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.

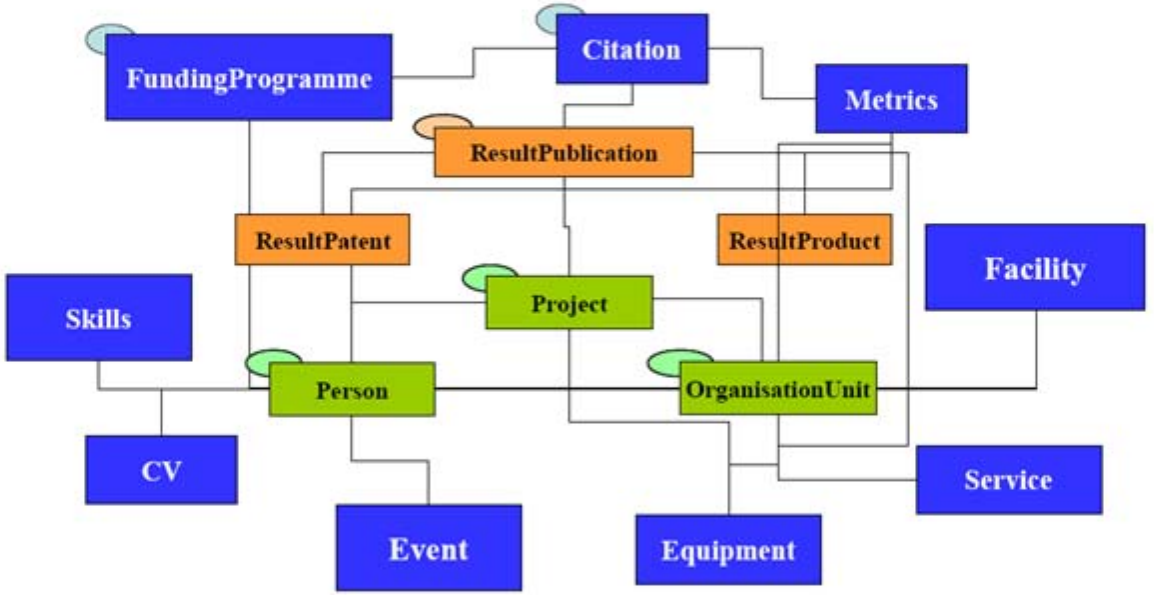


