

Topic Maps, NewsML and XML: Possible Integration and Implementations.
By Soelwin Oo.

Introduction

This paper will discuss how the integration of different Topic Map based technologies can lead to the development of powerful knowledge based resource retrieval systems.

The paper will discuss in detail the possible implementation for integrating a data resource that supports Topic structures with the knowledge embodied within a Topic Map. It will discuss this using examples of technology currently being developed by empolis illustrating the possible architecture of such a system and its potential real world use.

Finally, the paper will investigate the potential for further integration and scalability of the system with other Topic Map resources. More specifically, it will elaborate on the possible hurdles and pitfalls that may arise from the integration of data from multiple resources and the possible need for managing ontologies originating from different sources.

Topic Maps

Topic Maps allow the capturing of concepts and the creation of relationships between these particular concepts. Topics can represent real world entities or abstract concepts.

Through the concepts, Topic Maps allow the imposing of knowledge structure upon raw data. This data can then be navigated using the knowledge-based relationships that are within the corresponding Topic Map. Furthermore, this knowledge can be used to produce a more powerful method for resource retrieval. This ability is provided through the use of inference based queries that utilise the ability for intelligent information gathering enabled by the Topic Map structure.

NewsML

NewsML is a structured flexible framework based on XML developed by the IPTC (International Press Telecommunications Council) for electronic news based publication. It supports the representation of news items and the relationships between these news items in an XML based structure. Because NewsML possesses associated metadata concerning its news content, it provides the ability for having multiple representations of the same information along with provision for handling arbitrary mixtures of media types, languages and formats.

The prime interest towards NewsML within the scope of Topic Maps is that NewsML possesses metadata concerning Topics that provide the ontology of its news content. This 'news item ontology' puts forward an appropriate example for an opportunity to 'capture' concepts presented by an XML based format that supports Topic structures. Once the

base ontologies used within NewsML are present within a Topic Map, an application can process NewsML documents and present to the user the instances of the base ontologies that are associated with a NewsML document. This will then present a content driven approach for navigation of a Topic Map because the user's starting point will be the base ontologies instantiated by the NewsML document.

NewsML and Topic Map Integration

I will first discuss a potential method for transforming ontology data held within NewsML into data that can be captured within a Topic Map.

The main repositories for concepts within NewsML are its 'Vocabularies'. These 'Vocabularies' or 'Topic Sets' are XML NewsML documents that collect together related Topics.

The following shows an example of an IPTC Topic Set:

```
<TopicSet Duid="iptc.importance" FormalName="Importance">
  <Comment xml:lang="en">Relative significance of the metadata applied to a NewsComponent.</Comment>
  <Topic Duid="imp1">
    <TopicType Scheme="IptcTopicType" FormalName="Importance"/>
    <FormalName Scheme="IptcImportance">High</FormalName>
    <Description xml:lang="en">The metadata is very important.</Description>
  </Topic>
  <Topic Duid="imp2">
    <TopicType Scheme="IptcTopicType" FormalName="Importance"/>
    <FormalName Scheme="IptcImportance">Medium</FormalName>
    <Description xml:lang="en">The metadata is quite important.</Description>
  </Topic>
  <Topic Duid="imp3">
    <TopicType Scheme="IptcTopicType" FormalName="Importance"/>
    <FormalName Scheme="IptcImportance">Low</FormalName>
    <Description xml:lang="en">The metadata is of low importance.</Description>
  </Topic>
</TopicSet>
```

The above Topic Set represents the set of Topics concerned with the different levels of '*Importance*' one may wish to attach towards a particular subject. If we examine a particular Topic representation more closely, we see that it conveys the following properties for that Topic in question:

- Formal Name
- Description
- Topic Type

The NewsML excerpt below describes the properties of the Topic called '*High*':

```
<Topic Duid="imp1">
  <TopicType Scheme="IptcTopicType" FormalName="Importance"/>
  <FormalName Scheme="IptcImportance">High</FormalName>
  <Description xml:lang="en">The metadata is very important.</Description>
</Topic>
```

In order to capture the properties of this particular Topic within a Topic Map, one could represent the information in the following way.

