IPTC Standards DRAFT

NewsML 2 Architecture

Version 1.0

Model

Document Revision 12

IPTC address:
Postal mail: Royal Albert House, Sheet Street, Windsor, Berkshire, SL4 1BE, United Kingdom
Web: www.iptc.org
email: office@iptc.org

This project intends to use materials that are either in the public domain or are available by the permission for their respective copyright holders.
Permissions of copyright holder will be obtained prior to use of protected material.
All materials of this IPTC standard covered by copyright shall be licensable at no charge.

Document file name (+ Word file extension ".doc")**: DRAFT-NAR_1.0-spec-Model_12.doc**

Document URN: urn:iptc:std-draft:NAR:1.0:spec:Model:12
(More information on IPTC URNs in RFC 3937)

Specification Versioning History

Version	Issue Date	Approved by	Remark
1		IPTC Standards Committee	UNDER REVIEW

Document Revision History

Revision	Issue Date	Author (revised by)	Remark
Draft 1	2005-10-30	Laurent Le Meur	Initial draft, merging NSTR, NMDF and CoCo Models.
…			
Draft 11	2005-12-04	Laurent Le Meur	First public draft
Draft 12	2005-12-16	Laurent Le Meur	Second public draft

Acknowledgments

This documentation is the result of a team effort by members of the International Press Telecommunications Council, with input and assistance from other contributors.

The documentation was edited by Laurent Le Meur (Agence France Presse), and incorporates work by the following (ordered by family name): Mark Birbeck (x-port.net Ltd.), Dave Compton (Reuters), Jay Cousins (RivCom), Takahiro Fujiwara (EAST Co. Ltd.), Darko Gulija (HINA), Paul Harman (Press Association), Johan Lindgren (Tidningarnas Telegrambyrå), Jayson Lorenzen (BusinessWire), Stuart Myles (wsj.com), Michael Steidl (IPTC Managing Director), Ulf Wingstedt (CNET), Misha Wolf (Reuters) .

Table of contents

1   Introduction

1.1   Abstract

This document describes the **IPTC NewsML 2 Architecture (NAR) Model**, i.e. a conceptual model and the associated processing model. It describes the data types, properties and constructs shared by all new IPTC standards, the way metadata are expressed, the way items of news-related information are structured, managed and referenced in the NAR at a functional level.

This model is independent of the XML implementation of the NAR. It may be as easily implemented via object oriented software, in Java or C#. For this reason, UML diagrams illustrate the document, and the corresponding XMI (XML Metadata Interchange) files are available for software implementers.

This model is also compatible with the orientations of the Semantic Web. Therefore, the RDF model [RDF] associated with the metadata framework is added to this document.

The IPTC NewsML 2 Architecture Model builds on:

- The **NewsML 2 Business Requirements** [NML-BR] edited by the NewsML WG,

- The **EventsML Business Requirements** [EVT-BR] edited by the EventsML WG.

- The **News Metadata Framework Business Requirements** [NMDF-BR] edited by the NMDF WG.

- The **Implementation Guidelines for the IPTC Standards Architecture using W3C XML Schema** [NAR-IG] edited by Jay Cousins (RivCom) and Ulf Wingstedt (CNet) for the NAR WP.

- The current generation of IPTC standards, namely **NewsML 1.x**, **SportsML 1.x**, **NITF 3.x**.

- The work of other standards bodies, which have taken a similar modelling approach.

This document is complemented by the **IPTC NewsML 2 Technical Specification [NAR-TS]**, which details the XML implementation of the NAR, element by element, attribute by attribute. After reading the model document, implementers of the NAR shall use the Technical Specification as a reference for the development of NewsML 2 processors.

All IPTC NewsML2 specifications are complemented by the **IPTC NewsML 2 Architecture Glossary [NAR-GL]** which provides an extensive set of terms and definitions.

1.2   Status of this document

This document is under development by the IPTC News Architecture Working Party (NAR WP).

Comments from IPTC members which are intended to be visible to IPTC members only should be sent to the mailing list described at:

> http://groups.yahoo.com/group/iptc-news-architecture-dev/

Public comments should be sent to the mailing list described at:

> http://groups.yahoo.com/group/newsml-2/

Public versions of this document and of related IPTC documents are available at:

> http://www.iptc.org/dev/

1.3   Terminology

The document uses the following expressions:

"Contains" means **composition**; it means that an element of the model aggregates other elements or that an element has a particular attribute (e.g. an Any Item *Contains* a ManagementComponent).

"Is a kind of" means **specialization**; it means that an element of the model acquires properties from another element of the model (e.g. a News Item *is a kind of* Any Item). The IS concept usually doesn't appear in DTD oriented XML, but rather in W3C XML Schema work, where generalization and restriction of structures are supported.

This model makes use of several externally defined datatypes: **XS any**, **XS other**, **XS string** and **IRI** correspond to the datetypes xs:any, xs:other, xs:string and xs:anyURI defined in W3C XML Schema [XSD].

1.4   Cardinality

A property may:

- Be marked as *mandatory:* i.e. it must be present,

- Include the expression *collection of*: i.e. it can appear zero or several times,

- Include the expression *ordered collection of*: i.e. it can appear zero or several times and the order in which it appears is meaningful.

If no specific indication is given, all properties are **optional** and **not repeatable**.

1.5   Notation

a/ *Bold & italic* strings are used to identify class and property names. The names chosen here do not imply that the final XML element and attribute names will be identical.

b/ Property descriptions are identified by a black bullet point. Sub-property descriptions are identified by a white bullet point.

Example:

- *property*: description.
  - *Sub-property* : description.

c/ Temporary notes are in italics and coloured blue:

> e.g. *To be defined*

d/ Currently open questions are coloured orange:

> e.g. **@@ Isn't urn:newsml:xxx an improper term for this global class?**

> **We kindly ask users of this specification to discuss these questions during the current Experimental Phase.**

e/ Some properties support sub-properties. These are shown using an indented list item.

2   Overview of the NewsML 2 Architecture Model

2.1   A common model for all IPTC standards

News exchange covers not only the core news content, but also data that describe the content in an abstract way (i.e. metadata), information about how to handle news in an appropriate way (i.e. news management data), packaging of news related information and finally information about the news transportation or routing process itself (i.e. exchange data).

Having a strong background in developing and maintaining news exchange formats, the IPTC created the **IPTC NewsML 2 Architecture** as the most comprehensive and versatile way to move all types of data between media systems in order to make news exchange efficient and reliable.

This document is split in several sections:

- **Building blocks:** this section introduces a model made of reusable structures, i.e. data types, basic and aggregate components.

- **News Metadata**: this section defines the model chosen for the representation of metadata and labels.

- **News Structure**: this section defines an abstract class acting as a model for all managed items defined by the NAR. It also defines three standard classes derived from the abstract model: News Item, News Topic and News Package. These classes are available for all IPTC standards, when more specialized classes are not needed.

- **News Message**: this section defines a model proposed for the exchange of Items in professional news industry syndication networks.

The choice of a common model for all IPTC standards eases their adoption and the related work on a precise processing model guarantees a very high interoperability level between providers and consumers of news. When specific needs arise, the extensibility of the model allows news providers to use externally defined components of information, and plug them into the framework.

2.2   Items and components

The NewsML 2 Architecture model defines *Items* and *Components*.

Items are manageable objects, i.e. objects with a persistent, universally unique identity and a set of management properties associated with a precise processing model.

An *Any Item* class models this abstract notion; generic and specialized Items are derived from the abstract model as part of each IPTC standard. Items optionally contain components. Using common components, Items of all kinds share certain properties (i.e. metadata) and may be interconnected in a consistent way, thus creating a powerful web of newsworthy information.

Components represent pieces of information that have a precise meaning and processing model. Components are included in items; some of them are specific to one type of Item, but others are common to several types of items. Common components can range from basic (also called properties in this document) to aggregate structures depending on the scale of reuse envisioned.

A set of Common Components defines context-free components representing the **administration**, **description**, **rights**, **publication**, and **protection** (via a digital signature) of newsworthy information.

2.3   Generic and specialized Items

The NewsML 2 Architecture model defines several classes of generic usage, which are not associated with any specific IPTC standard, but rather potentially used by any standard.

These classes are globally termed Generic Items. The NAR currently defines:

▪ A *News Item* aims to convey news with the sense of the reporting of a topical event.

▪ A *Topic Item* aims to convey knowledge about concepts (named entities such as organisations or classification term such as news subjects).

▪ A *Package Item* facilitates the packaging of all kinds of items, from really simple constructs to the highly hierarchical structures created by some news providers.

The model of these classes is described in details in this document.

Specialized Items are Items which include a collection of components usually chosen from a library. When no common component fits a given need, an IPTC standard defines or recommends components optimized for the handling of Specialized Content (sports results, TV program guides, event coverage, etc...).

2.4   News Message

The NewsML 2 Architecture model defines a *News Message* as a way to facilitate the exchange of all kinds of items in a broadcast (or multicast) network.

The use of a News Message is totally optional in a news workflow for conveying NAR Items. Any other possible syndication protocol may be used in order to exchange Items in a news workflow.

2.5   Class diagram

The following diagram represents a partial class diagram of the NewsML 2 Architecture. The abstract Any Item class is at the top of the hierarchy. Three derived classes are defined in the core architecture: News Item, Topic Item and Package Item. Other classes will be defined by individual IPTC standard Working Groups.

Further specialization of the main classes allows for more detailed structures and processing.

Important note: The XML implementation of the specialized classes is different from what could be implied from this class diagram: a News Item is implemented as an XML element called <news:item/>, but a textual News Item is eventually implemented as a *typed* news:item rather than as a <text:item/>. The implementation of specific TopicItems is still to be decided at the time of this writing.
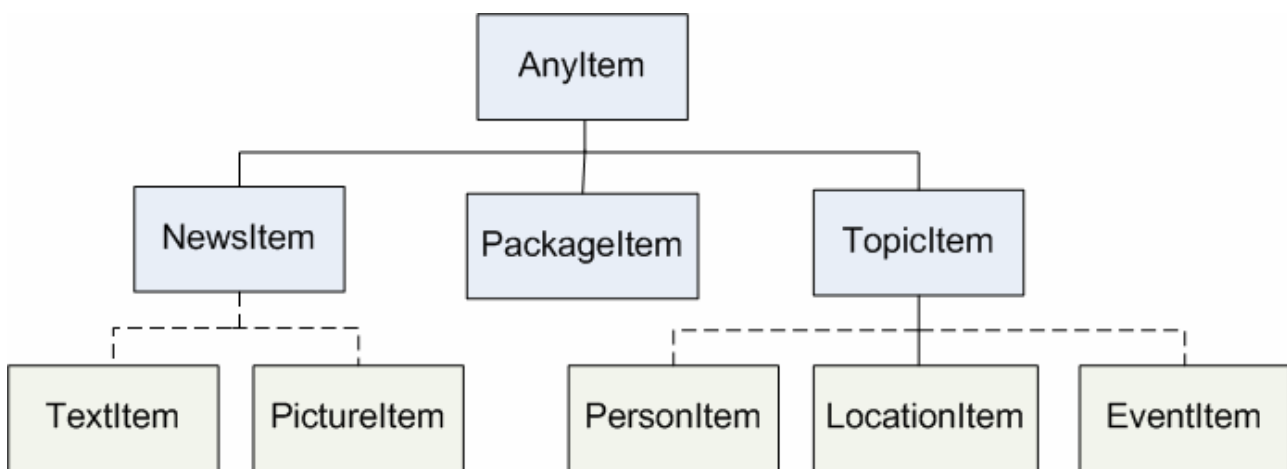
2.6 Overview diagram – Items and components

The following diagram illustrates the way the architecture is constructed. At the bottom of the diagram is a "real world" entity, like a car accident, a sports event, a famous athlete, a TV program. This entity is represented by some content. Sets of metadata are associated with this content. Some metadata components are shared between the different entity classes; other metadata components are specific to a given entity. This set of information is then wrapped in a managed container called an Item. Specific management properties indicate the precise class of the Item, and each Item gets a persistent, globally unique identifier.

