CERIF XML update

Jan Dvořák

Andrea Bollini

Thomas Vestdam

Outline

- The Goal
- Motivation
- Solution
- Tools and processes
- When?

CERIF XML update

- A novel approach to constructing the CERIF XML exchange format
- Goals:
  - Easier to understand & use
  - More in line with general expectations

Motivation

Current CERIF XML perceived as:

- – Not easy to embed in documents
- – Fragmented
- – Unsuitable for processing of longer streams
- – Not attractive for developers

Vestdam T., Plauborg B., Van Campe L.:
CERIF XML in large scale exchange of research information.
CRIS 2016
http://dspacecris.eurocris.org/handle/11366/495

Real world CERIF usage

- Anyone needs the full CERIF at once
  (in one CRIS)?
  - Very unlikely
- Everybody needs to support
  specific information interchange scenario(s)
  - Specific subsets of CERIF
  - <u>Profiles</u>
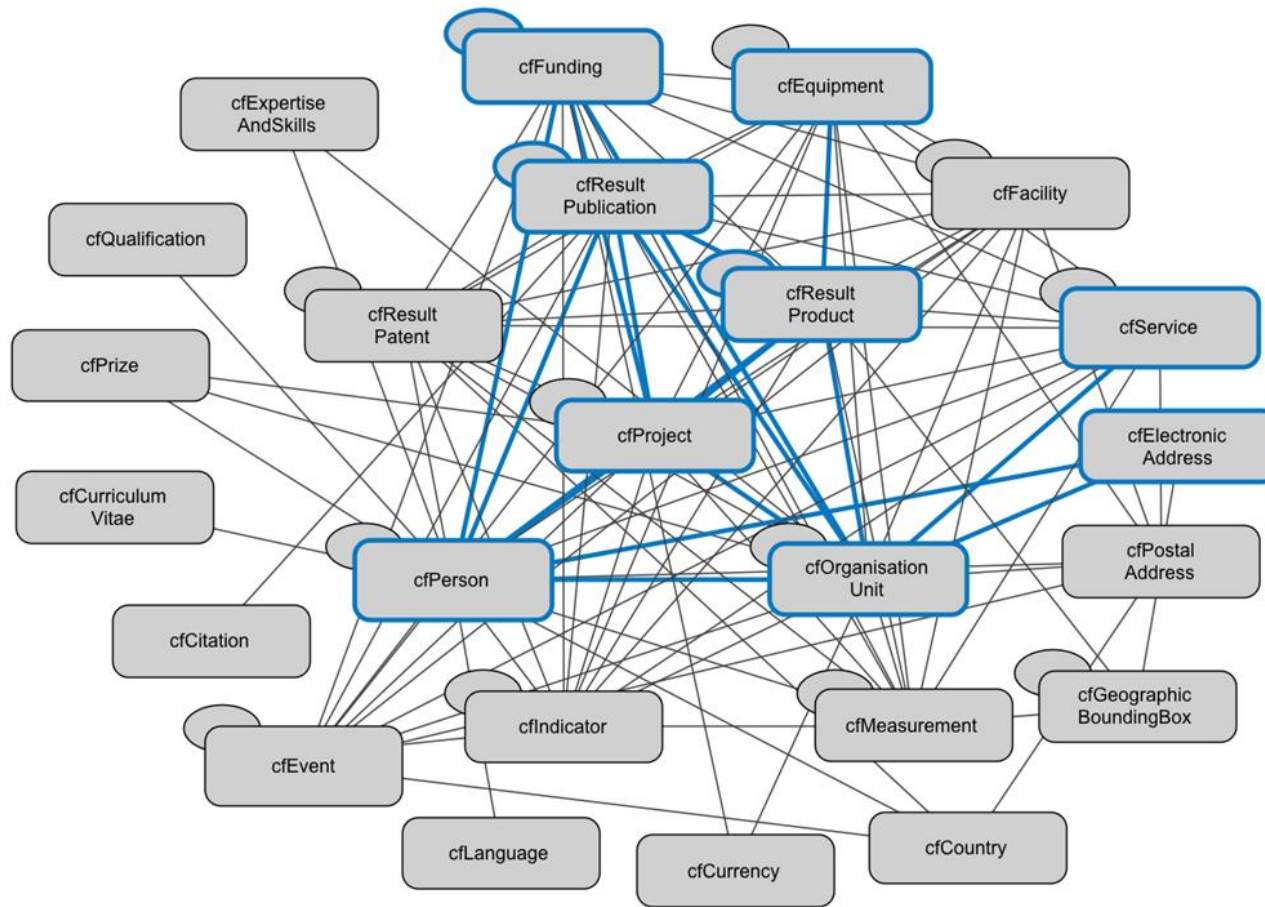
A CERIF Profile