FormalName

```
<FormalName Scheme="IptcImportance">High</FormalName>
```

The '*FormalName*' element above describes the name of the Topic. The name in question is '*High*' in the '*Scheme*' of '*IptcImportance*'. If we were to create this Topic within a Topic Map, we would create a Topic that possessed a *baseName* with a *baseNameString* represented by the character string '*High*' and this would be scoped by the Topic with the name of '*IPTCImportance*'.

For one to accomplish this, the Topic with the name '*IptcImportance*' must already exist. The XTM representation of the above Topic Map would appear as shown below:

```
<topicMap>
  <topic id="t-IptcImportance ">
    <instanceOf>
      <subjectIndicatorRef xlink:href="http://www.TopicMaps.org/xtm/1.0/index.html#topic" />
    </instanceOf>
    <baseName>
      <baseNameString>IptcImportance</baseNameString>
    </baseName>
  </topic>

  <topic id="t-High">
    <instanceOf>
      <subjectIndicatorRef xlink:href="http://www.TopicMaps.org/xtm/1.0/index.html#topic" />
    </instanceOf>
    <baseName>
      <scope>
        <topicRef xlink:href="#t-IptcImportance " />
      </scope>
      <baseNameString>High</baseNameString>
    </baseName>
  </topic>
</topicMap>
```

Here we see two Topics, one named '*IptcImportance*' that has its name in the unconstrained scope and a second Topic called '*High*' that has its name scoped by the first Topic '*IptcImportance*'.

The Topic named '*IptcImportance*' is represented within the Topic '*High*' by its unique Topic ID. This ID is portrayed as an attribute of the '*topicRef*' child-element of the '*scope*' element.

From this initial step, we now possess our first NewsML ontology Topic that possesses a *baseName* in a given scope.

Description

```
<Description xml:lang="en">The metadata is very important.</Description>
```

The '*Description*' element above provides a more verbose description of the NewsML Topic. In the context of a Topic Map, this description can be considered as another *baseName* for our current Topic. To distinguish it from other *baseNames* that the Topic may possess, we scope this *baseName* with a Topic called '*Description*'.

If we look more closely at the '*Description*' element, we see that it has the attribute '*xml:lang*' with the value of '*en*'. This indicates that the description for the Topic is in the language of English. To represent this within the context of our current topic, we would scope the name currently scoped by the Topic '*Description*' with a second Topic called '*xml:lang=en*'. This Topic called '*xml:lang=en*' would represent the concept of the language English.

With the addition of this second *baseName*, our Topic '*High*' will have the following *baseNames* as shown by the XTM representation below:

```
<topicMap>
  <topic id="t-High">
    <baseName>
      <scope>
        <topicRef xlink:href="#t-IptcImportance " />
      </scope>
      <baseNameString>High</baseNameString>
    </baseName>
    <baseName>
      <scope>
        <topicRef xlink:href="#t-Description " />
      </scope>
      <scope>
        <topicRef xlink:href="#t-xml:lang=en " />
      </scope>
      <baseNameString> The metadata is very important.</baseNameString>
    </baseName>
  </topic>
</topicMap>
```

Within the NewsML specification and K42 Topic Map Engine, only one Topic can possess a particular name within a particular scope. This unique name-scope pairing allows one to unambiguously specify a certain Topic from an individual name-scope pair.

This means that the above Topic can be referred to without ambiguity by specifying it using either of its names with their corresponding scoping Topics.

We now possess within our Topic Map a Topic representing the concept of *High Importance*.

TopicType

```
<TopicType Scheme="IptcTopicType" FormalName="Importance"/>
```

This element indicates that our Topic '*High*' is an instance of another Topic. The Topic, which is its Class, is referred to by its *FormalName* '*Importance*' which is scoped by the Topic '*IptcTopicType*'.