The Item (individual instances derived from "AnyItem")							
Item components		Packaging	General News report	Event description	Person description	Radio/TV Program Listing	etc.
Management Component	itemClass	Package	News	Topic	Topic	Listing	TBD
	contentClass	Composite Text Photo Still Graphic Animated graphic Audio Video …	Text Photo Still Graphic Animated graphic Audio Video	Event	Person	TV	TBD
Common Metadata Components	signature administrative descriptive rights publication links …	☑ ☑ ☑ ☑ ☑ ☑	☑ ☑ ☑ ☑ ☑ ☑	☑ ☑ ☑ ☑ ☑ ☑	☑ ☑ ☑ ☑ ☑ ☑	☒/☑ TBD ☒/☑ TBD ☒/☑ TBD ☒/☑ TBD ☒/☑ TBD ☒/☑ TBD	☒/☑ TBD ☒/☑ TBD ☒/☑ TBD ☒/☑ TBD ☒/☑ TBD ☒/☑ TBD
Optional Content Component	generic topic information …	-	-	☑	☑	-	☒/☑ TBD
		Group Structure = XML hierarchical structure of references of other Items	News Content Components. = XML markup; plain text; binary photo, graphic, audio, video	Event Content Component = XML marked up event information	Person Content Component = XML marked up person information	Program-Guide Content Component = XML marked up radio/TV program information	TBD
equivalent to … in the „Real World"		none					TBD

Version 4 (as of 2005-11-02)

Legend to the diagram:
TBD = To Be Defined (= not defined at the time of creating this diagram)

2.7   Conformance levels

Different Conformance Levels are defined in the model, each of them related to a level of complexity (at the conceptual and processing level) of the related Items. This feature adds **modularity** to the model.

The current model defines two conformance levels named "**core**" and "**power**". The core conformance level is focused on simplicity and interoperability. The power conformance level gives more flexibility to providers who choose it, but the recipient processors are more complex to program, and interoperability is lower than in the first case as not all recipients will implemented the power level.

A NewsML2 processor can therefore be labeled as supporting either "core" or "power" functionality.

A NewsML2 "power" processor must provide all of the "core" functionality, plus the functionally marked as "power conformance level feature" in this document.

As the "power" features are only an extension of the "core", a NewsML2 "core" processor can process "power level" Items, simply ignoring the information he's not able to process. As rights management is part of the power conformance level, a provider must be careful not to send to "core" processors rights-protected information.

The features belonging to the "power conformance level" are still discussed in the NAR WP. The current split will be reviewed during the experimental phase.

3   Building Blocks

3.1   Introduction to Common Components

The IPTC Common Components are used as building blocks of the IPTC NewsML 2 Architecture. They may be used directly, but they also may be extended or restricted in order to obtain more specialized components for use in a specific Item.

A common component is a generic piece of context-independent information that can be reused to construct business messages in different business contexts. A common component therefore represents a piece of information that has a precise meaning but is in essence context-free. Context free means that a piece of information is not semantically bound to a particular business context. It can be used in more than one context without changing its essential meaning and in each context where it is reused that essential meaning is not changed but rather the semantics of the usage context are added to its meaning in an additive process of semantic enrichment. Consequently, a common component can be used and combined with other common components to create larger context-free structures available for reuse. When used in a business-specific context, common components take on the business-specific semantics of their usage context.

Reuse enables the development of libraries of schema components ('architectural schema') that can be created for use in modular architectures in order to promote consistency of design and content structure.

Context neutral design of data structures – i.e. the granularity and semantic levels of the common components – creates data structures that can be reused across specific news domains either as-is or used as the building blocks of new structures. This design style gives extensibility to the architecture while maintaining the meaning of the individual common components wherever they are used. Making design independent of business context provides for this semantic interoperability as it allows common components to be reused and reassembled to create new structures where the original semantics of the components are enhanced by the specific business context of their reuse.

The risk of reuse is that a change to a schema component can have wide-ranging consequences, but if the components have been designed with an appropriate granularity then such change consequences can be managed. Successful reuse of data structures across specific news domains requires the design of the common component granularity to accommodate these considerations.

3.2   Levels of granularity of Common Components

The building blocks come in three levels:

1. Fine grained, called **datatype**. A datatype can be defined with different levels of complexity.

2. Medium, called **basic Component**. Represents a single piece of business information and uses an existing data type to provide its content model. It is capable of being used independently or as part of another component.

3. Coarse, called **aggregate component**. Represents a composite piece of information. It is a collection of properties that together is more than the sum of its constituent parts, and when taken as a whole carries its own unique business semantics. The properties composing the whole can be basic or aggregate components. An aggregate component may be designed so it supports an extension point where news providers can extend its usage.

Both level 2 and 3 constructions can be used directly in an Item.

Common components range from simple to aggregate structures, depending on the scale of reuse envisioned. Full component architectures like ebXML core components are considered overly complex for IPTC's requirements and a simplified hierarchy of data type, basic component, and aggregate component is sufficient to meet the IPTC requirements. Further, should the IPTC in the future decide to refactor to ebXML then such a simplified hierarchy provides a foundation on which such refactoring could be done.

The **data type** is the lowest level of common component. It has no specific business meaning or semantics of its own and only takes on business meaning when used as the data type of a basic component. A data type can range from a simple type to a complex type and can be used to represent restricted values. A data type can be specialised to create specific variants of it. A data type can be a:

- Simple data type – a basic data type. For example,
    - A primitive XML Schema data type, such as **XS String,**
    - A named primitive data type where some restriction has been imposed, such as Integer100Type where **XS Integer** has been restricted to a value range of 1 to 100.
- Complex data type – a basic data type extended to add further information in order to correctly represent the value. Such ancillary information takes the form of attributes. For example,
    - A LabelType which support mixed content and is extended to add an xml:lang attribute,

The **basic component** represents a single property or piece of business information and has specific business meaning. A basic component can be used as is, or can be used as part of an aggregate component, in which case it is one of a set of properties. As with a complex data type, a basic component can be qualified by ancillary data if required to complete it's meaning. The content model of a basic component could range from simple to complex according to its data type. For example,

- A slugLine element of data type `XS String`, with an additional separator attribute.

The **aggregate component** is a set of properties that represent a distinct piece of business information and has specific business meaning. An aggregate component can be composed of basic components and other aggregate components. For example,

- A headline element of data type LabelType, with rich text as value.
- A descriptive component, defined as a group of basic components, like title and subject,
- A content metadata component, which contains different groups of components, and allows markup from other namespaces to be used, for example Dublin Core metadata.

4   News Metadata

4.1   Introduction

The NAR defines several specific types of metadata - simple or complex - to be used as a model for the properties assigned to an Item.

These properties are metadata about an Item as a whole, about the content of the Item as a whole (e.g. administrative, descriptive metadata), or about specific renditions of the content (e.g. physical characteristics).

The agreed perception is that news content is a representation of real life facts and metadata is supplemental information about such content, but one has to be aware the same kind of information may act as content and as metadata. An example: structured information about a person may be metadata in a News Item (e.g. information about the creator), but may also be the content of a Topic Item describing this person. But be aware that these structures are not defined by the metadata types of this section, they are context-less common components.

As the NAR aims to be compatible with the Semantic Web of the W3C it is also compatible with their underlying technology, the Resource Description Framework (RDF). But the IPTC model and syntax implementation of metadata do not require any knowledge about RDF, any transformation from the IPTC metadata model to RDF will be done by means beyond the NAR. Transforms to RDF/XML will be achieved by using a mechanism like GRDDL [GRDDL], so that those who wish to convert the metadata to RDF triples can do so.

4.2   Properties with simple values

These are:

- **Date**: this is a date or a timestamp (XS Date or XS Datetime), as defined in [XSD].

@@ Partial dates: the need for partial dates has been expressed by the EventsML WG. Should we implement them in the NAR?

- **String**: this is a set of characters (XS String), as defined in [XSD].

- **Integer**: this is an integer (XS Integer), as defined in [XSD].

- **IRI**: this is an Internationalized Resource Identifier, as defined in [RFC987].

4.2.1 Power Conformance Level features

In a professional news workflow, metadata may be initialized in real time and later updated (e.g. before archiving the information). Therefore it may be needed to reference a piece of metadata locally in order to act on it individually.

Also, in such a collaborative environment, metadata is not always entered by one person at one time, but may be entered by different people, organisations or systems at different time. Therefore it may be needed to keep track of who is assigned the editing of specific properties, and when and by whom a property has been given a value.

All metadata properties – metadata with simple or coded values, or labels - share the following features, defined only at the Power Conformance Level:

- A *Local Identifier*: unique within the XML document and supported by all properties. Local references to this element may be expressed using this identifier. This feature may especially be used for partial updates.

- A *Creator*: if the supporting element is empty, the creator qualifier specifies which entity (person, organisation or system) will create the metadata. If the element is non-empty, it specifies which entity (person, organisation or system) has created the metadata.

- A *Creation Date*: the date (and, optionally, the time) when the metadata value was created or updated.

4.2.2 Associated RDF model

The *subject* of the statement is the current Item.

The *predicate* is represented by the property.

The *object* – i.e. the value of a simple metadata element - is an RDF literal.

Example: the provider of an item is defined as "iptc.org". This leads to the following triple, using the [N3] notation:

> @prefix newsml: <urn:iptc:std:newsml:2.0:xmlns>
>
> <> newsml:provider "iptc.org" .

The additional power conformance features - creator and creation date - are properties of the reified statement:

Example: the provider of an item is "iptc.org"; this was specified by an editor registered by the code "iptcmb:md" at "2005-11-11T08:00:00Z". This leads to the following additional triples:

> @prefix iptcmb: <http://iptc.org/members>
>
> {<> newsml:provider "iptc.org"}   newsml:creator iptcmb:md ;
>
>   newsml:created "2005-11-11T08:00:00Z" .

4.2.3 Value validation of simple properties

Validation of the value found in an instance is achieved simply via the XML Schema declaration of the element.

4.3   Properties with coded values

Coded properties have values taken from a well defined controlled vocabulary (taxonomy, thesaurus, classification system ...) which is called a "concept scheme" in the scope of the NAR. A scheme is a list of codes which must be managed by a specific body, this may be the IPTC, any other well known standardisation body, or a news provider.

These values are represented by a combination called a {scheme alias, code} pair, usually expressed as the value of the "code" attribute of a metadata property, and primarily intended to be consumed by processing software.

4.3.1 Introduction to {scheme alias, code} pairs

A {scheme alias, code} pair represents a term from a controlled vocabulary.

For compliance with the Semantic Web, a scheme takes the form of a URI, and the simple concatenation of the scheme URI and a code is also a valid URI.

As an example, in the IPTC scheme defined for news categories, which could be identified by the URI "urn:newsml:iptc.org:20001006:topicset.iptc-subjectcode:"([*]), the code "15000000" is associated with the concept of "Sport".

Another example, in the NASDAQ scheme, which could be unambiguously identified by the URI "urn:newsml:iptc.org:20001006:nasdaq-company:", the code "ADBE" is associated with the company known as "Adobe Systems Incorporated".

