TT NewsML

En beskrivning av TT Nyhetsbyråns användning av NewsML-G2 från IPTC.

Revisionshistoria

2015-09-01 Första versionen.

2015-11-13 Version 1.3

Om Ni har frågor om innehållet i det här dokumentet eller om TT:s sändformat, kontakta någon av följande personer:

Johan Lindgren	Jan Eriksson
Systemutvecklare	IT-chef
060-176815	08-6922600
johan.lindgren@tt.se	jan.eriksson@tt.se

Inledning

I samband med TT Nyhetsbyråns byte av redaktionellt system i februari 2015 lanserade TT också ett nytt leveransformat. Målsättningen med det nya formatet är att förbättra möjligheterna att paketera nyheter med olika media som text och bild. Samt att öka mängden metadata som levereras ihop med nyheterna.

Kunder kan välja att ta emot det nya formatet antingen som json eller som xml. Båda varianterna följder internationella standarder från IPTC.

News in JSON: https://iptc.org/standards/ninjs/

NewsML-G2: https://iptc.org/standards/newsml-g2/

Json, generellt: http://www.json.org/

XML, generellt: http://www.w3.org/XML/

IPTC-standarderna är omfattande och täcker många behov. TT:s egna versioner kallas därför TTNINJS respektive TT-NewsML.

Sida med detaljerad teknisk information: http://tt.se/spec

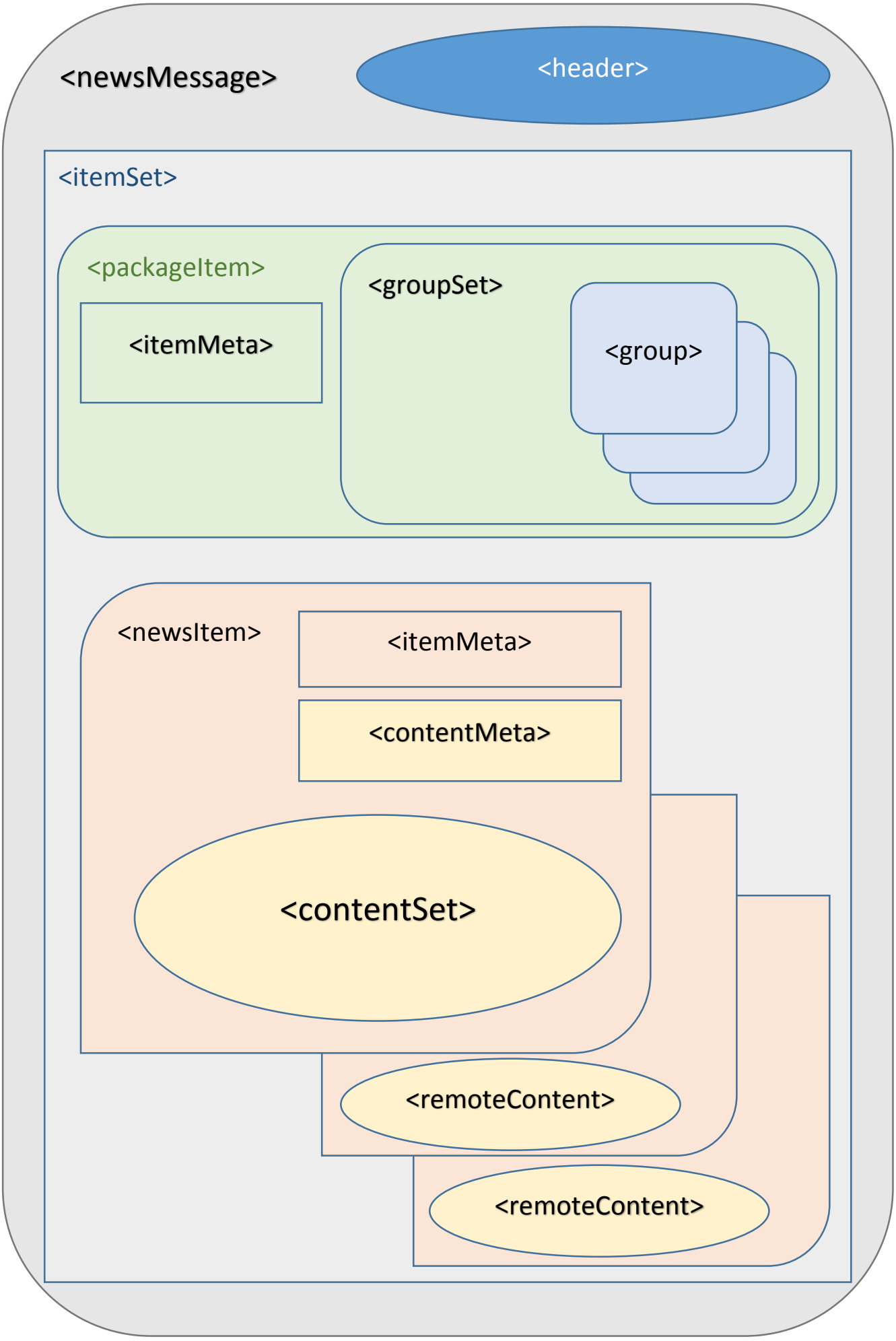
Sida om TTNewsML: http://tt.se/spec/newsml

Själva texten om nyheten finns både i json-varianten och xml-varianten som HTML5.