Figure 1: Some CERIF Entities and their Relationships

2.1 CERIF Conceptual Structure

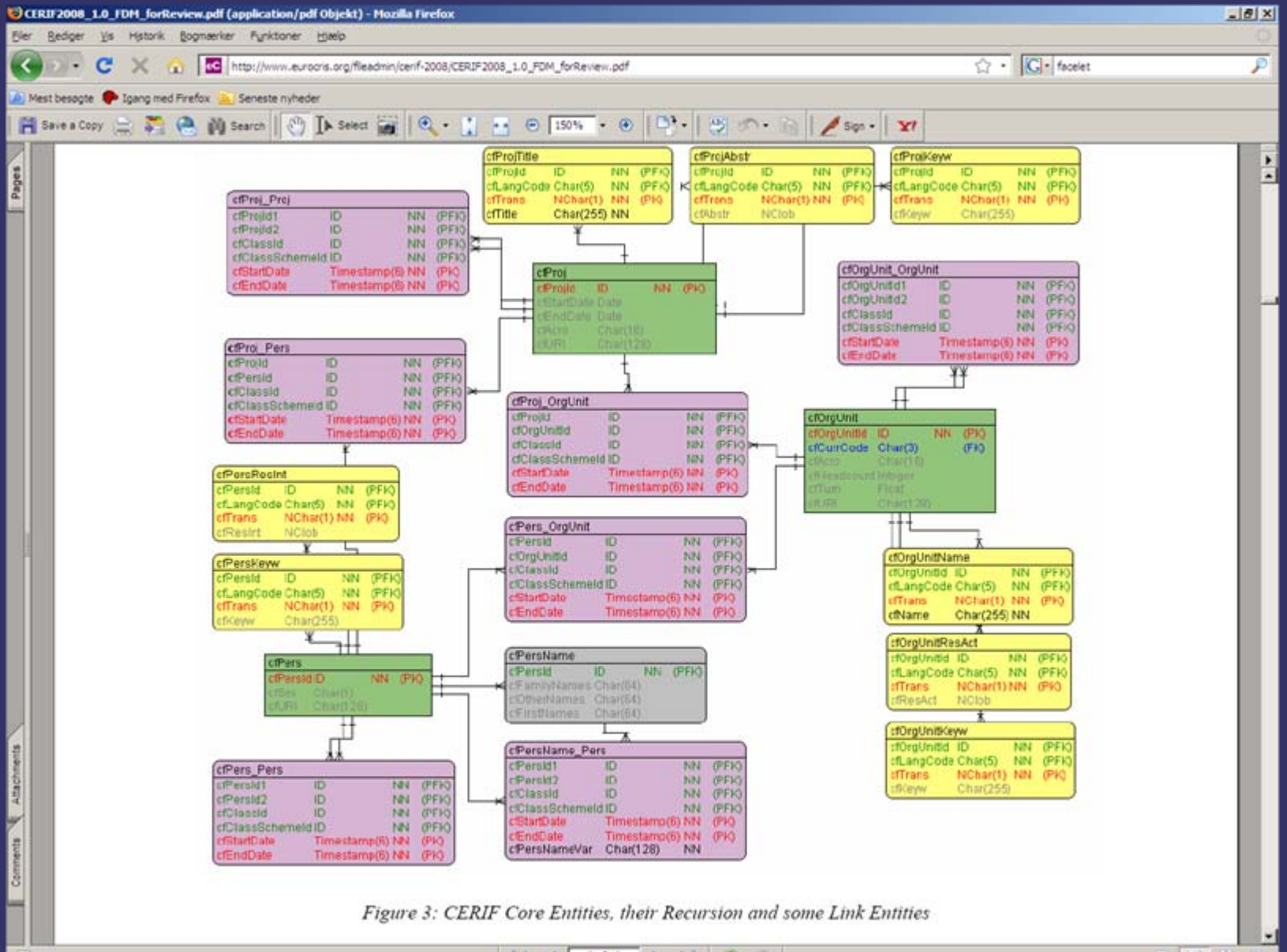
