Importance of metadata in the multimedia news delivery

Darko Gulija

HINA

Contents:

- Intoduction: the problem
- XML as a news delivery standard
- NewsML
  - – structure, metadata, design principles
- NewsML features
  - – unique identification, versioning, content identification, content selection
- Conclusion: the wider picture

Introduction: the problem

- Rapid growth of information quantity
- Decrease in information quality
- Metadata: data that describes information content
  - enables information processing without processing the content
  - problem: relevance of the metadata to the content

Introduction: the problem

- Characteristics for news delivery format
  - Open, platform independent, widely accepted and easy transferable
  - Able to include or reference arbitrary mixture of media types, languages and encodings
  - Reach and flexible metadata structure including provenance of the data and the content
  - Relationships and manageability of the data

XML as a news delivery standard

- Advantages of XML
  - – open, platform independent, widely addopted
  - – base for numerous W3C and industry standards
  - – unicode support: language transparency
  - – excelent in linking and referencing the data
  - – rich and flexible data structure
  - – document hierarchy corresponds to the data sturcture

XML as a news delivery standard

- Metadata attachment
  - **Attributes:**

```
<Content  type=“heading”>
Heading </Content>
```

  - **Elements:**

```
<Content>
        <Lang variant="en-us">
        English </Lang>
        <Cont>Heading</Cont>
</Content>
```

  - **ID/IDREF attributes**

```
<Content id=“CE001”>
Heading </Content>
<Metadata idref=“CE001”
type=“heading” />
```

  - **XPath/Xpointer**

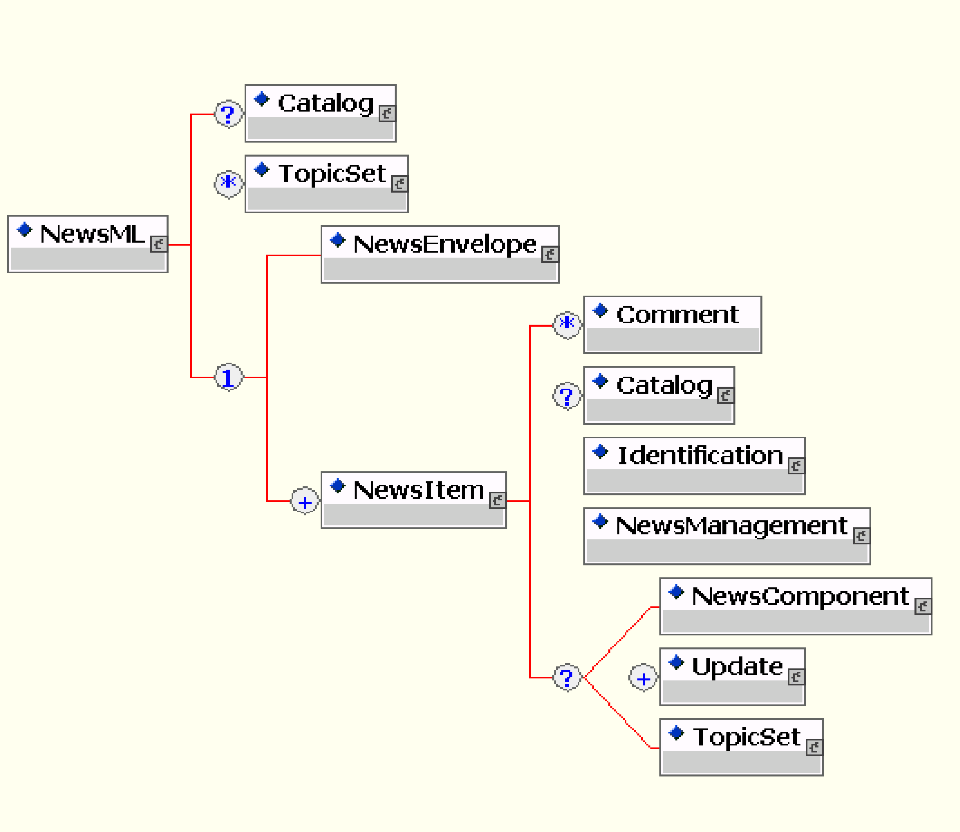
```
<Content> Heading </Content>
<Metadata ref=“../Content”
type=“heading” />
```

NewsML

- compact, extensible and flexible XML framework for news
- supports representation of electronic news items, metadata and relationships between them
- handles arbitrary media types, formats, languages and encodings
- support all stages of the news lifecycle
- allows insertion of provenance of metadata and news content

NewsML: structure

- NewsML hierarchy
  - NewsEnvelope = transport data
  - NewsItem = event (news)
  - NewsComponent = news object instance (text, photo, audio)
  - ContentItem = renderable content



