# Making models into reality

## The DDI Profile journey at CESSDA

*Darren Bell – UKDS, dbell@essex.ac.uk*
*Kerrin Borschewski – GESIS, kerrin.borschewski@gesis.org*

December 01, 2020 | EDDI 2020

metadata-office@cessda.eu
cessda.eu @CESSDA_Data

# About CESSDA ERIC

## About

CESSDA stands for Consortium of European Social Science Data Archives and ERIC stands for European Research Infrastructure Consortium.

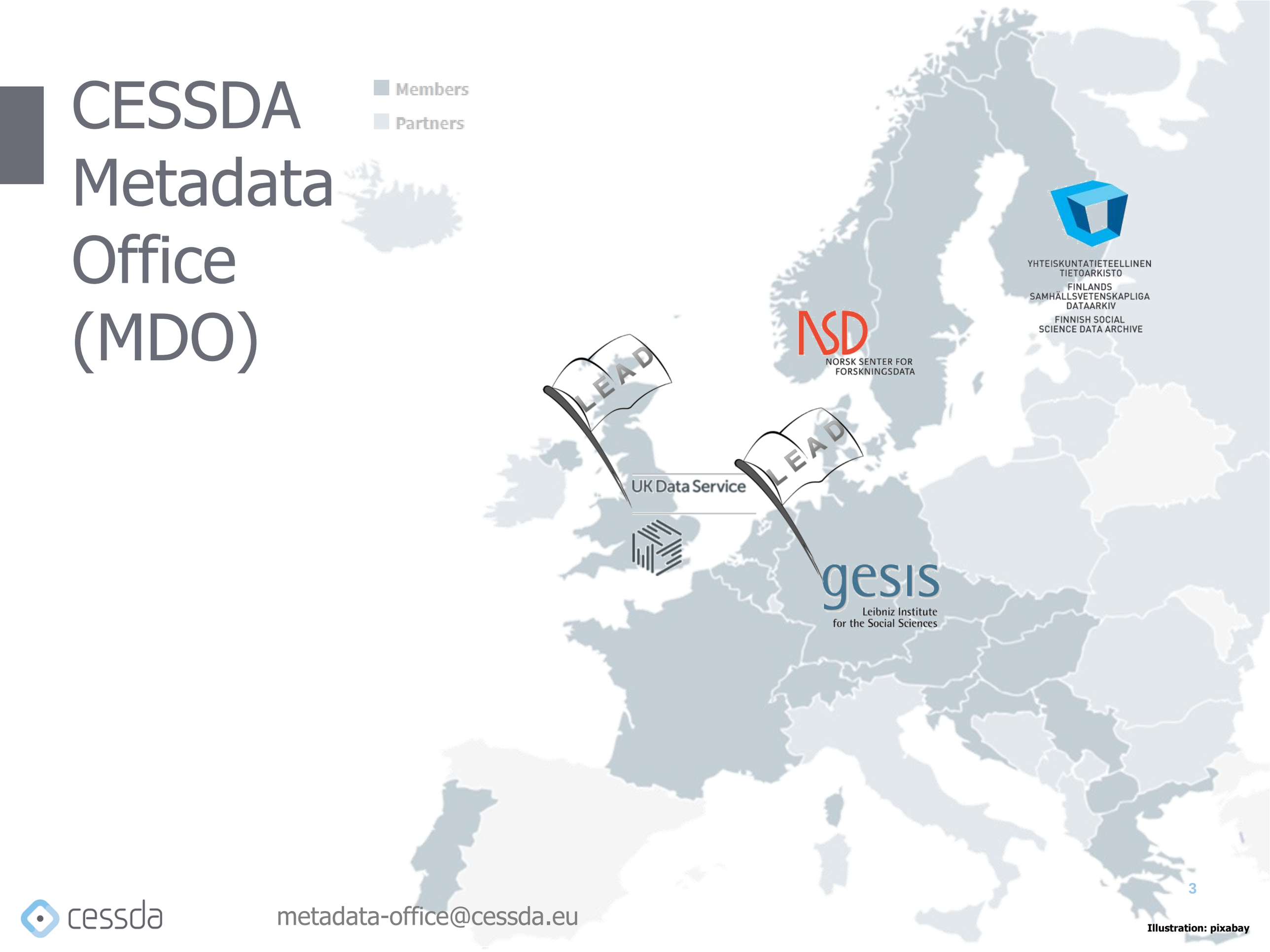
## Large-scale, integrated & sustainable data services

