I P T C

NEWS MANAGEMENT

A GUIDE TO NEWS MANAGEMENT WITH NEWSML

Version 0.9 DRAFT

Prepared by Stuart Myles

2 July 2002

NEWS MANAGEMENT

A GUIDE TO NEWS MANAGEMENT WITH NEWSML

INTRODUCTION

This is a guide to using NewsML to manage news items.

RELATED DOCUMENTS

1. NewsML Functional Specification
2. NewsML Version 1.01 DTD
3. http://www.newsml.org
4. http://www.iptc.org
5. http://www.iptc.org/site/NewsML/NewsMLTopicSets.zip for default Topic Sets.
6. NewsML News Agency Guidelines
7. STA9707 A Method for News Management using the ARM of IIM Version 3.0

DOCUMENT HISTORY

0.9 DRAFT – 2 JULY 2002 - Corrected based on comments from members. Added an appendix to explain the use of the Status TopicSet by news agencies.

DRAFT – 19 JUNE 2002 - First version circulated to IPTC membership for review.

WHAT IS NEWS MANAGEMENT?

Often, news providers need to modify a news object, which they have previously sent to a customer. For example, they may correct a headline, expand upon the body of a story or delete a piece of news altogether. This process of updating, deleting and modifying is known as “news management”. Different news providers may have different news management policies. IPTC’s NewsML standard provides sophisticated means for providers and their customers to implement a variety of procedures.

This document is a guide to the parts of NewsML that relate to news management. It is “non normative” – it does not change anything in the NewsML functional specification or related documents. The goal for this document is to illustrate how to implement news management rules using the NewsML standard.

WHAT PARTS OF THE NEWSML STANDARD RELATE TO NEWS MANAGEMENT?

Various parts of NewsML are relevant to news management. In order to modify a previously transmitted news object, a news provider needs to be able to refer to it uniquely. NewsML supports several identification schemes. There are also elements that provide information relevant to the management of a news item. In particular, NewsML supports elements that can be used to provide information about a news item's type, status, revision history and relationships to other news items. There are ways to automatically update a previous news item, along with a facility for news providers to pass arbitrary instructions to their customers. Below, these various capabilities are explored in detail.

KEYS

NewsML provides several ways for providers to identify news objects that have previously been sent to a customer. The NewsItem is the prime unit of news management in NewsML; it has a sophisticated means of formal identification. In addition, NewsML allows providers to supply keys to identify other news objects at the element level.

FORMAL INDENTIFICATION OF A NEWSITEM VIA THE <NEWSIDENTIFIER> ELEMENT

It must be possible to identify a NewsItem as it moves through the business workflow, and is transferred from place to place and from system to system. NewsML therefore requires NewsItems to have a globally unique identifier in the form of a NewsIdentifier element.

The NewsIdentifier has four component subelements – ProviderId, DateId, NewsItemId and RevisionId – and a PublicIdentifier, which concatenates all four components in a single string. The NewsIdentifier provides a globally unique identifier for a NewsItem. Providers must therefore ensure that no two NewsItems carry the same ProviderId, DateId, NewsItemId and RevisionId. If a NewsItem is re-created after a change in content, however slight, a new RevisionId should be allocated to the new version.

ProviderId

The ProviderId is a unique identifier for the news provider that produced the NewsItem. The content of the ProviderId element must be an Internet domain name that is owned by the provider at the date identified by the DateId element, or the name for the provider drawn from a controlled vocabulary identified by a URN specified in the Vocabulary attribute. This will ensure that the identity of the provider can be inferred unambiguously from the full NewsIdentifier.

DateId

The DateId is a date in ISO 8601 Basic Format (CCYYMMDD), where CCYY is a four-digit year number, MM is a two-digit month number and DD is a two-digit day number. Note that because the DateId is part of the formal identification of the NewsItem, it must remain the same through successive revisions of the same NewsItem. It does not represent the date of release of the current revision.

NewsItemId

The NewsItemId is an identifier for the NewsItem. The combination of the NewsItemId and the DateId must be unique among NewsItems that emanate from the same provider. Within these constraints, the NewsItemId can take any form the provider wishes. It may take the form of a name for the NewsItem that will be meaningful to humans, but this is not a requirement.