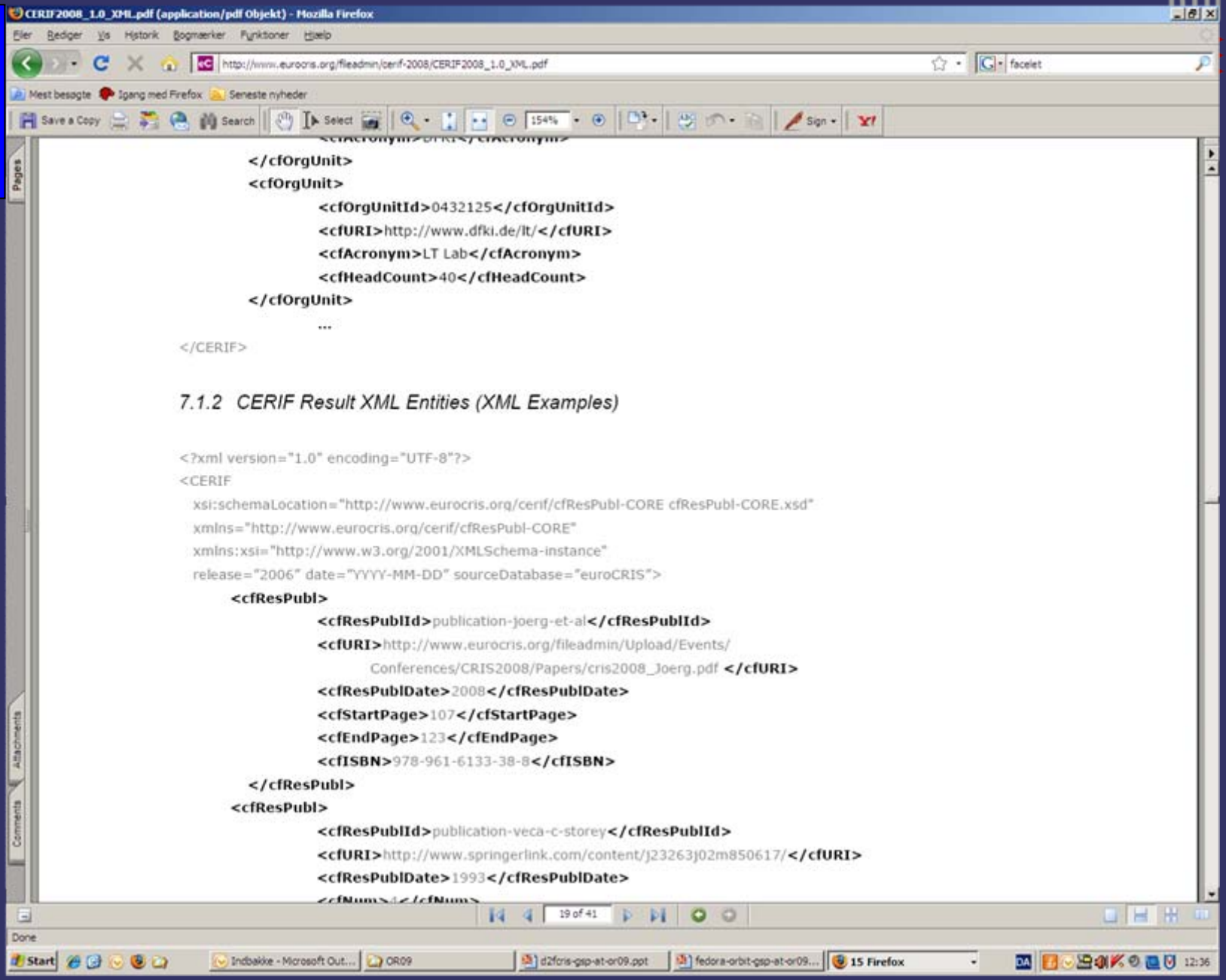
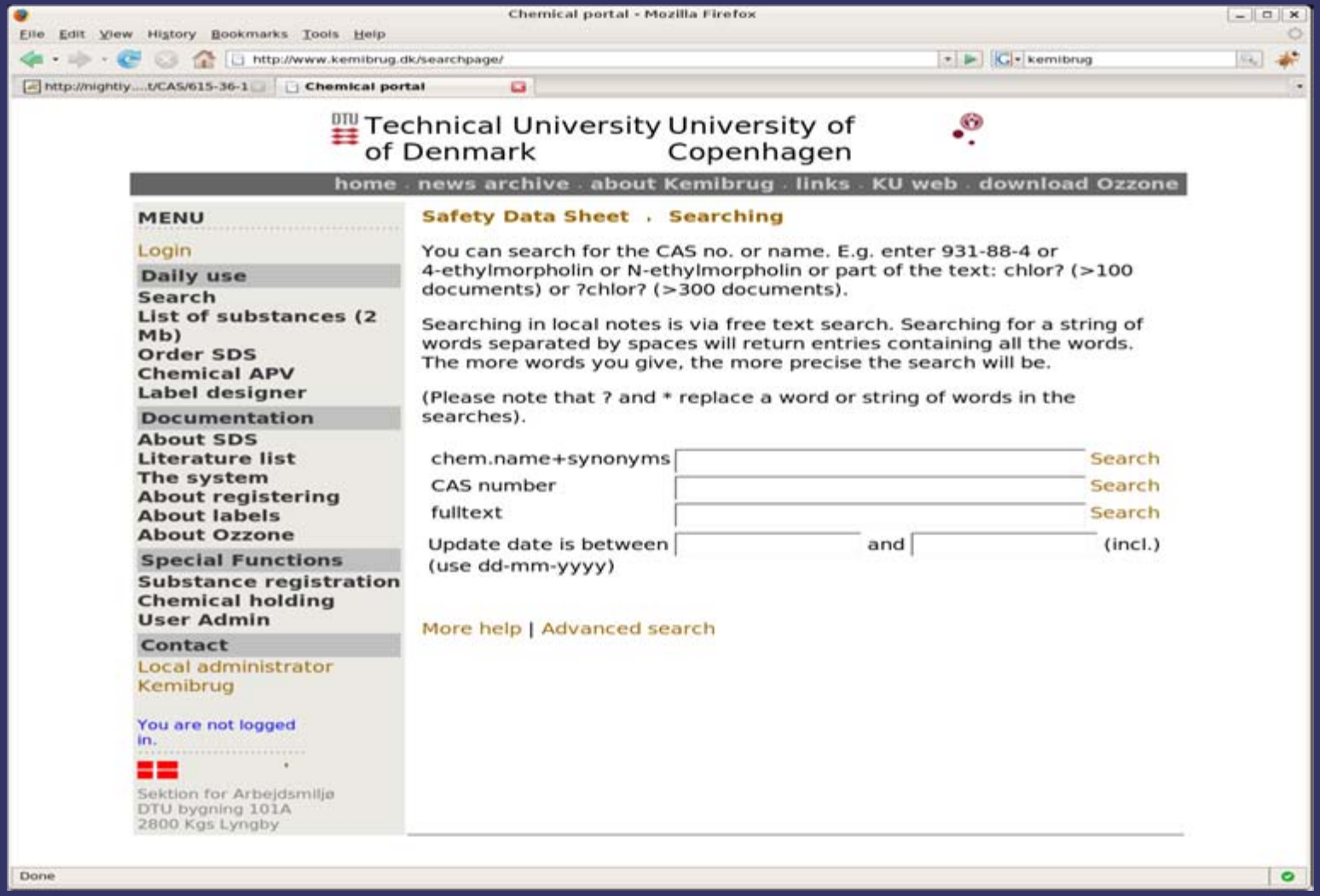