Information om TT:s HTML5: http://tt.se/spec/body_html5

TT har en github-samling med information, verktyg och moduler som kan användas för att dra nytta av TT:s material.

GitHub: http://ttab.github.io/


<newsMessage>
<header>
<itemSet>
<packageItem>
<itemMeta>
<groupSet>
<group>
<newsItem>
<itemMeta>
<contentMeta>
<contentSet>
<remoteContent>
<remoteContent>

Beskrivning

Den föregående skissen beskriver sammansättningen av NewsML. Själva nyhetsobjekten består av NewsItems. Deras övergripande struktur är likadan oavsett om de beskriver text, bild, video eller något annat nyhetsobjekt.

Innehållet kan finnas med inuti NewsML-objektet och finns då i ett ContentSet. Men det kan också vara en länk till data någon annanstans och då används RemoteContent.

PackageItem finns där för att länka ihop NewsItems. Alltsammans finns i en ItemSet och TT har valt att använda denna struktur även om det är ett NewsML-objekt som bara har en text i form av en ensam NewsItem. Detta för att bearbetningen ska bli likadan i grunden.

NewsMessage kan liknas vid kuvertet och Header är metadata om själva leveransen. Inte så aktuellt för själva innehållet. TT:s NewsML är UTF-8 kodat.

NewsItem

NewsItem har en rad attribut (guid="http://tt.se/media/text/151118-terrorfrankrikeuv10-185880" version="1" standard="NewsML-G2" standardversion="2.20" conformance="power") där guid är den viktiga att hålla rätt på.

De första elementen i NewsItem är katalog-referenser. IPTC:s NewsML-G2 använder mycket **QCodes**. Det står för Qualified Codes. Koderna består av ett prefix, ett kolon och så själva värdet. IPTC har definierat en grunduppsättning via sin katalog. Men precis som andra leverantörer har TT fyllt på med några egna. Och det är tanken med QCodes. Man kan ha värden från flera leverantörer men ändå kunna spåra dem och få veta vad de står för på ett kontrollerat sätt.

Därefter följer en grupp för rättighetsinformation. TT anger bara vem som har copyright till det aktuella innehållet. Men gruppen kan innehålla avancerad rättighetsinformation som ska kunna bearbetas maskinellt.

Nästa grupp kallas **itemMeta** och innehåller metadata om själva newsItem. Man håller isär det från contentMeta som är metadata om innehållet.

```
    <itemClass qcode="ninat:text"/>
    <provider qcode="nprov:TT"/>
    <versionCreated>2015-11-18T09:41:19+01:00</versionCreated>
    <pubStatus qcode="stat:usable"/>
    <edNote>Uppdaterade uppgifter om gripande, citat från boende i huset, samt med karta över Saint-Denis.</edNote>
    <link rel="irel:previousVersion" href="http://tt.se/media/text/151118-terrorfrankrikeuv9-185867"/>
```

itemClass talar om vilken typ av innehåll det är. *Text* och *Picture* är de vanligaste hos TT.

pubStatus är normalt *usable*, men kan exempelvis vara *replaced* om det är en version som ersatts av en nyare eller *commissioned* om det är innehåll beställt av en viss kund.

edNote är information om denna newsItem från TT till mottagarna.

Om den här versionen ersätter tidigare så finns länk till de tidigare version(er) som ersätts.

Så har vi **contentMeta** som består av metadata som beskriver innehållet i denna newsItem.

```
      <urgency>4</urgency>
      <contentCreated>2015-11-18T09:41:19+01:00</contentCreated>
      <language tag="sv"/>
      <subject type="cpnat:abstract" qcode="medtop:02000000"
typeuri="http://tt.se/spec/subref/1.0"><name>Brott, lag och rätt</name></subject>
      <subject type="cpnat:person" typeuri="http://tt.se/spec/person/1.0">
        <name>Abdelhamid Abaaoud</name></subject>
      <subject type="cpnat:organisation" typeuri="http://tt.se/spec/organisation/1.0">
        <name>Inter France</name></subject>
      <subject type="cpnat:place" typeuri="http://tt.se/spec/place/1.0"
literal="plc_saint-denisfrankrike">
        <name>Saint-Denis, Frankrike</name></subject>
      <subject type="cpnat:object" typeuri="http://tt.se/spec/object/1.0">
        <name>Stade de France</name></subject>
      <slugline>terror-frankrikeUV10</slugline>
      <headline>Två döda i räd mot terrormisstänkta</headline>
      <description role="drol:caption">Minst två personer har dödats i en polisräd mot en
lägenhet i Saint-Denis, norr om centrala Paris. En av dem är en kvinnlig självmordsbombare.
Insatsen uppges ha riktats mot den man som tros ha planerat fredagens terrorvåg. Polis
bekräftar för AFP att t</description>
```