The provider may optionally relate the values of the NewsItemId to a controlled vocabulary, which is invoked by the Vocabulary attribute. The value of the Vocabulary attribute may be an http URL or a NewsML URN, or the # character followed by the value of the Duid attribute may be an attribute of a TopicSet in the current document. The Scheme attribute, if present, serves to distinguish which of possibly multiple naming schemes in the controlled vocabulary in the one that governs the NewsItemId.

RevisionId

The RevisionId is an integer indicating which Revision of a given NewsItem this is. Any positive integer may be used, but it must always be the case that of two instances of a NewsItem that have the same ProviderId, DateId and NewsItemId, the one whose RevisionId has the larger value must be the more recent revision. A RevisionId of 0 is not permitted. The PreviousRevision attribute must be present, and its value must be equal to the content of the RevisionId element of the NewsItem's previous revision, if there is one, and 0 if the NewsItem has no previous revision. If the NewsItem contains an Update element or elements, then the Update attribute must be set to U. If the NewsItem consists only of a replacement set of NewsManagement data, then the Update attribute must be set to A. If neither of these is the case, then the Update attribute must be set to N.

PublicIdentifier

The PublicIdentifier element provides a public identifier (in the sense defined by the XML 1.0 Specification) for a NewsItem. This is the NewsML URN, and must be constructed as follows:

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

where {x} means "the content of the x subelement of the NewsIdentifier" and {x@y} means "the value of the y attribute of the x subelement of the NewsIdentifier", with the exception that if the Update attribute of the RevisionId element has its default value of N, it is omitted from the URN.

Note that the set of characters that can be included within a URN is limited. The allowed characters are specified by the Internet Engineering Task Force (IETF) in its Request For Comments (RFC) number 2141. This document is available at http://www.ietf.org/rfc/rfc2141.txt. Any character that is not within the permitted URN character set must be represented as a % character followed by the sequence of one to six bytes of its UTF-8 encoding, represented in their hexadecimal form. Thus, for example, the space character in a URN would appear as %20, and the % character itself would appear as %25. This mechanism does not cater for all Unicode or UTF-16 characters. Therefore, it is important not to include characters in a NewsItemId that cannot be encoded in UTF-8.

Note that the existence of this URN enables the NewsItem to be referenced unambiguously by pointers from other XML elements or resources. Within such pointers, if the RevisionId, its preceding : character and its following Update qualifier are omitted, then the pointer designates the most recent revision at the time it is resolved.

INFORMAL IDENTIFICATION OF A NEWSITEM

In addition to the formal NewsItem identification mechanisms, NewsML provides elements that are designed for use by humans. These informal identification mechanisms are not meant to be used within automated news management. Therefore, they are not discussed further here. Please refer to the NewsML Functional Specification for details.

IDENTIFYING ELEMENTS VIA THE DUID AND EUID ATTRIBUTES

Every element in a NewsML document - other than NewsIdentifier and its subelements - may optionally have a Duid (document-unique identifier) and/or an Euid (element-unique identifier) attribute, whose purpose is to enable pointers elsewhere in the document, or in other NewsML or XML documents, to refer to it. The use of identifier attributes gives global identification to the document.

The "Document-unique" identifier

The Duid is a "Document-unique Identifier". It must satisfy the rules for XML ID attributes: it must only contain name characters, and it must start with a name-start character (not a digit). Its value must be unique within any NewsML document. Every NewsML element type has Duid as an optional attribute. Combined with the Identifier element, providing a value for the Duid of any element in a NewsML document makes the element globally identifiable. The Identifier element gives global identification to the document, and the Duid provides local identification for the element within the document.

The "Element-unique" identifier

The Euid is an "Element-unique Identifier". Its value must be unique among elements of the same element-type and having the same parent element. Use of Euid attribute makes it possible to identify any NewsML element within the context of its local branch of the NewsML document tree. This makes it possible to copy, or include by reference, subtrees into new combinations in ways that would break the uniqueness of Duids (thereby forcing new Duids to be allocated), but still being able to retain the identity of each element. If Euids are maintained at every level, it is possible to identify, for example "The ContentItem whose Euid is abc within the NewsComponent whose Euid is def". Such identification patterns would be preserved even after "pruning and grafting" of subtrees.

