# Importance of metadata in the multimedia news delivery

## Introduction: the problem

Every real-life event is in its nature a multi media event. The reason for that is obvious: we use different media to capture portions of our senses because we lack the ability to present an event the same way, with all the impressions a real-life spectator could have experienced.

News tries to depict an event to its recipients as close to reality as possible. However, recipients differ very much among themselves, first in their interest of various aspects of an event, and second in their ability to receive and use different media formats.

Therefore, news providers are producing information about particular event in different media formats to let their customers choose what portion of it is most suitable for their needs. That raises the quality of service for the multimedia-enabled customers giving them richer experience of the event. For less demanding, or for customers with limited bandwidth, it creates an opportunity to choose particular format that best fits in their environment, without limiting anyone to access the information.

Internet, being inexpensive and a globally present communication channel, gave customers access to much more information sources than they used to have in the traditional environment. At the same time, by easing access to potential customers, fast spreading of Internet significantly increased the number of the information sources.

As a result, the number of information every recipient received or had access to grew rapidly, making it impossible to process all of them. For things to be even worse, it made quality and trustworthiness of the information lower as well. Finally, to decide whether some multimedia information was relevant for them or not, customers had to actually process the information completely (see the video or picture or hear the audio)

To solve the problem, various data formats introduced the concept of metadata: data that describes the actual information content, usually in a computer processable way. Metadata can describe many characteristics of the information content: its structure, format, subject matter, source, legal or other rights pertaining to it, relations to other information or contain any other data somehow relevant to it.

Metadata enables recipients to automatically classify, search, filter, relate or process in any other way information they have received without needing to actually process the content. Doing so, they help humans to limit the number of information he/she has to deal with, ideally only to the significant ones.

However, there is the question of relevance of metadata to the data it describes. What might seem irrelevant to the author, or the person who applied metadata, might be very important for a particular recipient, and vice versa. Furthermore, relying on metadata to retrieve the information makes metadata even more important than the actual content: information with wrongly attached metadata would never reach the recipients, however valuable its content may be.

News adds one additional aspect to the problem: speed. Old news is no news any more. Consequently, news data has to be built to provide for the fast processing of its content, without loosing any important information in the process. As the data comes from different sources and in different media, it is very important to have standard metadata structure as well.

Therefore, an ideal news data format must satisfy the following criteria:

1. Be open, platform independent, widely accepted and easy transferable over the variety of communication channels
2. Provide for inclusion or referencing of the content that contains arbitrary mixtures of media types, languages and encodings
3. Have rich and flexible metadata structure, including the information on provenance and relevance of the metadata to the information content
4. Support creating relationships between the data and the management of data over time

### XML as a multimedia news delivery standard

When looking for the technology of choice for the multimedia news delivery standard, XML certainly provides an excellent option.

XML is open, platform independent standard. Recently it has become one of the most used industry buzzwords, and the number of tools and platforms that support it rises literally every day.

Besides that, it is a base for numerous other W3C standards, like namespaces, XPath, XPointer, and XSLT (stylesheet translations), which altogether creates very powerful technology. In addition, the number of XML based industry standards grow very fast as well.

XML file is a textual file, what makes it humanly readable and easy to transfer whatever protocol or technology is being used to do it. With the full unicode support, XML is transparent to languages and national character encodings, yet still supporting traditional ones.

It has extremely rich capabilities to structure and describe data. Although completely open, it contains mechanisms to enforce proper structure of the data, according to the predefined, user definable rules.

XML does not directly allow inclusion of binary data, but with minimal precautions it can carry such data, either directly or encoded. Furthermore, it has excellent capabilities to reference the data instead of including it directly. With XPath and XPointer, XML allows very precise addressing of particular piece of data inside the information content or metadata part.