It is not mandatory that one maintains as an XML file a complete list of codes defined in a given scheme. In the NASDAQ example, this would be a loss of time. It is sufficient to have an unambiguous identifier for each scheme a provider wants to use, and that this identifier is known from the customers of the news feed this provider offers.

4.3.2 Compact URI Type (CURIE)

In order to manipulate {scheme alias, code} pairs is an efficient manner, a compact syntax is needed.

In all NAR XML instances, the scheme URI is replaced by a scheme alias, and the code is appended after a colon (":") character. Such syntax allows the use of a "scheme:code" pair as value of a single XML attribute.

This datatype is currently defined in an internal draft produced by the W3C RDF-in-HTML task force [CURIE]. The value space of the CURIE datatype is a set of {scheme, code} pairs which identify concepts. This is similar to the value space of the QName datatype, i.e a set of {namespace, local part} pairs which identify element or attribute names.

But QNames cannot be used for this purpose, because the local part of QNames cannot be a numeric, and the News industry and the Financial industry are full of taxonomies making use of numeric codes. They aren't alone in this aspect (consider ISBN and ISSN).

Note that:
- CURIEs allow any sequence of legal URI characters in the suffix, including, for example, digits only, dashes, slashes, etc...
- In the NAR implementation, CURIEs MUST have a non-empty prefix.

CURIEs can be viewed to a certain extend as short, lexical representations of URIs. Let's be careful: the mapping from a {scheme alias, code} pair to a URI is not bijective: a URI cannot be mapped back to a {scheme alias, code} pair, as the tuple is lost in the concatenation.

4.3.3 Concept Type

Some coded properties support the indication of the type of the concept used as property value. As an example, the subject of a news story may be a general category (e.g. sport or football), a business sector, a person or more precisely an artist or musician, an organisation, a location, an event ...

---

[*] This URN, which is based on the NewsML URN RFC 3085 but not fully compliant, is given as an example: the IPTC may choose another URI style, like "urn:iptc:nc:cat:" or "http://iptc.org/category/"

4.3.4 Uncontrolled values

It is not always possible to use a {scheme alias, code} pair as metadata value. As an example, few news organisations maintain a formal listing of their editors, and therefore using a {scheme alias, code} pair as value of the administrative "creator" property is not always possible. Also, the formal expression of a language is a {scheme alias, code} pair, e.g. "iso3066:en-US", but many systems will directly generate "en-US" as language value.

In order to fulfil this need, a large number of properties allow for uncontrolled values to be expressed as the value of a "content" attribute, instead of {scheme alias, code} pairs.

4.3.5 The catalog of scheme aliases and scheme URIs

Each news provider is free to use their own taxonomies, e.g. of subjects, entities, genres, etc. So there are no "well-known" scheme prefixes to hardwire.

The NAR therefore uses a solution found in schematron: a catalog is defined as a set of scheme declarations in use by a news provider. This catalog is included at the top of each Item. Due to the large number of schemes potentially used in a single Item, and knowing that bandwidth is very important to the News industry, the catalog may be stored in a central repository, and referenced from the Items which use it.

Scheme alias declarations are global to the Item in which they appear and cannot be overridden in a given Item.

Note: A natural solution would be to use xmlns declarations to declare scheme URIs; but xmlns declarations must be inline, therefore this solution is rejected.

4.3.6 Power Conformance Level features

Note: Some features are available on some elements only. The reason is that making them available on all elements would lead to unexpected results and interoperability issues. The technical specification gives details about the features supported by a given property.

4.3.6.1 Child Of and Same As

Some coded properties support specific sub-properties, such as "child of" and "same as".

The "child of" property is used to connect nodes in a directed graph, in order to form a taxonomy. Such information makes possible a hierarchical indexing of Items, and the constitution of *breadcrumbs*, i.e. the display of a hierarchy of tokens like "culture > literature > fiction" where the precise subject of a news story is "fiction".

A given concept may be identified by different codes in different schemes. The "same as" property is the expression of the semantic equivalence of two {scheme alias, code} pairs, allowing the indexing of an Item with multiple <scheme, codes> pairs identifying the same concept.

4.3.6.2 Supplemental information

Some coded properties support sub-properties specific to the type of concept used as value of the property, for example the name and email address of the creator of a story, or the full name of an organisation in different languages.

In some occasions, like in the case of a creator property, a provider will not publish the internal {scheme alias, code} pair associated with the person, but will still give to its users some information about this creator, such as its initials and its email address.

If a {scheme alias, code} pair is provided, then this persistent value takes precedence over any transient information provided in child elements.

4.3.6.3 Confidence

Some coded properties support a "confidence" sub-property, i.e. an indication of the confidence with which the metadata has been assigned.

Such feature is useful in case of automatic categorization or indexing.

4.3.6.4 Relevance

Some coded properties support a "relevance" sub-property, i.e. the relevance of the metadata to the piece of news to which it is attached.

A high relevance indicates that this piece of metadata truly expresses what the piece of news is about, while a low relevance indicates a low correlation between the metadata and the essence of the piece of news.

As an example,

4.3.6.5 Level of indirection (why)

A concept used as a property value may be:
- Directly extracted from the content by a tool and/or by a person, eg Paris or Sport.
- An ancestor of some other concept, e.g. the concepts France and Europe are ancestors of the concept Paris.
- Derived by look-up in some taxonomy/database, e.g. the concept Pharmaceutical Industry Sector may be derived from the concept Merck.

Therefore, some coded properties support a "why" sub-property as an indicator of the reason why the metadata has been included. The allowed values are detailed in [NAR-TS].

4.3.6.6 About

Some coded properties support an "about" sub-property, which explicitly specifies the subject being described. This feature is used when the subject being described is not the containing Item.

As an example, take the headline of a story stating that "While in New Orleans, *President George Bush* said that ...". A coded property outside the headline label may provide information about this "President George Bush". For this purpose, it will use the about sub-property to refer to this substring inside the headline. E.g. it may indicate that the president may be "George W. Bush" with a confidence of 60%, but may be "George Herbert Walker Bush" with a confidence of 40%.

4.3.7 Associated RDF model

The *subject* of the statement is usually (see below why) the current Item.

The *predicate* is represented by the property.

The *object* is an RDF URI Reference represented by a "code" attribute, or an RDF literal represented by a "content" attribute.

Example: the subject of an item is defined as "cat:01011001", the IPTC category code for "art > music > classical music". This leads to the following triple, using the [N3] notation:

> @prefix newsml: <urn:iptc:std:newsml:2.0:xmlns>
>
> @prefix cat: <urn:iptc:concepts:category:>
>
> @prefix ccls: <urn:iptc:concepts:contentclass:>
>
> <> newsml:subject   cat:01011001 .
>
> cat:01011001 newsml:type   ccls:category .

Some additional power conformance features – child of, same as, addition information - are properties of the object. For example classical music has a broader concept of music:

> cat:01011001 newsml:childOf cat:01011000 .

Other additional power conformance features – confidence, relevance, why - are properties of the reified statement. For example a categorization engine may state the a story is about classical music with a confidence of 70%, and that this subject is of medium relevance vis à vis the story:

> {<> newsml:subject   cat:01011001} newsml:confidence 70 ;
>
>   newsml:relevance 50 .

The about attribute modifies the subject of the statement. For example, if a story about a soccer match contains a sentence about the "Fifth Symphony , Beethoven", locally identified as "id5symph", it is possible to state that:

> <#id5symph>   newsml:subject   cat:01011001 .

4.4   Labels

Labels expose aspects of news as natural language strings. They are metadata assertions expressed as natural language strings intended to be consumed by human beings. They are typically displayed alongside the content of an Item or in place of Items in a list, providing a means of selection among them. Or they are those notes, comments or instructions created by a news provider for use by recipient editorial teams.

A Label contains:

- A *Language Indicator*: the language of the content of the label.

- A *Role*: a refinement of the semantics of the label.

- A *Media indicator*: an indication of the target media type(s), as defined by the Cascading Style Sheets specification [CSS].

- *Textual content*: the content of a label. At the core conformance level, only plain textual content is allowed.

4.4.1 Power Conformance Level feature: Rich text

At the power conformance level, all labels support rich text, i.e. text interspersed with some specific markup.

Therefore, at the power conformance level, a Label contains:

- A *Markup Indicator*: the information helps the processor checking the presence and the nature of markup in the content of the label.

The exact allowed markup is detailed in [NAR-TS].

@@ Why indicating the nature of the markup if only xhtml is defined?

5   News Structure

5.1   Introduction

This section defines an abstract class – Any Item - acting as a model for all managed items defined by the NAR.

It also defines three standard classes derived from the abstract model: News Item, News Topic and News Package. These classes are available for all IPTC standards, when more specialized classes are not needed.

5.2   Any Item

5.2.1 Description

Any Item is an *abstract class* defined as a template for all Items standardized by the IPTC. The different IPTC standards use Any Item as a blueprint for their Specialized Items.

Each Item contains:

- A *mandatory Schema Version*: the major-minor version of the XML schema specifying the Item .

- A *mandatory Item Identifier*: a persistent, universally unique identifier for the Item. This identifier has the form of a NewsML URN [RFC-3085], with the particularity that no revision identifier is indicated as part of the identifier.

- An *Item Version*: the version of the Item. Its default value is 1.

- A *Language Indicator*: the default language used along the textual content of the Item. Note that a different language indicator may be set on any given label (i.e. a property of type LabelType) included in the Item.

- A *Signature Component* (power conformance level only): a digital signature associated with the whole Item, or only parts of it. It is for example possible to sign each individual News Content Component of a News Item, using by its local identifier.

- A *mandatory Catalog*: a wrapper for scheme alias declarations.

- A *mandatory Item Metadata Component*: A set of properties associated with the Item. See below.

Note: The conceptual model keeps the Item identifier and the version identifier as two separate properties. RFC 3085 being under revision at the time of writing, the final format of the Item Identifier cannot be formally agreed yet.

5.2.2 UML diagram associated with the Any Item
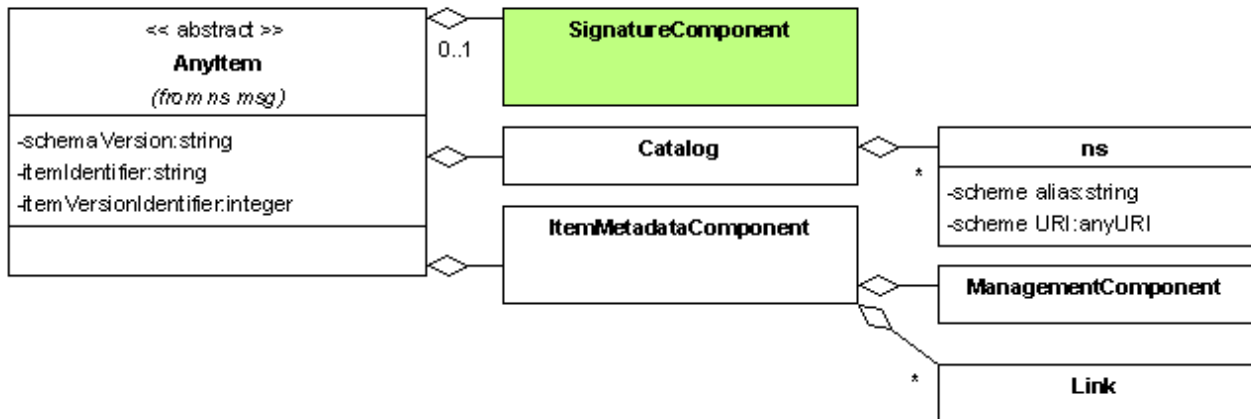
Note: in the UML diagrams, components defined at the power conformance level only are in green.

