

# ***Fedora Content Modeling at DTU Library***

***Fedora-EU Meeting, Oxford, 8 December 2009***

***Gert Schmeltz Pedersen  
DTU Library, Technical Information Center  
@ DTU, Technical University of Denmark***

***Funded partly by the CAMMP Project***

# Three Use Cases

- CRIS/CERIF (for research databases)
  - Current Research Information Systems
  - Common European Research Information Format
- Chemical Portal
  - Substance registration, risk assessment, ...
- CAMMP (backend for service provision)
  - Converged Advanced Mobile Media Platforms

***Fedora Content Modeling: How? Why? Alternatives?***

***Approach for “sound” CMA/ECM (like 3NF for RDB)?***

# *Three Use Cases background and specs*

- CRIS/CERIF

Entity-Relationship Model -> RelDB design, XML serialization

~ 100.000 entities

- Chemical Portal

Law, best practices, ... -> XML Schemas

~ 10.000 chemicals

- CAMMP

TV-Anytime Broadcast and On-line Services -> XML Schemas

~ 1.000.000 – 10.000.000 objects

***In common: XML orientation***

## 2. The CERIF 2008 – 1.0 Model

To reduce the complexity of the model towards a better understanding, this introduction and specification document follows a conceptual structure. The conceptual structure allows for different perspectives and views when talking about parts of the model and enables the emphasis to particular model features. This conceptual structure is only a virtual structure and as such not inherent in the physical data model, and therefore, also not incorporated in the physical SQL scripts. It is used for organizing this document and considered an instrument that supports the comprehension of the CERIF model.