XML organizes its information objects into a hierarchical tree of nodes, called elements. Each element may contain textual content, other elements, or some mixture of them. That makes XML very powerful in structuring the data, because the structure of the XML document corresponds to the structure of the data itself, and document hierarchy creates relations between the data at the various levels of hierarchy.

In addition, each XML element may contain one or more attributes: unstructured name-value pairs that describe some properties of the element. Although

attributes may carry only textual content, there are some attribute types that provide for special functions, like unique element identification (ID), references to other elements (IDREF and IDREFS), inclusion of foreign content (ENTITY) and so on.

### Attaching metadata to the information content

Hierarchical organization of the document makes it very easy to attach metadata to the information content. Furthermore, it allows attaching metadata only to the relevant part of it instead of only to the whole document (as in most of the traditional, particular media oriented standards).

If metadata consists of plain, unstructured data, it may simply be put as attributes of the elements that contain the information content. The attribute name identifies which property of the information content particular metadata describes. Using attributes offers the advantage of specifying the legal values particular data may have.

```
<ContentElement type="heading" language="en">Heading</ContentElement>
```

If the metadata has its own structure, it could be mapped into elements and connected with the information content by simply placing it into the same tree branch as the data it relates to. Elements could be used to place non-structured metadata as well, offering the possibility of introducing the structure in the future, or simply for design purposes.

```
<Content>
      <ElementType>heading</ElementType>
      <Language variant="en-us">English</Language>
      <ContentElement>Heading</ContentElement>
</Content>
```

Besides the document hierarchy, various other techniques could be used to attach the metadata to the information it belongs to.

One of the possibilities is to use ID / IDREF type attributes. Attribute of the type ID must be unique inside the whole XML document. IDREF type attribute must contain value of an ID attribute somewhere in the document, and the IDREFS one provides for referring to more than one value.

Uniquely naming logical parts of information content with ID type attributes is always a wise practice, for it enables their addressing, even from other documents.

```
<ContentElement id="CE001">Heading</ContentElement>
<Metadata idref="CE001" type="heading" language="en" />
```

The advantage of this approach is that it does not require the parser to support nothing more than XML 1.0. That widens the option of tools that could be used, because some of them do not support other XML based standards. The disadvantage is that it requires a generation of unique names and changing the original content. It diminishes the readability of the document as well.

Another technique being used quite often is XPath / XPointer. XPath is a standard aimed to enable addressing different parts of an XML document. It operates on the tree model of the document and has operators to traverse the tree and select nodes based on various conditions. XPointer enables pointing to various parts of XML document from outside, usually through hyperlinks. It uses most of the capabilities of XPath and adds to it only a few additional operators.

Metadata elements could have attributes or child elements containing XPath or XPointer expressions that select the information content they describe. It is very flexible technique, because it does not require any modification of the content itself and even allows metadata to be placed in the separate document, completely outside the original content.

```
<ContentElement>Heading</ContentElement>
<Metadata ref="../ContentElement" type="heading" language="en" />
```

## NewsML: an XML based news interchange standard

Considering all these advantages, it is no wonder that XML has been chosen as a base for the next generation news exchange standard.

