NewsML-G2 and the IPTC News Architecture G2

Laurent Le Meur
Head of Médialab, AFP
News Architecture WP chairman, IPTC

News industry response to content fragmentation

"… We need to be 'the provider of tools.' This means promoting open standards and interoperability, which will allow a diverse set of consumer-creators to combine disparate types of content."

(Tom Glocer, Reuters CEO, to the FTW, March 06)

Objectives of the IPTC

- International Press Telecommunications Council
- Develop and maintain technical standards for improved news exchange
- Standardize the structure of editorial content
  - Textual news (text markup)
  - Media content (alternative renditions of content)
  - Specialized data (events, sports results & stats)
- Standardize effective ways to package news
- Standardize news metadata
  - Allowing more effective management / filtering & searching
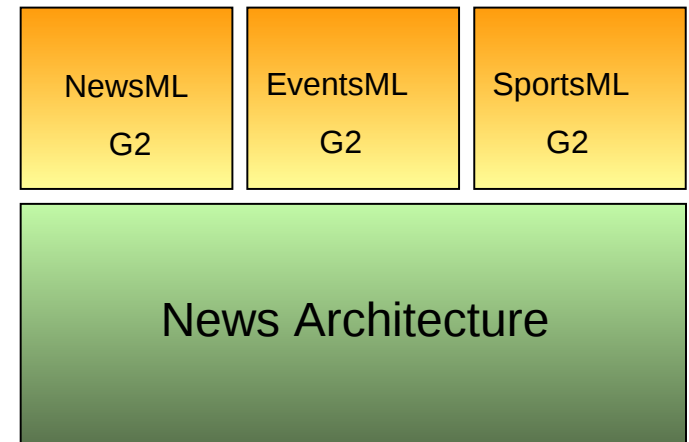 - Using language independent controlled vocabularies

IPTC - some history

IPTC 7901	1979	Text messages
NITF	v2 in 1997 v3.4 in 2007	SGML/XML Text markup
IIM	v4 in 1997	Binary file format, metadata IPTC Headers in Photoshop
IPTC Core for XMP	2004	Transfer of metadata values from the IPTC Headers to the XMP framework
NewsML 1	v1.0 in 2000 V1.2 in 2004	XML media-independent, structural framework for multi-media news.

IPTC G2-standards

- The IPTC is building standards for the news industry:
  - **NewsML-G2**: representation and management of general news = text stories, pictures, audio & video clips, or multimedia stories.
  - **EventsML-G2**: representation and management of events.
  - **SportsML-G2**: representation and management of sports results and statistics.

- All based of the same News Architecture G2

NewsML G2	EventsML G2	SportsML G2
News Architecture		

The NAR in brief

- The News Architecture defines:
  - Building blocks = data types and reusable components
  - News metadata = different sets of metadata
  - News Structure = different variants from a common base class
  - News Message = an optional exchange envelope

- Main goals are:
  - Interoperability between providers and consumers
  - Extensibility
  - Clarity and compactness of the syntax
  - Ease of storage
  - Ease of processing for relational or object mapping
  - New focus on "semantic" capabilities
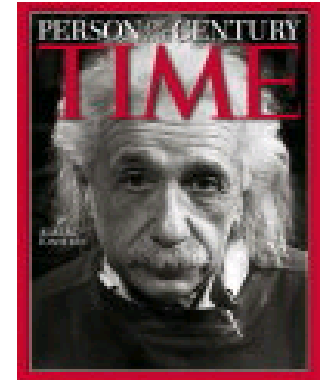 - Use of the latest XML technologies

NAR class hierarchy

Conformance Levels

- *Everything should be made as simple as possible, but not one bit simpler*
  (Albert Einstein)

- Tension
  - On one side (Core conformance level), search for
    - Simplicity (the KISS principle)
    - Usability, flexibility
    - Pragmatism
    - Interoperability by using a least common denominator
    - Web compatibility
  - On the other side (Power conformance level), search for:
    - Satisfaction of high profile news providers
    - Great rigour in the structure and processing
    - Control of content creation and usage

A model for news

newsItem

identification

Report on the fact or event

Metadata associated with content (contentMeta)

