<!DOCTYPE NewsML SYSTEM "http://ww.afp.com/dtd/NewsMLv1.0.dtd" [
    <!ENTITY % nitf SYSTEM " http://ww.afp.com/dtd /nitf-2-5.dtd">
    %nitf;
]>
```

You can see that the dtds have been copied to our web site to avoid a bottleneck on the iptc site.

As usually practiced in the XML community, AFP asks its customers to deactivate the NewsML document validation in their parser in production mode, or if not possible to copy in the reception system the NewsML DTD and modify in the received documents - before validation - the URL that references the DTD, so the validation could be local.

Two main advantages to this :
- no Internet access to the DTD at *each* file opening,
- no access problem to the AFP site.

Note : AFP could later decide to suppress this string in production mode.

Catalog

- NewsML/Catalog/@Href : sets the URL of the file that contains the controlled vocabularies used by AFP. This file is on-line on the AFP web site : most of the vocabularies are the standard IPTC vocabularies, found on the IPTC web site; but Format, Property, Role, Language are customized AFP vocabularies.

News envelope

- NewsML/NewsEnvelope/DateAndTime : transmission date. Like other dates, follows the ISO8601 compact format, **UTC based**, using the Zulu (Z) syntax. Example : 20012805T1000Z is equivalent to 20012805T1000+0000,

News item

The news item contains three parts : identification, management and a component.

News item identification

- NewsItem/Identification/NewsIdentifier/ProviderId : provider domain name (afp.com),
- NewsItem/Identification/NewsIdentifier/DateId : creation date of the document,
- NewsItem/Identification/NewsIdentifier/NewsItemId : item identifier, unique for the given provider and date,
- NewsItem/Identification/NewsIdentifier/RevisionId : item revision (by default 1) ; AFP doesn't use item revision at he moment,
- NewsItem/Identification/NewsIdentifier/RevisionId/@PreviousRevision : previous revision (by default 0),
- NewsItem/Identification/NewsIdentifier/RevisionId/@Update : update flag (always N) ; AFP currently doesn't use this feature,
- NewsItem/Identification/NewsIdentifier/PublicIdentifier : news item URN : this is the compact representation of the previous identification elements,
- NewsItem/Identification/NameLabel : item name (slug),

News item managment

- NewsItem/NewsManagement/NewItemType/@FormalName : item type (News |Data |Advisory),
- NewsItem/NewsManagement/FirstCreated : creation date,
- NewsItem/NewsManagement/ThisRevisionCreated : last revision date,
- NewsItem/NewsManagement/Status/@FormalName : status (Usable |Canceled),

News item component

The news component contains newslines and administrative metadata.

- NewsComponent/NewsLines: textual information, directly publishable ; see below,
- NewsComponent/AdministrativeMetadata : administrative metadata ; see below,

NewsLines describe some directly publishable textual information :

- NewsLines/HeadLine : item title,
- NewsLines/CopyrightLine : AFP short copyright ,

The AdministrativeMetadata element defines some factual information :

- NewsComponent/AdministrativeMetadata/Provider/Party/@FormalName : provider (always AFP).

This basic document has no content in the NewsComponent. Now let's see a more complex document.

2.2 Simple multimedia document

The simplest multimedia document we can think of is a wire picture : it aggregates a text component (the caption), an a binary component (the raw picture).

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE NewsML PUBLIC "urn:newsml:iptc.org:20001006:NewsMLv1.0:1" "http://www.afp.com/dtd/NewsMLv1.0.dtd" [
    <!ENTITY % nitf SYSTEM "http://www.afp.com/dtd/nitf-2-5.dtd">
    %nitf;
]>
<NewsML>
    <Catalog Href="http://www.afp.com/dtd/AFPCatalog.xml"/>
    <NewsEnvelope>
        <DateAndTime>20010705T123552Z</DateAndTime>
    </NewsEnvelope>
    <NewsItem>
        <Identification>
            <NewsIdentifier>
                <ProviderId>afp.com</ProviderId>
                <DateId>20010705</DateId>
                <NewsItemId>010705102423.s27th00q</NewsItemId>
                <RevisionId PreviousRevision="0" Update="N">1</RevisionId>
                <PublicIdentifier>urn:NewsML:afp.com:20010705:010705102423.s27th00q:1</PublicIdentifier>
            </NewsIdentifier>
            <NameLabel>MODE</NameLabel>
        </Identification>
        <NewsManagement>
            <NewsItemType FormalName="News"/>
            <FirstCreated>20010705T102425Z</FirstCreated>
            <ThisRevisionCreated>20010705T102425Z</ThisRevisionCreated>
            <Status FormalName="Usable"/>
        </NewsManagement>
        <NewsComponent>
            <NewsLines>
                <HeadLine>Mode: la haute couture défile à Paris la semaine prochaine</HeadLine>
                <DateLine>FRANCE (AFP) </DateLine>
                <CopyrightLine>© 2001 AFP</CopyrightLine>
            </NewsLines>
            <AdministrativeMetadata>
                <Provider>
                    <Party FormalName="AFP"/>
                </Provider>
                <Creator>
                    <Party FormalName="JEAN-PIERRE MULLER"/>
                </Creator>
            </AdministrativeMetadata>
            <NewsComponent>
                <Role FormalName="Caption"/>
                <ContentItem>
                    <MediaType FormalName="Text"/>
                    <Format FormalName="bcNITF2.5"/>
                    <DataContent>
                        <p>les deux frères stylistes Vartan et Guevork Tarloyan posent en compagnie d'un mannequin
portant un chemisier de leur collection, le 21 juin 2001 à Paris, avant les présentations de haute couture qui auront lieu du
07 au 11 juillet 2001 à Paris.</p>
                    </DataContent>
                </ContentItem>
            </NewsComponent>
            <NewsComponent>
                <Role FormalName="Preview"/>
                <ContentItem Href="SGE.SDI86.050701102222.photo00.default-384x256.jpg">
```

```
                    <MediaType FormalName="Photo"/>
                    <Characteristics>
                        <Property FormalName="Width" Value="384"/>
                        <Property FormalName="Height" Value="256"/>
                    </Characteristics>
                </ContentItem>
            </NewsComponent>
        </NewsComponent>
    </NewsItem>
</NewsML>
```