THE <NEWSMANAGEMENT> ELEMENT

The NewsManagement element provides information relevant to the management of a NewsItem: information about a NewsItem's type, history and status, as well as its relationship to other NewsItems, and any special instructions to be applied to it or additional properties it might have.

NEWSITEMTYPE

The NewsItemType element contains an indication of the type of a NewsItem. The value of the FormalName attribute is a formal name for the news-item type. Its meaning and permitted values are determined by the controlled vocabulary identified by the Vocabulary and Scheme attributes.

FIRSTCREATED

The date and, optionally, time at which a NewsItem was first created, expressed in ISO 8601 Basic Format.

THISREVISIONCREATED

The date and, optionally, time at which the current revision of a NewsItem was created, expressed in ISO 8601 Basic Format.

STATUS

This element indicates the current status of a NewsItem. The value of the FormalName attribute is a formal name for the Status. Its meaning and permitted values are determined by a controlled vocabulary.

STATUSWILLCHANGE

Advance notification of a status change that will automatically occur at the specified date and time. For example, an item with a Status of "embargoed" might have a StatusWillChange element stating that the status will become "usable" at a specified time. This is equivalent to announcing in advance the time at which the embargo will end and the item will be released. Within StatusWillChange, the required FutureStatus element indicates the status the NewsItem will have at a specified future date. The value of the FormalName attribute is a formal name for the status. Its meaning and permitted values are determined by a controlled vocabulary. The required DateAndTime element indicates, using ISO 8601 Basic Format, the date or date and time at which the status will change.

URGENCY

An indication of the urgency of a NewsItem. The value of the FormalName attribute is a formal name for the Urgency. Its meaning and permitted values are determined by a controlled vocabulary.

REVISIONHISTORY

A pointer to a file containing the revision history of the NewsItem. The provider may choose whatever syntax and structure they like for this file.

DERIVEDFROM

A reference to an NewsItem from which this one is derived. The NewsItem attribute identifies the relevant NewsItem. Its value can be an http URL or a NewsML URN.

ASSOCIATEDWITH

A reference to a NewsItem with which this one is associated (for example, a series of articles, or collection of photos, of which it is a part). The NewsItem attribute identifies the relevant NewsItem.

Its value can be an http URL or a NewsML URN as described in the comment to PublicIdentifier. The Comment can be used to indicate the nature of the association.

INSTRUCTION AND REVISIONSTATUS

An instruction from a news provider to the recipient of a NewsItem. A special case of Instruction is an indication of the effect the current revision of a NewsItem has on the status of any previous revisions of the NewsItem that may still be on the recipient's system. In this case, it will contain one or more RevisionStatus elements. Otherwise, the value of the FormalName attribute is a formal name for the Instruction, and its meaning is determined by a controlled vocabulary.

A RevisionStatus element indicates the status that previous revisions now has as a result of the release of the current revision. The optional Revision attribute is an integer, equal to the RevisionId of the revision in question. If it is not present, then the status applies to ALL previous revisions, without exception.

THE <UPDATE> ELEMENT

The Update element is used in NewsML to indicate how to modify a previously-published NewsItem. It allows NewsComponents to be deleted or replaced; it also allows news providers to insert a NewsComponent either before or after another NewsItem. If a NewsItem is modified in these ways, the provider must use a RevisionId that is a higher number than it was previously, and the PreviousRevision attribute should be equal to the previous version’s RevisionId.

NewsManagment and Identification elements cannot be modified using the Update element; instead, a news provider must issue the NewsItem under the current revision number, with only the Identification and NewsManagement elements present. This will replace the previous Identification and NewsManagement elements in their totality.

THE <INSTRUCTION> ELEMENT

The optional and repeatable Instruction element contains an instruction from a news provider to a recipient of a NewsItem. A special case of Instruction is an indication of the effect of the current revision of a NewsItem has on the status of any previous revisions of the NewsItem that may still be in the recipient’s system. In this case, it will contain one or more RevisionStatus elements. Otherwise, the value of the FormalName attribute is a formal name for the instruction. Its meaning and permitted values are determined by the controlled vocabulary identified by the Vocabulary and Scheme attributes.