Management metadata (itemMeta)

Fact or event

News Item

Courtesy ATC

newsItem

- **<newsItem>**

- **Metadata**
  - Management properties
  - Administrative and descriptive properties

- **Different content renditions**
  - Text, image, sound, sports data, event structure ...
  - With associated physical characteristics
  - Goals:
    - Multi-resolution display
    - Cross-media publishing

Identity & Versioning

- A persistent and universal identity (guid) is key for:
  - Managing properly news items
  - Creating links between news items
- Any persistent globally unique identifier URI format can be used
- One document managed by one "desk" -> one identity
- A version number is mandatory for the audit of:
  - Addition or modification of metadata
  - Corrections
  - Small additions to the content (!= large editorial updates)
- Ex:
  ```
  <newsItem guid="tag:corp.net,2007:article1" version="5">
      <xxxx/>
  </newsItem>
  ```

Example - xhtml text

```
<newsItem xmlns="http://iptc.org/std/newsml/2006-10-01/">
    <catalogRef />
    <itemMeta />
    <contentMeta />
    <contentSet>
        <inlineXML contentType="application/xhtml+xml">
                <html xmlns="http://www.w3.org/1999/xhtml">
                    <head><title>Alert</title></head>
                    <body>It just happened.</body>
                </html>
        </inlineXML>
    </contentSet>
</newsItem>
```

Any text markup in the proper namespace.
IPTC provides NITF
W3C provides XHTML

Example - picture

```
<newsItem >
    <itemMeta />
    <contentMeta />
    <contentSet>
        <inlineData rendition="rnd:thumbnail"
            encoding="enc:base64"
            contentType="image/jpeg"
            width="80" height="80">
            234234FEZ…R2323</inlineData>
        <remoteContent rendition="rnd:fullres"
            contentType="image/jpeg"
            href=http://x.net/p.jpg
            size="1234000"/>
    </contentSet>
</newsItem>
```

Wrapper for *alternative* renditions of news content

Inline Data: plain text or base64 encoded data

Remote content

Multi-rendition items

A thumbnail cannot always be automatically processed from the original picture.

More: AFP info on TV5.fr

Linking news

- Links may be set from an Item (newsItem …) to another Item or Web Resource.

en.wikipedia.org/wiki/Tony_blair

LONDON (AFP) -
Outgoing British
Prime Minister
Tony Blair …

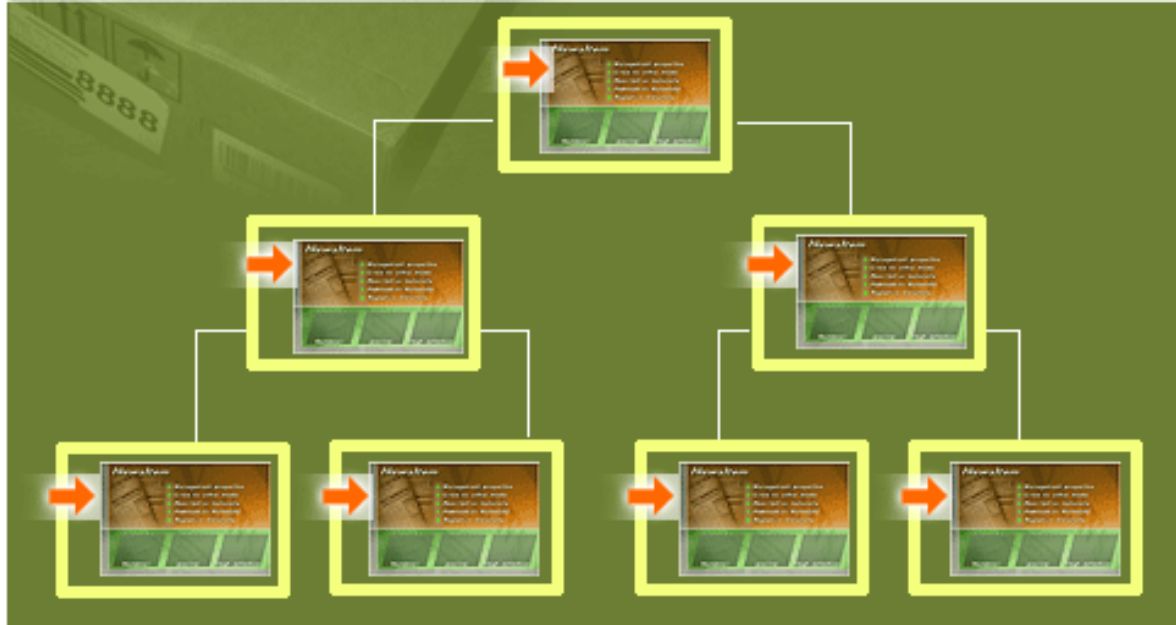
LONDRES (AFP) – Le Premier
Ministre Britanique sortant Tony
Blair …