Aims:

- Promoting results of social science research
- Supporting national & international research & cooperation

https://www.cessda.eu/

# CESSDA Metadata Office (MDO)






cessda

metadata-office@cessda.eu

Illustration: pixabay

# CESSDA Metadata Model (CMM)




Illustration: pixabay

# From theory to the technical implementation

| No. | Element | Definition | Status (Mandatory / Recommended / Optional) | Standardized/Controlled content for this element (if DDICV: http://www.ddialliance.org/controlled-vocabularies) (or ISO / Thessaurus | Occurence | DDI 3.2 Element [remark: the X-Paths are only preliminary and exemplary mappings. They are supposed to help the understanding of the meaning of the elements. However, the final technical implementation of the elements will be up to the CESSDA SP. There will no constraint to adopt them] |
|---|---|---|---|---|---|---|
| **COMPLETE LIST OF ELEMENTS** | | | | | | |
| Information on Study: | | | | | | |
| 1 | Study | [no metadata element]<br><br>Information on the study/studies | M | | 1 | |
| **Information on Study: Bibliographic Information** | | | | | | |
| 1.1 | Bibliographic Information | [no metadata element] | M | | 1 | |
| 1.1.1 | Study ID | Identifier of the study according to DDI 3.2 structure | M (for DDI3.2) | | 1-2 | Either URN or triple Agency, ID, Version is mandatory in DDI3.2<br>URN:<br>ddi:DDIInstance/s:StudyUnit/r:URN<br>Triple Agency, ID, Version:<br>ddi:DDIInstance/s:StudyUnit/r:Agency<br>ddi:DDIInstance/s:StudyUnit/r:ID<br>ddi:DDIInstance/s:StudyUnit/r:Version |
| 1.1.2 | | | | | 1-n | |
| 1.1.2.1 | Type | | | | | |
| 1.1.3 | | | | | 1-n | |
| 1.1.3.1 | Language of Study Title | The language of the content of the element | M (for DDI3.2) | Use ISO 639-1 (Language Code) | 1 | ddi:DDIInstance/s:StudyUnit/r:Citation/r:Title/r:String/@xml:lang |
| 1.1.3.2 | Translation Status of Study Title | Is the content of the element translated? | R | true, false | 0-1 | ddi:DDIInstance/s:StudyUnit/r:Citation/r:Title/r:String/@isTranslated |
| 1.1.4 | Subtitle | Subtitle of the study | O | | 0-n | ddi:DDIInstance/s:StudyUnit/r:Citation/r:SubTitle/r:String |
| 1.1.4.1 | Language of Subtitle | The language of the content of the element | M | Use ISO 639-1 (Language Code) | 1 | ddi:DDIInstance/s:StudyUnit/r:Citation/r:SubTitle/r:String/@xml:lang |
| 1.1.4.2 | Translation Status of Subtitle | Is the content of the element translated? | R | true, false | 0-1 | ddi:DDIInstance/s:StudyUnit/r:Citation/r:SubTitle/r:String/@isTranslated |
| 1.1.5 | Alternative Title | Alternative title of the study | O | | 0-n | ddi:DDIInstance/s:StudyUnit/r:Citation/r:AlternateTitle/r:String |
| 1.1.5.1 | Language of Alternative Title | The language of the content of the element | M | Use ISO 639-1 (Language Code) | 1 | ddi:DDIInstance/s:StudyUnit/r:Citation/r:AlternateTitle/r:String@xml:lang |
| 1.1.5.2 | Translation Status of Alternative | Is the content of the element translated? | R | true, false | 0-1 | ddi:DDIInstance/s:StudyUnit/r:Citation/r:AlternateTitle/r:String@isTranslated |
| 1.1.6 | Funding Information | [no metadata element] | O | | 0-n | |
| 1.1.6.1 | Funding Agency Reference | Reference to the funding agency (the agency which funded the described entity) | O | | 0-n | Either URN or triple Agency, ID, Version and the type of referenced object are mandatory in DDI3.2<br>URN:<br>ddi:DDIInstance/s:StudyUnit/r:FundingInformation/r:AgencyOrganizationReference/r:URN<br>Agency, ID, Version:<br>ddi:DDIInstance/s:StudyUnit/r:FundingInformation/r:AgencyOrganizationReference/r:Agency<br>ddi:DDIInstance/s:StudyUnit/r:FundingInformation/r:AgencyOrganizationReference/r:ID<br>ddi:DDIInstance/s:StudyUnit/r:FundingInformation/r:AgencyOrganizationReference/r:Version |