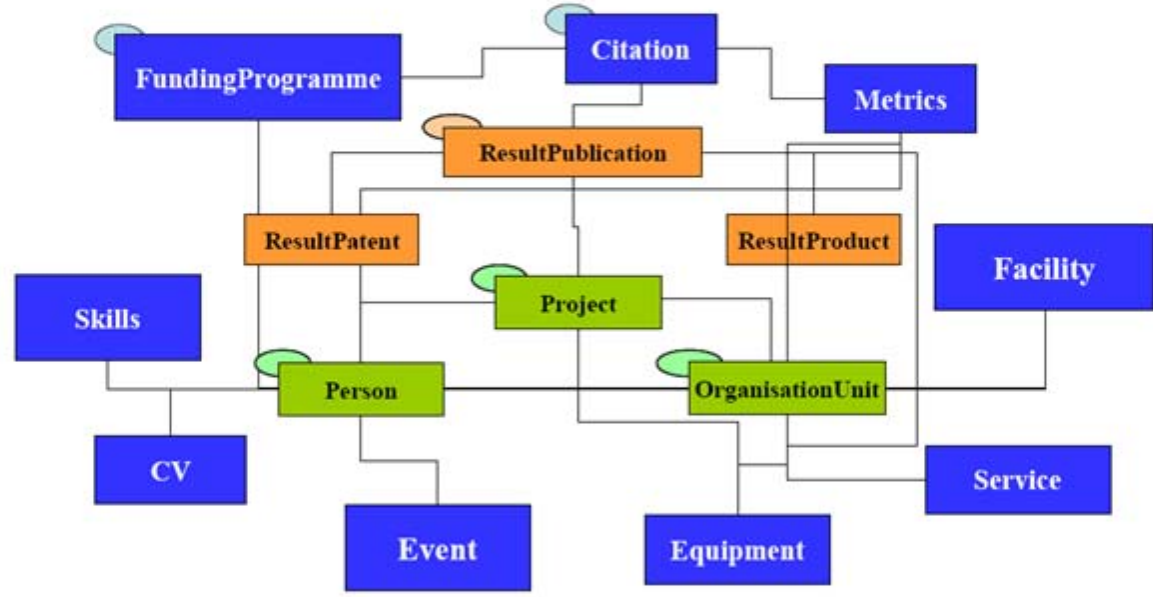
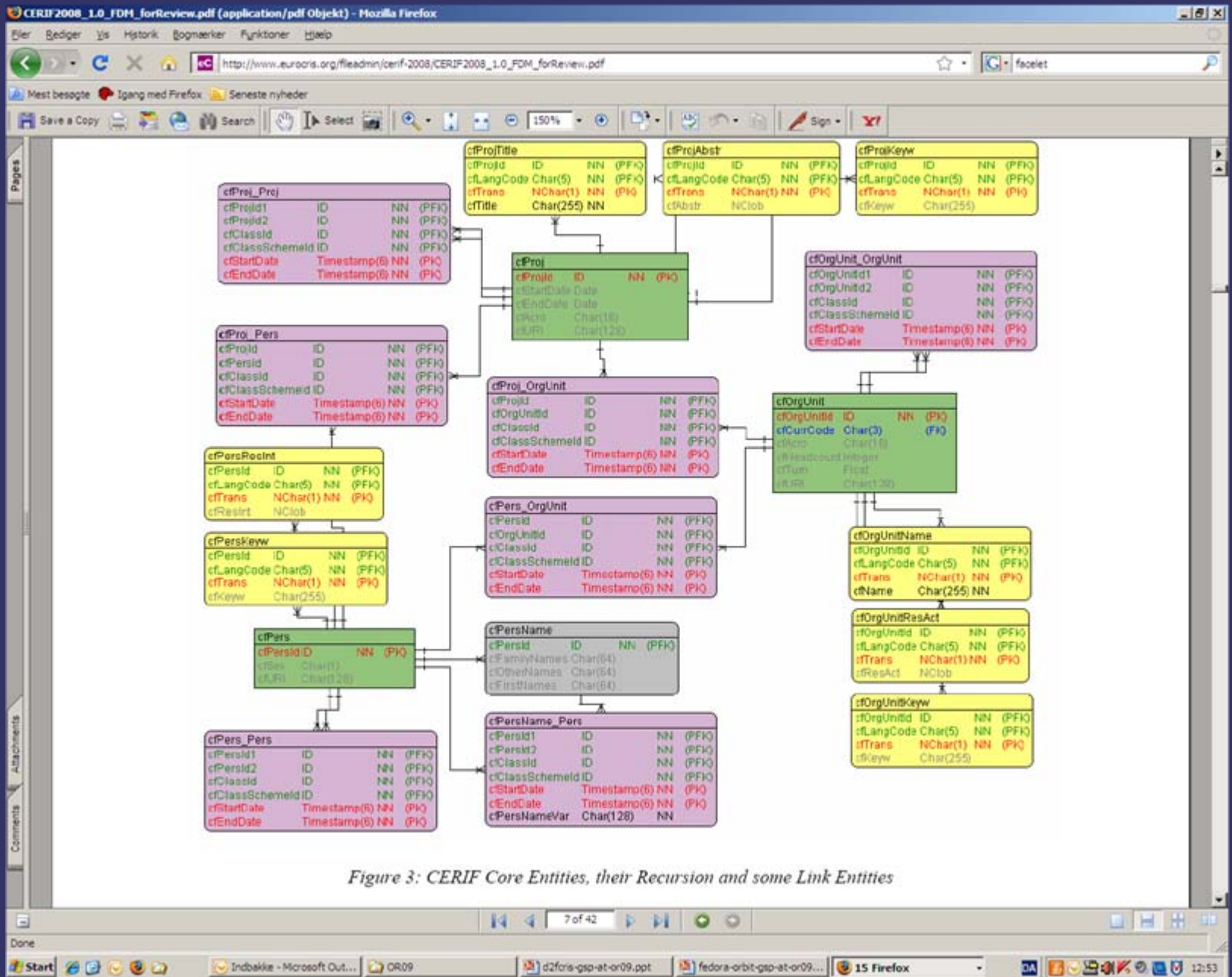


Figure 1: Some CERIF Entities and their Relationships

### 2.1 CERIF Conceptual Structure



# XML Data Exchange for Interoperability

CERIF2008\_1.0\_XML.pdf (application/pdf Objekt) - Mozilla Firefox

http://www.eurocris.org/fileadmin/cerif-2008/CERIF2008\_1.0\_XML.pdf

```
</cfOrgUnit>
<cfOrgUnit>
  <cfOrgUnitId>0432125</cfOrgUnitId>
  <cfFURI>http://www.dfki.de/lt/</cfFURI>
  <cfAcronym>LT Lab</cfAcronym>
  <cfHeadCount>40</cfHeadCount>
</cfOrgUnit>
...
</CERIF>
```