Package Item

Courtesy ATC

Package Item

<packageItem >

   <groupSet root="g1">
      <group id="g1">
           <groupRef>g2</groupRef>
           <itemRef href=" tag:afp.com,2007:article "/>
      </group>
      <group id="g2"/>
   </groupSet>
</packageItem>

Wrapper for groups and item references

Group of groups and item references

Group reference

Item reference

Exchange (optional)

Management metadata

itemClass	item class
provider	content provider
versionCreated	date last version was created
firstCreated ?	date the item was first created
embargoed ?	date the embargo ends
pubStatus ?	p

Example - Management

```
<newsItem guid="tag:corp.net,2007:article1" version="5">
    <itemMeta>
        <itemClass qcode="icls:text" />
        <provider literal="News Agency" />
        <versionCreated>2007-05-01T10:00:00Z</versionCreated>
        <embargoed>2007-05-01T12:00:00Z</embargoed>
        <role qcode="irol:urgent" />
        <service qcode="svc:mainFeed" />
        <title>Urgent story</title>
        <edNote>embargoed until twelve UTC</edNote>
    </itemMeta>
    <contentMeta />
    <contentSet/>
</newsItem>
```

Few properties required

Administrative (Core)

urgency ?	editorial urgency
contentCreated ?	date of creation of the content
contentModified ?	date of last modification
located *	location of origin of the content
infoSource *	source of information
creator *	creator of the content
contributor *	contributor in the creation

Descriptive (Core)

language *	of the <u>content</u>
subject *	what the content *is about*
genre *	what the content *is* –> form / nature
slugline *	~ set of tags
headline *	brief and snappy introduction
dateline *	free-text, place/date of origin
by *	free-text, author/creator
description *	free-form, textual description

Example – content meta

```
<newsItem >
    <itemMeta />
    <contentMeta>
        <contentCreated>2007-05-17</contentCreated>
        <located qcode="frcity:Paris" />
        <creator literal="Laurent Le Meur" />
        <language>en-GB</language>
        <subject qcode="cat:13022000"/>
        <subject qcode="xml:NewsML-G2"/>
        <slugline>XTech-NewsML</slugline>
        <headline>NewsML-G2 presented at XTech 2007</ headline>
        <description>During XTech 2007 …</ description>
    </contentMeta>
    <xxxx/>
</newsItem>
```

All properties optional

NAR, DC and RDF/XML

- DC is generic, loosly defined, defines metadata about a « resource », implemented as either pure XML or RDF/XML

- Avoid the « Frankenstein monster » effect

- RDF/XML is verbose, XML schema validation is difficult, reification is needed when qualifers are added to properties.

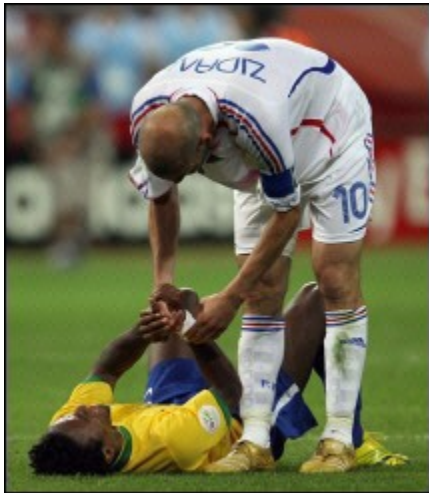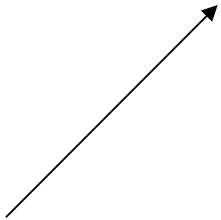- Solution: mappings will be provided

QCodes

- Controlled vocabulary:
  - The value of a property is a code in a scheme
  - A term identifies a concept, is language independent
- Ex:
  - {IPTC subject news codes, 13022000}
- Expressed in a compact manner, like QNames, but digits allowed as first character of the code:
  - cat:13022000
- The prefix corresponds to a URI
  - A catalog defines the mappings
- Can be translated to URIs (SemWeb identifiers) by simple concatenation.