THE REVISIONSTATUS ELEMENT

The RevisionStatus element indicates the status that previous revisions now have as a result of the release of the current revision. The optional Revision attribute is an integer, equal to the RevisionId of the revision in question. If it is not present, then the status applies to all previous revisions, without exception.

HOW CAN NEWS MANAGEMENT BE IMPLEMENTED WITH NEWSML?

A news provider and its subscriber may implement news management in a number of different ways, using the facilities of NewsML. The most basic level is – essentially – no news management. For example, each time the news provider issues a set of NewsItems, the subscriber discards the previous set and replaces it with the new one. A more sophisticated scenario involves the news subscriber maintaining an archive of previously-published NewsItems. The news provider may issue new NewsItems, replace old NewsItems or delete old NewsItems in their entirety. The most sophisticated level of news management allows inserting, replacing and deleting parts of NewsItems – as well as the facilities provided at the "lower" levels of news management.

NO ARCHIVE – REPLACING ONE NEWSML DOCUMENT WITH ANOTHER

In this scenario, the news subscriber does not maintain a news archive. The news provider publishes a collection of NewsItems (e.g. a NewsML document with top ten headlines and news articles). Each time the news provider publishes the NewsML document, the news subscriber discards the previous document and replaces it with the new one.

PROCESSING REQUIREMENTS

The news subscriber does not need to track the public identifiers for NewsItems – there is no need to try to match up subsequent NewsItems with previous NewsItems. The subscriber *must not* archive NewsItems, since the provider will not provide updates or deletes. The provider is also not required to use consistent NewsItemIds.

The news provider does not need to "version" NewsItems per se. To publish a new version of a NewsItem, the news provider simply sends the entire new NewsItem; the provider does not need to indicate that this NewsItem is an update. To "delete" a NewsItem, the provider issues a new collection of NewsItems that omits the one that needs to be deleted.

"WRITE THROUGH" – REPLACING AND DELETING COMPLETE NEWSITEMS

In this scenario, the news subscriber maintains a news archive, i.e. a set of previously-published NewsItems. The news provider may issue subsequent NewsItems that replace or delete entries in the subscriber's archive. The subscriber must modify the news archive to reflect the changes specified by the provider.

PROCESSING REQUIREMENTS

The news subscriber must replace and delete NewsItems in its archive, using the NewsItemId to identify which NewsItem to modify. The entire NewsItem must be deleted or replaced. The subscriber does not need to track element identifiers (euids or duids).

The news provider must track the NewsItemId and RevisionId associated with each NewsItem that it publishes. It does not need to track (or provide) element identifiers – euids and duids – in this scenario.

When publishing an update to a NewsItem, the provider must use a RevisionId that is a higher number than it was previously, and the PreviousRevision attribute should be equal to the previous version's RevisionId. The entire NewsItem is published, incorporating all changes that may have

been made, and the value of the Update attribute of the RevisionId element is set to "N". The subscriber must replace the entire NewsItem in its archive with the new revision. If the PreviousRevision is set to 0 then this is a new NewsItem to be added to the archive.

When deleting a NewsItem, the provider must use a NewsItem that will only contain the complete Identification and NewsManagement elements, and nothing else. The content of the RevisionId element should be identical to that of the original NewsItem, with the value of its Update attribute set to "A". The Status element will indicate that the current status of the NewsItem is cancelled. (The news subscriber and the news provider will have previously agreed upon the controlled vocabulary to be used to indicate the status of a NewsItem. The IPTC provides a standard TopicSet for Status, which should be used by news agencies). The news subscriber must delete the entire NewsItem from its archive.

UPDATING, DELETING AND REPLACING PARTS OF NEWSITEMS

In this scenario, the news subscriber maintains an archive of previously published NewsItems. The news provider may issue subsequent NewsItems that update, delete or replace parts of NewsItems – at the element level – or that may replace NewsItems in their entirety. The subscriber must, therefore, track NewsItemIds for entire NewsItems and duids and euids for their constituent elements.