### 7.1.2 CERIF Result XML Entities (XML Examples)

```
<?xml version="1.0" encoding="UTF-8"?>
<CERIF
  xsi:schemaLocation="http://www.eurocris.org/cerif/cfResPubl-CORE cfResPubl-CORE.xsd"
  xmlns="http://www.eurocris.org/cerif/cfResPubl-CORE"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  release="2006" date="YYYY-MM-DD" sourceDatabase="euroCRIS">
  <cfResPubl>
    <cfResPublId>publication-joerg-et-al</cfResPublId>
    <cfFURI>http://www.eurocris.org/fileadmin/Upload/Events/
      Conferences/CRIS2008/Papers/cris2008_Joerg.pdf </cfFURI>
    <cfResPublDate>2008</cfResPublDate>
    <cfStartPage>107</cfStartPage>
    <cfEndPage>123</cfEndPage>
    <cfISBN>978-961-6133-38-8</cfISBN>
  </cfResPubl>
  <cfResPubl>
    <cfResPublId>publication-veca-c-storey</cfResPublId>
    <cfFURI>http://www.springerlink.com/content/j23263j02m850617/</cfFURI>
    <cfResPublDate>1993</cfResPublDate>
    <cfNum>4</cfNum>
```




# Chemical portal

The screenshot shows a Mozilla Firefox browser window displaying the Chemical portal website. The browser's address bar shows the URL <http://www.kemibrug.dk/searchpage/>. The website header includes the DTU logo and the text "Technical University of Denmark" and "University of Copenhagen". A navigation bar contains links for "home", "news archive", "about Kemibrug", "links", "KU web", and "download Ozone".

**MENU**

- Login
- Daily use
- Search
- List of substances (2 Mb)
- Order SDS
- Chemical APV
- Label designer
- Documentation
- About SDS
- Literature list
- The system
- About registering
- About labels
- About Ozone
- Special Functions
- Substance registration
- Chemical holding
- User Admin
- Contact
- Local administrator
- Kemibrug

You are not logged in.

 Sektion for Arbejdsmiljø  
DTU bygning 101A  
2800 Kgs Lyngby

**Safety Data Sheet** | **Searching**

You can search for the CAS no. or name. E.g. enter 931-88-4 or 4-ethylmorpholin or N-ethylmorpholin or part of the text: chlor? (>100 documents) or ?chlor? (>300 documents).

Searching in local notes is via free text search. Searching for a string of words separated by spaces will return entries containing all the words. The more words you give, the more precise the search will be.

(Please note that ? and \* replace a word or string of words in the searches).

chem.name+synonyms  [Search](#)

CAS number  [Search](#)

fulltext  [Search](#)

Update date is between  and  (incl.)  
(use dd-mm-yyyy)

[More help](#) | [Advanced search](#)