Flexible properties

```
<subject literal="politics" />
<subject qcode="subj:03002000"/>
<subject qcode="subj:03002000">
    <name xml:lang="en">earthquake</name>
</subject>
<subject type="ctyp:person" qcode="frpers:j.chirac">
    <name>Jacques Chirac</name>
    <definition>President of xxxx</definition>
</subject>
<creator>
    <name>John Doe</name>
    <contactInfo><address>xxx</address>/<contactInfo>
<creator>
```

Extract information DNA

- Concepts/Entities that are subject of news
  - Thematic categories
  - People
  - Organizations
  - Geopolitical Areas
  - Points of Interest
  - Events
  - Products or artefacts

Definition of a concept

- Concept identifier (preferred and alternates)
- Type (person, organization …)
- Name, definition and note in free text
- Thesaurus relationships (NT, BT, RT)
- Specific properties of a concept
- Plus properties specific to each type of concept
  - Person: born, died, contact information
  - Organisation: founded, dissolved, contact information
  - Geopolitical area: gps coordinates, altitude
  - …

Concept Item

ConceptItem
→ Management properties
→ Links to other Items
→ Content metadata
→ Rights Metadata
→ Publication Metadata
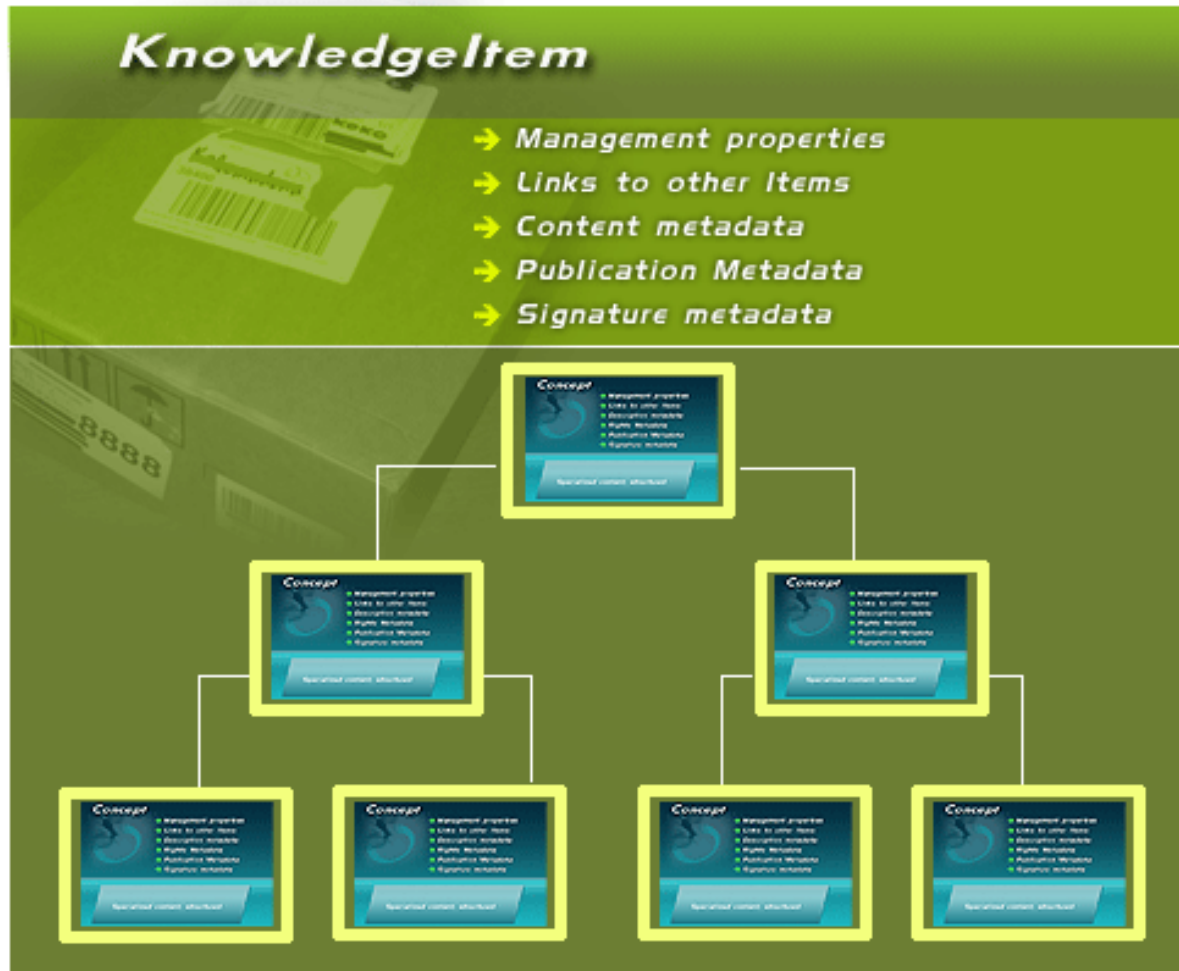→ Signature metadata