The XTM representation of our Topic '*High*' would then appear as shown below:

```
<topicMap>
  <topic id="t-High">
    <instanceOf>
      <topicRef xlink:href="#t-Importance" />
    </instanceOf>

    <baseName>
      <scope>
        <topicRef xlink:href="#t-IptcImportance" />
      </scope>
      <baseNameString>High</baseNameString>
    </baseName>

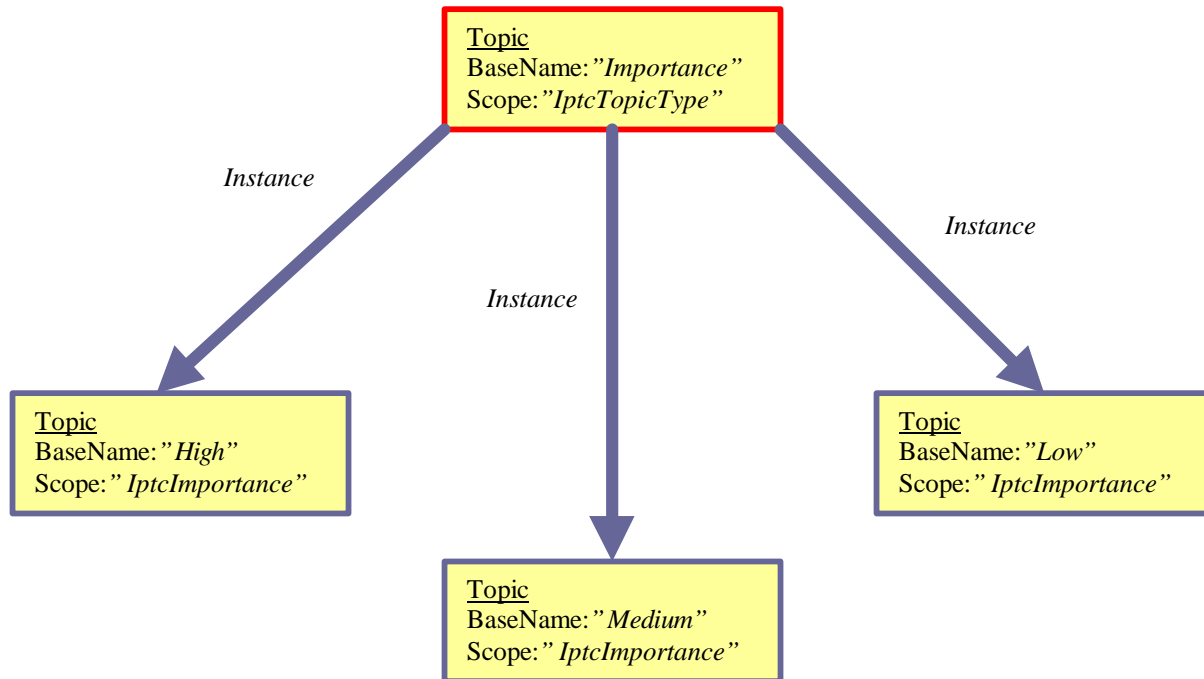
    <baseName>
      <scope>
        <topicRef xlink:href="#t-Description" />
      </scope>
      <scope>
        <topicRef xlink:href="#t-xml:lang=en" />
      </scope>
      <baseNameString> The metadata is very important.</baseNameString>
    </baseName>
  </topic>
</topicMap>
```

The Type Topic is represented above by its unique Topic ID. This ID is portrayed as an attribute of the '*topicRef*' child-element of the '*instanceOf*' child-element of the Topic '*High*'.

We now possess our first NewsML Topic represented within our Topic Map. If we process the rest of the Topics within the Topic Set '*Importance*', we will produce a small set of Topics that are uniquely identifiable and organised into class/instance relationships where all the Topics will be of Topic Type '*Importance*'.

Hence we can traverse this relationship to retrieve Topics by inferring their relation to other Topics. For example, if we start with the Topic '*High*', we can query its Type and see that it is of Topic Type '*Importance*'. If one then retrieves all the Topics of Type '*Importance*', we can obtain a set of Topics that represent the different types of '*Importance*' that can be used to describe a subject.

This relationship is shown in the following diagram:



The other Topic Set vocabularies within NewsML are organised into such TopicType/Instance hierarchies and each Topic Set is grouped into similarly related groups of Topics.

If we apply the same process to the other Topic Sets as we have done with the Topic Set ‘*Importance*’, we obtain a Topic Map populated by all the Ontology Topics within NewsML.

