NewsML Support Working Party
NewsML 1.2 - "Expert Zone"
Multilingual news

Laurent Le Meur
AFP Multimedia development Team
laurent.lemeur@afp.com

(written from excerpts of a mail exchange on the NewsML mailing group, between Shum Kwan, Khalid Shahen, Mischa Wolf, Jo Rabin, Daniel Rivers-Moore, David Megginson, Irvine Levine, Edward Middleton ; **modified after IPTC Aarhus, Denmark, meeting in June 2003**)

What is the standard use of xml:lang?

xml:lang is used to indicate the language of character data contained within the element that includes this attribute (ie the language of XML PCDATA, for example the language of a metadata tag).

XML 1.0 (Second Edition) Section 2.12 Language Identification says:
"In document processing, it is often useful to identify the natural or formal language in which the content is written. A special attribute named xml:lang may be inserted in documents to specify the language used in the contents and attribute values of any element in an XML document. "

An errata (http://www.w3.org/XML/xml-V10-2e-errata#E11) explains that xml:lang values are taken from RFC 3066.

RFC 3066 codes are defined as

- 2-letter subtags interpreted as ISO 639 language codes (or 3-letters subtags from ISO 639 part 2, or i for IANA-defined registrations, or x for private use)
- and optionally a dash followed by
- 2-letter subtags interpreted as ISO 3166 alpha-2 country codes (or 3 to 8 letters IANA codes).

Examples : 'en', 'en-US', 'en-GB', 'fr-FR', 'fr-CA', 'zh-Hans', 'zh-Hant', 'mas', 'i-klingon' etc...

In NewsML, xml:lang can be found in the following list of elements:

- NewsItem
- NewsComponent
- RightsMetadata/Copyright/CopyrightHolder and CopyrightDate
- RightsMetadata/UsageRights/ UsageType, Geography, RightsHolder, Limitation,s StartDate and, EndDate
- DescriptiveMetadata/Location *
- [NewsLines=] HeadLine, SubHeadLine, ByLine, ByLineTitle, DateLine, CreditLine, CopyrightLine, RightsLine, SeriesLine, SlugLine, NewsLineText
- Comment
- Topic/Description

* Note: the presence of xml:lang in Location could be a NewsML dtd mistake, as location values should be defined as controlled vocabularies. What is users opinion on this issue?

What is the definition of DescriptiveMetadata/Language?

As the NewsML Functional specification states: "The Language element indicates the, or a, language used in a content item. The value of the FormalName attribute is a formal name for the Language element. Its meaning and permitted values are determined by the controlled vocabulary identified by the Vocabulary and Scheme elements."

Note: this NewsML Language element is equivalent to the Dublin Core Language element, which is also used to describe the language of the resource.

Using the RFC3066 as the base for a controlled vocabulary providing the language formal names is clearly a very good practice. IPTC does provide an IETF language TopicSet which includes RFC3066 compatible language codes used by the IPTC membership.

Note about character encoding:

Each XML document must be in a single character encoding. Consequently, unless the NewsML document is encoded using Unicode (e.g. UTF-8), "en-US" and "zh-HK" stories must be placed in separate documents.

Could we use a special ContentItem Characteristics element?

The DTD states that Characteristics element is to be used to describe physical properties of content. It is arguable that language is not a physical property. This would especially be the case where the content if of a non-textual nature - e.g. audio. The Language information clearly applies to the item as INFORMATION OBJECT and so it is part of the NewsComponent wrapper, in addition to anything that may need to be said about them AS DATA OBJECTS, at the ContentItem level (as at the NITF text layer).

So how do we set the language of news content ?

Let's consider a news item - written in Unicode - that carries news in several languages. The clearer case is when the news item is constituted of several information objects (NewsComponents), each of them being written or spoken in a certain language (multiple translation).

The agreed approach is to use the **NewsComponent / DescriptiveMetadata / Language** element to indicate the language of the resource (content) being described by each inner NewsComponent.

This indication is only given at the NewsComponent level. The language information given in a NewsComponent is inherited by all the inner NewsComponents, and only superceded by a redefinition of the language value in an embedded NewsComponent.

How do we compose translated news?

Using the above rule, it is easy to embed news and its translation in one NewsItem, using two NewsComponents.