Done

Fedora Administrator - fedoraAdmin@localhost:8080

File Tools Window Help

Object - CAS:615-36-1

Properties | Datastreams

register	ID	register
RELS-EXT	Control Group	Internal XML Metadata
DC	State	Active
E-section	Versionable	Updates will create new version
summaryB	Created	2009-10-21T13:22:29.017Z
H-section	Label	register
C-section	MIME Type	text/xml
B-section	Format URI	
RS-section	Alternate IDs	
G-section	Fedora URL	http://localhost:8080/fedora/get/CAS:615-36-1/register
D-section	Checksum	DISABLED none
F-section		
New...		

```

<register created="2004-07-06, 10:28:44" creator="136" lastmodified="" lastmodifier="" modified="2006-09-06, 10:31:06"
modified-date="2006-09-05" modifier="Kirsten" public="0" release_eng="2006-09-05" status="3" under_revision="0">
<CASno>615-36-1</CASno>
<title>2-Bromoaniline</title>
<synonyms>
<synonym>2-Bromoaniline</synonym>
<synonym>2-Bromo benzenamine</synonym>
</synonyms>
<productdata xml:lang="da"/>
<productdata xml:lang="en"/>
<revision>
<text xml:lang="da"/>
<text xml:lang="en"/>
</revision>
<links>
<link/>
</links>
<propertylist>
<moleculeformula><sub>6</sub></sub><sub>6</sub></sub></moleculeformula>
<moleculerweight>172,04</moleculerweight>
<boilingpoint xml:lang="da">229°C</boilingpoint>
<boilingpoint xml:lang="en"/>
<meltingpoint xml:lang="da">29-32°C</meltingpoint>
<meltingpoint xml:lang="en">32°C</meltingpoint>
<density xml:lang="da">1,578 g/cm³</density>
<density xml:lang="en">1,578 g/cm³</density>
</propertylist>
<substitutions>
<TLV xml:lang="da"/>
<TLV xml:lang="en"/>
<flashpoint xml:lang="da">&gt;66°C</flashpoint>
<flashpoint xml:lang="en">&gt;66°C</flashpoint>
<vapourpressure xml:lang="da">0,04 mm; 25°C</vapourpressure>
<vapourpressure xml:lang="en"/>
<octanol-waterpartitioncoefficient xml:lang="da">2,11</octanol-waterpartitioncoefficient>
<octanol-waterpartitioncoefficient xml:lang="en"/>
</substitutions>
<warnings xml:lang="da"/>
<warnings xml:lang="en"/>
</register>

```



Project Explorer

- XsltProcFormGenerator.java
- dk.dtu.dtic.k2.configuration
  - Configuration.java 208 1
  - JndiConfiguration.java 20
- dk.dtu.dtic.k2.exception
- dk.dtu.dtic.k2.servlet
  - ControllerServlet.java 23
  - DeliverXmlServlet.java 1
- dk.dtu.dtic.k2.servletfilter
- dk.dtu.dtic.k2.storage
  - FedoraDao.java 236 11/
  - FileStorageDao.java 236
  - IStorageDao.java 236 1
- dk.dtu.dtic.k2.util
- src/main/resources
  - xml
  - xsd
    - A-section.xsd 90 10/8/05
    - B-section.xsd 90 10/8/05
    - C-section.xsd 90 10/8/05
    - D-section.xsd 90 10/8/05
    - E-section.xsd 90 10/8/05
    - F-section.xsd 90 10/8/09
    - G-section.xsd 90 10/8/05
    - H-section.xsd 90 10/8/05
    - I-section.xsd 90 10/8/09
    - register.xsd 90 10/8/09
    - RS-section.xsd 90 10/8/0
    - RS-text.xsd 90 10/8/09 1
    - section-text.xsd 90 10/8/
    - summaryB.xsd 90 10/8/
    - typedefs.xsd 6 9/18/09 5
  - xsl
  - log4j.properties 203 11/23/
- src/test/java

```

1<?xml version="1.0" encoding="UTF-8"?>
2<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
3
4  <xs:import namespace="http://www.w3.org/XML/1998/namespace" schemaLocation="http://www.w3.org/2001/xml.xsd"/>
5
6  <xs:include schemaLocation="typedefs.xsd"/>
7
8  <xs:complexType name="linkType">
9    <xs:sequence>
10     <xs:element name="caslink" minOccurs="0" maxOccurs="1"/>
11     <xs:complexType>
12       <xs:sequence>
13         <xs:element name="identifier" type="xs:anyURI" minOccurs="0"/>
14       </xs:sequence>
15     </xs:complexType>
16   </xs:element>
17   <xs:element name="link" minOccurs="0" maxOccurs="unbounded">
18     <xs:complexType>
19       <xs:sequence>
20         <xs:element name="identifier" type="xs:anyURI" minOccurs="0"/>
21       </xs:sequence>
22     </xs:complexType>
23   </xs:element>
24 </xs:sequence>
25 </xs:complexType>
26
27 <xs:complexType name="revisionType">
28   <xs:sequence>
29     <xs:element type="languageDependentStringType" name="text" minOccurs="0" maxOccurs="2"/>
30     <xs:element name="date" type="xs:string" minOccurs="0"/>
31   </xs:sequence>
32 </xs:complexType>
33
34 <xs:element name="register">
35   <xs:complexType>
36     <xs:sequence>
37       <xs:element name="CASno" type="xs:string" minOccurs="1" maxOccurs="1"/>
38       <xs:element name="title" type="xs:anyType" minOccurs="1" maxOccurs="1"/>
39       <xs:element name="synonyms" minOccurs="0" maxOccurs="1">
40         <xs:complexType>
41           <xs:sequence>
42             <xs:element name="synonym" type="xs:anyType" minOccurs="0" maxOccurs="unbounded"/>
43           </xs:sequence>
44         </xs:complexType>
45       </xs:element>
46       <xs:element name="productdata" type="languageDependentAnyType" minOccurs="0" maxOccurs="2"/>
47       <xs:element name="revision" type="revisionType"/>
48       <xs:element name="links" type="linkType"/>