5.2.3 Signature Component

A digital signature is a unique seal placed on data. It is very difficult to forge and assures that any change made to the signed data cannot go undetected.

The NAR supports the model and syntax defined by the W3C in [XMLDSIG], and introduced by the following: "XML Signatures provide *integrity*, *message authentication*, and/or *signer authentication* services for data of any type, whether located within the XML that includes the signature or elsewhere".

The NAR model excluded two functionalities defined by the W3C's XML-Signature Processing Recommendation. These are: "Signed content included within an XML Signature Construct " and "Detached Signatures".

 Therefore this component offers the following features:
- A Signature is "enveloped" (the Signature Component is contained within the Item being signed).
- The Signature cannot be "enveloping" (it cannot sign content found within the signature itself)
- A Signature cannot be "detached" (a detached Signature Component would not be contained within the Item being signed and could be external to the containing document).
- A Signature shall sign the Item containing the Signature Component or child Components of the Item containing the Signature.
- A Signature shall not be related to Items and Components external to the enclosing document (via references).

5.2.4 Catalog

The Item Metadata Component contains:

- *A collection of Namespace entries*: a declaration of the equivalence between a scheme alias and a scheme URI .

5.2.4.1 Namespace entry

Each Namespace Entry in the catalog defines:

- A *scheme alias*: a short string used by the provider as a replacement for a scheme URI.

- A *scheme URI*: the full URI this prefix represents.

5.3   Item Metadata Component

The Item Metadata Component contains:

- A *Management Component*: a set of properties associated with the management of the Item.

- *A collection of Link*s: a named relationship linking the current Item to a target Item (or web resource).

5.3.1 Management component

The management component is a set of properties directly associated with the management of the Item:

- A *mandatory Item Class*: the nature of the Item, in accordance with the namespace the Item is defined in.

- A *mandatory Content Class*: the nature of the content, i.e. the type of content it conveys.

- A *collection of* '*Instance Of' Relationship*s: an indication that this Item is an instance of the specified recurring report (also known as a fixture).

- A *Provider*: the entity responsible for the management of the Item.

- A *Date Item Created*: the date and time on which the first version of the Item was created.

- A *Date Item Modified*: the date and time on which the current version of the Item was modified (i.e. revised).

- A *Date Item Released*: the date and time on which the provider's system intends to transmit the current version of the Item to the target audience.

- A *Date Embargo Ends*: the date and time before which all versions of the Item are embargoed.

- A *Date Item Retired*: the date and time after which all versions of the Item are retired.

- A *Publish Status*: the publishing status of the Item.

- A *Conformance Level*: the conformance level the Item is compliant with.

- A *collection of Editorial Signal*s: An indication that the content is of particular interest.

- A *Generator Tool*: the name of the software tool used to generate the Item.
  - A *version*: a qualifier which specifies the version of the tool.

- A *Profile Name*: the name of the structural template (aka profile) used for the generation of the Item.
  - A *version*: a qualifier which specifies the version of the profile.

- A *File Name*: the recommended file name for this Item.

5.3.1.1.1 <u>Power Conformance Level features</u>

- A *collection of Alternative Locations*: a pointer to a location where an alternative representation of the Item can be found.
  The constraints of time and space as well as the demands of specific applications and delivery environments make it necessary to support different physical representations of the same Item. This element points to a location from where one can pull a backup of this instance, possibly with a different physical representation.
  - o A *representation*: a qualifier which specifies the representation available at this location.
- An *Original Identifier*: The original non-RFC3085 identifier of the content.
  - o A *type*: a qualifier which specifies the context within which the alternative identifier has been allocated. For example the name of a system or of a URI scheme.
- A *collection of Alternative Identifier*s: an alternative non-RFC3085 identifier of the content.
  - o A *type*: a qualifier which specifies the context within which the alternative identifier has been allocated.

@@ Is this feature really needed in the standard, or should we left it to provider defined extensions?

5.3.2 Links

The Link Component offers a generic mechanism for linking items within the NAR framework as well as creating links from Items to other Web resources.

Links have several usages in the Item Header:

- **Navigation links** allow for navigation from an Item to another related item or a Web resource. Example: a News Item representing a section of a transcript (a "take" in the news industry) may be linked to the previous and next takes; an article about a person may be linked to this person's biography.

- **Derivation links** allow for the expression of a parent/child relationship. Example: a News Item representing an article may be linked to the article it is a translation of; a wrap-up may be linked to the previous stories used as source material for the article; a processed picture may be linked to its source picture.

- **Attachment links** allow for the attachment of an Item to the current one, so the processor may retrieve it when some of its content is needed along with the content of the current Item. Example: a NewsItem representing an illustrated story may be linked to the News Items representing the pictures that appear in the story content.

Links have an additional usage in the Package Item:

- **Composition links** are used in a Package Item to aggregate the Items included in the package.

All types of links share the same technical structure, and their semantic is refined via a relationship indicator.

5.3.2.1 Identification of the linked resource

A Link component contains:

- A *mandatory Target Resource Identifier*: the Item Identifier of the target Item, or the locator of the target Web resource.

- A *Target Item Version*: the version of the target Item identified by a Target Item Identifier. By default, the latest revision is retrieved when the link is activated.
  Absent if the target is a Web resource.

- A *Relationship Indicator*: the identifier of the relationship between the current Item and the target resource. In the absence of this indication, the relation is a simple association.

- A *Link Title*: a short human-readable name representing the link and displayed to the users.

5.3.2.2 Characteristics of the linked resource

In order to ease the processing of a link, some common characteristics of a target resource may be added to the linking information.

Such feature is especially useful if the target on the link is a Web resource. Note that if the target resource is an Item, these characteristics must not be confused with the characteristics of the content of the Item, which are properties available at the power conformance level, as defined in 3.1.4.

A Link component additionally supports:

- A *ContentType* of the target resource.

- A *Size* of the target resource.

5.3.2.3 Validity period

At the power conformance level, additional properties supported by a Link are:

- A *Valid From*: the date (and, optionally, the time) *before* which the link is not valid.

- A *Valid To*: the date (and, optionally, the time) *after* which the link is not valid.

@@ Is it really a power feature?

5.3.2.4 Metadata about the target resource

At the power conformance level, supplemental metadata extracted from the target resource may be added to the linking information. Such feature is especially useful if the target on the link is an Item.

@@ Should we restrict this feature to links pointing to Items.

Such information is not constrained by the model. It may be part of:

- The Target *Item Metadata*: e.g. Publish Status, Alternative Location ...

- The Target *Content Metadata*: e.g. Intended Audience, Subject, Genre …

- The *Characteristics* of the content: e.g. Size, Content Type, Format, or specific characteristics like the Height and Width of a picture. Different sets of characteristics may be provided, corresponding to specialized content components.

5.3.2.5 Processing model

5.3.2.5.1 Retrieving the linked resource

In order to retrieve the linked resource, a recipient should:

1. Check whether the target resource (either Item or Web resource) is available in a local repository (most NAR objects are stored in the recipient cache, so they are usually retrieved directly using the *Target Resource Identifier*, and not from the provider's site via a URL).

2. Check whether the scheme of the *Target Resource Identifier* allows some direct retrieval of the target resource via the Web (e.g. if the scheme is http: or ftp:)

Then, at power conformance level:

3. Check whether *Target Resource Identifier* contains an implicit resolution mechanism (e.g. DOI, PURL etc).

4. Check whether there is the *Alternative Location* exposed from the *Item Header.* This information may complement the Target Item Identifier and provide an immediate URN resolution mechanism. Multiple locations may be given, as allowed in the Management Component. In such case the processor will use the role qualifier and URL scheme for choosing the most appropriate resource.

5. Signal an error or ignore the link.

5.3.2.5.2 Processing the link instance

When used in the Item header, the processing model of a link instance depends on its Relationship Indicator.

The processor must discard without warning the information it is unable to process (i.e. follow a "must ignore" rule).

In the case of Navigation links the processor should display the link title as supplemental information to the final user and enable easy navigation to the target resource.

In the case of Derivation links the processor should enable a user to navigate to the original resource and could also undertake more specific actions such as such as stopping the distribution of all Items derived from a Cancelled Item.

In the case of Attachment links the processor may pre-fetch the target resource, because the Item content will need the resource to display correctly. A usual use-case is of an illustrated article represented as a News Item, where the illustrations are also represented as News Items.

In the case of Instance links the processor should be able to dynamically find all the Items that point to the same resource.

When used in the scope of Package Item, link is always a Composition link and the Relationship Indicator property defines the role of the target Item within the group.

Composition links are used in Package Items as a mechanism for grouping various Items without embedding their content. Such links may contain some information about the target Items, therefore allowing a user to decide whether to follow the link. It is up to the processor to decide how to present the package to the final user and when to retrieve the actual content.

5.3.2.5.3 *Checking metadata associated with the content of an Item*

A link does not enable direct access to the content of an Item, only to the Item as a whole.

But in some occasions, it is useful to get some properties of the content of an Item, in order to check if this content is usable or not. Therefore, content metadata may be added to the linking information.

For example, a mobile user will follow links leading to Items containing small resources only (because of cost or time constraints). On the other hand, a picture magazine user will follow links leading to Items containing high quality pictures only.

In the case of a News Item, as a provider may not know which content component might be useful for a given recipient, the main properties of all content components should be exposed. The amount of included information is a trade off between usability and size of the Item.

Pointing at the latest version of an Item while exposing content metadata may lead to unwanted display or selection criteria if these metadata were subsequently modified; therefore only the stable content properties should be exposed in a link.

5.3.3 Additional components

The class deriving from Any Item class may support any number of additional components, representing content or metadata (i.e. data about content).

Most metadata components are shared between standards, and are thus defined as common components. Content-oriented components are usually not shared between standards and thus are rarely defined as common components.

5.4   Content Metadata Component

The most usual metadata components are the administrative, descriptive, rights and publication components, which represent information relative to the content of an Item.

They are all wrapped in a global container, ready for use by different classes of Items: in particular, this component is used by the News Item, Topic Item and Package Item.

The creators of other classes of Item may prefer to use individual sub-components of this macro-component if some part of it makes no sense in the new construct.

