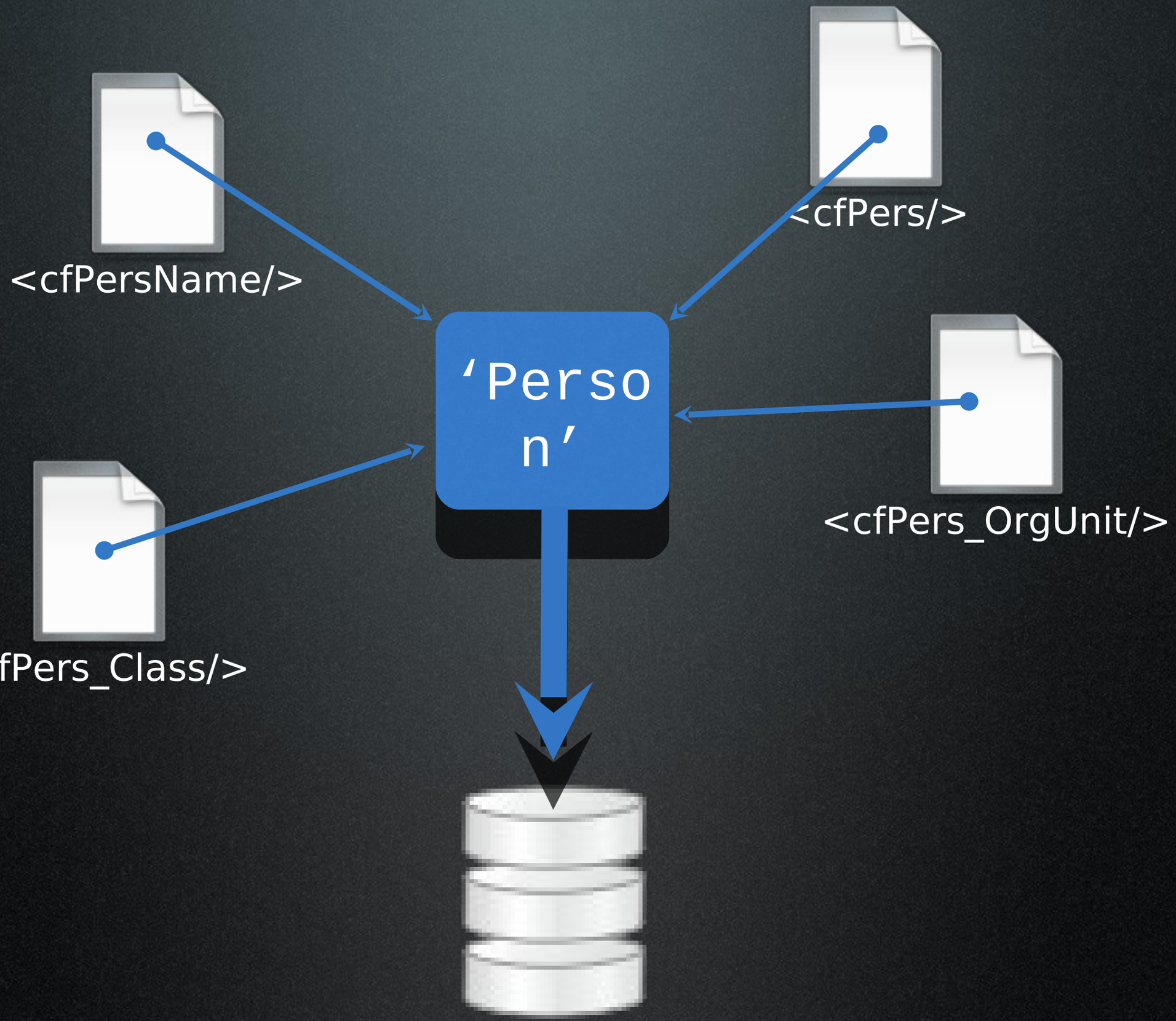
Streamlining the CERIF XML Data Exchange Format Towards CERIF 2.0

Brigitte Jörg, Innovation Support Centre, UKOLN, University of Bath
Jan Dvořák, InfoScience Praha
Thomas Vestdam, Atira A/S

Streamlining the CERIF XML Data Exchange Format Towards CERIF 2.0

The situation before CERIF 1.4 (1)

- 1:1 mapping from ER-model to the XML format
- One XML file per entity (type)
  - Every entity
- One namespace per entity-”file”
  - All had the same root element


‘Perso
n’
<cfPersName/>
<cfPers/>
<cfPers_OrgUnit/>
<cfPers_Class/>

The situation before CERIF 1.4 (2)