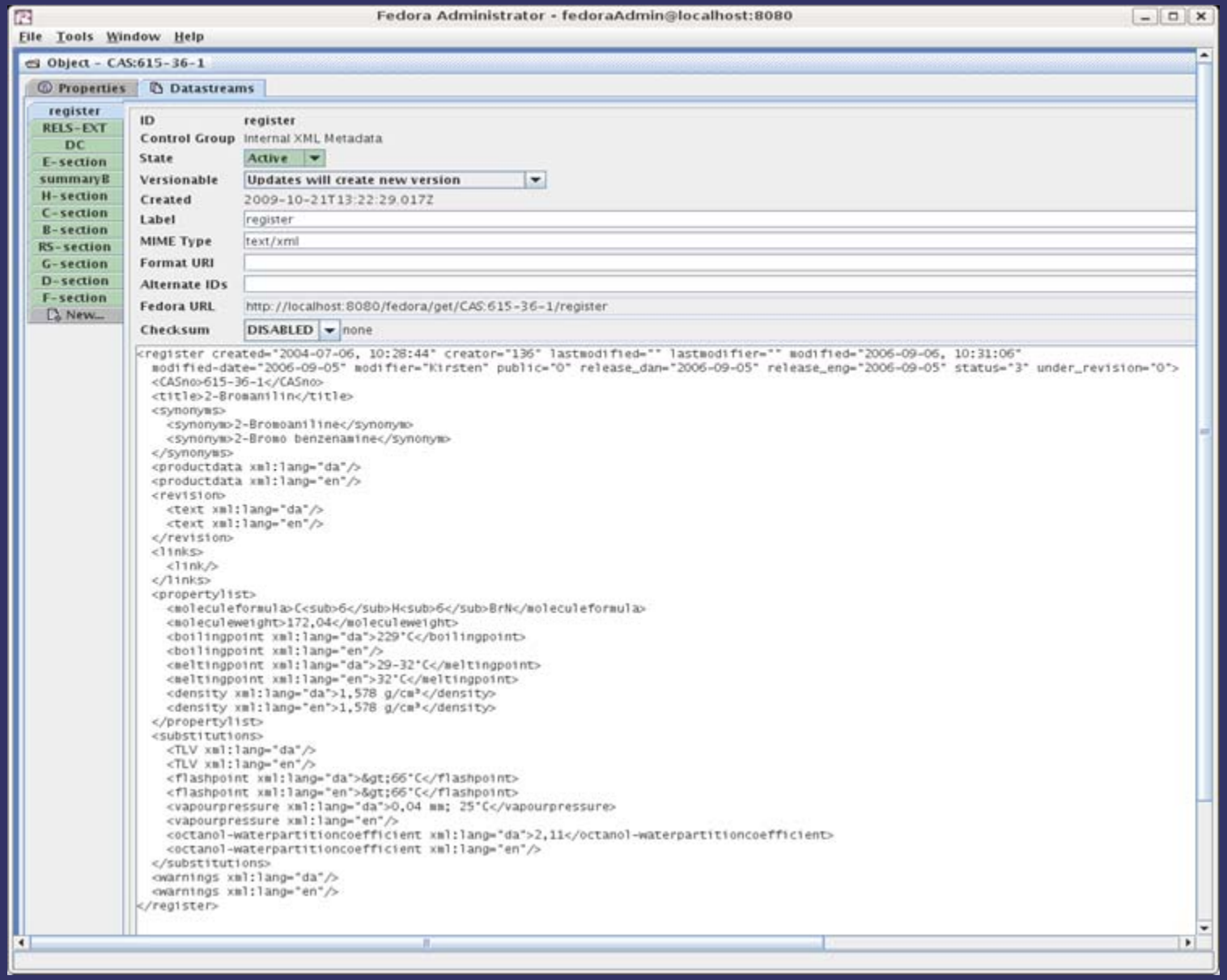
Figure 3: CERIF Core Entities, their Recursion and some Link Entities