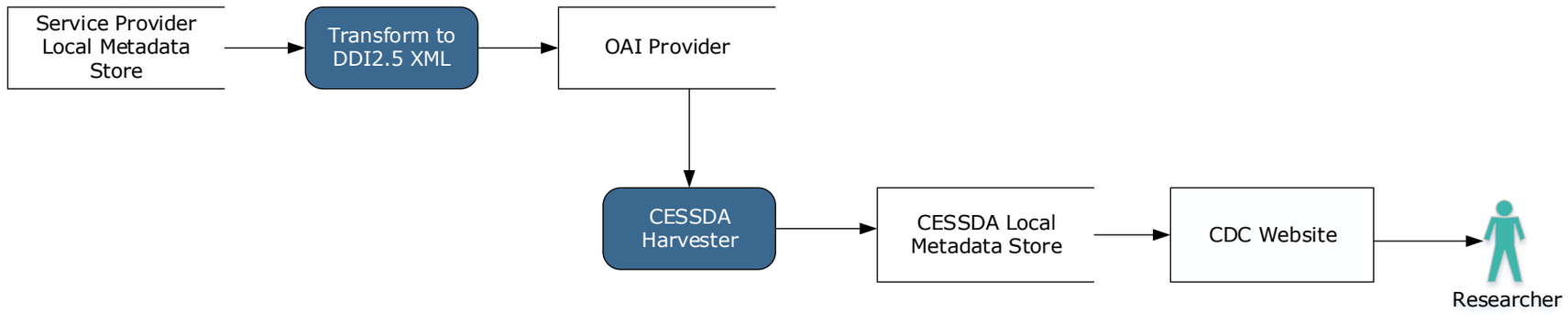
**Excel** → **XSD** → **UML** → **Profiles**



cessda metadata-office@cessda.eu

# CMM-XSD

- Problem statement
  - We want to ingest valid, well-formed metadata from multiple Service Providers into products with different sets of metadata (EQB, CDC are the obvious examples)
  - We want to build a simple, stable toolchain to achieve this
  - We want formalised, maintainable descriptions of the artefacts used to support that toolchain



# CMM-XSD

- Excel is great for information gathering but not suitable for schema design and implementation  (lacks semantics, difficult to do multi-dimensional or hierachical structures).

- Brief in 2019 was to build out a “CESSDA XSD” file with ALL elements captured so far.  Less clear what the dataflow and toolchain should be.

- Essentially, transcribed contents of existing Excel sheet into XSD schema.  Mappings specified to DDI2.x or DDI3.x as <xs:appinfo> annotations