We already described the news item ; let's focus on the NewsComponent.

News component

The main NewsComponent is now more complex :

- NewsComponent/NewsLines: textual information, directly publishable ; see below,
- NewsComponent/AdministrativeMetadata : administrative metadata ; see below,
- NewsComponent/NewsComponent : component (caption and picture) ; see below.

News lines

- NewsLines/HeadLine : item title,
- NewsLines/DateLine : origin of the information (date and place),
- NewsLines/CopyrightLine : AFP short copyright ,

Administrative metadata

- NewsComponent/AdministrativeMetadata/Provider/Party/@FormalName : provider (always AFP).
- AdministrativeMetadata/Creator/Party/@FormalName : name of the creator,

Caption

The first NewsComponent (text) is the caption :

- NewsComponent/Role/@FormalName : component role (Caption),
- NewsComponent/ContentItem :content ; see below,

The ContentItem element of the caption includes :

- ContentItem/MediaType/@FormalName : component type (Text),
- ContentItem/Format/@FormalName : component format (bcNITF2.5),
- ContentItem/DataContent/p : caption content ; there is at the moment no tags inside the single caption paragraph.

Picture

The other NewsComponent contains :

- NewsComponent/Role/@FormalName : component role (Preview),
- NewsComponent/ContentItem :content, see below,

The ContentItem element describes the illustration content :

- ContentItem/@Href : content file URL,
- ContentItem/MediaType/@FormalName : component media type (Photo …),
- ContentItem/Characteristics/Property[@FormalName="Width"]/@Value : image height,
- ContentItem/Characteristics/Property[@FormalName="Height"]/@Value : image width.

2.3 The text markup

AFP text components are marked-up using the NITF standard (again an IPTC standard, currently in version 2.5, see www.nitf.org).

An *enriched text* includes :

- paragraphs,

- illustration placeholders,
- preformatted string, sub-titles, tables, lists and other structural elements,
- organization codes.

« Illustrations » are photos, graphics and other media objects included in a document. A document can include several such objects ; their recommended location is given before a paragraph tag, using a local identifier that leads to another NewsML component identified by the same identifier attribute ; a recommended style can be given for each illustration.

As seen in the previous example, AFP chose to use exclusively the 'body.content' part of NITF to describe text, and to support a subset of NITF focused on structure, stripping layout markup. All metadata are set at the NewsML level.

NITF being declared in the doctype, and DataContent being of type ANY, **any** NITF element can be put as a child of the DataContent element. Using the 'body.content' element would be a waist of space, and doesn't mean much in this NewsML environment, so we use directly the elements found inside the body.content element.

This subset of NITF is indicated by the format value used in the *ContentItem/Format/@FormalName* element, which is set to "**bcNITF2.5**" (bc stands for 'body content'). It would be great if other implementers were using the same approach ;-). The *MimeType* element is not used in our implementation.

