Centrum voor Wiskunde en Informatica

Bringing NewsML2 into the Semantic Web

Raphaël Troncy

George Anadiotis

raphael.troncy@cwi.nl

K-Space

Passepartout

ITEA

Why Bother with Metadata?

- A News agency is a content provider
  - Content (stories, photo, video, etc.) are assets
- Metadata add value to these assets as they provide human and machine readable information about them
- Metadata is much more than just a bunch of keywords added at the end of the chain so the customer can find your image
- Metadata covers all information about an asset, which enables machines to do smart things with your assets

2

Why Bother with Semantics?

- High quality *semantic* multimedia metadata enables:
  - Easy exchange of news items
  - Semantic search of particular news items
  - Delivery of personalized news content to customers
    - Interactive browsing in a news archive
    - Cross-modality: packaging the news stories, photos, graphics, audio, videos
    - For different end-user platforms (mobiles, PC, handhelds, etc.)

3

IPTC Metadata Standards

- Metadata "fields"
  - Informal definition and guidelines to use the field according to its semantics
  - e.g. "Date Created": content creation date ≠ digital representation creation date

Property name: **Creator**
User interface label: Creator
Description: Contains preferably the name of the person who created the content of this news object, a photographer for photos, a graphic artist for graphics, or a writer for textual news. If it is not appropriate to add the name of a person the name of a company or organisation could be applied as well.
Note(s): Aligning with IIM notions IPTC Core intents to have only one creator for this news object despite the underlying XMP property dc:creator allows for more than one item to be included. If there are more than one item in this array the first one should be considered as the IPTC Core Creator value.

XMP Schema specifications:

XMP Category: External XMP Value Type: Seq ProperName
XMP Path: dc:creator/*[1]

4

IPTC Metadata Standards

- Metadata "values"
  - Expressed as *controlled* vocabularies (standardization bodies)
  - A vocabulary is composed of terms (flat list, taxonomy organization)
  - IPTC has defined 28 sets of multilingual News Codes
    - NewsCodes use numeric strings = language agnostic
    - Ex: Subject ≈ 1300 terms, 3 levels hierarchy in 4 languages
    - NewsCodes Viewer application View
- XML Wrapper
  - Metadata embedded in a photo: XMP
  - Metadata stored in a separate file: NewsML

Problem: XML and Semantic *)

うかを検出するために、文書の完全性を保証することです。しかしながら多くのアプリケーションは、XML 文書にまず署名をし、その後文書を改変することで、その文書の一部を暗号化しようと考えています。復号化変換では、署名の確認に先立ち、文書を改変前の状態に戻し、文書のどの部分を復号化すればよいかをデータ受信者に通知します。

業界リーダーや暗号の専門家らの幅広い支持とともに、既に実装もされている XML Encryption

W3C の XML Encryption ワーキンググループによってまとめられた実装及び相互運用性報告書に示されているように、数多くのアプリケーションや他の仕様が既に XML Encryption を利用しています。特に、配送データのセキュア化が必要な Web サービス仕様群が本仕様の利用を進めています。また多くの企業が XML Encryption の実装についてその支持と計画を表明しています。

XML Encryption は、Baltimore Technologies、BEA Systems、DataPower、IBM、Microsoft、Motorola、ジーゲン大学、Sun Microsystems、VeriSign の各 W3C 会員と個人技術者とで構成される、W3C の XML Encryption ワーキンググループによって策定されました。

World Wide Web Consortium [W3C] について

W3C は、Web の発展と相互運用性を確保するための共通のプロトコルを開発することにより、Web の可能性を最大限に引き出すべく設立されました。W3C は、アメリカ合衆国マサチューセッツ工科大学計算機科学研究所 (MIT/LCS)、フランス国立情報処理自動化研究所 (INRIA)、及び日本の慶應義塾大学がホスト機関として共同運営にあたっている国際産業コンソーシアムです。コンソーシアムにより提供されるサービスには、開発者及び利用者のための World Wide Web に関する豊富な情報、新技術を応用した様々なプロトタイプやサンプルアプリケーションの開発などが挙げられます。現在までに、450近くの組織がコンソーシアムの会員となっています。詳しくは http://www.w3.org/ をご参照下さい。

<News>
<Subject>...</Subject>
<References>...
</References>
<Testimonial>...
</Testimonial>
<Presentation>...
</Presentation>
</News>

subject

references

testimonial

company presentation

⇒ **Need for formal semantics for the content**

*) adapted from Frank van Harmelen

Problem: interoperability

- Different management applications may label the same field differently
  - e.g. Creator / By-Line (Author) / Author / By-Line
- The informal semantics (guidelines) of the various metadata fields prevent an automatic validation of their use

⇒ **Need for formal semantics for the structure**

7

Role of the Semantic Web

- "Oh no! Not yet another metadata standard!" Like we don't have enough of them already:
  - EXIF, Dublin Core, VRA Core, IPTC Core, XMP, MPEG-7, Creative Commons, ... ?
- But again: No single standard can cover all metadata needs
- SW is a framework that could make existing metadata standards and tools interoperable ... and make them interoperable with the rest of the Web!