Urgency är nyhets prioritet där lägre siffra är mer prioriterat. Så 1 betyder FLASH.

Subject kan det finnas många av. De förekommer i olika typer och förekommer ofta flera av samma typ. I det här exemplet finns bara en av varje. Lite speciellt är cpnat:place om den har en literal av typen plc_. Då finns geo-koder till denna plats i ett eget element efter contentMeta:

```
    <assert literal="plc_saint-denisfrankrike">
      <type qcode="cpnat:geoArea"/>
      <name>Saint-Denis, Frankrike</name>
      <geoAreaDetails>
        <position latitude="48.936181" longitude="2.357443"/>
      </geoAreaDetails>
    </assert>
```

I contentMeta finns också slugg, rubrik och en beskrivning som kan vara en inledande del av texten eller en bildtext.

```
<contentMetaExtProperty type="ttext:profile" literal="PUBL"/>
<contentMetaExtProperty type="ttext:representationtype" literal="complete"/>
<contentMetaExtProperty type="ttext:job" literal="104462"/>
<contentMetaExtProperty type="ttext:webprio" literal="2"/>
<contentMetaExtProperty type="ttext:charcount" literal="1890"/>
<contentMetaExtProperty type="ttext:originaltransmissionreference" literal="185880"/>
```

De ovanstående elementen är sätt i NewsML att lägga till metadata som inte har en egen namngiven plats. De här är exempel på sådant som TT har lagt till.

Profile anger om innehållet kan publiceras (*PUBL*) eller om det är informationsmaterial (*INFO*).

Representationtype kan vara *complete* eller *incomplete*. Vilket ska tolkas som att innehållet står för sig själv eller inte.

Job är ett internt värde som kan användas för att koppla ihop separata nyheter om samma händelse.

Webprio anger om nyheten ingår i TT:s online-tjänst och den prio den har där. Kan skilja från urgency tidigare.

Charcount är antalet tecken om det är en text.

Originaltransmissionreference är id-numret som också är en del av objektets uri.

Slutligen innehåller en newsItem ett **contentSet**. Det kan vara innehåll inlagt direkt i NewsML-objektet (**contentXML**) eller så är det länkar till extern data (**remoteContent**). Generellt kan man säga att text finns inlagt medan bilder eller annan visuell data finns länkad.

Text presenteras som HTML5. Det finns inget officiellt xml-schema för HTML5, varken från W3C eller från IPTC. Validerar man en TTNewsML så kommer systemet att klaga på att det inte finns något att validera innehållet med. TT lovar dock att den HTML5 som levereras är korrekt XML. Se mer om HTML5 längre fram.

Ett exempel på **remoteContent** för en bild:

```
<remoteContent href="https://beta.tt.se/media/text/151118-terrorfrankrikeuv10-185880/a000_NormalHires.jpg" contenttype="image/jpeg" rendition="rnd:highRes" size="3240095" width="4662" height="2838">
  <remoteContentExtProperty type="ttrend:variant" literal="Normal"/>
  <remoteContentExtProperty type="ttrend:usage" literal="Hires"/>
</remoteContent>
```

Den har en **href** som pekar på själva bilden samt några attribute och tt-specifik metadata som beskriver bilden.

PackageItem

```
<packageItem guid="http://tt.se/media/text/151118-terrorfrankrikeuv10-185880-pack" version="1" standard="NewsML-G2" standardversion="2.20" conformance="power" xml:lang="sv">
  <catalogRef href="http://www.iptc.org/std/catalog/catalog.IPTC-G2-Standards_24.xml"/>
  <catalogRef href="http://tt.se/spec/catalog/catalog.tt-g2.1_0.xml"/>
  <itemMeta>
    <itemClass qcode="ninat:text"/>
    <provider qcode="nprov:TT"/>
    <versionCreated>2015-11-18T09:41:19+01:00</versionCreated>
    <pubStatus qcode="stat:usable"/>
    <edNote>Uppdaterade uppgifter om gripande, citat från boende i huset, samt med karta över Saint-Denis.</edNote>
  </itemMeta>
```

I NewsML är **packageItem** att betrakta som ett objekt som kan uppdateras för sig. Därför har den sin egen **guid** och egen metadata, samt en egen katalog-referens. För tillfället, i TT:s leveranser, så kommer dock inte packageItems att uppdateras för sig. Utan om något behöver ändras så uppdateras hela NewsML-objektet.