```

Design Source

Markers Properties Servers Data Source Explorer Snippets History Console Search Error Log JUnit

SVN

Writable Smart Insert 35 : 1

Outline Task List

Directives

- http://www.w3.org/2001/xml.xsd {f}
- typedefs.xsd

Elements

- XsltProcFormGenerat
- dk.dtu.dtic.k2.configurat
- Configuration.java 20
- JndiConfiguration.java
- dk.dtu.dtic.k2.exception
- dk.dtu.dtic.k2.servlet
  - ControllerServlet.java
  - DeliverXmlServlet.jav
- dk.dtu.dtic.k2.servletfilt
- dk.dtu.dtic.k2.storage
  - FedoraDao.java 236
  - FileStorageDao.java
  - IStorageDao.java 23
- dk.dtu.dtic.k2.util
- src/main/resources
  - xml
  - xsd
    - A-section.xsd 90 10/
    - B-section.xsd 90 10/
    - C-section.xsd 90 10/
    - D-section.xsd 90 10/
    - E-section.xsd 90 10/
    - F-section.xsd 90 10/
    - G-section.xsd 90 10/
    - H-section.xsd 90 10/
    - I-section.xsd 90 10/8
    - register.xsd 90 10/8/
    - RS-section.xsd 90 10/
    - RS-text.xsd 90 10/8/
    - section-text.xsd 90
    - summaryB.xsd 90 1
    - typedefs.xsd 6 9/18/
  - xsl
  - log4j.properties 203 11

Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://nightly.k2.cvt.dk/k2-0.3.4-SNAPSHOT/app/edit/CAS/615-36-1

Google

# Kemibrug

Logged in as gspe | [Logout](#)

Browse Search Register Help About

## Edit CAS: 615-36-1

Register A B C D E F G H I RS Summary B

### Oprettelse af ny brugsanvisning

Udarbejdet af	136
Oprettelsesdato	2004-07-06, 10:28:44
Senest revideret af	
Forrige version	
Aktuel ændring af	gspe
Seneste version	2006-09-05
Senest åbnet	2009-11-30T13:09:37
Dansk frigivelse	2006-09-05
Engelsk frigivelse	2006-09-05

Under revision  Ja  Nej  
 Offentlig  Ja  Nej

Status

CAS- eller Pnr.   
 Navn

### Synonymer

Synonym					2-Bromoaniline
Synonym					2-Bromo benzenamine

### Leverandør

### Revision

### Links til internetressourcer

### Fysisk kemisk data

Produktnr.	<input type="text"/>
Bruttoformel	<input type="text" value="C&lt;sub&gt;6&lt;/sub&gt;H&lt;sub&gt;6&lt;/sub&gt;BrN"/>
Molvægt	<input type="text" value="172.04"/>

### Kogepunkt

Done

# CAMMP backend for service provision



The screenshot shows a Mozilla Firefox browser window with the address bar containing <https://cammp.imi.aau.dk/wiki/ProjectDesc>. The page title is "CAMMP: Next generations mobile – when the media converge". The main content consists of several paragraphs of text describing the project's goals, funding, and organizational structure.