5.4.1 UML Diagram associated with the Content Metadata Component
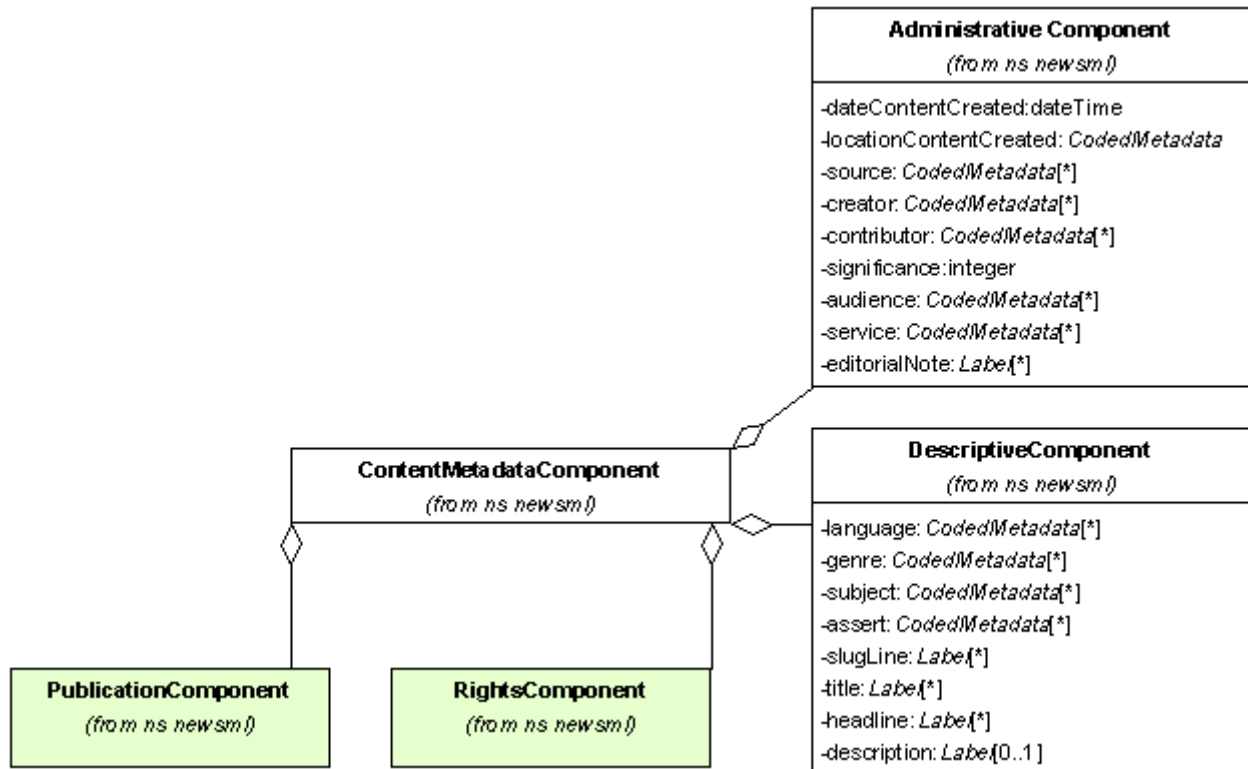
5.4.2 Administrative Component

The Administrative Component is a set of properties associated with the administrative facet of content, i.e. data that cannot be inferred from "consuming" (reading, listening to, watching) the content:

- A *Date Content Created*: the date on which the content was created.

- A *Location Content Created*: the location at which the content was created.

  This information especially applies to news, and may also be expressed as free text in the "dateline" of a story. A dateline may be found at the start of a news content component (i.e. the data content of a News Item), and represents a location and date of content creation, plus the name of the news provider.

  The location at which the content was created can be determined differently for different types of content:

  - Text: The practices of news providers either identify the location the content relates to or the location the content was created by a reporter or a writer. For some providers, if a correspondent is resident in town A but writes about an event in town B where the provider has a bureau, the name of town A or B is used, depending on the provider's policy, which should be available as a written document.
  - Photo: The location of origin of content is the place shown in the photo image.
  - Graphics: The location of origin of content should be the editorial office from where this graphics is produced.
  - Audio / video: In the case of raw footage the location of origin of the content should be the place of event, if people can be heard/are shown from different places, then the news provider can decide by its own policy, but this policy should be available as written document.

- Package: The location of origin of content should be the editorial office from where the items were packaged together.

- A *collection of Sources of Information*: an individual and/or company or organisation that originated some information used to create or enhance the content.

  - A *role*: a qualifier which specifies the role the entity plays in the creation of the content.

- A *collection of Creators*: an entity which created the content, preferably the name of a person (e.g. a photographer for photos, a graphic artist for graphics, or a writer for textual news).

  - A *role*: a qualifier which specifies the role the entity plays in the creation of the content.

- A *collection of Contributors*: an entity which modified or enhanced the content. Preferably the name of a person (e.g. a caption writer for photos).

  - A *role*: a qualifier which specifies the role the entity plays in the creation of the content.

- A *Significance to Audience*: the default significance of the content to the intended audience.

- A *collection of Intended Audiences*: an intended audience of the content.

  - An *exclusion*: a qualifier which indicates that this audience must be excluded (i.e. a logical NOT).

  - A *significance*: a qualifier which indicates that the expected significance of the content for this specific audience.

- A *collection of Services*: an editorial service to which an Item is assigned to by its provider. The values of this property are defined by each provider, and are often associated with the notion of a desk or a feed. Some examples are a "french wire" service, an "international picture" service or a "mobile news" service.

- A *collection of Editorial Notes*: a note addressed to the editorial people processing the content.

5.4.3 Descriptive Component

The Descriptive Component is a set of properties associated with the descriptive facet of content, i.e. data that can be inferred from "consuming" (reading, listening to, watching) the content:

- A *collection of Languages*: the language (or one of the languages) of the content.

- A *collection of Genres*: the nature, intellectual or journalistic characteristics of the content.

- A *collection of Subjects*: what the content is about.

- A *collection of Assert*: an assertion about a concept.

- A *collection of Slug Line*: a sequence of tokens describing the content, which get increasingly specific towards the end of the sequence.

  - A *Slug Separator*: a single character acting as a separator.

- A *collection of Titles*: a short human-readable name for the content. This label supports optional role, media and language attributes. Such a name is usually long-lived, for example, the common name of a person, organisation or location.

- A *collection of Headline*s: a narrative introduction to the content. This label supports optional role, media and language attributes. Such an introduction to the content is usually news-related, for example, the headline of a news story.

- A *collection of Description*s: a free-form textual description of the content. This label supports optional role, media and language attributes. This may be a copy of the first paragraph of a news story, the abstract of an event, or a computer-generated summary of a long feature document. A caption of an image may include details about the entities included (eg people, events and locations).

@@ What are the Language, Genre and Subject of a Package Item?

@@ What are the Language, Genre and Subject of a Topic Item? Should the subject be the concept identifier, and the genre be the concept type?

@@ Headline/Description: Are the semantics of these properties so different that two elements are needed? Should we rather implement the Headline as a variant of Description with a specific role?

5.4.4 Publication Component

The Publication Component is a set of properties associated with the publication of content:

This component is still to be defined.

5.4.5 Rights Component

The Rights Component is a set of properties associated with the ownership and usage rights of the content.

This component is still to be defined.

If no Usage rights are specified, then no specific restrictions on use of the content beyond contractual ones are being asserted.

5.4.6 Provider extensibility

The Content Metadata Component may be extended by via provider-defined aggregate components. These properties are also considered metadata (information about the Item content).

6   News Management

This section defines the processing model associated with an Item. It focuses on the areas where interoperability is a strong requirement.

6.1   Processing the status of an Item

6.1.1 Definition of the different status values

In this section, Item should be taken as Item and/or its content.

6.1.1.1 Publishing Status

The IPTC makes these values normative for the exchange of Items between a provider and its customers.

Usable: The Item MAY be published without restriction.

Withheld: Until further notice, the Item MUST NOT be published or used under any circumstances. If the Item has been published the publisher MUST take immediate action to withdraw or retract it.

Canceled: The Item MUST NOT be published or used under any circumstances. If the Item has been published the publisher MUST take immediate action to withdraw or retract it.

6.1.1.2 Status qualifiers

Embargoed: The Item MUST NOT be published until the date and time specified by the provider.

Retired: The Item MAY be published and accessed, but new references to this Item SHOULD NOT be created after the date and time specified by the provider.

Note: A Topic Item which is retired should include an additional Editorial Note, with the following content: "Retired – shall not be assigned after yyyy-mm-dd".

6.1.1.3 Publication property

Expired: The Item SHOULD NOT be published, and new references to this Item SHOULD NOT be created after the date and time specified by the provider. Depending on contractual conditions access to this Item may be forbidden.

6.1.2 Use cases

6.1.2.1 Use cases associated with a status of Withheld

Use case 1: a provider distributes a story as a news Item (version1). At a later stage he learns that there may be a problem with the information included in the Item. He sends a new version of the news Item (version 2) with a status equal to Withheld. All recipients systems must display a warning on this news Item, and recipient publishers must postpone the publication of the information contained in the news Item until further notice. The news provider can then choose between two solutions: one is to bring the news Item back to the Usable status, the other is to set the status to Canceled (both end-up with version 3).

Use case 2: an eCommerce system proposes a large collection of illustrated articles managed as news items. The publisher managing the system sees that the information included in a news Item (version 1) is not up to date anymore, and decides to hide this Item from its customers until it is properly revised. He set then its status to Withheld (version 2), edits the news Item and set its status back to Usable (version 3).

6.1.2.2 Use cases associated with a status of Expired

Use case: an eCommerce system proposes a large collection of illustrated articles managed as news items. The publisher managing the system sets up an expiration date for each news Item entering the system (version 1); when this date comes, the news Item is automatically hidden from the customers of the platform. The publisher is warned that a news Item expired. He is able to edit the news Item and move the expiration date to a future date, so that the customers of the platform can get access to it again.

6.1.3 Diagram

This depicts the state transition diagram reflecting the ways in which the *status* values are intended to be used. Thus, upon creation of an Item, allowed statuses are "Usable" and "Withheld". It is possible to withhold a "Usable" document; it is possible to release a 'Withheld' document; it is possible to cancel a "Usable" or 'Withheld' document. Once an Item has had its status set to "Canceled", it has reached a final state.
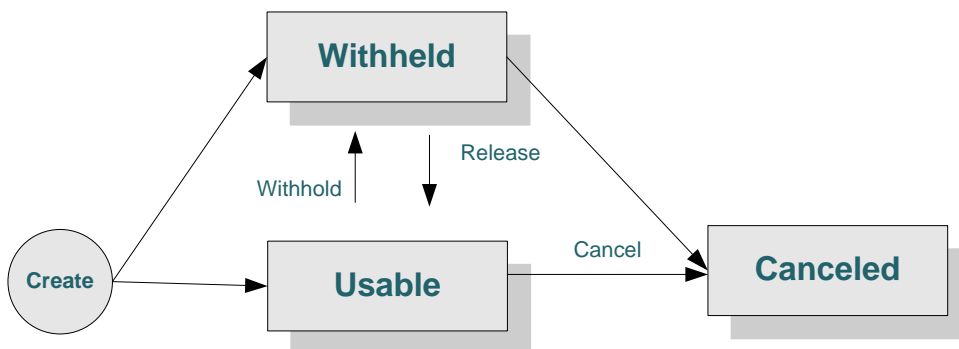
Embargoes are managed via the *dateEmbargoEnds* property. This property is optional; if present, the date and time it contains must be compared with the current date and time before the Item is used. The Item must not be published before this time. The embargo overrides the usability of the Item, inferred from its status.

Retirement is managed via the *dateRetired* property. This property is optional; if present, the date and time it contains must be compared with the current date and time before the Item is used. The Item must not be referenced after this time. The retirement overrides the usability of the Item, inferred from its status.

6.1.4 Processing model on the recipient side

Here is the processing model on the recipient side:

> Test status = "Canceled":
>> The Item must not be used, ever. Any usage of the Item must be prohibited, if needed by the way of alerts.

Else: next

Test status = "Withhold":

The Item must not be used until further notice. Any usage of the Item must be prohibited, if needed by the way of alerts.

Else: next

Test status = "Usable":

Test dateEmbargoEnds is not past => live status = Embargoed

Else: The Item is usable and may be safely published

7   News Item

7.1   Description

A *News Item* aims to convey news with the sense of the reporting of a news event. Its content is gathered by journalists, presented with a journalistic style, and updated according to the progression of the story.

A News Item may convey this information in any media type or format. It can also convey more structured news information, e.g. information about companies, sports events and general events, in instances when this information is related to an event or fact.

Typical characteristics of a News Item are:

- o   Short term interest for its content: news items are volatile, and people's interest in them fades as time passes ("nothing is older than yesterday's news").

- o   It is expressed via a set of alternative renditions of some media content.

- o   It will usually be updated only for a short period of time, as long as the covered event evolves, and then may be archived.

- o   It refers to an arbitrary set of concepts and entities.

- o   It may be associated with other News Items or web resources via typed links.