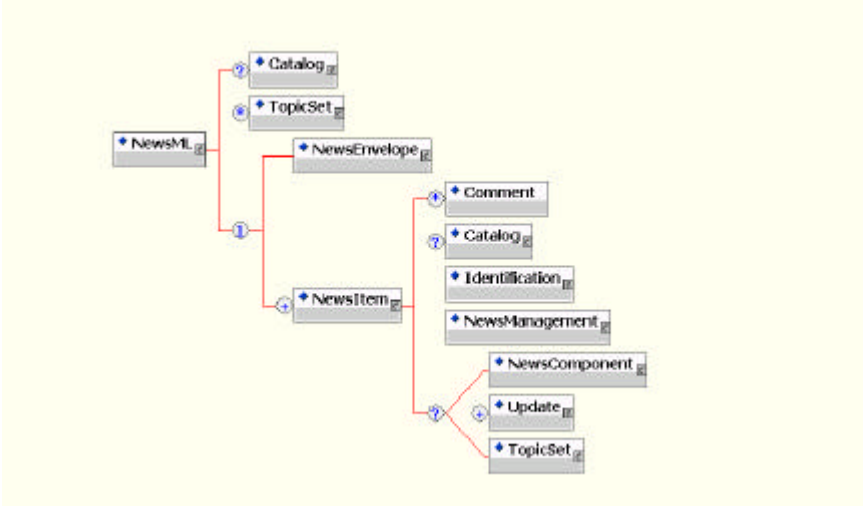
For quite a long time the IPTC (International Press Telecommunications Council, the international news industry association that creates news exchange standards) had requirements from its members for the standard that would enable delivery of multimedia content regardless of media format, language, or transport layer. After a year of very intense work, the new standard, called NewsML was adopted in September 2000 in Amsterdam.

As the specification says: "NewsML is a compact, extensible and flexible framework for news, based on XML and other appropriate standards and specifications. It supports the representation of electronic news items, collections of such items, the relationships between them, and their associated metadata. It allows for the provision of multiple representations of the same information, and handles arbitrary mixtures of media types, formats, languages and encodings. It supports all stages of the news lifecycle and allows the evolution of news items over time. Though media independent, NewsML provides special mechanisms for handling text. It allows the provenance of both metadata and news content to be asserted."

A NewsML document is an XML document, which must be valid with respect to the NewsML Document Type Definition (DTD).

### Structure of the NewsML document

The NewsML element is the root element of the complete NewsML document. At least, it must contain a NewsEnvelope and one or more NewsItem elements.



The NewsEnvelope element contains information about how the NewsML document is being used within a business workflow or contractual relationship between news provider and receiver. It identifies only the transmission process, not the content carried by the transmission.

A NewsItem element represents the news: a managed set of information representing a point of view, at a given time, on some event or events. It has globally unique identifier and NewsManagement element that provides for manageability. In addition, it may contain NewsComponent representing the actual content, Update element(s) that modify a previous revision of the same NewsItem, or a TopicSet with the controlled vocabulary (explained later).

The NewsComponent is a container for news objects. News often brings together multiple data objects, for example, a text story and a photograph with the caption. Further, it is often necessary to bring together multiple complete stories and handle them as a coherent collection.

NewsComponent serves to handle this complexity by specifying structural relationships between news objects, identifying the role of news objects in relation to one another and ascribing metadata to them. Therefore, NewsComponent may contain other NewsComponents and NewsItems or point them through URN or URL.

Finally, NewsComponent may contain renderable content in the ContentItem elements. A ContentItem must carry some raw data, contained inline within a DataContent element, or a pointer to it. The DataContent element may be wrapped in one or more Encoding elements, indicating how it has been encoded. If a pointer is used, NewsML document is to be interpreted exactly the same way as if the data were included inline.

### Metadata assignment

Metadata is assigned to the various levels of NewsML hierarchy, attributing metadata to the content it most closely describes.