- Entities & attributes:
  - Profile $\subset$ CERIF
- Semantic vocabularies:
  - Profile – more specific
  - Sources: CERIF & beyond
- Integrity constraints:
  - Profile $\supset$ CERIF

Profile data is CERIF

Producers know what to include

Consumers know what to expect

Example of a CERIF Profile:
The OpenAIRE Guidelines for CRIS Managers

https://doi.org/10.5281/zenodo.17065

CERIF Profile Requirements

- Easy to construct and share
- Effective to use
- Remain CERIF
  - through the links to the CERIF Data Model

CERIF Profile Artefacts

- Description

- Examples

- XML Schema

    - with CERIF annotations

Example

```xml
X *openaire_cerif_xml_example_publications.xml ☒

188 <Publication xmlns="urn:xmlns:eu:openaire:cris-mgr-guidelines-1.1-EXPERIMENTAL" id="4123451">
189     <Title xml:lang="en">The Data Model of the OpenAIRE Scientific Communication e-Infrastructure</Title>
190     <Type xmlns="http://openaire.eu/cerif/vocab/OutputTypes#">ConferenceProceedingsArticle</Type>
191     <PublicationDate>2012-11-30</PublicationDate>
192     <Volume>343</Volume>
193     <StartPage>168</StartPage>
194     <EndPage>180</EndPage>
195     <Identifier type="http://openaire.eu/cerif/vocab/IdentifierTypes#DOI">10.1007/978-3-642-35233-1_18</Identifier>
196     <Authors>
197         <Author>
198             <Person id="2123452"><!-- Paolo Manghi -->
199                 <PersonName>
200                     <FamilyNames>Manghi</FamilyNames>
201                     <FirstNames>Paolo</FirstNames>
202                 </PersonName>
203                 <Identifier type="http://openaire.eu/cerif/vocab/IdentifierTypes#ORCID">http://orcid.org/0000-0001-7291-3210
204             </Person>
205             <Affiliation>
206                 <OrgUnit id="312346"><!-- CNR -->
207                     <Acronym>CNR</Acronym>
208                 </OrgUnit>
209             </Affiliation>
210         </Author>
211         <Author>
212             <Person id="2123451"><!-- Nikos Houssos -->
213                 <PersonName>
214                     <FamilyNames>Houssos</FamilyNames>
215                     <FirstNames>Nikos</FirstNames>
216                 </PersonName>
217                 <Identifier type="http://openaire.eu/cerif/vocab/IdentifierTypes#ORCID">http://orcid.org/0000-0002-5277-285X
218             </Person>
219             <Affiliation>
220                 <OrgUnit id="312347"><!-- EKT -->
221                     <Acronym>EKT</Acronym>
222                 </OrgUnit>
223             </Affiliation>
224         </Author>
```