To designate a text component, AFP uses the *ContentItem/MediaType/@FormalName* set to the value "Text".

Example fragment :

```
<NewsComponent>
    <ContentItem>
        <MediaType FormalName="Text"/>
        <Format FormalName="bcNITF2.5"/>
        <DataContent>
            <p>France achieved 3.4 percent growth in 2000, up from a previous announcement of 3.3 percent
announced initially, the statistical office INSEE said on Friday.</p>
            <p>It maintained its forecast for 2001 at 2.3 percent.</p>
            <p> … </p>
            <p> em/jmy/djw  </p>
        </DataContent>
    </ContentItem>
</NewsComponent>
```

2.3.1 NITF elements

The DataContent element directly contains the enriched text elements :

- p : paragraph, can contain '**a**' and '**org**' tags (see below) ; the 'lead' is the first paragraph in the content,
- hl2 : sub-title,
- ol : ordered list of 'li' elements,
- table : table,
- pre : preformatted string,
- media : placeholder for an illustration.

2.3.1.1 Hyperlinks

Hyperlinks – an <a> tag – are present inside a paragraph. They are usually used to link a document to an external web site ; other classes of hyperlinks could be used in some AFP products.

a attributes :

- href : target URL

- name : link description
- title : title of the target content
- class : hyperlink class (website …).
- style : hyperlink style.

Web site hyperlink

This class of hyperlink points to a Web site page, which content is considered as complementary of the current document.

The 'class' attribute is set to 'webSite' ; 'href' points to the site ; 'name' gives the site name ; the tag content describes the link.

Such links are often put in a specific paragraph, at the end of the enriched text (a link 'box'); the specific role of this paragraph is given by the 'class' attribute, set to the 'links' value.

Example :

```
<NewsComponent>
  <ContentItem>
    <MediaType FormalName="Text" />
    <Format FormalName="bcNITF2.5" />
    <DataContent>
       <p> … </p>
       <p> … </p>
       <p class="links">
           <a class="webSite" href="http://www.iptc.org " name="IPTC web site">More
           information about IPTC</a>
       </p>
    </DataContent>
  </ContentItem>
</NewsComponent>
```

2.3.1.2 Organization codes

Organization codes – the <org> tag – are present inside a paragraph. The 'org' tag contains the name of an organization / company, and supports two mandatory attributes, 'idsrc' and 'value' ; it can contain also a set of <alt-code> tags, that add alternates values.

org and alt-code attributes :

- idsrc : code vocabulary (ISIN or SICOVAM),
- value : organization code.

Example :

```
<NewsComponent>
  <ContentItem>
    <MediaType FormalName="Text" />
    <Format FormalName="bcNITF2.5" />
    <DataContent>
       <p>
        <org idsrc="ISIN" value="FR0000130650" >
          DASSAULT SYSTEMES
          <alt-code idsrc="SICOVAM" value="13065" />
        </org>
       </p>
    </DataContent>
  </ContentItem>
</NewsComponent>
```

2.3.1.3 Sub-titles

Sub-titles – the <hl2> tag – are present before a paragraph. Their use in AFP products should growth in the future.

Example :

```
<NewsComponent>
  <ContentItem>
    <MediaType FormalName="Text" />
    <Format FormalName="bcNITF2.5" />
    <DataContent>
      <hl2>Growth in France</ hl2>
      <p>France achieved 3.4 percent growth in 2000, up from a previous announcement of 3.3 percent announced initially, the statistical office INSEE said on Friday.</p>
      <p> … </p>
    </DataContent>
  </ContentItem>
</NewsComponent>
```

2.3.1.4 Tables and lists

The NITF table uses the HTML syntax, and imposes the use of a 'tbody' element.

Example :

```
<NewsComponent>
  <ContentItem>
    <MediaType FormalName="Text" />
    <Format FormalName="bcNITF2.5" />
    <DataContent>
        <table>
            <tbody>
                <tr>
                    <td>01</td>
                    <td>Michael Schumacher</td>
                    <td>Deutschland</td>
                    <td>Ferrari</td>
                    <td>1:20,447</td>
                </tr>
                <tr>
                    <td>02</td>
                    <td>Mika Häkkinen</td>
                    <td>Finnland</td>
                    <td>McLaren</td>
                    <td>1:20,529</td>
                </tr>
                <tr>
                    <td>03</td>
                    <td>David Coulthard</td>
                    <td>Schottland</td>
                    <td>McLaren</td>
                    <td>1:20,927</td>
                </tr>
            </tbody>
        </table>
    </DataContent>
  </ContentItem>
</NewsComponent>
```

Note :
We hope that tbody won't be mandatory anymore in NITF3.0 tables, and will follow XHTML rules.