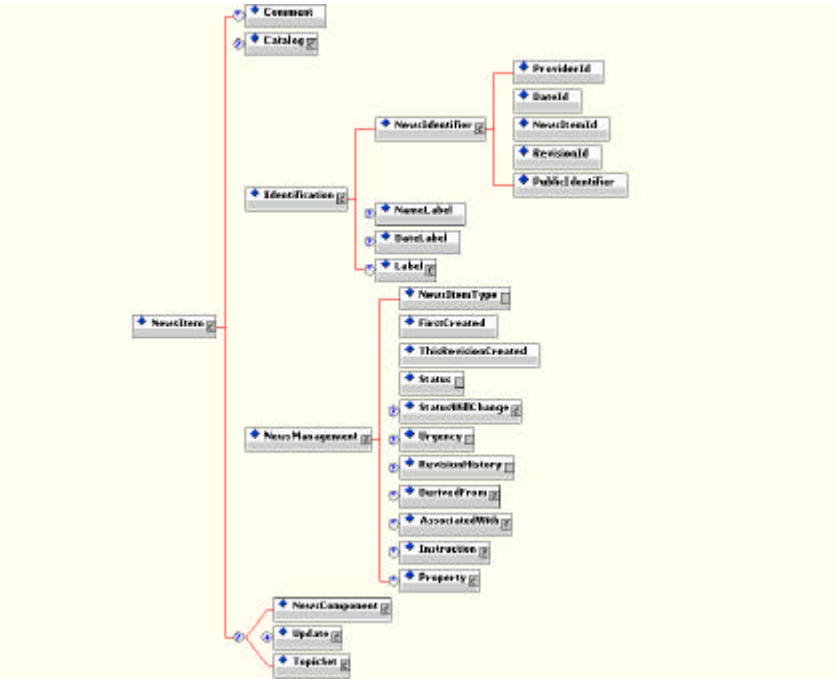
Most of the elements that carry metadata may as well contain the attributes that specify who, when and why assigned it and information about the importance of the particular metadata and the confidence in the assignment. That enables recipients to implicitly judge the quality of the metadata and act accordingly.

NewsEnvelope directs transmission of the news content. It must contain date and time of the transmission and may contain the service and product that the transmitted information belongs to, information about the news provider and receiver and the transmission ID.

NewsItem contains Identification and NewsManagement metadata elements.

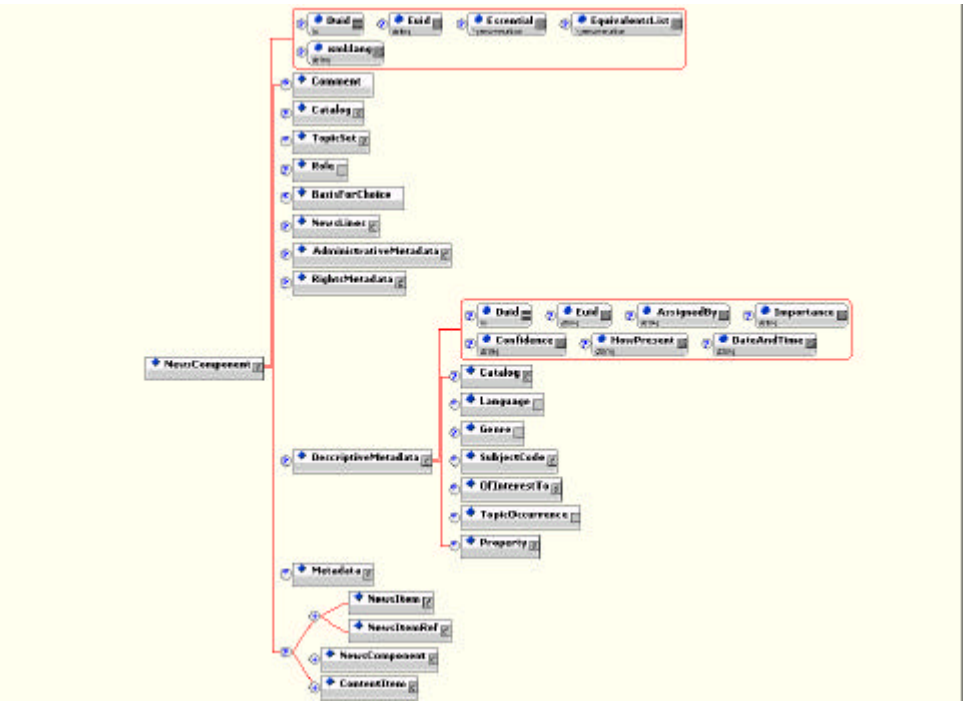
Identification element contains URN, a public identifier for the NewsItem in the sense defined by the XML 1.0 Specification, enabling the NewsItem to be referenced unambiguously by pointers from other XML elements or resources. Beside of formal identification of the NewsItem, Identification element may contain various Label elements. Their sole purpose is to provide a convenient way for human users to identify a particular NewsItem in informal exchanges or as a part of a user interface.