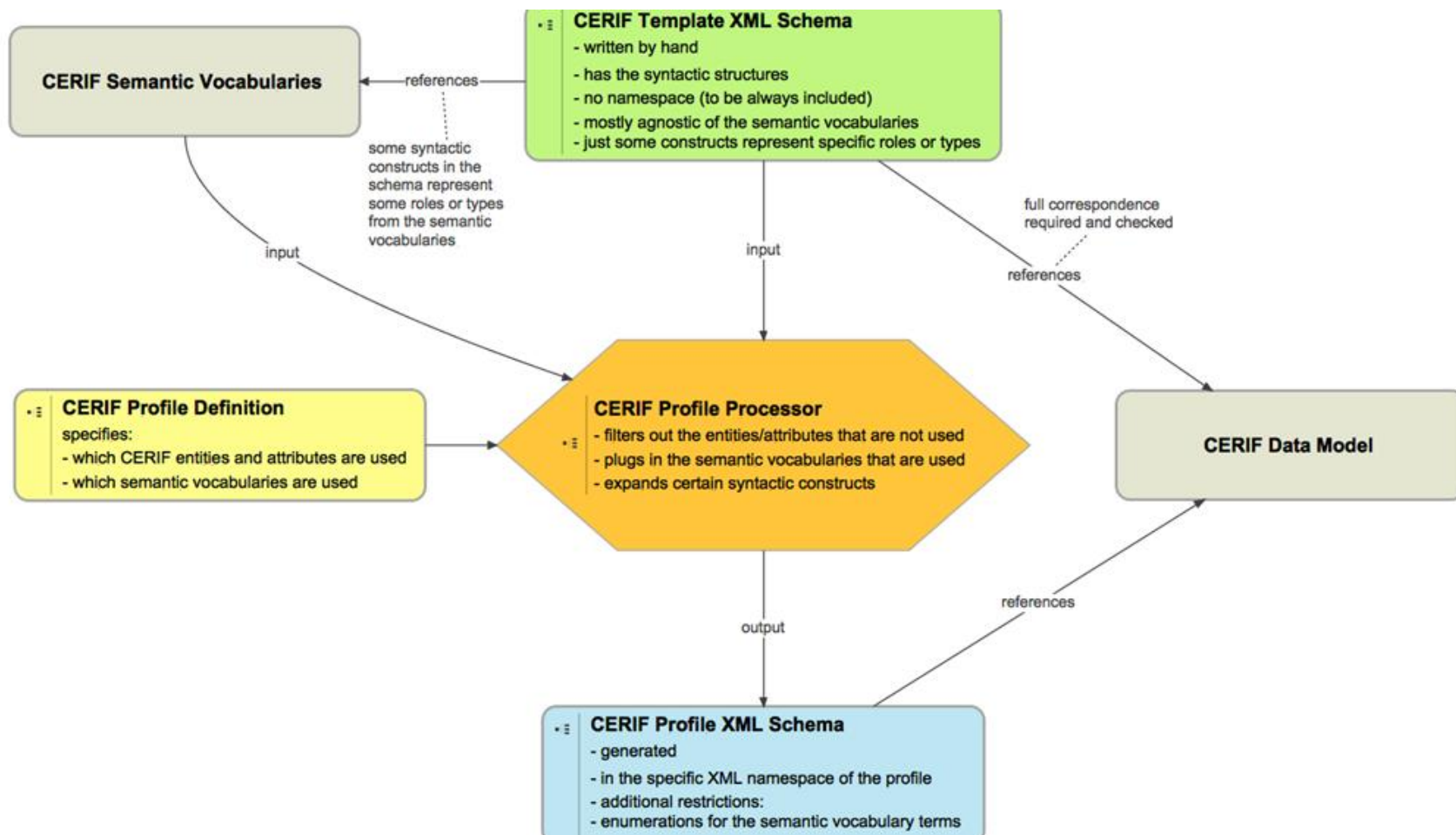
```
1  <?xml version="1.0" encoding="UTF-8" ?>
2  <!-- A fictitious project funding award notice based on a real-world project, see http://www.narcis.nl/research/RecordID/OND1
3  <FundingAwardNotice xmlns="urn:xmlns:nl:pop-project:cerif-profile-1.0-EXPERIMENTAL" xmlns:xsi="http://www.w3.org/2001/XMLSche
4      <Project id="13045b7f-ceb0-4a3f-a82d-82c19e1994e7">
5          <Title xml:lang="en">Televising Locality: Industrial, historical and socio-cultural investigations of local televisio
6          <Identifier type="http://openaire.eu/cerif/vocab/IdentifierTypes#ReferenceNumber">OND1336115</Identifier>
7          <StartDate>2010-01-01</StartDate>
8          <EndDate>2015-11-30</EndDate>
9          <Funders>
10             <Funder>
11                 <OrgUnit id="87a63334-0943-4669-a3a4-151826c45c97">
12                     <Acronym>NWO</Acronym>
13                     <Name xml:lang="en">Netherlands Organisation for Scientific Research</Name>
14                     <Name xml:lang="nl">Nederlandse Organisatie voor Wetenschappelijk Onderzoek</Name>
15                     <FundRefId>http://dx.doi.org/10.13039/501100003246</FundRefId>
16                 </OrgUnit>
17             </Funder>
18         </Funders>
19         <Consortium>
20             <Secretariat>
21                 <OrgUnit id="e77efdc0-b5f7-4851-8273-ce0fbefc4938">
22                     <Acronym>ICON</Acronym>
23                     <Name xml:lang="en">Institute for Cultural Inquiry</Name>
24                     <Name xml:lang="nl">Instituut voor Cultuurwetenschappelijk Onderzoek</Name>
25                     <PartOf>
39                 </OrgUnit>
40             </Secretariat>
41             <Collaboration>
42                 <OrgUnit id="7276968b-cb6e-4e41-bb86-61bf5054b00f">
43                     <Acronym>HUIZINGA</Acronym>
44                     <Name xml:lang="en">Huizinga Institute - The Netherlands Graduate School for Cultural History</Name>
45                     <Name xml:lang="nl">Huizinga Instituut Interuniversitaire Onderzoekschool voor Cultuurgeschiedenis</Name>
46                     <PartOf>
54                 </OrgUnit>
55             </Collaboration>
56         </Consortium>
57         <Team>
58             <Supervisor>
59                 <Person id="5bdbdd3e-369a-41ba-8ab4-e35f55a58d3c">
60                     <PersonName>
61                         <FamilyNames>Arps</FamilyNames>
62                         <FirstNames>Ben</FirstNames>
63                     </PersonName>
```