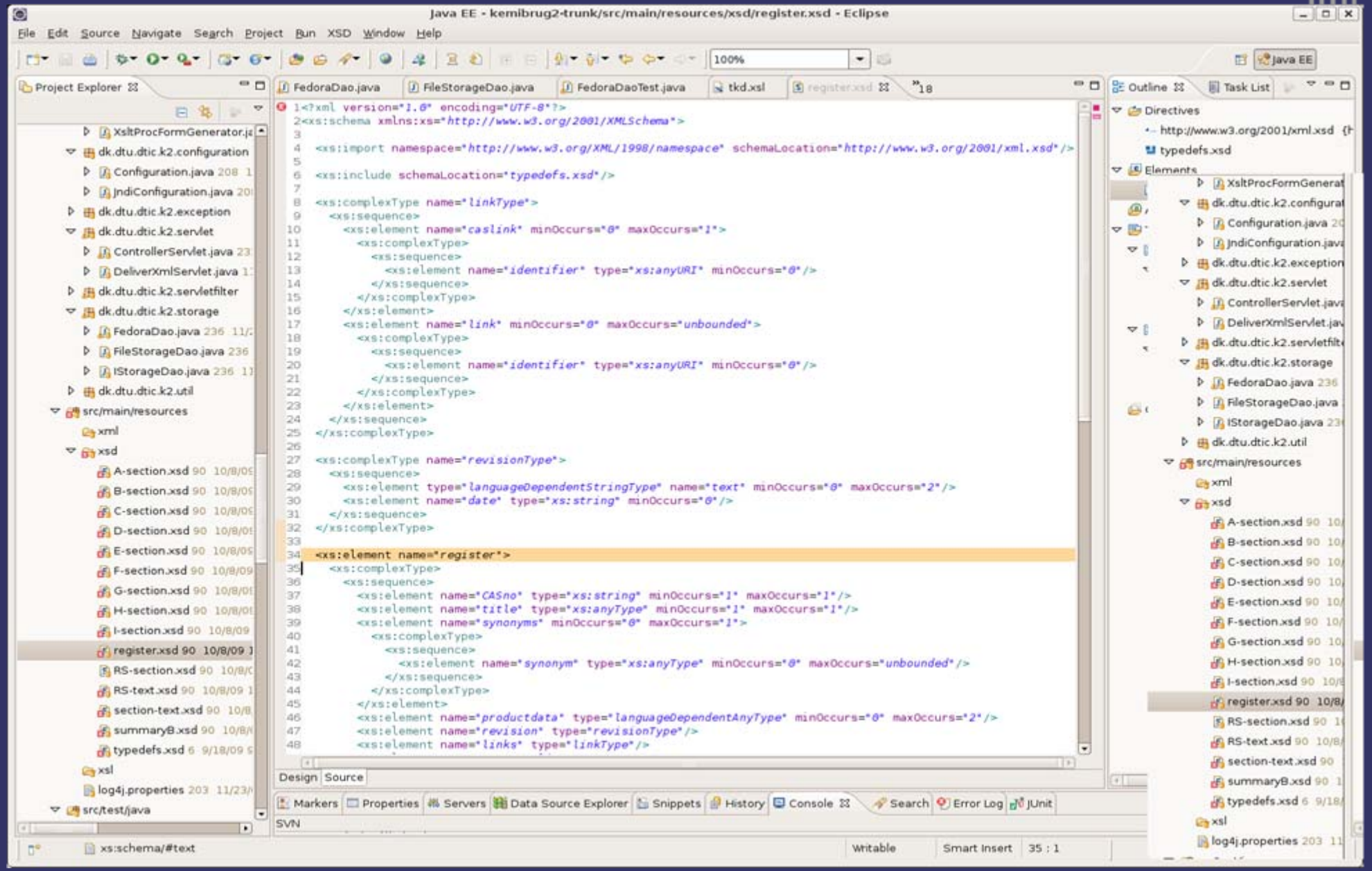
XML Data Exchange for Interoperability

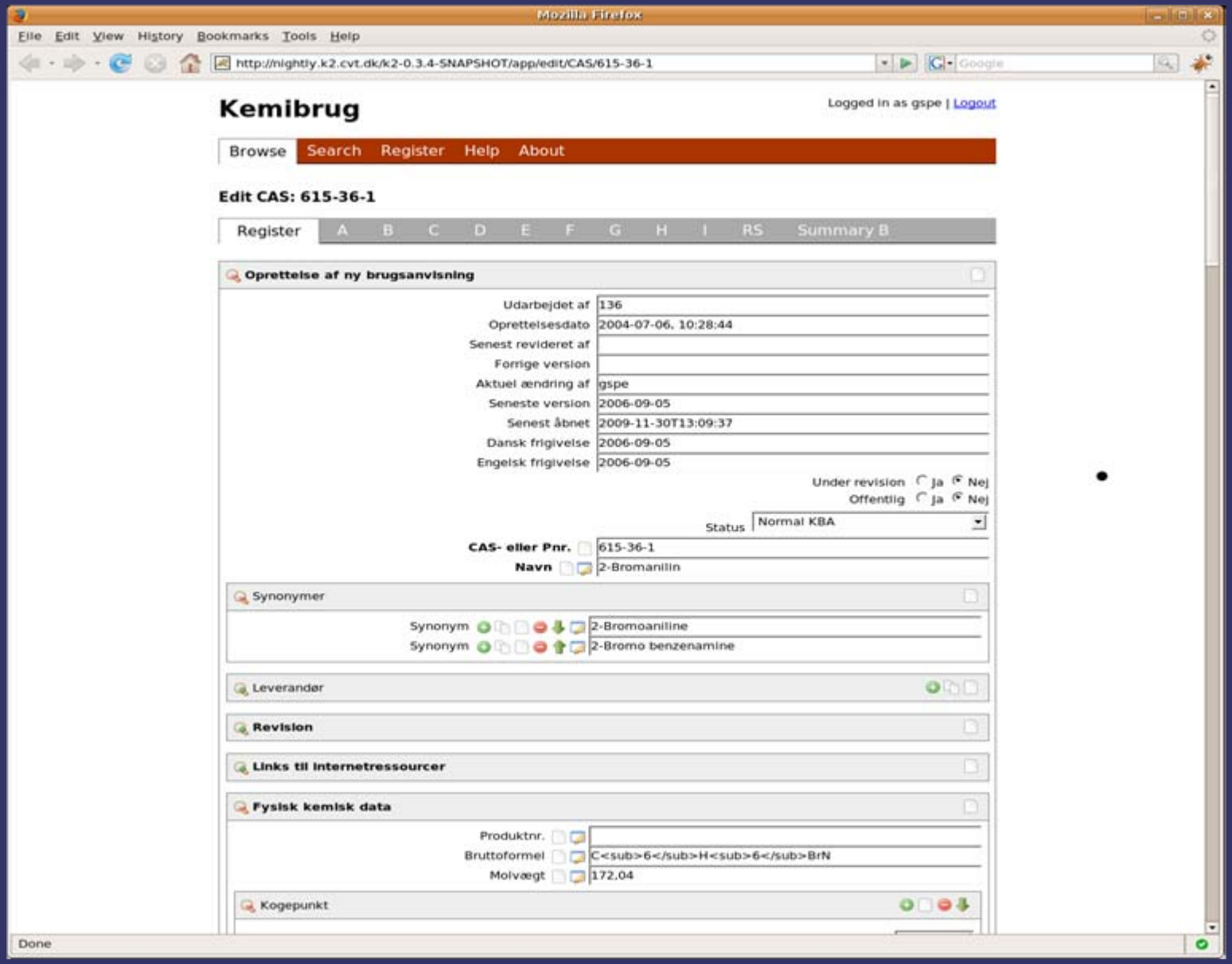


Chemical portal




Fedora Administrator - fedoraAdmin@localhost:8080
File Tools Window Help
Object - CAS:615-36-1
Properties
Datastreams
register
RELS-EXT
DC
E-section
summaryB
H-section
C-section
B-section
RS-section
G-section
D-section
F-section
New...
ID register
Control Group Internal XML Metadata
State Active
Versionable Updates will create new version
Created 2009-10-21T13:22:29.017Z
Label register
MIME Type text/xml
Format URI
Alternate IDs
Fedora URL http://localhost:8080/fedora/get/CAS:615-36-1/register
Checksum DISABLED none
<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_dan="2006-09-05" release_eng="2006-09-05" status="3" under_revision="0">
<CASno>615-36-1</CASno>
<title>2-Bromanilin</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>C₆H₆BrN</moleculeformula>
<moleculeweight>172,04</moleculeweight>
<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">>66°C</flashpoint>
<flashpoint xml:lang="en">>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>


Java EE - kemibrug2-trunk/src/main/resources/xsd/register.xsd - Eclipse
File Edit Source Navigate Search Project Run XSD Window Help
Project Explorer
FedoraDao.java
FileStorageDao.java
FedoraDaoTest.java
tkd.xsl
register.xsd
Outline
Task List
Directives
Elements
Design
Source
Markers
Properties
Servers
Data Source Explorer
Snippets
History
Console
Search
Error Log
JUnit
SVN
Writable
Smart Insert
35 : 1


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
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 Normal KBA
CAS- eller Pnr. 615-36-1
Navn 2-Bromanilin
Synonymer
Synonym 2-Bromoaniline
Synonym 2-Bromo benzenamine
Leverandør
Revision
Links til internetressourcer
Fysisk kemisk data
Produktnr.
Bruttoformel C₆H₆BrN
Molvægt 172,04
Kogepunkt
Done

CAMMP backend for service provision

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.