## CAMMP: Next generations mobile – when the media converge

CAMMP is a prestigious mobile media platform funded by the Danish Advanced Technology Foundation (Højteknologifonden). The project will run for a 4-year period (starting in 2008) and includes the main research institutions and companies in Denmark dealing with converged mobile media platforms and services.

The convergence between the Internet, digital TV and radio, and 3G mobile technologies will lead to many new possibilities. CAMMP will investigate and uncover the potential in the new converged infrastructure, which will change the known media as radio and TV by combining them with user-generated content and interaction between content providers and users.

The research will define new business models and new value chains for next generation mobile services. The new platform has focus on technological innovation in the industry and on strengthening of research and education.

The Danish National Advanced Technology Foundation has granted the project 22 million kroner.

The platform will be organised in 7 Work packages. The research and technological development will be performed in five dedicated Work Packages (WPs) structured in a matrix organisation, depicted below. WP1-5 are the research and development (R&D) WPs, where WP2,3 and 5 will develop and evaluate new technical solutions while WP1 and 4 are shaping and modifying these solutions in a user centric process. The user requirements, the regulatory and standardization environment and the business logic are truly integrated in the technical development in an iterative process as it is further illustrated in the figure. Taking the point of departure in the user requirements related to a fairly general conception of the very broad potentials of the involved technologies, the 'shaping and modifying' WPs are throughout the project involved in directing the technical solutions towards the relevant service area. Furthermore, all the partners are assigned key roles in this according to their competencies.