The Fragmentation Problem

- CRISPOOL
  - Transformation from CERIF XML was complex
- Exchange in general (web-service)
  - Format should be concise
  - Format should “support” streaming

The solution (CERIF 1.4 XML)

- One XML Schema and one namespace
  - One vocabulary
  - One file
- Make extensive use of xml-types
  - Generalisation, composition and reusability
- Allow embedding
  - Multilingual Entities
  - Link Entities

Examples (1) - Multilingual Entities

```
<cfProj>
<cfProjId>2476d360-9e21-4107-84a9-e676b7e5334d</cfProjId>
<cfStartDate>2004-03-27T00:00:00</cfStartDate>
<cfEndDate>2012-03-27T00:00:00</cfEndDate>
<cfTitle cfLang=”en” cfTrans=”o”>English original project
title</cfTitle>
<cfTitle cfLang=”fr-CA” cfTrans=”h”>Traduit titre
français</cfTitle>
</cfProj>
```

Examples (2) - Link Entities

```
<cfOrgUnit>
<cfOrgUnitId>fe0ab199-bb97-4e20-b109-
2371b82d773a</cfOrgUnitId>
<cfName cfLang=”en” cfTrans=”o”>organization name</cfName>
   <cfProj_OrgUnit>
      <cfProjId>2476d360-9e21-4107-84a9-
e676b7e5334d</cfProjId>
      <cfClassId>classification-uuid</cfClassId>
      <cfClassSchemeId>classification-scheme-
uuid</cfClassSchemeId>
      <cfStartDate>2004-03-27T00:00:00</cfStartDate>
   </cfProj_OrgUnit>
</cfOrgUnit>
```

Examples (3) - Link Entities

```
<cfProj>
    <cfProjId>2476d360-9e21-4107-84a9-e676b7e5334d</cfProjId>
    <cfTitle cfLang=”en” cfTrans=”o”>English original project
title</cfTitle>
    <cfProj_OrgUnit>
        <cfOrgUnitId>fe0ab199-bb97-4e20-b109-
2371b82d773a</cfOrgUnitId>
        <cfClassId>classification-uuid</cfClassId>
        <cfClassSchemeId>classification-scheme-
uuid</cfClassSchemeId>
        <cfStartDate>2004-03-27T00:00:00</cfStartDate>
    </cfProj_OrgUnit>
</cfProj>
```

Examples (4) - Unary Link Entities

```
<cfOrgUnit>
<cfOrgUnitId>fe0ab199-bb97-4e20-b109-
2371b82d773a</cfOrgUnitId>
<cfName cfLang=”en” cfTrans=”o”>organization name</cfName>
<cfProj_OrgUnit> ... </cfProj_OrgUnit>
<cfOrgUnit_Class>
<cfClassId>classification-uuid</cfClassId>
<cfClassSchemeId>classification-scheme-uuid</cfClassSchemeId>
<cfStartDate>2004-03-27T00:00:00</cfStartDate>
</cfOrgUnit_Class>
</cfOrgUnit>
```

Examples (5) - In symphony

```
<?xml version="1.0" encoding="UTF-8"?>
<CERIF ... >
<cfOrgUnit> ... </cfOrgUnit>
<cfPers> ... </cfPers>
<cfPers_OrgUnit> ... </cfPers_OrgUnit>
<cfClassScheme>
<cfClass> ... </cfClass>
...
</cfClassScheme>
</CERIF>
```

Examples (6) - the namespace

urn:xmlns:org:eurocris:cerif-1.4-0

- urn - the URN scheme
- xmlns - the realm of XML namespaces
- org:eurocris - the responsible organisation
- cerif - the name of the product
  - 1.4 - the release of the product
  - 0 - the release version of the XML exchange format

Additional Advantages

- Derived from CERIF ER Model
  - Auto generated based on a set of rules
- Backwards compatible
  - Just change the namespace in old files

Possibilities for improvement

- There are more options for embedding
  - Addresses and names
- Stream processing can still be difficult
  - You do not necessarily have names and links at hand when you have an entity at hand
  - Should we utilize redundancy?
  - Can we have rules for ordering of elements?

Future directions (1)

- Semantics
  - Classification schemes as imports (or in-line)
    - Validation rules?
- Context information in root element
  - "Specification" of how the XML message should be interpreted

Future directions (2)

- Use-case specific CERIF XML versus Canonical CERIF XML
  - Example in context of publications
    - The name of an author on a publication
    - Orderings (authors, etc.)
- Could convey actual semantics
- Requires transformations

Conclusion

- We now have a "modern" XML format
  - Greater flexibility and scalability with well-defined object aggregations as called for
- CERIF XML 1.4 focus more on exchange
  - ... at least that is very useful for us vendors
- We have some ideas for the future, but please share your ideas with us as well