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

File Edit View Go Help

Previous Next 31 of 127 200%

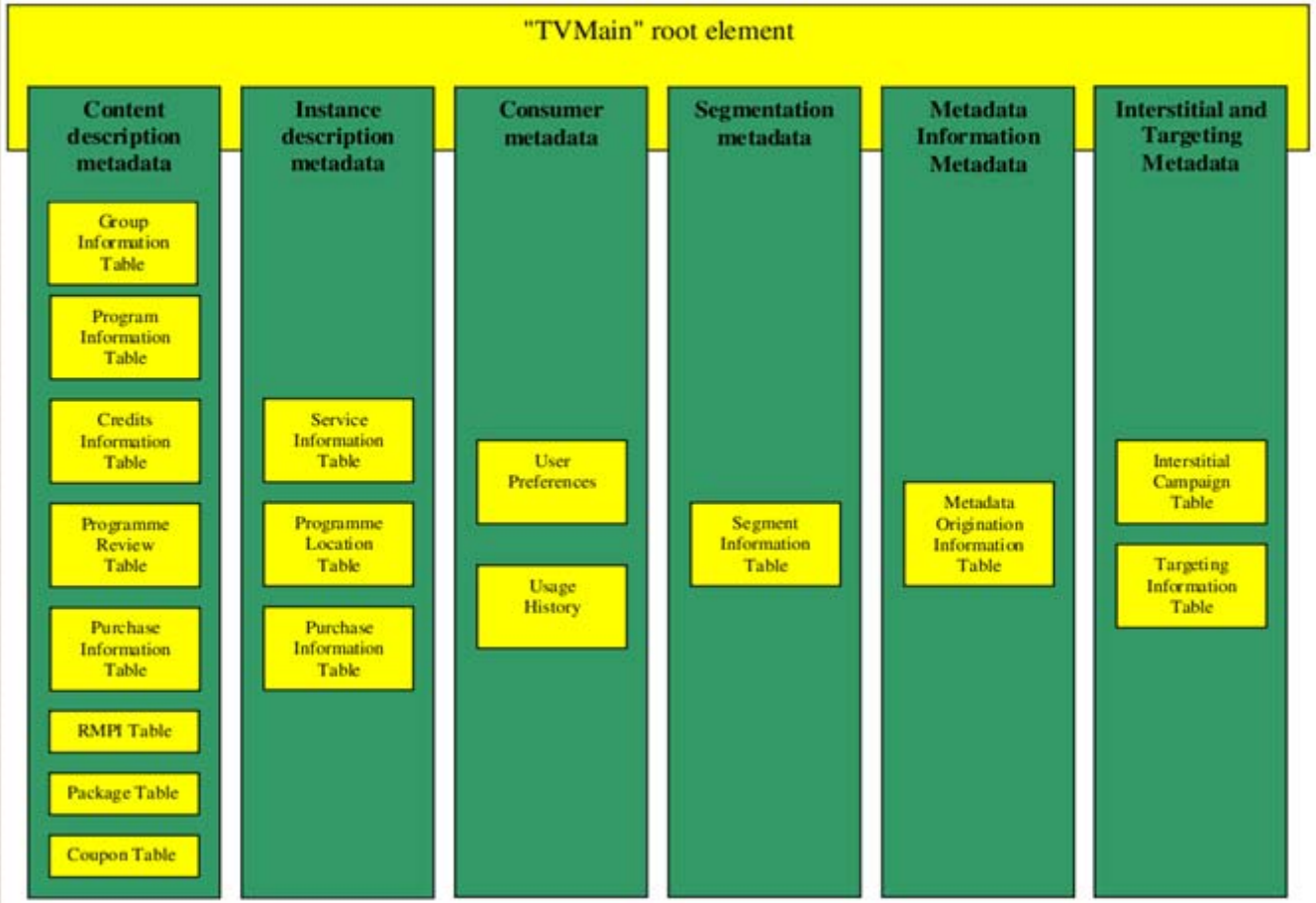


Figure 7: *TV-Anytime* documents with "TVA Main" as a root element

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%

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>
```

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 40 of 127 175%

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="ProgramGroupTypeType" 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.

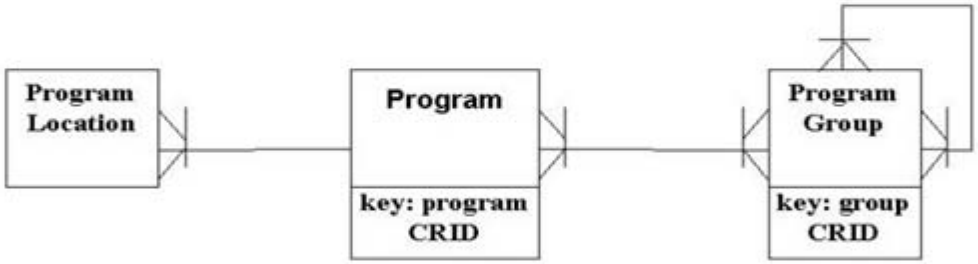


Figure 8: Programme descriptions related to group descriptions through the CRID

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)