NewsML: metadata

- NewsEnvelope
  - TransmissionID, SentFrom, SentTo, DateAndTime, NewsService, NewsProduct, Priority
- NewsItem
  - Identification:
    - Formal Identification: NewsIdentifier
      - Contains URN (PublicIdentifier)
    - Informal Identification: NameLabel, DateLabel, Label
  - NewsManagement:
    - NewsItemType, FirstCreated, ThisRevisionCreated, Status, StatusWillChange, Urggency, RevisionHistory, DerivedFrom, AssociatedWith, Instruction, Property

NewsML: metadata

- NewsComponent
  - Content selection:
    - Role, BasisForChoice, @EquivalentsList, @Required,@xml:lang
  - Content description:
    - AdministrativeMetadata: FileName, SystemIdentifier, Provider, Creator, Source, Contributor, Property
    - RightsMetadata: Copyright, UsageRights, Property
    - DescriptiveMetadata: Language, Genre, Subject, OfInterestTo, TopicOccurence, Property, Metadata
  - NewsLines - publishable metadata:
    - HeadLine, ByLine, DateLine, CreditLine, CopyrightLine, RightsLine, SeriesLine, SlugLine, KeywordLine, NewsLine

NewsML: metadata

- ContentItem
  - MediaType, Format, MimeType, Notation
  - Characteristics
- Metadata provenance
  - enables judging the metadata quality
  - may be included in most of the metadata elements
    - @AssignedBy, @Importance, @Confidence, @HowPresent, @DateAndTime

NewsML: design principles

- RULE: use the most basic XML feature
  - 3 criteria for external standards:
    - formal ratification, tool support, public understanding
  - Metadata attachment throuth document structure
  - References through fragment identifiers (*#Duid*)
  - XPath for defining targets
- Not used:
  - Namespaces: eliminate validation
  - RDF: uses Namespaces and lacks tool support

NewsML features

- Unique identification
  - – Every NewsItem has a globally unique identifier
    - – urn:newsml:{ProviderId}:{DateId}:{NewsItemId}:{RevisionId}{RevisionId@Update}
      - » urn:newsml:hina.hr:20000101:H9261234:1N
  - – Element identification: *Duid* and *Euid* attributes
    - – <ContentItem Duid="CI001">
      - » urn:newsml:hina.hr:20000101:H9261234:1N#CI001
    - – <ContentItem Euid="CI001" >
      - » #xpointer(//ContentItem[@Euid="CI001"])

NewsML features

- Versioning and correction management
  - – referencing the previous version
    - » <NewsItem><NewsManagement>
      <DerivedFrom NewsItem=“urn:newsml:hina.hr:20000101:1234” />
  - – sending full updates
    - » <NewsItem><Identification>
      <ProviderId>hina.hr</ProviderId><DateId>20000101</DateId>
      <NewsItemId>1234</NewsItemId>
      <RevisionId PreviousRevision=“1” Update=“N”>2</RevisionId>
      <PublicIdentifier>urn:newsml:hina.hr:20000101:1234:2N
      </PublicIdentifier></Identification>.....
      <ContentItem> UPDATED CONTENT</ContentItem></NewsItem>

NewsML features

- Versioning and correction management
  - sending incremental updates
    - » <NewsItem><Identification>
      <ProviderId>hina.hr</ProviderId><DateId>20000101</DateId>
      <NewsItemId>1234</NewsItemId>
      <RevisionId PreviousRevision="2" Update="U">3</RevisionId>
      <PublicIdentifier>urn:newsml:hina.hr:20000101:1234:3U
      </PublicIdentifier></Identification>.....
      <Update><Replace DuidRef="#CI001">
      <ContentItem>REPLACED CONTENT></ContentItem></Replace>
      </Update></NewsItem>
    - Problem: how to request the missing copy (for update)