Sample 1:

```
<?xml version="1.0" encoding="UTF-8"?>
<NewsML>
    <NewsEnvelope><DateAndTime>20030303T000000Z</DateAndTime></NewsEnvelope>
    <NewsItem>
        <Identification/>
        <NewsManagement/>
        <NewsComponent EquivalentsList="yes">
            <BasisForChoice>DescriptiveMetata/Language</BasisForChoice>
            <AdministrativeMetada>
                <Provider><Party FormalName="AFP"/></Provider>
            </AdministrativeMetada>
            <DescriptiveMetada>
                <SubjectCode><Subject FormalName="01000000"/></SubjectCode>
            </DescriptiveMetada>
            <NewsComponent>
                <DescriptiveMetada>
                    <Language FormalName="en-US"/>
                </DescriptiveMetada>
                <ContentItem Href="http://www.test.org/english-audio.mp3">
                    <MediaType FormalName="Audio"/>
                </ContentItem>
            </NewsComponent>
```

```
            <NewsComponent>
                <DescriptiveMetada>
                    <Language FormalName="fr-FR"/>
                </DescriptiveMetada>
                <ContentItem Href="http://www.test.org/french-audio.mp3">
                    <MediaType FormalName="Audio"/>
                </ContentItem>
            </NewsComponent>
        </NewsComponent>
    </NewsItem>
</NewsML>
```

In this sample, the main NewsComponent of the NewsItem holds the metadata shared by two NewsComponents (in the sample, Provider and SubjectCode). Each of those NewsComponents holds audio content, and the spoken language is given via the Language element.

This approach holds if the content is some text embedded in the NewsComponent:

Sample 2:

```
<NewsML>
    <NewsEnvelope><DateAndTime>20030303T000000Z</DateAndTime></NewsEnvelope>
    <NewsItem>
        <Identification/>
        <NewsManagement/>
        <NewsComponent>
            <AdministrativeMetada>
                <Provider><Party FormalName="AFP"/></Provider>
            </AdministrativeMetada>
            <DescriptiveMetada>
                <Language FormalName="en-US"/>
                <SubjectCode><Subject FormalName="01000000"/></SubjectCode>
            </DescriptiveMetada>
            <ContentItem>
                <MediaType FormalName="Text"/>
                <DataContent>english story</DataContent>
            </ContentItem>
        </NewsComponent>
    </NewsItem>
</NewsML>
```

How do we set the language of content metadata?

As seen before it would not be appropriate to use xml:lang to give information about content that is referred to rather than included directly, as it can be the case in a NewsComponent.

In NewsML, xml:lang indicates the language of the resource *description* (language dependant metadata like DescriptiveMetadata and RightsMetadata; NewsLines, Comment and Topic Description).

This indication can be given at the NewsItem or NewsComponent level, or it can be given at the level of the descriptive element. The language information given in an xml element is inherited by all the inner xml elements, and only superceded by a redefinition of the language value in an embedded element.

Consider, for example, a NewsML package with an English description of an Italian videoclip: @xml:lang would be set to "en-US" rather than "it-IT", Language would be set to "it-IT".

Sample 3:

```
<NewsML>
```

```
    <NewsEnvelope><DateAndTime>20030303T000000Z</DateAndTime></NewsEnvelope>
    <NewsItem>
        <Identification/>
        <NewsManagement/>
        <NewsComponent xml:lang="en-US">
            <Comment>english comment</Comment>
            <AdministrativeMetada>
                <Provider><Party FormalName="AFP"/></Provider>
            </AdministrativeMetada>
            <RightsMetadata>
                <UsageRights><UsageType>allowed usage type</UsageType></UsageRights>
            </RightsMetadata>
            <DescriptiveMetada>
                <Language FormalName="it-IT"/>
                <SubjectCode><Subject FormalName="01000000"/></SubjectCode>
            </DescriptiveMetada>
            <NewsLines>
                <HeadLine>US english title</HeadLine>
            </NewsLines>
            <ContentItem Href="http://www.test.org/italian-video.mpg">
                <MediaType FormalName="Video"/>
            </ContentItem>
        </NewsComponent>
    </NewsItem>
</NewsML>
```

In this sample, NewsComponent/@xml:lang defines the language of Comment, HeadLine and UsageType.

It is good practice to define at the NewsComponent level the language of the metadata used in the NewsComponent, so multilingual news items that contain translated metadata and translated content can be accurately described.