- These can be transformed into HTML style documentation

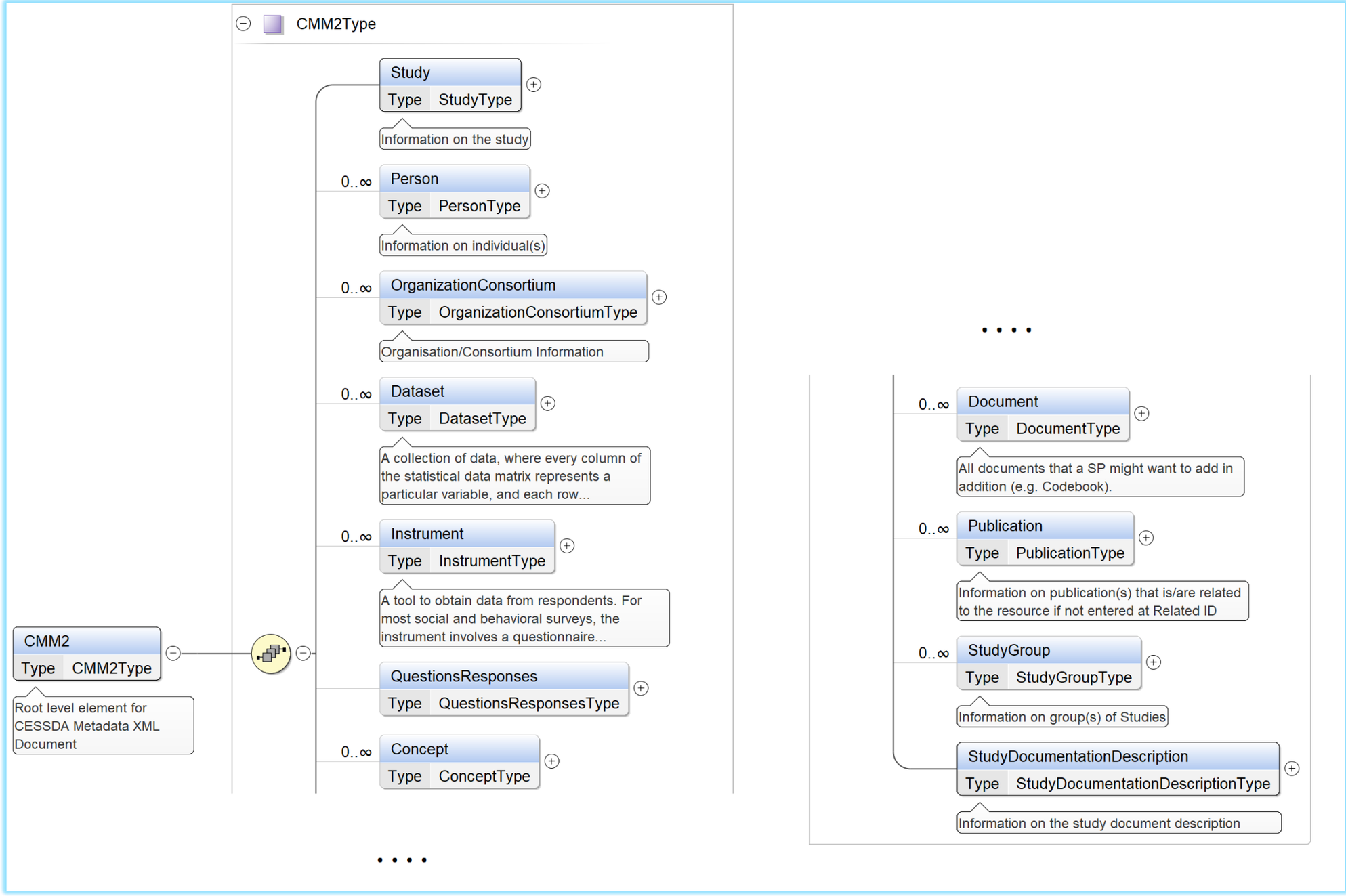
# CMM-XSD

## Example for Analysis Unit element

```
<xs:element id="CMM_01.03.05.00.00.00 " name="UnitOfAnalysis"
    type="ReferencedVocabType_UnitOfAnalysis" minOccurs="0" maxOccurs="unbounded">
    <xs:annotation>
        <xs:documentation xml:lang="en">Describes the entity being analyzed in the study
            or in the variable. The unit sampled - type of units in the dataset, e.g.
            individual, organization</xs:documentation>
        <xs:appinfo source="cmm2xpath">/CMM2/Study/MethodicalInformation/UnitOfAnalysis</xs:appinfo>
        <xs:appinfo source="ddi32xpath">/ddi:DDIInstance/s:StudyUnit/r:AnalysisUnit</xs:appinfo>
        <xs:appinfo source="usage">RECOMMENDED</xs:appinfo>
    </xs:annotation>
```

# CMM-XSD

## Top-level structure in XSD



# UML Model

- Problem with XSD is that it's a schema, not a model.
- No XML document will ever be generated that has to conform to the CESSDA XSD Schema.
- The CESSDA Metadata Model is essentially a documentary artefact that describes what elements across multiple dialects of DDI are supported by CESSDA
- XSD gives us a head start in terms of an object model and in terms formalizing objects, their  properties and attributes