Examples of News Items are a news report, a picture, the graphical illustration of some event, a video clip or an illustrated biography.

A News Item is a kind of Any Item.

As such, a News Item supports management properties, administrative and descriptive metadata, and it may be associated with News Items or web resources via typed links. Providers may extend the information a News Item handles by implementing custom properties. At the power conformance level, a News Item supports supplemental properties relating to rights, publication and protection of the information it contains.

In addition to the information shared by all items deriving from Any Item, a News Item contains:

- A *Content Metadata Component*: this container wraps administrative, descriptive, rights and publication components.
  This component is described in detail in a previous section.

- A *News Content Set:* this container wraps a set of alternate renditions of the Item content. See below.

Providers may extend the information a News Item handles by implementing custom properties.

All specific components of a News Item are optional. This allows for a lightweight and progressive representation of information, suitable in the case of linkblogs for example. Nevertheless, some properties will be recommended in the guidelines, e.g. the News Item title and creator.

7.2   Illustration

This diagram below represents a News Item. At the top are sets of metadata. The News Item contains three alternative renditions of a picture: a thumbnail, a preview and a high-definition image.
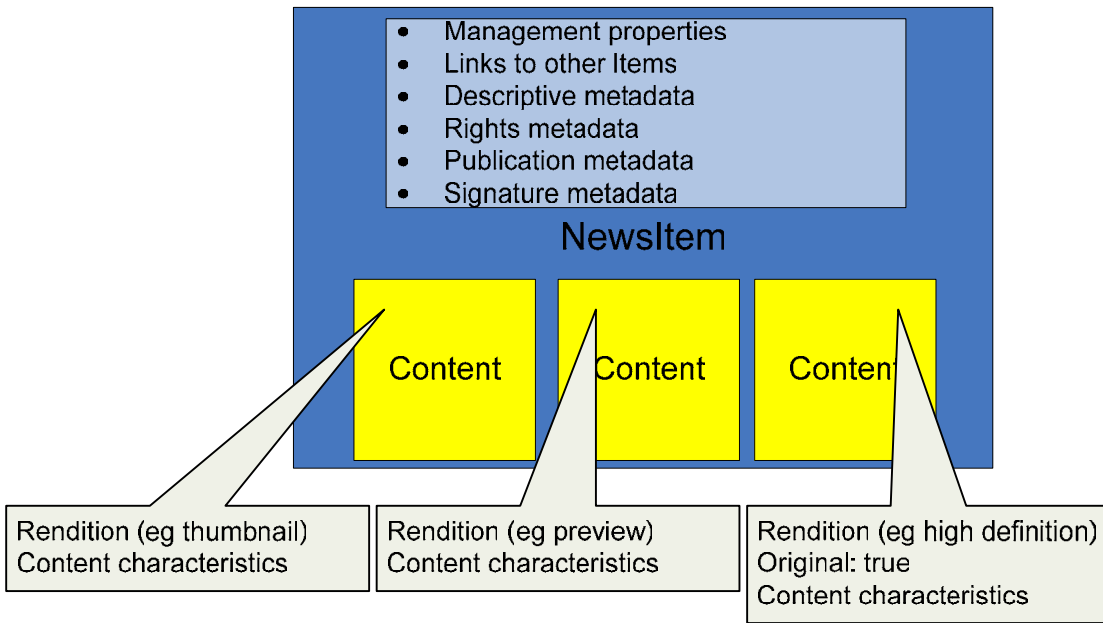
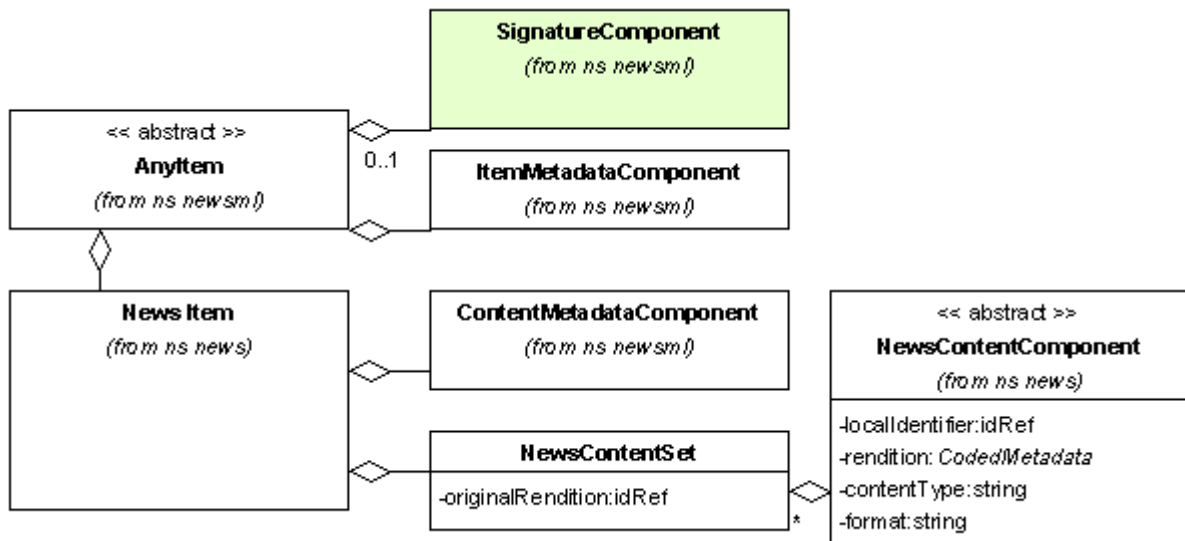## 7.3   UML diagram associated with the News Item

Diagram 2: News Content Component

7.4   News Content Set

A News Content Set contains:

- An *Original Rendition Reference*: a local reference to the original piece of content, from which all renditions have been derived. For example, the high-definition rendition of a picture from which a thumbnail has been processed will be referenced using this Original reference.

- *A collection of News Content Component*s: each component wraps one alternative rendition of the same content. Alternative renditions are different technical representation of the same news content. All alternative renditions can be derived from an "original rendition" by a software processor. For example: images in different resolutions, graphics as vector format or bitmap, text with different kinds of markup (eg. NITF or PDF). See below.

7.5   News Content Component

Content may be included by value or by reference, and useful characteristics are represented along with such content, in order to facilitate its processing.

News Content Component is an *abstract class* defined as a model for all types of content components. Such content may belong to any XML language capable of expressing generic or specialized news information, e.g .NITF, XHTML, SportsML or XBRL.
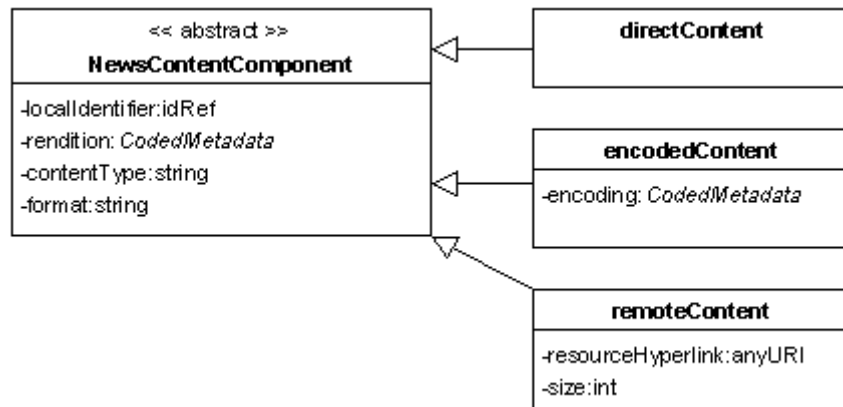
A News Content Component contains:

- A *Local Identifier*: the local identifier of the content component. Used especially in association with the Original Rendition Indicator.

- A *Rendition*: the specific rendition of content this component supports. This property helps the processor choose between alternative content components. Thus a picture may have pieces of content rendered as "thumbnail" or "preview", a text Item may contain an "sms", a "web" and a "print" rendition; values may be extended by individual providers.

- A *Content Type*: an IANA (Internet Assigned Numbers Authority) MIME type. This type applies to the content before any technical encoding needed to make the data xml-compliant (e.g. base64).

  Note that Content Type and Content Class (from the Item Metadata Component) are complementary. Content Class indicates the nature of the Item's content, but not the format of the components it contains: an Item can be a 'photo' with a gif thumbnail and a jpeg preview; a 'graphic' Item may contain a gif (bitmap) component and an eps (vector) component.

- A *Format*: a refinement of a generic Content Type (e.g. 'application/xml' or 'text/plain'), when no precise Content Type exists*.*

7.5.1 UML diagram associated with the News Content Component

7.5.2 Direct News Content

Direct News Content is a kind of News Content Component.

This component holds XML content or plain text, which is directly embedded in the component.

The XML language of the content is identified by the Content Type described above. The root element of this structure must be the root element of the language.

A special case is plain text, which is supported as well by this element, and which MUST be identified by the "text/plain" Content Type.

7.5.3 Encoded News Content

Encoded News Content is a kind of News Content Component.

Such content has been encoded before being embedded in the News Item structure. Binary content, like images, audio clips or even pdf or Word documents may be exchanged this way, but it is strongly recommended to use this structure for small assets only.

An Encoded Content Component contains:

- An *encoding*: it specifies the exact encoding applied to the content before inclusion in the component.
  At the core conformance level, the encoding is constrained to 'base64', and no specific attribute is needed.
  At the power conformance level, the encoding is free, and the encoding value must be set accordingly.

7.5.4 Remote News Content

Remote News Content is a kind of News Content Component.

Such content is not inline, but rather kept remote and referenced via a locator. Therefore the supporting element is empty.

A Remote Content Component contains:

- A *Resource Hyperlink*: the locator of the asset representing the content, as a URL.

- A *Size*: the size in bytes of the remote content.

7.5.5 Content Characteristics

Furthermore, specific content get specific characteristics, defined by specialized IPTC standards, like the Height and Width associated with a picture, the Word Count associated with a news story, or the Duration of an audio clip.

7.5.6 Power Conformance Level features

At the power conformance level, specific rights may be associated with each individual content component. For that purpose, the Rights Component, supported in the Content Metadata Component, may have a section dedicated to a specific News Content Component.

The mechanism is similar to the digital signature mechanism when applied to sub-parts of the Item.

7.6 Usage by an IPTC standard

An IPTC standard must use the News Item class as it is. When doing so is not sensible, a specialised Item must be derived from the Any Item class.

7.7 Provider extensibility

A News Item can be extended via provider-defined components added to the Content Metadata Component.

Content Characteristics can also be extended via provider defined properties. Examples of such properties are Adobe XMP or DIG35 metadata.

8   Topic Item

8.1   Description

A *Topic Item* aims to convey knowledge about concepts (named entities such as organisations or abstract notions such as news subjects). Typically a Topic Item itself holds only rather short and structured information about the topic and about its relationships with other concepts; it can refer to some external resources to support more information about the concept.

A Topic Item can be seen as a "hub" of information about a given concept.

We will introduce in the model the idea of TopicItem as a vehicle for descriptions of concepts, with parallel TopicItems existing, from diverse providers, about the same concept (same concept identifier).

Typical characteristics of a Topic Item are:

- o   Long term interest for its content.

- o   It will usually be updated infrequently but over a long period of time, as long as the covered topic evolves.

- o   It focuses on a single concept or entity.

- o   It is referred to by a large number of other resources (especially News Items and other Topic Items).

Examples of Topic Items are: a set of information about a person, a location, an organisation, a business sector, an animal, any object made by human work etc...

A Topic Item is a kind of Any Item.