PROCESSING REQUIREMENTS

Upon receipt of a NewsItem from the news provider, the news subscriber must check the Update attribute of the RevisionId. If it is set to U, then the subscriber must process the Update element or elements contained in the NewsItem, updating the NewsItem in its archive. If this it set to A, then the subscriber must replace the NewsManagement data of the archived NewsItem. If it is set to N and the PreviousRevision is 0, then this is a new NewsItem, which may be added to the archive. If it is set to N and the PreviousRevision is greater than 0, then this NewsItem replaces the entire previous revision in the archive.

To modify one or more subelements of NewsManagement and/or Identification, without any change to any other parts of the NewsItem, then the content of the RevisionId element should be identical to the original one, the value of its Update attribute should be set to "A", and the NewsItem should contain the complete Identification and NewsManagement elements, incorporating any changes, and nothing else.

If any other part of the NewsItem is modified in any way, the provider must use a RevisionId that is a higher number than it was previously, and the PreviousRevision attribute should be equal to the previous version's RevisionId. There are two choices:

- The entire NewsItem is published, incorporating all changes that may have been made, and the value of the Update attribute of the RevisionId element is set to "N".

- The NewsComponent subelement of the NewsItem is not included in the new document, but in its place, one or more Update elements are provided, indicating the modifications that have been made, and the value of the Update attribute of the RevisionId element is set to U.

The Update element indicates a modification to an existing NewsItem. This can be an insertion, replacement or deletion. Note that the Update element cannot be used to modify the

NewsManagement or Identification element or any of their descendants. Modification to these parts of the NewsItem can be made by issuing the NewsItem under the current revision number, with only the Identification and NewsManagement elements in their totality. An Update element contains any number of subelements of the following kinds:

- Delete
- Replace
- InsertBefore
- InsertAfter

It is the responsibility of the recipient to generate a new copy of the NewsItem on their system, by applying the Update instructions to the previous revision of the NewsItem, which they should already have, or be able to request from the provider. To generate the new revision of the NewsItem, each subelement of each Update element is processed in turn, in the order in which they occur. The value of each subelement's DuidRef attribute should match the Duid of an element in the previous revision. This is the element to which the instruction applies. In the case of Delete, the identified element is omitted from the revised NewsItem. In the case of Replace, the identified element is replaced by the content of the Replace element. In the case of InsertBefore, the content of the InsertBefore element is added to the revision in front of the identified element. In the case of InsertAfter, the content of the InsertAfter element is added to the revision afer the identified element.

SPECIAL INSTRUCTIONS AND CHANGING THE REVISION STATUS

As explained above, a news provider may pass arbitrary instructions to a news subscriber via the Instruction subelement of the NewsManagement element. The meaning and permitted values are determined by a controlled vocabulary that must be agreed upon by both parties in advance. As part of an Instruction, a news provider may supply one or more RevisionStatus elements. These allow the provider to indicate the status that previous revisions now have as a result of the release of the current version. The optional Revision attribute allows the provider to specify a particular RevisionId, otherwise the status applies to all previous revisions of the NewsItem.

APPENDICES

APPENDIX 1: NEWS AGENCY USE OF THE STATUS CONTROLLED VOCABULARY

The IPTC has created a Status controlled vocabulary for use by news agencies. (See http://www.iptc.org/site/NewsML/NewsMLTopicSets.zip for the complete set of NewsML TopicSets. See also the NewsML News Agency Guidelines). This consists of the following states:

- Usable – The NewsItem and its content may be published without restriction.
- Withheld - Neither the NewsItem nor its content may be published until further notice.
- Embargoed - Neither the NewsItem nor its content may be published until released for publication by the provider.

- Cancelled - 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.

Their use is illustrated by the accompanying graphic. This depicts the state transition diagram reflecting the ways in which the Status values are intended to be used. Thus, upon creation of a NewsItem, allowed values of the Status element are “usable”, “withheld” and “embargoed”. Once a NewsItem has had its Status set to “cancelled”, it has reached a final state. When a NewsItem has a Status of “usable”, “withheld” or “embargoed”, it may be changed to any other Status.

Within the StatusWillChange element, the values of the FutureStatus element should also conform to the same state transition rules. In other words, the Status and FutureStatus of a NewsItem should be an allowed change.