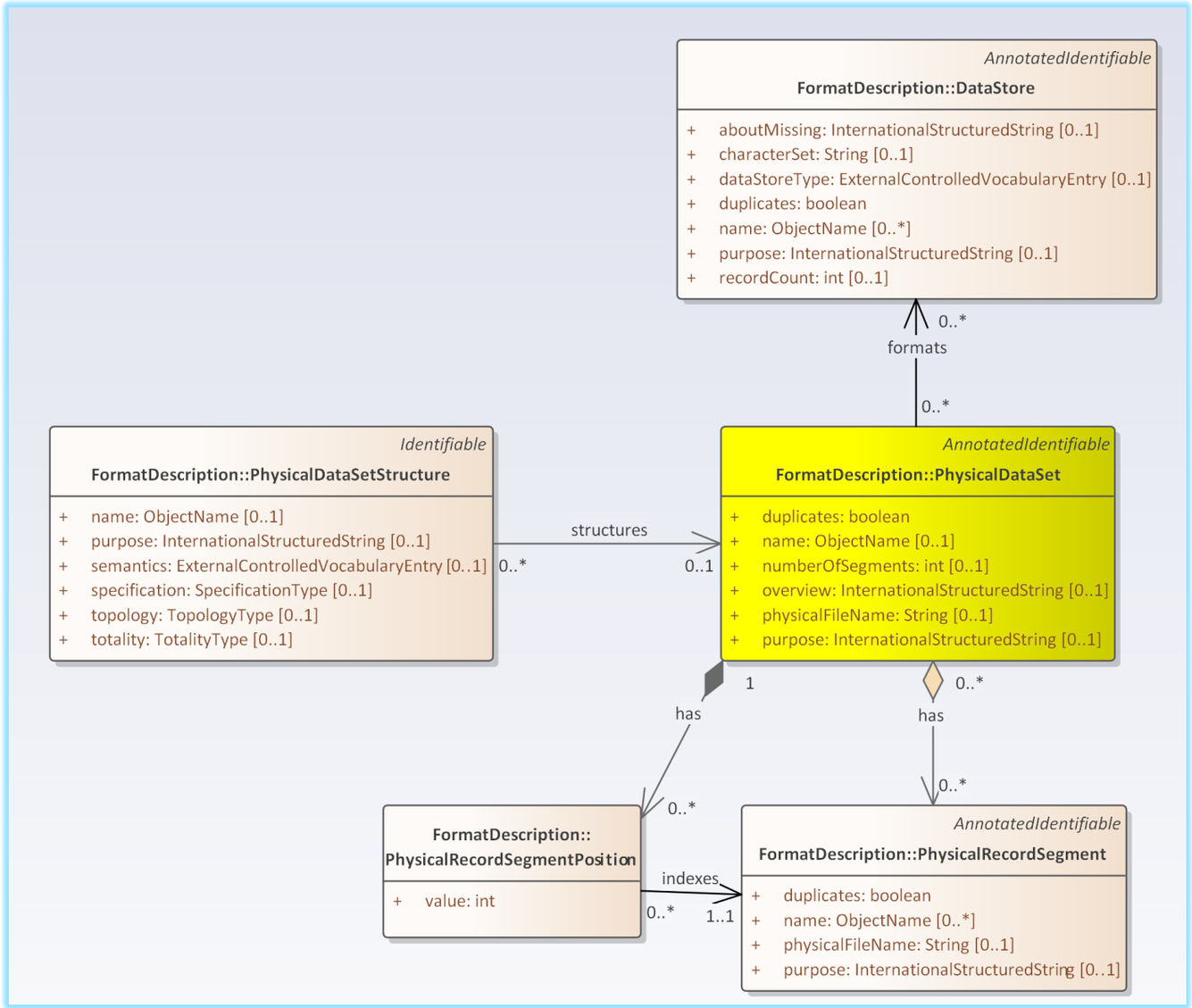

cessda metadata-office@cessda.eu

# UML Model

- “Classes” which represent objects e.g. Study, Question, Instrument and so on
- (Very) loosely analogous to modelling tables in a database but with a much higher degree of abstraction
- Each “class” can have attributes e.g. a Study class might have an **attribute** “Title”. It might have a **relationship** “*contains*” to the “Variable” class.
- Diagrams themselves stored in XML Metadata Interchange (XMI) format – portability
- Well-established toolchain in DDI (http://cogsdata.org/docs/) allowing generation of HTML documentation from XMI
- Classes can be reused (“trace relationships”) from other models