As such, a Topic Item supports management properties, administrative and descriptive metadata, and it may be associated with News Items or web resources via typed links. Providers may extend the information a Topic Item handles by implementing custom properties. At the power conformance level, a Topic Item supports supplemental properties relating to rights, publication and protection of the information it contains.

In addition to the information shared by all items deriving from Any Item, a Topic Item contains:

- A *Content Metadata Component*: this container wraps administrative, descriptive, rights and publication components.
  This component is described in detail in a previous section.

- A *Topic Content*: this component contains properties specific to the concept.
  See below.

All specific components of a Topic Item are optional. Nevertheless, some properties will be recommended in the guidelines, e.g. the Topic Item title, creator, or concept identifier.

8.2   Illustration

The following diagram represents a Topic Item. At the top are sets of metadata, and among them is a Content Class indicating the nature of the topic (e.g. 'person', 'location', 'category', 'business sector'). Below is a structured set of information.
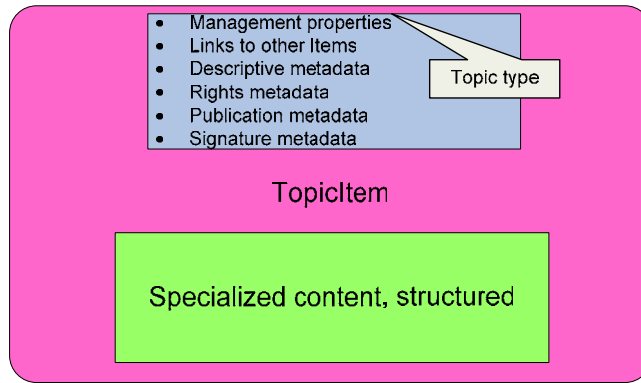
8.3   UML diagram associated with Topic item
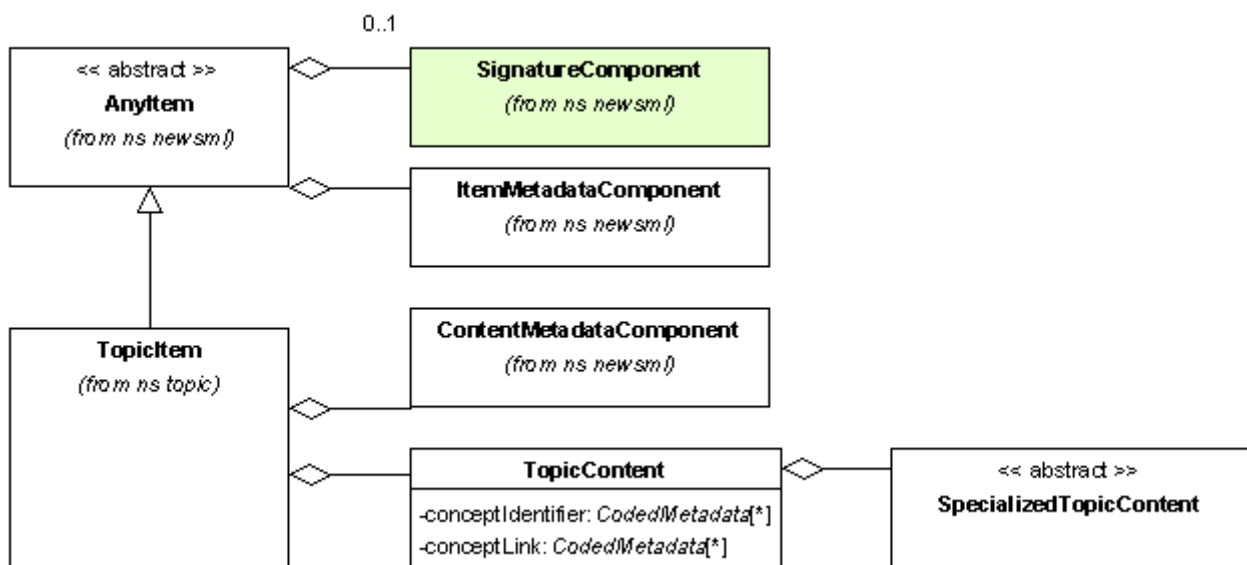
8.4   Item Metadata Component specific values

The Content Class from the Item Metadata Component (as defined in the Management Component, see NMAN-TS) identifies the type of content the Topic Item supports: its values may be 'person', 'organization', 'location', 'event', 'artefact', 'businessSector', 'theme' etc.

8.5   Topic Content

The Topic Content component supports a set of properties specific to the concept.

A Topic Content contains:

- *A collection of Concept Identifier*s: each identifier is a persistent, unambiguous identifier for the concept. It takes the form of a CURIE.

- A *collection of Concept Link*s: this is the reference to an associated concept. See below.

- *Specialized Topic Content*: some specific content associated with the concept. See below.

@@ One of our requirements (4.8.2) is to be able to request "a complete collection of TopicItems belonging to a given scheme". How shall we implement it?

@@ Should we add a concept type (so that a person has 'person' as content class and 'artist' or 'musician' as concept type)?

@@ Is a CURIE the proper format for Concept Identifier, or should it be a full URI?

8.5.1 Concept Link

A Concept Link contains:

- A *Target Concept Identifier*: the identifier of the target concept. It takes the form of a CURIE.

- A *Relationship Indicator*: a qualifier which specifies the name of the relationship between the two concepts. Example values are: 'broader' and 'narrower' (the base of a taxonomy), 'synonym' and 'preferredTerm' (in a thesaurus), or more specific values like 'hasParticipant' (btw an event and a person), 'hasManager' (btw an organisation and a person), 'hasCreator' (btw an artefact and an organisation).

- A *collection of Title*s: a short description of the relation between the source and the target concepts.

8.6   Specialized Topic Content

A Topic Item may contain any XML structure, defined as its content. This fairly flexible structure – even at the Core Conformance Level - allows for the inclusion of constructs representing information about a person, an organisation, a location etc.

Named entities are associated with specific content. The content of a TopicItem describing a person must be a Person Component; the content of a TopicItem describing an Event must be an Event Component, and so-on. Therefore the type of content expected in a TopicItem is directly related to its concept type.

Abstract concepts, like categories, usually don't support specific content, as enough information is provided in the generic TopicItem headers.

8.6.1 Event Component

This component is still to be defined.

8.6.2 Person Component

This component is still to be defined.

8.6.3 Organisation Component

This component is still to be defined.

8.6.4 Location Component

This component is still to be defined.

8.6.5 Contact Information Component

This component is still to be defined.

8.6.6 Postal Address Component

This component is still to be defined.

8.7   Usage by an IPTC standard

An IPTC standard must use the Topic Item class as it is, including the optional Topic Header properties. When doing so is not sensible, a specialised Item must be defined directly out of the Any Item class.

8.8   Provider extensibility

A Topic Item can be extended via provider-defined components added to the Content Metadata Component.

9   Package Item

9.1   Description

A *Package Item* facilitates the **packaging** of all kinds of items, from really simple constructs to the highly hierarchical structures created by some news providers.

Typical characteristics of a Package Item are:

- o   It provides some structure to news related information, and is expressed via a hierarchy of items.

- o   The items found in a Package Item stay independent from the package: they can be managed individually, and the package keeps only references to them.

- o   Its content is of medium term interest

Examples of Package Items are a collection of pictures, the "top ten" list of news items, an unordered set of news items relative to the same event, the representation of a news paper section of page.

A Package Item is a kind of Any Item.

As such, a Package Item supports management properties, administrative and descriptive metadata, and it may be associated with other Package Items, News Items or web resources via typed links. Providers may extend the information a Package Item handles by implementing custom properties. At the power conformance level, a Package Item supports supplemental properties relating to rights, publication and protection of the information it contains.

In addition to the information shared by all items deriving from Any Item, a Package Item contains:

- •   A *Content Metadata Component*: this container wraps administrative, descriptive, rights and publication components.
  This component is described in detail in a previous section.

- •   A *Group Set:* a hierarchical set of groups of items.
  See below.

9.2   Illustration

The following diagram represents a Package Item. It contains references to a News Item, a Topic Item and another Package Item.
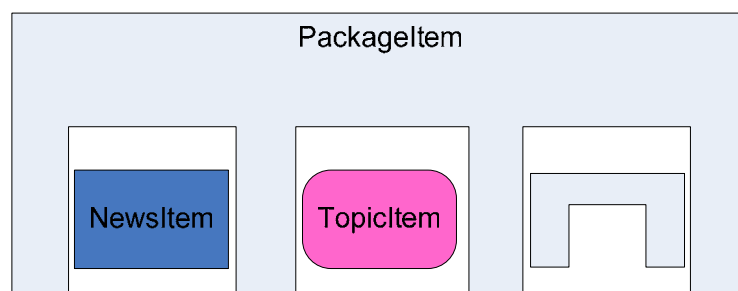
9.3   UML diagram
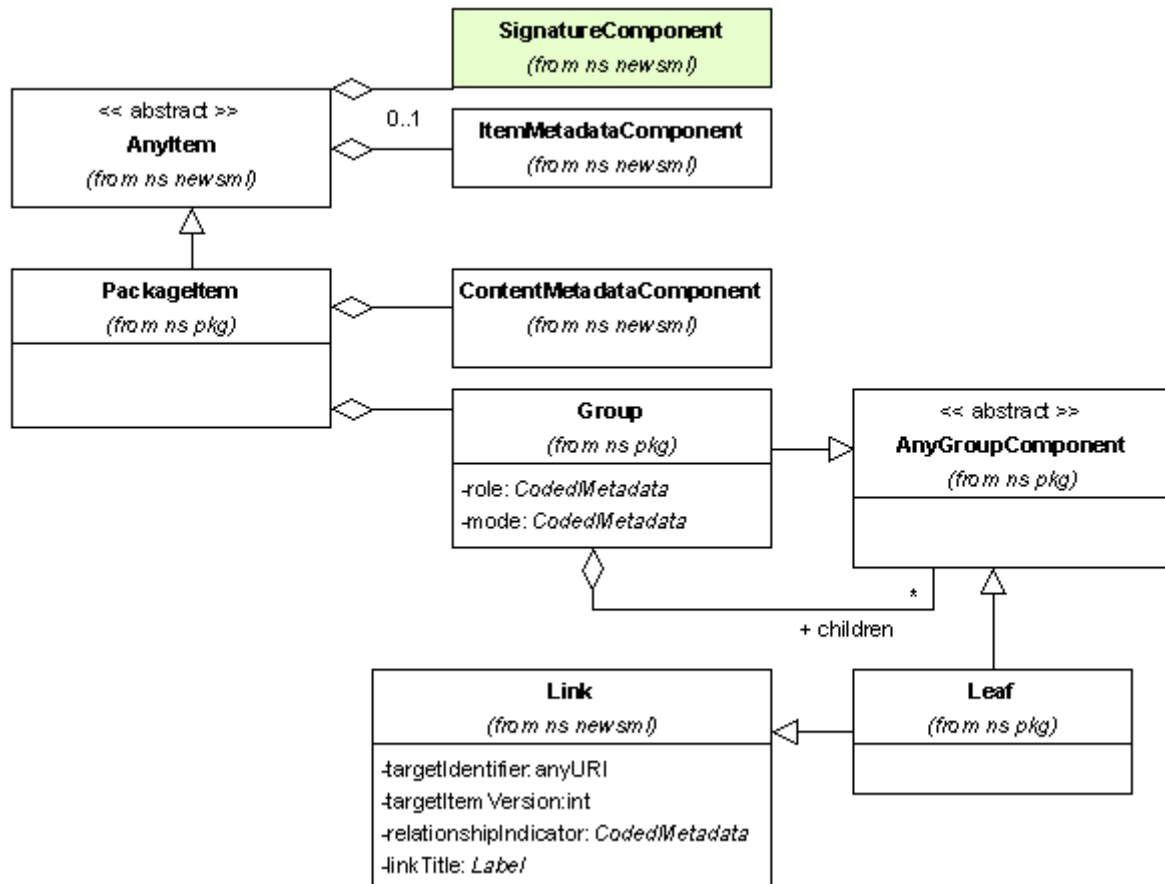
This structure is logically recursive. Note that the implementation of this model makes it a flattened structure: all groups are structurally sibling and each group holds references to other groups, as a way to avoid syntactic nesting.