PackageItem innehåller också **groupSet**-objektet som kan sägas vara limmet som håller ihop delarna i ett NewsML-objekt. Delarna heter **group**, vilket kan vara lite konstigt namn när de bara har en beståndsdel.

Det intressanta att hålla utkik efter är den group som har **role=”group:main”**. Den pekar på huvudobjektet i hela NewsML-objektet.

Så arbetsgången blir:

a) Leta reda på group:main
b) Ta fram uri till det objektet (residref=”xx”).
c) Matcha det med en newsItems guid-attribut.

```
  <groupSet root="root">
    <group id="root" role="group:main">
      <itemRef residref="http://tt.se/media/text/151118-terrorfrankrikeuv10-185880"/>
      <groupRef idref="media"/>
    </group>
    <group id="media" role="group:package" mode="pgrmod:bag">
      <itemRef residref="http://tt.se/media/image/5D788841DE6E438D8F4530E36F6519BF">
        <altId type="tt:associd">a000</altId>
      </itemRef>
      <itemRef
```

```
residref="http://tt.se/media/image/0879451A9A83461E8EE571B7491720C2">
          <altId type="tt:associd">a001</altId>
        </itemRef>
        <itemRef
residref="http://tt.se/media/image/99B107B361644A59B9F6D88DF9427D22">
          <altId type="tt:associd">a002</altId>
        </itemRef>
      </group>
    </groupSet>
```

I xslt-sammanhang kan man då använda något liknande:

Plocka in uri till en variable:

```
<xsl:variable name="mainuri"
select="newsMessage/itemSet/packageItem/groupSet/group[@role =
'group:main']/itemRef/@residref"/>
```

Variabeln kan sedan användas för att hämta önskade delar:
`newsMessage/itemSet/newsItem[@guid = $mainuri]/contentMeta/headline`

Innehållet som HTML5:

```
<inlineXML contenttype="text/html">
<html>
        <head>
                <title>Två döda i räd mot terrormisstänkta</title>
        </head>
        <body>
                <article>
                        <section data-charcount="1890">
                        <h1>Två döda i räd</h1>
                        <div class="dat">
                        <span class="vignette">Terrorism</span>
                        <span class="source">TT</span></div>
                        <h4><p>Minst två personer har dödats i en polisräd mot en lägenhet i Saint-Denis, norr om centrala Paris.</p></h4>
                        <div class="bodytext">
                        <p>Polis bekräftar för AFP att två personer dödats i insatsen.</p>…
                        <blockquote>Jag vaknade av en explosion, berättar vittnet.</blockquote>
                        <h2>Mellanrubrik</h2>
                        </div>
                        <div class="byline">Mattias Areskog/TT</div>
                         <figure>
                        <img src="https://beta.tt.se/media/text/151118-terrorfrankrikeuv10-185880/a000_CroppedThumbnail.jpg" alt="Antiterrorräd i Saint-Denis, norr om centrala Paris." data-assoc-ref="a000" data-uri="http://tt.se/media/image/5D788841DE6E438D8F4530E36F6519BF" data-preview-src="https://beta.tt.se/media/text/151118-terrorfrankrikeuv10-185880/a000_NormalPreview.jpg"/>
                        <div class="byline">François Mori/AP/TT</div>
```

<figcaption>Antiterrorräd i Saint-Denis, norr om centrala Paris.</figcaption>

</figure>

</section>

</article>

</body>

</html>

</inlineXML>

Det ovanstående exemplet är en kraftigt förkortad version. Huvudtexten finns i en **section** som innehåller **h1** för rubriken och en **div** med **vinjett** och **källa**.

Har texten en ingress så finns den i en **h4**-sektion.

Själva brödtexten finns i en **div** med typen **bodytext**. Brödtexten kan innehålla en blandning av **p**-element, **blockqoute** för citat och **h2** för mellanrubriker. Det kan också förekomma listor och tabeller enligt vanlig html-kodning.

Under brödtexten finns **byline** i en **div**, ifall texten har en byline. Eventuella bildkopplingar finns med en **figure** per bild.

Textdelen kan också ha **asides** för faktarutor och notiser. Det ovanstående är ingen komplett beskrivning av TT:s html5 utan ett exempel. Lämpligt är att titta på ett antal texter från TT av olika typ och storlek för att få en samlad bild över de alternativ som kan förekomma.