# UML Model

## Example from DDI-CDI



# DDI Profiles

- CMM UML Model under development.  This is a useful abstraction as a documentary artefact

- BUT it does not solve the problem of validating and improving Service Providers' metadata for ingest into CESSDA products on a case by case basis

- DDI Profiles are a well-established mechanism available in DDI3x to create "profiles" with clear semantics and extensibility for machine processing instructions.

- Flexibility: Can create profiles for multiple products targeted at all flavours of DDI.

- Biggest barrier to adopting this was that there was no programmatic library to validate an XML document against an XML profile.  We wrote one.

# DDI Profiles

- An **XSD schema** document (itself in XML format) defines ALL permissible elements and attributes in an XML document.
- A **“DDI profile”** document (again in XML format) defines the SUBSET of required elements from the superset defined in the XSD schema.
- Additionally, you can extend a profile with your own semantics e.g. whether an element is “Recommended” or “Optional” for example.
- When a DDI XML metadata document needs validating and checking, you can (1) use the XSD to check things like structure, element types and cardinality and then (2) use the “profile” to check that all required elements for that particular use case are present
- For example , for CESSDA Data Catalogue. @xml:lang is not mandatory in the DDI 2.5 schema but we can specify that it be mandatory in the profile.


cessda metadata-office@cessda.eu

# What a DDI profile looks like

## Header

```
<?xml version="1.0" encoding="UTF-8"?>
<!--_____________________________________________________________________
Copyright: 2020 CESSDA Metadata Office
Licence: This document is issued under a CC BY licence (https://creativecommons.org/licenses/by/4.0/)
Author: Darren Bell - UK Data Archive
Created: 2019-10-02
Version: v1.0.2
Change summary from v1.02: Adjusted version number to align with CESSDA Semantic Versioning practices
Last modified 2020-10-20 Darren Bell

_____________________________________________________________________
-->
<pr:DDIProfile xmlns:r="ddi:reusable:3_2" xmlns:xhtml="http://www.w3.org/1999/xhtml"
    xmlns:pr="ddi:ddiprofile:3_2" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="ddi:ddiprofile:3_2 https://ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/ddiprofile.xsd"
    versionDate="2020-10-20">
    <r:Agency>CESSDA</r:Agency>
    <r:ID>CDC DDI25 PROFILE</r:ID>
    <r:Version>1.0.2</r:Version>
    <r:VersionResponsibility>Darren Bell - UKDA</r:VersionResponsibility>
    <pr:DDIProfileName>
        <r:String>CESSDA DATA CATALOGUE (CDC) DDI2.5 PROFILE</r:String>
    </pr:DDIProfileName>
    <pr:XPathVersion>1.0</pr:XPathVersion>
    <pr:DDINamespace>2.5</pr:DDINamespace>
    <pr:XMLPrefixMap>
        <pr:XMLPrefix/>
        <pr:XMLNamespace>ddi:codebook:2_5</pr:XMLNamespace>
    </pr:XMLPrefixMap>
    <pr:XMLPrefixMap>
        <pr:XMLPrefix>xsi</pr:XMLPrefix>
        <pr:XMLNamespace>http://www.w3.org/2001/XMLSchema-instance</pr:XMLNamespace>
    </pr:XMLPrefixMap>
```