NewsManagement provides information about NewsItem's type, history and status, as well as its relationships to other NewsItems, and any special instructions to be applied to it or additional properties that it may have.




Perhaps the most important one is the Status, for it specifies the state of NewsItem's lifecycle and its current usability. For example, except of being usable, NewsItem may as well be embargoed, withheld, cancelled or killed. The status of a particular NewsItem may change over time with the evolvement of the event it represents.

NewsComponent carries most of the metadata connected with the information content.




First, it specifies role of a particular news object representation (text, picture, audio, etc) and its relations to the other ones contained in the same NewsItem.

Second, it associates various types of metadata recognized by the NewsML with the information content.

AdministrativeMetadata carries information about the provenance of a NewsComponent. The RightsMetadata element specifies the rights pertaining to it and any relevant usage rights that have been granted by the copyright holder to other parties.

The DescriptiveMetadata describes the content of a NewsComponent. Its subelements indicate the NewsComponent's genre, subject, target audience, any languages that it may use, and give information about any people, places, organizations, countries or other real-world things alluded to in the piece, or to whom the piece is relevant in any way.

The generic Metadata element provides for the metadata types not recognized by the NewsML, adding to the extensibility of the standard.

Finally, NewsComponents may include NewsLines element, whose purpose is to provide a human readable (publishable) representation of certain aspects of the metadata. It includes Headline, ByLine, DateLine, CreditLine, RightsLine, CreditLine, KeywordsLine etc. They can be used just for publishing, or to increase natural language search capabilities.

In the end, ContentItem, beside of carrying the data, describes its physical characteristics. It may specify a MimeType, Format, MediaType, Notation, or some other properties, such as size, in more general Characteristics element. Their purpose is to help determine the system requirements needed to handle the data before the content has been interpreted.

### Design techniques

To make the standard implementable as soon as possible, it was decided that it should use the most basic XML way to accomplish some task. For some standard or specification to be used, it must have satisfied the three important rules:

- Formal ratification
- Cross platform tool support
- Public awareness and understanding

Therefore, NewsML elements are referenced using only fragment identifiers: Duid attribute preceded by the "#" sign. The reason for that was to require the usage of only a XML parser to resolve them.

Where there is a need for external references, URN's are always either accompanied with the URL's that specify the location or replaceable by them. That helped usually complicated URN resolution and eliminated a need for external catalogs or other non-standard tools.

XPath was used only where it was necessary to select various parts of subtrees, what would be impractical to do with Duids. XPath is stable, ratified and widespread standard and its inclusion did not jeopardize ease of NewsML implementation.

There was huge amount of discussion whether to use Namespaces and RDF.

The NewsML does not directly use Namespaces, although it does not ban their usage either. The reason is the fact that, for the time being, use of Namespaces makes validation impossible, and the conclusion was that validation was too important to be excluded by design.

RDF, beside of using namespaces, was felt to be too complex to implement and was lacking tools support and public understanding. Since RDF today is mainly limited to Dublin Core metadata, it seems like it was a good decision.

Metadata is attached to the content using the document structure. That enabled easy processing and assigning of the metadata only to the content it described.

To provide for closed coding schemes, or controlled vocabularies, NewsML has chosen an implementation based on ISO 13250 Topic Maps standard. Since XML Topic Map standard (XTM) was in very early phase, it was not mentioned directly, although NewsML implementation was consistent with it.

These design principles resulted with the first free Java toolkit being published less than a year after the standard and with a few already running implementations by major world agencies.

## NewsML Features

### Unique identification

A NewsML feature that most of the other features rely on is the capability of uniquely addressing every NewsML document and element. The feature works on two levels: on the XML document and on the element level.