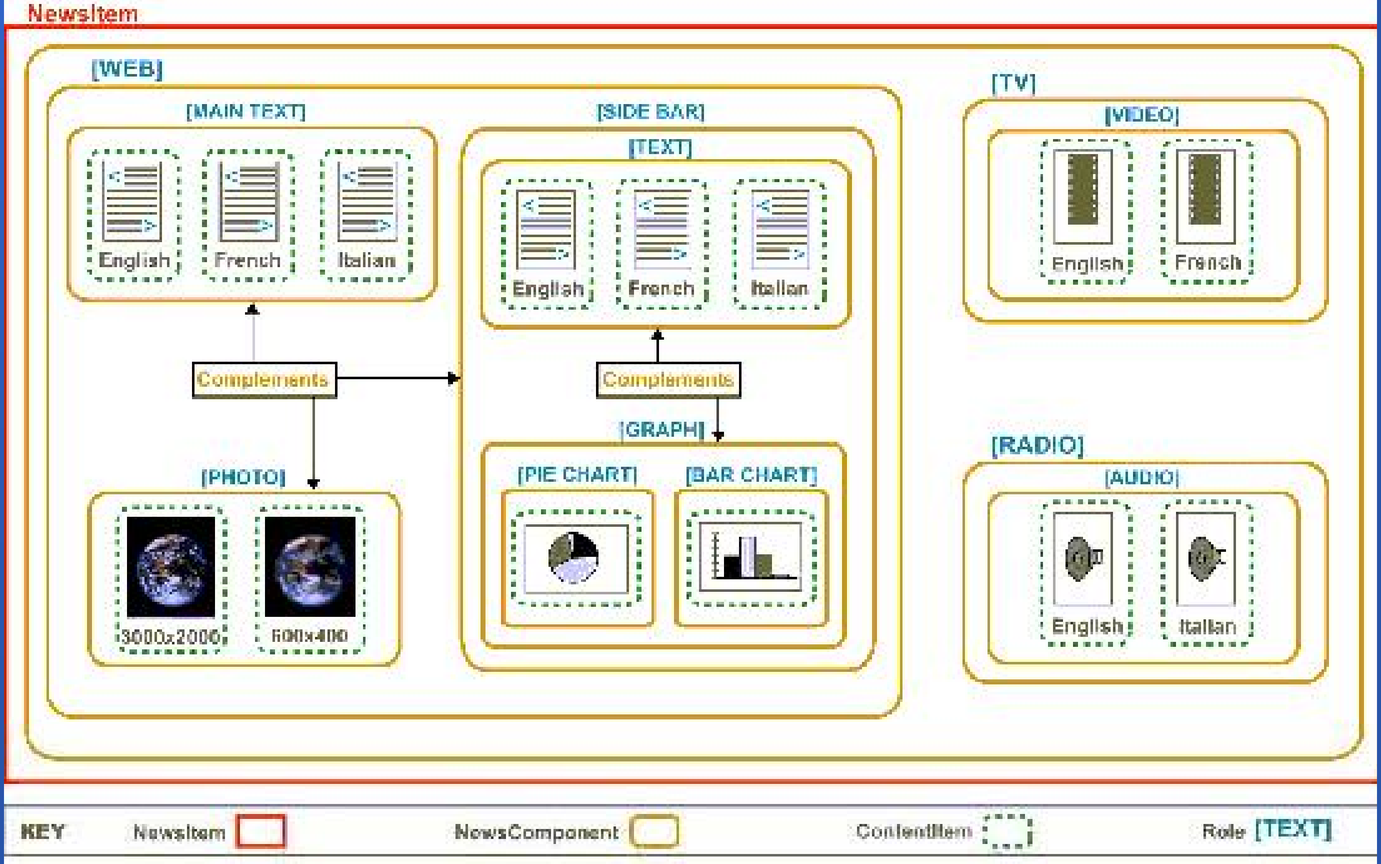
NewsML features

- Content identification and controlled vocabularies

  -
    ```
    <TopicSet Duid=“iptc.subject” FormalName=“Subject”>
    <Topic Duid=“sr15000000”>
        <TopicType Scheme="IptcTopicType" FormalName=”Subject"/>
        <FormalName Scheme="IptcSubject">15000000</FormalName>
        <Description xml:lang=“en “>Sport </Description></Topic>
    </TopicSet>
    ```

  -
    ```
    <ContentItem><Catalog><Resource>
        <Urn>urn:newsml:iptc.org:20001006:IptcSubjectCodes</Urn>
        <DefaultVocabularyFor Scheme=“IptcSubject” Context=“Subject” />
    </Resource></Catalog,>  ....
     <DescripteveMetadata  AssignedBy=“HINA” Confidence=“High”>
        <SubjectCode><Subject FormalName=“15000000”/></Subject>
    </DescriptiveMetadata>.....
    ```

NewsML features



NewsML features

- Choosing the right content

```
» <NewsComponent EquivalentsList=“Yes”>
   <BasisForChoice>./Role/@FormalName</BasisForChoice>
   <NewsComponent EquivalentsList=“No”>
     <Role FormalName=“WEB”>
     <NewsComponent EquivalentsList=“Yes” Essential=“Yes”>
       <Role FormalName=“MAIN TEXT”>
       <BasisForChoice>./ContentItem/@xml:lang</BasisForChoice>
       <ContentItem xml:lang=“en”>English content</ContentItem>
       <ContentItem xml:lang=“fr”>French content</ContentItem>
     </NewsComponent>
  ……
```

Conclusion: the broader picture

- NewsML is only an envelope format
  - Transport: ICE | HTTP | FTP | SMTP | SOAP
  - Envelope: NewsML
  - Content: NITF | existing multimedia formats | industry specific XML standards
    - ICE: XML-over-HTTP request/response protocol

Conclusion: the broader picture

- NewsML as a general purpose data wrapper
  - XML based: easy transfer and numerous tools
  - rich and flexible metadata structure
  - powerful linking capabilities and rich content models
- Importance of metadata
  - gives meaning to the content, enables its automatic processing and improves its usefulness

Conclusion

Content and its relationships are the essence of the Internet:

METADATA RUNS THE INTERNET