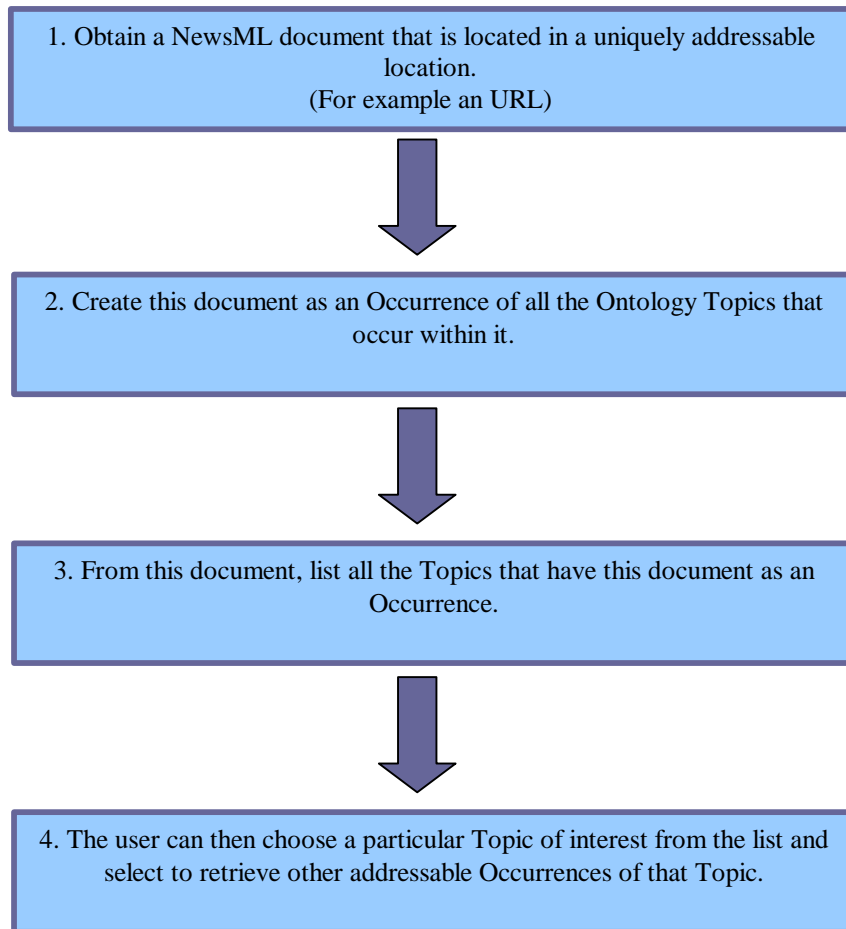
NewsML and Topic Map Implementation

This process of transforming Ontology data held within NewsML into Topics is a rather straightforward process. In order to obtain useful implementation from this set of Topics, one must create Topic Associations between Topics. This will allow us to convey some form of relationship between Topics representing a framework of knowledge portrayed by particular sets of associations.

Currently our sets of NewsML Topics are only related by TopicType/Instance relationships and this occurs only within each Topic Set as one Topic Set is unrelated to another.

Even with just this type of inter-Topic relationship, we can implement some useful form of resource retrieval. First, because Topics can possess Occurrences of themselves represented as addressable Resources, we can relate to individual NewsML articles as occurrences of the Ontology Topics that appear within them.

This is possible because NewsML documents possess metadata that convey the subject matter or Topics present within the news article. If we use this information, we can use the Topic Map to retrieve meaningful sets of related resources in the following way:



In this way, a user is presented with a content driven approach for navigation of a Topic Map. This is because the user's starting point will be the base ontologies instantiated by the NewsML article and the user may now explore the list of instantiated Topics along with their associated Occurrences that represent sets of related resources.

Implementation of Topic Association driven Resource retrieval

As mentioned previously, by using Topic Associations, we can create useful relationships between Topics that capture some form of useful knowledge. These Topic Associations can then be used to extract useful sets of resources.

I will give an example of how one may create a '*filtering mechanism*' using Topic Associations that can be used to organise the Occurrences of NewsML Ontology Topics into useful groups or *Channels*.