Every news (NewsItem element) has a globally unique identifier that identifies it as it moves through its lifecycle. The identifier always remains the same, no matter what happens with the news: whether it is being updated, corrected, canceled or expired.

News identifier consists of four parts: ProviderId, DateId, NewsItemId and RevisionId. All four parts are additionally concatenated together in the PublicIdentifier element in the form of an URN, as described in the RFC 3085.

```
urn:newsml:{ProviderId}:{DateId}:{NewsItemId}:{RevisionId}{RevisionId@Update}
```

On the element level, almost every NewsML element may contain two attributes: Duid and Euid.

Duid has to be unique in the whole NewsML document. Combined with the URN, it allows globally unique addressing of that element.

Euid has to be unique among elements of the same type that have the same parent. Use of the Euid makes it possible to identify any NewsML element within the context of the local branch of the NewsML document tree. This enables using part of the document in the new combinations that would break the uniqueness of the Duid's while still being able to retain the identity of each element.

If Euids are maintained at every level, XPointer expressions may be used to identify the elements based on the content they are contained within. It also eases reuse of the individual NewsML components, either by inclusion or by reference.

### Versioning and correction management

Since news report about live events, it is very important to follow how the events evolve over time. In case of an error or misrepresentation, it is even more important to notify users and stop the spreading of false information as soon as possible.

NewsML offers several ways to accomplish this task.

The easiest, but also the least powerful, is to simply send complete content each time something changes and reference the previous version via DerivedFrom subelement of the NewsManagement metadata.

The biggest disadvantage of this approach is that, in case of an error, it does not offer any error notification to the receiving system, so the correction management is left completely to the human recipient. However, this is the traditional way news business operates.

The more *NewsML-way* to send new revisions is to send complete content in each revision, with the same NewsIdentifier as the previous one, only changing the RevisionID part.

Beside of updating users with the new content, it would also correct the errors in the previous revisions, since the receiving system would always supply the users with the newest revision it has. It is also possible to invalidate some or all of the previous revisions using the Instruction sublement of the NewsManagement metadata, thereby preventing further spreading of the eventual error.

The most powerful, but also the method that puts the highest burden on the receiving system, is to send only incremental updates.

Content updates are performed by sending the NewsItem with only Update elements that contain InsertBefore, InsertAfter, Replace or Delete subelements. All those subelements, beside of eventually carrying a new content, have an DuidRef attribute pointing to the element in the previous revision that insertion, replacement or deletion is to be performed on.

It is up to the receiving system to apply the Update instructions to the NewsItem and generate a new copy of it, exactly as it has been sent in its entirety. That shows the power of assigning Duids to each NewsML element.

However, in distributed environments, special care should be taken when using the update mechanism, since there is no guarantee that the receiving system has successfully received or is still keeping the previous revision of a NewsItem that is being updated. Therefore, there has to be a way for receiving system to request the copy from the provider, which is not possible in broadcast only environments.

Because of that, especially in the text-based services, it is sometimes more convenient to simply send full updates, especially since it adds to the system robustness too.

### Content identification and controlled vocabularies

For some content to be searchable, it has to be marked up consistently. Since in the newsroom environment content is often multilingual and search mechanism has to be fast and efficient, simple text search cannot be used as a primary retrieval tool. So, there is a need for closed and, if possible, language neutral coding schemes that identify various aspects of the content.

NewsML offers such a tool through the *controlled vocabularies* mechanism (TopicSet elements). Its goal is to enable connection of various coding schemes with the real-world things, or ideas, they code.

NewsML TopicSets consist of the Topic elements. Each idea, expressed through the Topic element, contains one or more Description subelements that identify which individual thing it is, and one or more FormalName subelements with the codes. The FormalName element may as well have a Scheme attribute to indicate that it belongs to a particular naming scheme.

