News Metadata Framework (NMDF)

Misha Wolf
News Standards Manager, Reuters
Chair, News Metadata Framework WG

24 October 2005   IPTC meeting, Milan   Slide 1

NMDF Goals

- Support a spectrum of needs, from lightweight to heavyweight
- Be highly interoperable, while allowing proprietary "black holes"
- Be easy to understand and use
- Be compact

24 October 2005   IPTC meeting, Milan   Slide 2

Metadata Values

- Simple values (eg integer, date or string)
- Scheme-code pairs representing a concepts (eg "iso4217" and "USD")
- Structured values (eg <person>)
- Supplementary information
- Labels

24 October 2005   IPTC meeting, Milan   Slide 3

Simple Values

- <significance>80</significance>
- <created>2005-10-24</created>
- <slugLine separator="-">
  sports-boxing-tyson
  </slugLine>

24 October 2005   IPTC meeting, Milan   Slide 4

Scheme-Code Pairs

- A scheme is a taxonomy
- It may be very simple (a flat list) or very sophisticated (a thesaurus with typed relationships between nodes)
- Each scheme is identified by a URI

24 October 2005   IPTC meeting, Milan   Slide 5

Scheme-Code Pairs

- As URIs are long and quite difficult to use in XML, for each scheme URI we define a scheme alias
- For example: "iso4217" → "http://www.iptc.org/schemes/iso4217#"

24 October 2005   IPTC meeting, Milan   Slide 6

Scheme-Code Pairs

- Each node in a taxonomy is represented by a code (eg "3", "USD", "Interview")
- Each node is represented by a URI, obtained by appending the code to the scheme URI, eg:

  http://www.iptc.org/schemes/iso4217#USD

Scheme-Code Pairs

- QNames are Qualified Names
- Defined by Namespaces in XML
- An example of the QName syntax is "<news:item>"
- Unfortunately, we can't use QNames for scheme alias definition as …

Scheme-Code Pairs

- NameStartChar ::= ":" | [A-Z] | "_" | [a-z] | [#xC0-#xD6] | [#xD8-#xF6] | [#xF8-#x2FF] | [#x370-#x37D] | [#x37F-#x1FFF] | [#x200C-#x200D] | [#x2070-#x218F] | [#x2C00-#x2FEF] | [#x3001-#xD7FF] | [#xF900-#xFDCF] | [#xFDF0-#xFFFD] | [#x10000-#xEFFFF]
- NameChar ::= NameStartChar | "-" | "." | [0-9] | #xB7 | [#x0300-#x036F] | [#x203F-#x2040]

Scheme-Code Pairs

- So these are illegal QNames:

  nc:15000000 (sport)

  nc:15001000 (aero and aviation sport)

  nc:15001001 (parachuting)

  nc:15001002 (sky diving)

  …

Scheme-Code Pairs

- This is why we have adopted CURIEs (Compact URIs)
- <ns
  prefix="nc"
  uri="http://www.iptc.org/NewsCodes#"
  />

Structured Values

- For example:
- <person>

  <givenName>Marilyn</givenName>

  <familyName>Monroe</familyName>

  <pic>🖼</pic>
- </person>

Supplementary Information

- @creator and @created
- @confidence and @relevance
- @type
- @why
- <title>
- <sameAs> and <childOf>

Syntax

```
<subject code="nc:15062000"/>
  creator="…" created="…" type="…"
  confidence="…" relevance="…" why="…">
  <title>…</title>
  <sameAs code="…">
  <childOf code="…">
</subject>
```

Grouping Concepts

- Women's +100m+Swimming+Final
- Positive+pre-announcement + Citigroup
- Microsoft's share price + moved up
- the Clintons (Bill + Hillary)

Grouping Concepts

```
<subject>
  <bag>
      <bit code="nc:15000002"/>
      <bit code="nc:15062000"/>
  </bag>
  <title>Women's Swimming</title>
</subject>
```

Labels

- ```
<title creator="…" created="…" role="…"
markup="…" media="…" xml:lang="en">
The <abbr title="International Press
Telecommunications Council">IPTC </abbr>
(<abbr xml:lang="fr" title="Comité
International des Télécommunications de
Presse">CITP</abbr>)
announces major new standard!
</title>
```

## Labels

International Press Telecommunications Council
Comité International des Télécommunications de Presse

- The IPTC (CITP) announces major new standard!