8

NewsML2 and the SW

- Common basis
  - Distributed resources (news item) globally and uniquely identified => URI
  - Use of shared and controlled vocabularies
- Natural switch and numerous benefits
  - Better control of NewsML2 descriptions (logical consistency check)
  - Enhanced search of News topic (logical inferences)
  - Intelligent presentation – Semantic interfaces
  - Unified news management – Semantic CMS

9

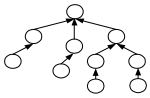
Use Case scenario

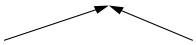
```
<newsItem schema="0.7" version="2">
  ...
  <itemMeta>
    <contentClass code="ccls:photo" />
    ...
  </itemMeta>
  <contentMeta>
    <infoSource literal="AFP" />
    <locCreated code="city:Kathmandu">
      <broader code="ctry:NEP" />
    </locCreated>
    <subject code="cat:01001000" type="ctyp:politics">
      <title>King</title>
    </subject>
    <description>
Nepal's King Gyanendra attended a Hindu festival in Kathmandu, his first
public appearance since being stripped of most of his powers by
parliament last month.
    </description>
  </contentMeta>
  ...
</newsItem>
```

Use Case scenario

Q: News about the *leader* of the *country Nepal* ?

The King Gyanendra of Nepal





Head State ⇔ and (King
(oneOf country Nepal, NL, ...))
Head Government ⇔ and (Prime Minister
(oneOf country Nepal, NL, ...))

The Prime Minister
Girija Prasad Koirala

11

What we have done?

- Creation of a News domain ontology in OWL
  - Based on the UML model specifications of NewsML2
- Online conversion service
  - Mapping of the IPTC NewsCodes into various SKOS thesaurus
  - Transforming dynamically the NewsML2 (XML) descriptions in its equivalent RDF counterpart
    - Using to the NewsML ontology
    - Linking to the SKOS IPTC NewsCodes

http://newsml.cwi.nl/

12

What is the added value?

- Example: A "normal" day in AFP
- Dataset
  - 200 NewsML2 stories, 35 photos (original size + thumbnails) + 35 NewsML2 descriptions
  - Covering various subjects:
    - A military drill for dealing with contaminations (toxic, nuclear or biological) - Photo
    - A regular meeting of the French cabinet - Photo
    - A strike in New Caledonia - Photo
    - A protest made on the Arch of Triumph in Paris, related to the Iran nuclear crisis - Photo
    - A wine makers protest - Photo
    - A meeting between the French president and Israeli prime minister - Photo
    - A senator's publicity pictures - Photo

13

Example 1: reasoning on the content

- Find all related news about "Nuclear"

Nucléaire → Military drill (NBC)

Nuclear → Iran nuclear crisis → *Arc de Triomphe* protest

Iran nuclear crisis → Chirac – Elmer summit

14

Example 2: reasoning on the structure

- Find photos of Y for which the author is X ?
- What the NewsML ontology provide ?
  - *slugline* and *headline* are *metadata properties*, whose values are *Basic Components*
  - *creator* and *contributor* are *authors*
  - history of the description (versioning)
- No need to know the NewsML structure to answer the query

15

What to do with the RDF data?

- Various tools that are able to digest RDF data and provide a unified view of these data
  - FOAF Viewer http://xml.mfd-consult.dk/foaf/explorer/
  - SIMILE project http://simile.mit.edu/piggy-bank/
- /facet: A Browser for Heterogeneous Semantic Web repositories
  - Faceted browser paradigm (*Flamenco*)
  - Provide a view on any RDF dataset

16

Conclusion

- Methods and conversion tools for bringing NewsML in the SW (RDF - compliant)
- Added-value:
  - Enhance search of news items (logical inferences on the structure and the content)
  - Enhance presentation of news items
    - Semantic media interfaces
    - Discover relations between Items / Topics / Packages
  - Semantic Content Management System
    - Keep track of provenance information

17

Future Work

- Making the use case scenario REAL!
  - Needs data: photos, videos, graphics, audio, textual stories !
- Implement interfaces for:
  - Browsing a News archive
  - Rendering the search results
- Establishing links between NewsML and other vocabularies
  - IPTC News Codes *versus* domain ontologies
  - NewsML *versus* DC, EXIF, MPEG-7, etc.

18

NewsCodeViewer

back