It is an error for there to exist two Topics in the same TopicSet that have the same FormalName with the same Scheme attribute. It is therefore possible to use a TopicSet as a controlled vocabulary, to ascertain the meaning of any given formal name.

Additional TopicSets may be included by reference and merged with the original one. The merging of topics need not be physically performed by the system, but the meaning of the data is the same as of the merging were actually performed.

Topics are referenced in the content through their FormalName and eventually a Scheme. TopicSet that particular Topic belongs to is resolved by pointing it in the Vocabulary attribute of the element that references it or by using a Catalog element.

Any of the main structural NewsML elements may have a Catalog with Resource and/or TopicUse elements. Catalog serves to specify resources some NewsML element use, default vocabularies for certain elements and to notify usage of specific Topics somewhere in the content.

Each Resource element identifies (through an URN and/or one or more URL's) an external resource, usually a TopicSet, that is being used within the content of the element that contains a Catalog. Resource may also contain an XPath pattern that selects elements or attributes that referenced TopicSets act as a default vocabulary for.

Catalog may also announce that a certain Topic is being used within the content of the element that contains the Catalog. That is especially practical with the non-XML content or the content that is being referenced from the NewsML, since it is not always possible to attach metadata directly to the content itself.

Although NewsML standard does not explicitly define any controlled vocabularies, there is a referential set being published together with the standard to increase the interoperability of various implementations.

One of the most important is the IPTC Subject Coding Scheme, a comprehensive coding system with more that 200 codes on three levels that cover subject matter, genre and type of the content. It is being used consistently in all the IPTC standards and has proven useful even outside the news environment.

### Choosing the right content

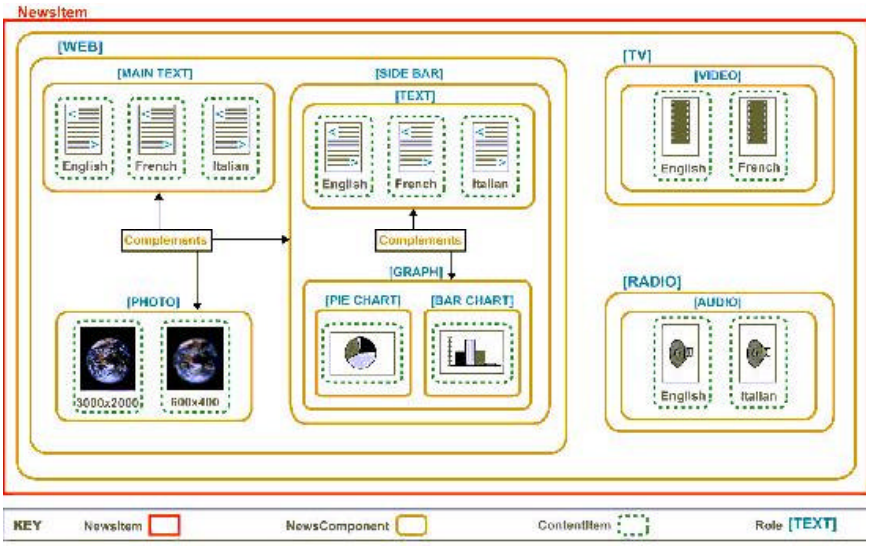
One of the most important NewsML features is its ability to carry and group together various media encoding of the same event. For example, news may consist of a picture, audio file, and the text part describing the background of the event.

Recipients must be able to discover the role of each part in the whole and to decide which one of them best suits their needs.

To provide for that, besides containing other NewsComponents and NewsItems, each NewsComponent may have several attributes and subelements that help to make the decision.

Essential attribute indicates whether the provider considers this NewsComponent essential to the meaning of the whole. Role subelement specifies the role this NewsComponent plays, taking the values from the controlled vocabulary published by the provider. EquivalentsList states whether or not the other news objects contained inside this NewsComponent are equivalent to one another in content and/or meaning, and BasisForChoice contains XPath pattern that suggests the data on the basis of which a choice between the equivalent items can be made.