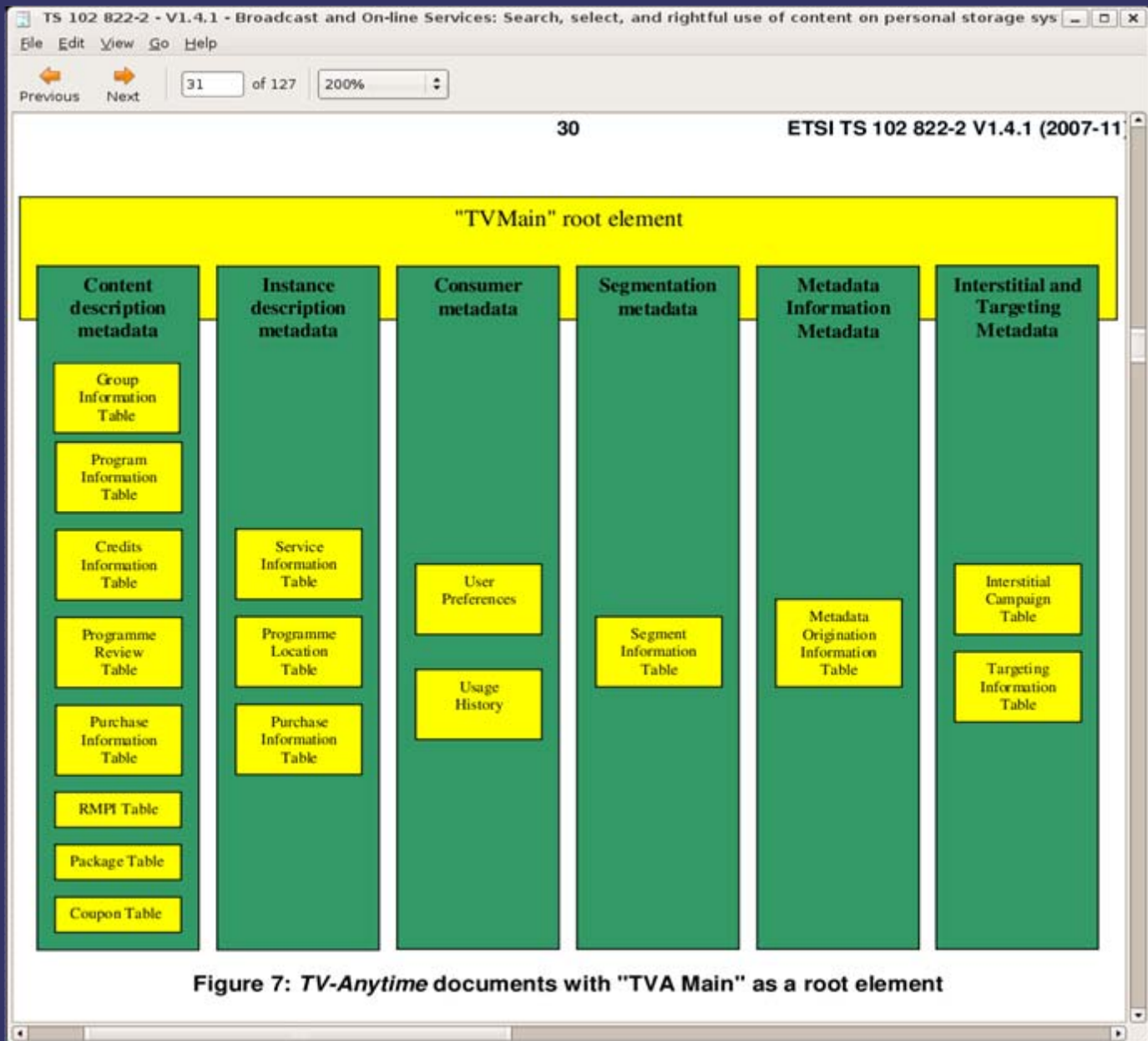


Figure 7: TV-Anytime documents with "TVA Main" as a root element

TS 102 822-2 - V1.4.1 - Broadcast and On-line Services: Search, select, and rightful use of content on personal storage systems (\*TV\*)

File Edit View Go Help

Previous Next 35 of 127 200%

35 ETSI TS 102 822-2 V1.4.1 (2007-11)

## 6.5 TV-Anytime document structure

The following example illustrates the structure of a valid *TV-Anytime* document:

```
<TVAMain xmlns="urn:tva:metadata:2007"
  xmlns:mpeg7="urn:tva:mpeg7:schema:2005"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="urn:tva:metadata:2007_schemas/tva_metadata_3-1_v141.xsd"
  version="03"
  xml:lang="en"
  publisher="..."
  publicationTime="2001-04-05T21:00:00.00+01:00">
  <CopyrightNotice>...</CopyrightNotice>
  <ProgramDescription>
  <ProgramInformationTable>...</ProgramInformationTable>
  <GroupInformationTable>...</GroupInformationTable>
  <ProgramLocationTable>...</ProgramLocationTable>
  <ServiceInformationTable>...</ServiceInformationTable>
  <CreditsInformationTable>...</CreditsInformationTable>
  <ProgramReviewTable>...</ProgramReviewTable>
  <PurchaseInformationTable>...</PurchaseInformationTable>
  </ProgramDescription>
  <UserDescription>
    <UserPreferences>...</UserPreferences>
    <UsageHistory>...</UsageHistory>
  </UserDescription>
</TVAMain>
```

Many of the elements are optional, so the following examples are also valid documents:

```
<TVAMain version="03" xml:lang="en" publisher="..." publicationTime="...">
  <CopyrightNotice>...</CopyrightNotice>
  <ProgramDescription>
  <ProgramInformationTable>...</ProgramInformationTable>
  </ProgramDescription>
</TVAMain>
```

```
<TVAMain version="03" xml:lang="en" publisher="..." publicationTime="...">
  <CopyrightNotice>...</CopyrightNotice>
  <ProgramDescription>
  <GroupInformationTable></GroupInformationTable>
  </ProgramDescription>
</TVAMain>
```

## Publish

A content service provider will publish a CRID that represents a programme series, and CRIDs that represent the constituents of that programme series. The same or different service provider will publish metadata that describes this series and its constituent episodes. The same or different service provider will publish location resolution data that describes where and when the constituent episodes of this series may be acquired. The series may be available from multiple content service providers.

In this example we will use a comedy show "Fox" which has two episodes. The included XML snippets show an almost minimal way to describe this show and its episodes. Three metadata tables are needed to describe the relations for the Fox show. The GroupInformation table that holds information for all episodes of Fox and two ProgramInformation tables that contain information for the different episodes.