2.3.1.5 Preformatted text

Some data formatted via a text editor have a presentation based on space entries (those are usually table-like texts). To be able to exchange them correctly in xml, the spaces must be kept as is.

The <pre> tag is set for this purpose.

Example :

```
<NewsComponent>
  <ContentItem>
```

```
    <MediaType FormalName="Text" />
    <Format FormalName="bcNITF2.5" />
    <DataContent>
        <pre>
        01  Michael Schumacher    Deutschland    Ferrari    1:20,447
        02  Mika Häkkinen         Finnland       McLaren    1:20,529
        03  David Coulthard       Schottland     McLaren    1:20,927
        </pre>
    </DataContent>
  </ContentItem>
</NewsComponent>
```

2.3.1.6 Illustration placeholders

The <media> element shows the recommended location for an illustration ; such a location is set between two paragraphs. The illustration is described in details in a <NewsComponent> element, outside the text part. The 'data-location' attribute identifies the component (as an XML fragment identifier). The component gets a local id, and contains the object metadata, including its URL.

1. DataContent/media/@media-type : illustration type (image | audio | video | other),
2. DataContent/media/@style : style of the illustration (position along the text),
3. DataContent/media/media-reference/@data-location : pointer to the component ('fragment identifier', beginning with a '#'),
4. DataContent/media/media-reference/@mime-type : mime type of the illustration (this mandatory attribute is empty at the moment),

Example :

```
<NewsComponent>
  <ContentItem>
    <MediaType FormalName="Text" />
    <Format FormalName="bcNITF2.5" />
    <DataContent>
       <media media-type="image" style="leftSide">
          <media-reference mime-type="" data-location="#photo0" />
       </media>
       <p>
         text …
       </p>
    </DataContent>
  </ContentItem>
</NewsComponent>
<NewsComponent Duid="photo0">
  <ContentItem Href="….jpg">
    <MediaType FormalName="Photo" />
  </ContentItem>
</NewsComponent>
```

Notes :

1. We hope that the media element will be simplified in NITF3.0, as its use in a NewsML environment isn't optimized at the moment.
2. The media-type vocabulary (image | audio | video | other) is different from the NewsML MediaType vocabulary in the IPTC TopicSet (Text | Graphic | Photo | Audio | Video |Animation), used by AFP. We hope that NITF3.0 will follow the new IPTC vocabulary.

2.4 AFP Controlled vocabularies

AFP did customizes some IPTC controlled vocabularies.

2.4.1 Roles

AFP publishes a list of roles in its Catalog, via the file “topicset.afp-role.xml”.

Current roles are :

- **Main** : Principal component.
- **Supporting** : Additional information that amplifies the original theme.
- **Links** : Links to related information.
- **Thumbnail** : A news-component substitute, smaller than the original, used for convenience.
- **Quicklook** : A news-component substitute, smaller than the original, used for convenience ; bigger than a thumbnail.
- **Preview** : A partial (possibly random) rendition of a news-component, such as a segment of video or audio ; if an image, rendered full screen.
- **Caption** : Text providing information about the related content.
- **Abstract** : A summary or synopsis featuring the most important points of a news-component.
- **Comments** : Comments on the NewsItem.
- **Description** : A description of the NewsItem.

2.4.2 Formats

AFP publishes a list of formats in its Catalog, via the file “topicset.afp-format.xml”.

Current formats are :

- **NITF2.5** : News Industry Text Format version 2.5.
- **bcNITF2.5** : body content portion of NITF 2.5.
- **NITF3.0** : News Industry Text Format version 3.0.
- **bcNITF3.0** : body content portion of NITF 3.0.
- **XHTML** : Extensible HyperText Markup Language
- **JPEG** : Join Photographic Experts Group.
- **GIF** : Graphical Interchange Format.
- **PNG** : Portable Network Graphics.
- **EPS** : Encapsulated Postscript File
- **PDF** : Portable Document Format (Adobe)

2.4.3 Properties

AFP publishes a list of properties in its Catalog, via the file “topicset.afp-format.xml”.

Current properties are :

- **Width** : Width of an image in pixels.
- **Height** : Height of an image in pixels .
- **Country** : Originating country of an information.
- **Area** : Originating region or state of an information.
- **City** : Originating city of an information.
- **Sublocation** : Originating sublocation of an information.
- **Keyword** : Keyword related to the information.