The following figure illustrates the complexity that such a structure enables. It shows a single NewsItem comprising three NewsComponents that tell the same story for WEB, for TV and for RADIO. Though, it is only a logical view. Each of the ContentItems may reference the content instead and supply the receiving system only with the metadata that would help it to choose among them and save the bandwidth.




NewsML can carry other XML-based formats as content. They can be included either via Namespace resolution, by declaring their DTDs in the internal subset, through the entity declarations, or simply as CDATA content. Usage of CDATA is discouraged in the implementation guidelines, since the XML parser does not interpret its content. Preference of the other three options depends on the Namespace support and the complexity of the content markup.

### Conclusion: the broader picture

Since NewsML is only an envelope format, it does not specify any transport protocol or how to encode or mark up the content.

Here is an outline of the currently available technologies for those two parts:

| Transport | Envelope | Content |
|---|---|---|
| ICE<br>HTTP<br>FTP<br>SMTP<br>SOAP | NewsML | NITF<br>Existing multimedia formats<br>Industry specific XML standards |

The Information and Context Exchange (ICE) is an XML-over-HTTP transport protocol that consists of the request/reply exchanges. It defines both the transport and business rules aspects of content syndication, like content offerings, delivery

schedules, subscriptions activation etc. It is the technology of choice for some of the largest content providers today, and because it is XML based, it fits perfectly with the NewsML concept.

NewsML recommends NITF for the text content format since NITF is an XML specification especially designed for the news content markup. Nevertheless, it can carry content in any other format, XML based or not, regardless of its origin or type.

Although designed for the news industry, NewsML is a general-purpose envelope format. Being an XML file, it is very easily transferred over various networks and communication protocols. XML base and its design constraints gave it a wide option of tools as well.

NewsML itself is not efficient for carrying binary data. Nevertheless, it's rich repertoire of content models and powerful linking capabilities make it good choice for multimedia content too, especially with the constant improvements in the network bandwidth and compression techniques.

It has very powerful and flexible metadata structure. Besides the existing metadata, controlled vocabularies mechanism is able to adopt different coding schemes, and generic metadata elements enable invention of the new metadata types. It's built-in support for the document lifecycle makes it suitable also for the in-house and public documentation purposes.

Considering all that, it is no wander that NewsML has found wide adoption among the largest news providers, despite of being brought only a year ago. It has gained a lot of interest in the wider arena as well.

Due to its openness and flexibility, NewsML is an excellent option for a general-purpose content wrapper. Its metadata, independent of the content type, provides for most of the features that the Internet content delivery needs today, and strong support from the major content providers guarantees it's further development.

It is the metadata that gives the meaning to the content, enables its automatic processing and improves its usefulness. Only a metadata-rich content may fulfill the needs of the time and reach the users on time. Since the content and its relationships are the essence of the Internet we know today, it wouldn't be wrong to say - metadata runs the Internet.

References:


1. IPTC, NewsML 1.0 Functional Specification, http://www.iptc.org/site/NewsML/NewsMLSpec.htm, 2000
2. W3C, Extensible Markup Language 1.0, http://www.w3.org/TR/REC-xml, 1998
3. W3C, XML Pointer Language (XPointer), http://www.w3.org/TR/xptr, 1998
4. W3C, XML Path Language (XPath), http://www.w3.org/TR/xpath, 1999
5. Group of authors, Professional XML, Wrox Press Ltd., 2000
6. IDEAlliance, The Information and Content Exchange (ICE) Format and Protocol, http://www.w3.org/TR/NOTE-ice, 2000
7. IPTC, News Industry Text Format, http://www.nitf.org/dtd.htm, 2000
8. IPTC, IPTC Subject Reference System, http://www.iptc.org/site/subject-codes/brochuresrs.html, 2001