9.4   Group Set

A Group Set represents a tree of Items. A specific Group Component acts as the root of the tree. It contains other Group Components and Items, in any order. The model follows a composite pattern, used to compose objects into tree structures to represent part-whole hierarchies. This pattern lets clients treat individual objects and compositions of objects uniformly.

9.5   Group Component

A Group Component contains:

- A *mandatory Group Role*: the part this group plays within its container (which is also a Group). The Group acting as root of the tree needs no explicit role.

- A *Group Mode*: an indication whether the elements in the group are "complementary and unordered", "complementary and ordered" or "alternatives" By default the group is "complementary and unordered".

- o Complementary and Unordered
  To be used for any kind of supporting content that does not require a sequence to be specified.
- o Complementary and Ordered
  To be used for any kind of content which must be displayed or consumed in a particular sequence, expressed by the order of the child elements of the group. The sequence could be a ranking. The semantics of the role attribute value determine the required processing.

- o Alternatives
  To be used if a group contains equivalent pieces of content (eg translations of the same news story into different languages). The recipient may pick one or more of these.

- A *collection of Group Component*s and *Link*s: The model allows Group Components and Links to be interleaved. Each link aggregates an external Item (or a web resource) to the package, and optionally indicates the relationship between the group and the target Item. Optionally, with each link, some 'hints' about the target Item are given, i.e. additional properties extracted from the target Item. This component is described in details in a previous section.

Implementers shall have a clear understanding of the difference between the aggregation (aka composition) links expressed as content of the package, and the navigation links a Package Item may support in its header.

9.6   Provider extensibility

A Package Item can be extended via provider-defined components added to the Content Metadata Component.

10 News Message

10.1   Description

A *News Message* facilitates the exchange of all kinds of items by any kind of IT transmission, especially in a broadcast or multicast network.

The use of a News Message is totally optional in a news workflow. Alternatively, items may be exchanged using SOAP, WebDAV, ICE, the Atom Publication Protocol (using Atom feeds, and items as payload of an Atom entry) or any other possible syndication protocol.

It may be useful for a recipient to store the information conveyed by a message, but this is not mandatory. Usually the messaging information will be maintained separately from the information relative to the contained items.


A News Message contains:

- A *Message Header*: A set of properties facilitating the exchange of items.

- An *Item Set:* A set of items. All Item classes derived from AnyItem – e.g. News Items, Topic Items and Package Items – can be exchanged this way.


10.2   Illustration

Here is an example of a message which contains several items. This message contains
- Three Topic Items: one may describe a concert (i.e. an event), the second a performer, the third the location of the concert.
- Two News Items: one may be a report of the event, the second a picture taken during the concert.
- One Package Item: it may reference two pictures and a video clip of the performer.

© 2005 International Press Telecommunications Council | www.iptc.org
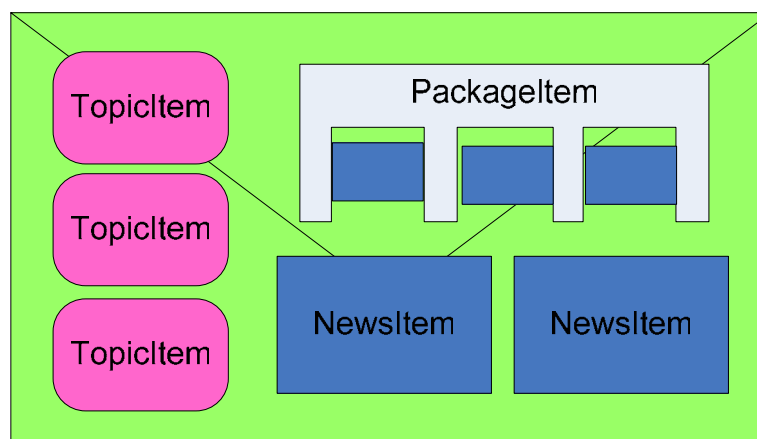
10.3   UML diagram
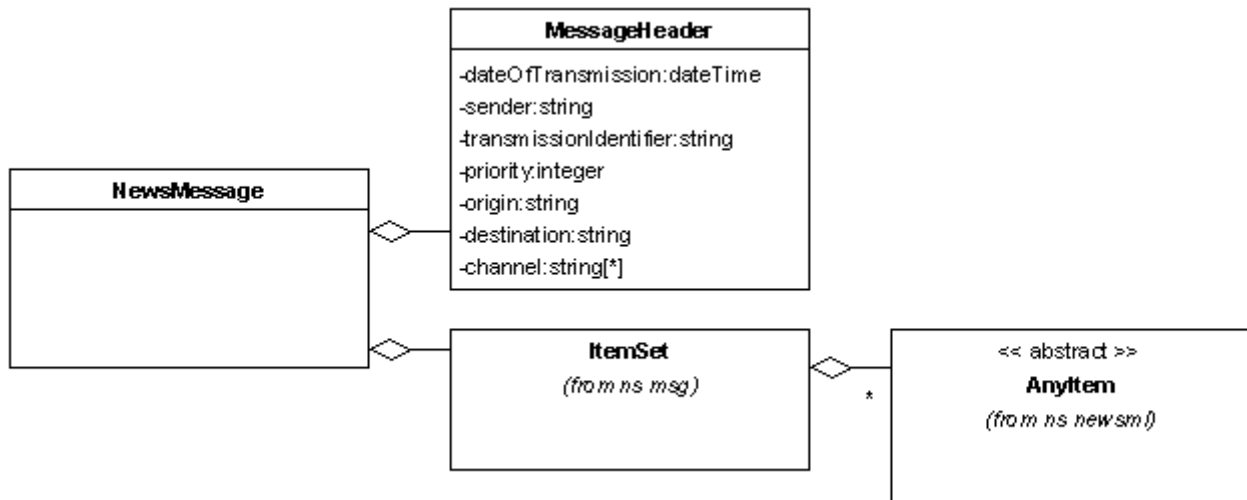
10.4   Message Header

The Message Header supports a set of properties facilitating the exchange of items.

A Message Header contains:

- A *mandatory Date of Transmission*: the date and time of transmission of the message. In case of retransmission it is not required to update this date.

- A *Sender:* the sender of the items, which may be an organisation or a person. The structure of this string is not specified by the IPTC. Best practice is to identify a sender by its domain name.

- A *Transmission Identifier*: The transmission identifier of the message. No two News Messages sent by the same sender on the same date can have the same identifier. In case of retransmission it is not required to update this identifier. The structure of this string is not specified by the IPTC.

- A *Priority*: the priority of transmission.

- An *Origin*: the point of origin of the transmission of the message. The structure of this string is not specified by the IPTC.

- A *Destination*: the point(s) of destination of the message. The structure of this string is not specified by the IPTC.

- A *collection of Channel*s: the transmission channel used by the message. A channel identifier is used to provide recipients with information on which select, route, or otherwise handle the content of the message. The structure of this string is not specified by the IPTC.

10.5   Item Set

An Item Set contains:

- A *collection of Item*s: The XML representation of an Item is directly included in the Item set.

The model assigns no significance to the order of Items within the News Message.

10.6   Provider extensibility

A News Message can be extended via provider defined properties. These properties represent contain message handling information, and are added to the Message Header.

11 References

11.1   IPTC documents

NML-BR	IPTC NewsML 2 Business Requirements http://newsml.org/dl.php?fn=NewsML_2.0-spec-BusinessRequirements_1.pdf
EVT-BR	IPTC EventsML Business Requirements http://www.iptc.org/download/dliptc.php?fn=EventsML/1.0-draft/specification/EventsML_1.0_spec_BusinessRequirements_4.pdf
NMDF-BR	IPTC News Metadata Framework Business Requirements http://www.iptc.org/dev/
NAR-TS	IPTC NewsML 2 Architecture Technical Specification http://www.iptc.org/dev/
NAR-GL	IPTC NewsML 2 Architecture Glossary http://www.iptc.org/dev/
NAR-IG	Implementation Guidelines for the IPTC Standards Architecture using W3C XML Schema NAR_1.0-doc-ArchitectureImplementationGuidelines_2, *not public*

11.2   Other references

XSD	W3C XML schema http://www.w3.org/XML/Schema
CSS	Cascading Style Sheets, level 2 revision 1 http://www.w3.org/TR/CSS21
CURIE	Internal draft produced by the W3C RDF-in-HTML task force http://www.w3.org/2001/sw/BestPractices/HTML/2005-10-27-CURIE
GRDDL	Gleaning Resource Descriptions from Dialects of Languages (GRDDL) http://www.w3.org/2004/01/rdxh/spec
N3	Notation 3 http://www.w3.org/DesignIssues/Notation3.html
RDF	Resource Description Framework (RDF) http://www.w3.org/RDF/
Wikipedia	"the free encyclopedia that anyone can edit". http://en.wikipedia.org/wiki

12 Change Log

Draft 1&2:
- Merge of the NSTR Model plus the CoCo model plus the NMDF model.

Draft 3:
- Addition of Administrative Component and Descriptive Component.
- Inclusion of agreed changes, from A.Karben's proposals
- Addition of a list of open issues.
- Addition of the model of Labels.

Draft 4:
- Addition of the model of Links.
- Addition of the model of Signature.
- Addition of the model of Formal Types.

Draft 5:
- Change in the order of descriptive metadata.
- Update of the Link description.
- Date Item Generated -> Date Item Created

Draft 6:
- Add processing model for status.
- Treat MS comments part 1 (from 8/11)
- Add entity components
- Add open issues

Draft 7:
- Item header and content header named back metadata component.
- News Metadata introduction updates (-> MS)
- Formal Metdata Type renamed Code Metadata Type, introduction updated (-> MS)
- Labels: Alternative description provided by MS
- Item Management Alternative Location: description updated, role added
- Item Link updated
- News Content Component updated after confcall
- Person, Location, Contact Info component updated (for discussion)
- LLM016 issue added

Draft 8:
- Topic Item: concept identifier simplified (as a CURIE, with no separate scheme/code)
- News Message: Item Set added, conforming to agreed modifications (AK), channel multiple (as NewsProduct was in NewsML1).
- Package Item: Group Reference added, to better reflect the model.

Draft 9:
- Intended Service renamed Service
- ItemMeta: Editorial Warning repeatable

- Descriptive Metadata: title modified, slugline suppressed and headline added
- Package Item: Group Reference suppressed, to better reflect the *logical* model.
- Open issues erased for this doc. To be maintained separately.

Draft 10:
- Original Identifier moved
- Definition of Alternative identifier updated
- "item" changed for "Item".
- Update of the Link model

Draft 11:
- Clean-up for public distribution

Draft 12:
- Warning renamed to signal
- (5.3.2) Instance Links removed
- CURIE is not a W3C note, but an internal draft from the W3C RDF-in-HTML task force
- (4.3.2) addition of the mechanism in use for resolving CURIEs (extracted from the RDF/A Primer, Mark Birbeck).
- Correction of the package item model and UML diagram.
- Additional clean-up

=== END of document ===