Specialized content, structured

Courtesy ATC

Example - Concept Item

```
<conceptItem >
    <catalogRef />
    <itemMeta>
    <itemClass qcode="ctyp:geoArea" />
    </itemMeta>
    <contentMeta />
    <concept>
    <conceptId qcode="cty:Madrid" />
    <type qcode="ctyp:geoArea" />
    <name>Madrid</name>
    <facet rel="gfct:geoType" qcode="gtyp:city" />
    </ concept>
</conceptItem>
```

> concept component, of type geo area

Knowledge Item

Courtesy ATC

Example - Knowledge Item

```
<knowledgeItem >
    <catalogRef />
    <itemMeta />
    <contentMeta />
    <conceptSet>
        <concept><type qcode="ctyp:person"/>…</concept>
        <concept><type qcode="ctyp:person"/>…</concept>
        <concept><type qcode="ctyp:organisation"/>…</concept>
        <concept><type qcode="ctyp:emotion"/>…</concept>
    </conceptSet>
</knowlegeItem>
```

Wrapper for concepts

Concept component
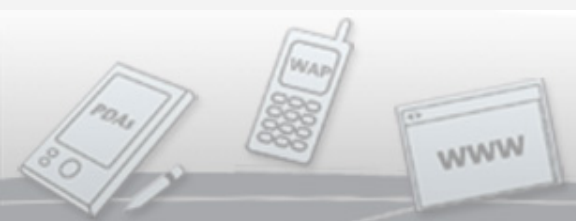
Much more for Power users

- Digital signature
- Different rights on different parts
- More management and administrative properties
- Editing attributes on metadata properties
- Refinement of the semantics of several properties
- Metadata with composite values
- More hints about metadata values
- Levels of confidence & relevance on metadata values
- Extended markup on labels
- Time validity attributes on links between NAR Items
- Etc…

The NAR model

- Built on a strict « object » model
  - UML representation
    - C# or Java, open source implementation
    - Automatic mapping between NAR Items and relational database.

- W3C XML Schemas
  - Common: Framework.xsd
  - News Objects: NewsItem.xsd, PackageItem.xsd, ConceptItem.xsd, KnowledgeItem.xsd
  - Exchange message (option): NewsMessage.xsd

- Studying an OWL representation

Roadmap

- News Architecture stable
  - Refinement on the processing model development and specification document
  - Working on G2 controlled vocabularies
- Working on finalizing IPTC-G2 standards
  - NewsML-G2
  - EventsML-G2
  - SportsML-G2
  - Experimental phase during summer 2007
- Preparing specification papers
  - Stable drafts should be ready in October 2007

Thank you for your time

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