The link between the group and the episodes is made by the content referencing system: if the Group CRID "://hbc/foxes/all" is put to the resolution engine in the PDR, it will come back with both programme CRIDs. The link between programmes and the group is being made by the <memberOf> element in the ProgramInformation table.

```
<ProgramDescription>
  <ProgramInformationTable>
    <ProgramInformation programId="crid://hbc.com/foxes/episode1">
      <BasicDescription>
        <Title type="main">
          The one where Fox jumps in the Potomac
        </Title>
        <Synopsis length="short">
          Fox goes to Washington and jumps in the Potomac
        </Synopsis>
      </BasicDescription>
      <MemberOf xsi:type="EpisodeOfType" crid="crid://hbc.com/foxes/all" />
    </ProgramInformation>
    <ProgramInformation programId="crid://hbc.com/foxes/episode2">
      <BasicDescription>
        <Title type="main">
          The one where Fox drowns in the Lake of Geneva
        </Title>
        <Synopsis length="short">
          Fox goes to Geneva and tries to climb the fountain
        </Synopsis>
      </BasicDescription>
      <MemberOf xsi:type="EpisodeOfType" crid="crid://hbc.com/foxes/all"/>
    </ProgramInformation>
  </ProgramInformationTable>
  <GroupInformationTable>
    <GroupInformation groupId="crid://hbc.com/foxes/all" ordered="true" numOfItems="2">
      <GroupType xsi:type="ProgramGroupType" value="show"/>
      <BasicDescription>
        <Title type="main">
          All episodes of Foxes ever
        </Title>
        <Synopsis length="short">
          More Foxes than you can handle
        </Synopsis>
      </BasicDescription>
      <MemberOf xsi:type="MemberOfType" crid="crid://hbc.com/comedy/all"/>
    </GroupInformation>
  </GroupInformationTable>
</ProgramDescription>
```



TS 102 822-2 - V1.4.1 - Broadcast and On-line Services: Search, select, and rightful use of content on personal storage systems

File Edit View Go Help

Previous Next 34 of 127 200%

that package components are matchable to usage environment characteristics).

## 6.4 Documents related through the CRID

Parts of a *TV-Anytime* document are related through the CRID. Metadata may be distributed across many *TV-Anytime* documents, but it is always possible to relate appropriate pieces through CRIDs.

### 6.4.1 Grouping

Programmes can belong to groups, and groups can belong to other groups. Linking programme descriptions with group descriptions using CRIDs reflects this relationship in the metadata, again, which is illustrated in figure 8.

```

graph LR
    PL[Program Location] --- P[Program]
    P --- PG[Program Group]
    PG --> PG
    subgraph P_metadata [ ]
        P --- PKP[key: program CRID]
    end
    subgraph PG_metadata [ ]
        PG --- PKG[key: group CRID]
    end
  
```

**Figure 8: Programme descriptions related to group descriptions through the CRID**

ProgramInformation elements are related to GroupInformation elements through the memberOf or episodeOf elements, e.g. the memberOf element contains a group CRID e.g. Foxes Episode 11 is a member of the Foxes group, which is a group that aggregates all episodes of Foxes. This supports the feature where a viewer can say "I like this. What is it? Are there more programmes like this?" By navigating up to the group the viewer may discover that the group is a member of another group and so forth. The higher one goes in the tree the more general the concepts become, e.g. moving from a specific episode of Foxes, to all episodes of Foxes, to all comedy shows, to all shows.

# *Three Use Cases – Fedora content model*

- CRIS/CERIF

  - Entity-Relationship Model -> Content model per entity, rel

  - Entity-Relationship Model -> RDF Triples

- Chemical Portal

  - XML Schemas -> One content model, many xml datastreams

- CAMMP

  - XML Schemas -> Many content models, many xml datastreams

  - XML Schemas -> RDF Triples

# Three Use Cases – Alternatives

- Relational database
  - + for simple field values, transaction oriented, SQL
  - longer text fields, bib-type queries, transformation to/from XML
- XML database
  - + flexible, Xquery
  - ? performance, utilities, bib-type queries
- Fedora
  - + all MIME-types, XML-oriented, RDF, bib-type queries, utilities
  - + CMA, ECM
  - ? approach for “sound” CMA/ECM (like 3NF for RDB)