The process for this would involve the following:

- Create a Topic Association Template called 'Channel'
- Create the following members to populate this template:

Member 1

Role >> **Channel Topic** to be played by Topic Type >> **Channel Topic**

Member 2

Role >> **Ontology Topic** to be played by Topic Type >> **Ontology Topic**

The possible arcs between the members are:

from Channel Topic to Ontology Topic: includes the topic of

For example:

Sport Channel <includes the topic of> Football

from Ontology Topic to Channel Topic: is assigned to

For example:

Football <is assigned to> Sport Channel

- We can now instantiate this ‘Channel’ Topic Association Template by associating the respective ‘Channel Topics’ with their respective NewsML Ontology topics:

For Example:

Associated with ‘Sport Channel’:

*Football
Basket Ball
Cricket
Sporting Competition
Etc...*

Associated with ‘Business Channel’:

*Nasdaq
Company
FT Index
Etc...*

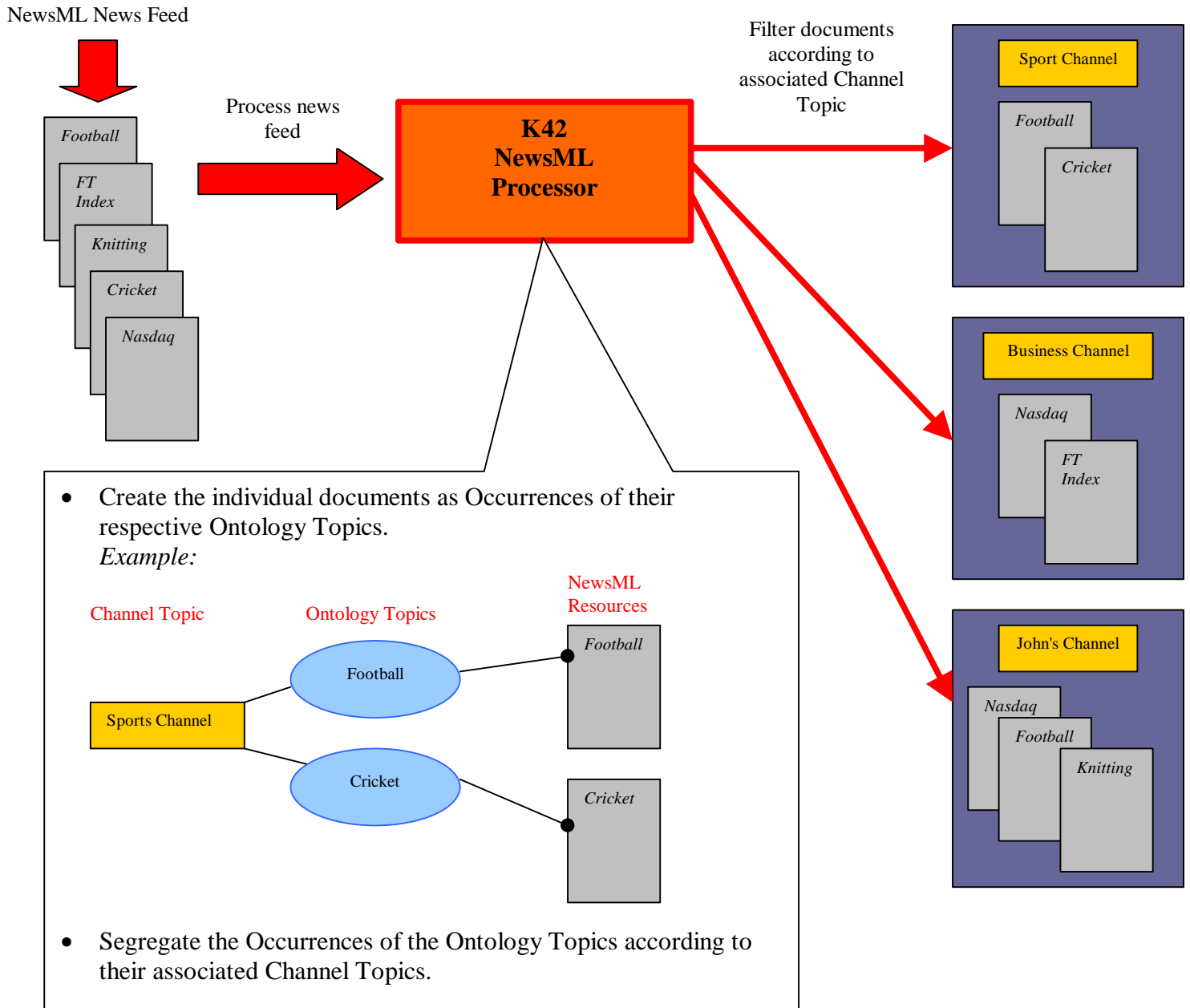
Associated with ‘John’s Channel’:

*Nasdaq
Football
Knitting
Etc...*

Once these associations are created they can be used to filter incoming NewsML documents as follows:

- Assuming all NewsML ontologies have been captured as Topics and organised into ‘Channels’, one can use a NewsML processor to parse the incoming NewsML documents and note the occurrences of the specific ontology topics within the documents.
- The documents will then be created as occurrences of all the Ontology Topics that occur within them.
- According to the ‘Channel Topic’ that these ontology topics are associated with, the document will be copied to a location segregated for documents representing the ‘Channel’ that the instantiated ontology topics are associated with.
- Once the segregation has been performed, a user can ‘view’ individual ‘Channels’ that are represented by sets of NewsML documents that include subject matter associated with the particular predefined ‘Channels’. A particular set of NewsML ‘Channel Documents’ can then be displayed by alternately presenting the latest set of documents within that channel in order to display a ‘News Ticker’ of that channel.

This process is shown in the diagram below:



This illustrates just one example of how Topic Map Technology can be used to retrieve data and resources. With the introduction of additional Topic Associations and Topics, it is possible to create richer queries that harness the knowledge embodied by the many relationships that exist between Topic Map Objects.

Integration of Multiple Topic Map Ontologies

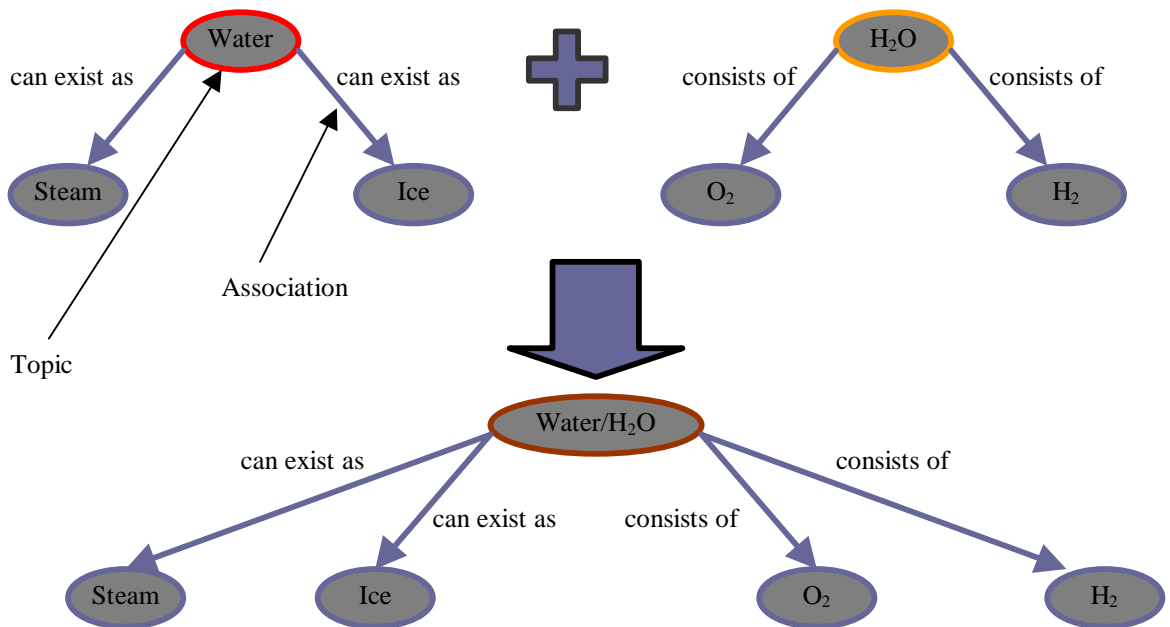
NewsML permits the creation of additional custom vocabularies that can be published along with individual NewsML documents. These vocabularies may contain additional Topics that may not be included within the official set of IPTC Ontology Vocabularies.