CERIF Profile Workflow

CERIF Template XML Schema
- written by hand
- has the syntactic structures
- no namespace (to be always included)
- mostly agnostic of the semantic vocabularies
- just some constructs represent specific roles or types

CERIF Semantic Vocabularies

references

some syntactic constructs in the schema represent some roles or types from the semantic vocabularies

input

full correspondence required and checked

references

CERIF Profile Definition
specifies:
- which CERIF entities and attributes are used
- which semantic vocabularies are used

input

CERIF Profile Processor
- filters out the entities/attributes that are not used
- plugs in the semantic vocabularies that are used
- expands certain syntactic constructs

CERIF Data Model

output

references

CERIF Profile XML Schema
- generated
- in the specific XML namespace of the profile
- additional restrictions:
- enumerations for the semantic vocabulary terms

Benefits

- For users:
  - Readable
  - Understandable
  - Closer to the domain
- For developers:
  - Tool-friendly
  - Streamable

CERIF development update

More like an open source software project
With feature requests, releases, ...

Using proper infrastructure:
GitHub, Wiki, Issue tracker

Using structured governance and roles:
stable members, contributors, etc.

Next Steps

Applications:
- Pilot in NL
- Snowball
- OpenAIRE

CERIF TG:
- Work out the tooling
- Implement the new working mode

Architecture TG:
- Review CERIF API to support profiles

When?

Help by joining the TG!

Embrace the updated CERIF XML
for new projects

The Future

Several CERIF profiles

All of them derive from the same model

Used
    by many parties
    around the world
    in many research information interchange
    scenarios