Sample 4:

```
<NewsML>
    <NewsEnvelope><DateAndTime>20030303T000000Z</DateAndTime></NewsEnvelope>
    <NewsItem>
        <Identification/>
        <NewsManagement/>
        <NewsComponent EquivalentsList="yes">
            <BasisForChoice>@xml:lang</BasisForChoice>
            <AdministrativeMetada>
                <Provider><Party FormalName="AFP"/></Provider>
            </AdministrativeMetada>
            <DescriptiveMetada>
                <SubjectCode><Subject FormalName="01000000"/></SubjectCode>
            </DescriptiveMetada>
            <NewsComponent xml:lang="en-US">
                <Comment>multilingual sample</Comment>
                <NewsLines>
                    <HeadLine>US english title</HeadLine>
                </NewsLines>
                <RightsMetadata><UsageRights><UsageType>allowed usage
type</UsageType></UsageRights></RightsMetadata>
                <DescriptiveMetada>
                    <Language FormalName="en-US"/>
                </DescriptiveMetada>
                <ContentItem Href="http://www.test.org/english-video.mpg">
                    <MediaType FormalName="Video"/>
                </ContentItem>
            </NewsComponent>
            <NewsComponent xml:lang="fr-FR">
                <Comment>exemple multilingue</Comment>
                <NewsLines>
                    <HeadLine>titre français</HeadLine>
                </NewsLines>
                <RightsMetadata><UsageRights><UsageType>type d'utilisation
accordée</UsageType></UsageRights></RightsMetadata>
```

```
                <DescriptiveMetada>
                    <Language FormalName="fr-FR"/>
                </DescriptiveMetada>
                <ContentItem Href="http://www.test.org/french-video.mpg">
                    <MediaType FormalName="Video"/>
                </ContentItem>
            </NewsComponent>
        </NewsComponent>
    </NewsItem>
</NewsML>
```

In this sample, NewsComponent/@xml:lang defines the language of Comment, HeadLine and UsageType. Language defines the video content language.

How do we compose a multilingual description of some news?

We can also describe a unique ContentItem with alternative descriptive elements, in several language; to do this, the xml:lang attribute must be set at the descriptive elements level, as in :

Sample 5:

```
<NewsML>
    <NewsEnvelope><DateAndTime>20030303T000000Z</DateAndTime></NewsEnvelope>
    <NewsItem>
        <Identification/>
        <NewsManagement/>
        <NewsComponent>
            <Comment xml:lang="en-US">multilingual sample</Comment>
            <Comment xml:lang="fr-FR">exemple multilingue</Comment>
            <AdministrativeMetada>
                <Provider><Party FormalName="AFP"/></Provider>
            </AdministrativeMetada>
            <RightsMetadata>
                <UsageRights><UsageType xml:lang="en-US">allowed usage type</UsageType></UsageRights>
                <UsageRights><UsageType xml:lang="fr-FR">type d'utilisation
accordée</UsageType></UsageRights>
            </RightsMetadata>
            <DescriptiveMetada>
                <Language FormalName="en-US"/>
                <SubjectCode><Subject FormalName="01000000"/></SubjectCode>
            </DescriptiveMetada>
            <NewsLines>
                <HeadLine xml:lang="en-US">US english title</HeadLine>
                <HeadLine xml:lang="fr-FR">titre français</HeadLine>
            </NewsLines>
            <ContentItem Href="http://www.test.org/english-audio.mp3">
                <MediaType FormalName="Audio"/>
            </ContentItem>
        </NewsComponent>
    </NewsItem>
</NewsML>
```

Here an English spoken audio clip is described in US English and French in the same NewsComponent.

How do we describe multilingual content, that is content expressed in more than one language?

Note that the specification specifically allows there to be several Language elements. So if a given NewsComponent contains more than one language, this can be covered by using multiple Language elements.

Hence there would seem to be a number of cases to consider:

a) NewsComponent contains single language element and single content item - implies that the content item is in the language stated
b) NewsComponent contains multiple language elements and a single content item - implies that the content item contains all the languages cited
c) NewsComponent contains single language element and multiple content items - implies that they are all in that language
d) NewsComponent contains multiple language elements and multiple content items - this would seem to mean that all the content items are in all the languages.

How do we tag multilingual text content?

Textual content gives a finer way to tag multilingual content.

Let's consider a text mainly written in English, but with a sentence in French. NITF defines a "lang" element and a "lang" attribute. This "lang" attribute can be found in the "body" element, and indicates the language of the target audience; an embedded string in a different language (as a quote) can be tagged with a "lang" element, or an element that holds a "lang" attribute.

For more information on this subject, please look at NITF documentation.

Sample 6:

```
<body lang="en-US"><body.content>
    <p>will the "french fries" be renamed "freedom fries" ? <q><lang lang="fr-FR">"c'est ridicule, et de toute façon les
    frites sont d'origine belge"</lang></q> told us a french citizen.</p>
</body.content></body>
```

--- end of document ---