With the creation of these custom vocabularies, it is not terribly difficult to see the potential for *duplication* of Topics. The context for the term *duplication* does not refer to the names Topics possess but the concepts that they embody.

For example, the concept of 'water' is universally understood but is referred to by many names depending on one's current geographical local. Within a Topic Map, one would represent the Topic that embodies the concept of 'water' by attributing to it many *baseNames* that translate the term 'water' into many languages. These *baseNames* will then be scoped by the respective language Topics.

Now envisage another Topic Map that contains a Topic that embodies the concept of 'water'. This Topic may be referred to by the term ' H_2O ' and describes the concept of 'water' from a scientific point of view. These two Topics refer to the same concept but from differing viewpoints. If we *merged* these two Topics that refer to the concept of 'water', we will eventually possess one Topic that embodies all the concepts of the previous two Topics. A significant benefit of this would be to merge the relationships and associations possessed by both Topics into one focal point. This then provides a richer network of knowledge as associations and relations that were previously unconnected now have a means of connecting to each other.

This example can be shown below:



One of the major obstacles accompanying merging of Topic Map Ontologies arises because of the difficulty in identifying duplicate Topics or concepts that may exist within more than one Topic Map. The following describes some possible solutions to this problem:

- **Unique Topic ID** - Within a Topic Map, a Topic is attributed a unique ID. This can be used to unambiguously identify and refer to that particular Topic. If two Topics have identical IDs even though their baseNames may differ, they can be regarded as referring to the same Topic and the two Topics can be merged.

Unfortunately, this assumption can only be true if the unique ID scheme is respected. The authors of both Topic Maps must be aware of the specific IDs of each other's Topics and assign IDs to their Topics accordingly.

- **Unique basename-scope pairs** - I have mentioned previously that within the Topic Map, only one Topic can possess a particular name within a particular scope. This unique name-scope pairing allows one to unambiguously specify a certain Topic from an individual name-scope pair. This will allow us to identify identical Topics in two separate maps and merge the Topics.

Unfortunately, this assumption also requires that the authors of both Topic Maps are aware of each other's basename-scope naming schemes and match up their Topics accordingly.

- **Name/ID Mapping Table** - An alternative solution is to use a *mapping table* that can translate between sets of Topic naming schemes. This allows one to match up Topics according to their translation on the mapping table and merge them according to the matched translation.

Unfortunately, this method requires one to possess such a mapping table that is capable of translating between different ontologies. The problem lies within the task of creating such a table that comprehensibly covers all possible terms and Topics.

- **Common Reference Ontology** - This method is similar to the previous solution where by all Topic Map authors adhere to a globally official standard ontology of Topics. This allows all Topic Map authors to produce maps populated by standard Topics that will provide an easy mechanism for Topic matching and merging.

Unfortunately, this solution also requires the generation of a comprehensive Ontology that covers all the possible Topics that a Topic Map creator may require. Also the problem of Topic matching still arises if a Topic Map author wishes to use his or her own custom Topic ontologies.

As one can see, there is no panacea for merging multiple Topic Map ontologies. Until such a solution appears, there will always be an obstacle surrounding the merging of multiple ontologies, hindering the ability to scale knowledge maps across multiple Topic Maps.

Conclusion

I have illustrated the simple steps required to process Ontology Topics represented in an XML format into Topics that can be represented within a Topic Map.

Once these Ontology Topics have been captured within a Topic Map, I have shown how one may provide a simple method of resource retrieval by using simple Occurrence relationships within the Topic Map. Additionally, I have shown an example of a more powerful means of resource retrieval and filtering using Topic Association relationships.

These few examples show how one can create simple yet powerful mechanisms using Topic Map Technology to filter and extract data and resources. Also, one must not forget that useful information can be equally gleaned from navigation within the Topic Map. This is afforded because the relationships and associations within a Topic Map capture knowledge that provides a semantic framework for the individual concepts that are represented by the Topics.

Finally, I have illustrated the potential benefit of further integration of multiple Topic Map resources. More specifically, I have elaborated on the possible hurdles and pitfalls that may arise from the integration of data from multiple resources and the possible need for managing ontologies originating from different sources.