**Profile version**

**Profile title**

**Profile is for DDI2.5 documents**

**This profile document conforms to the DDI3.2 Profile XSD schema**

# What a DDI profile looks like

- Element detail

AuthEnty element XPath

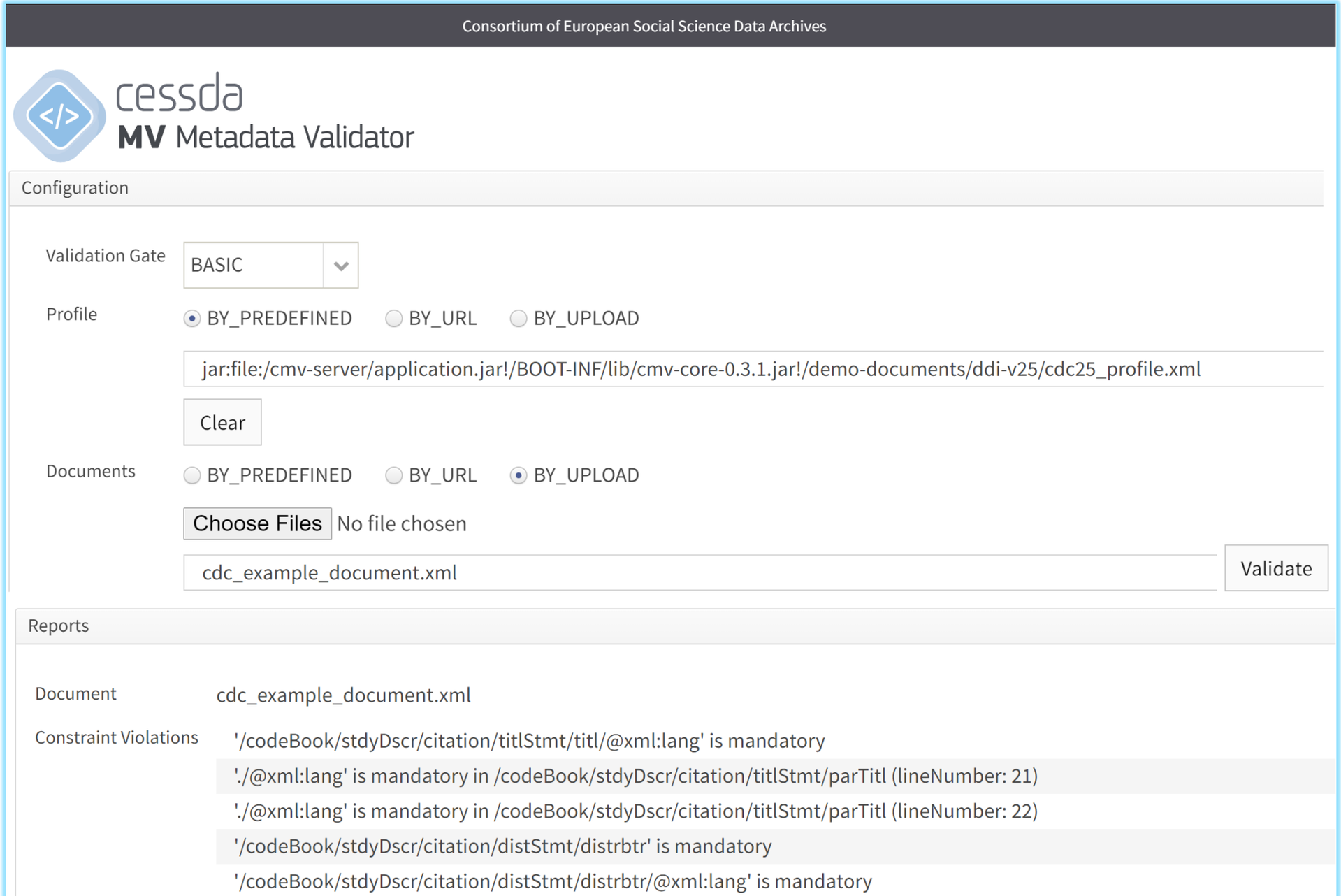
```
<!--*****************************************************************************-->
<!--PRINCIPAL INVESTIGATOR / CREATOR-->
<!--*****************************************************************************-->
<pr:Used xpath="/codeBook/stdyDscr/citation/rspStmt/AuthEnty" isRequired="false">
    <r:Description>
        <r:Content>Required: Recommended</r:Content>
        <r:Content>ElementType: Content element</r:Content>
        <r:Content>ElementRepeatable: Yes</r:Content>
        <r:Content>Usage: Principal investigator Person OR Institution</r:Content>
        <r:Content>CDC UI Label: Creator</r:Content>
    </r:Description>
    <pr:Instructions>
        <r:Content><![CDATA[
        <Constraints>
            <RecommendedNodeConstraint/>
        </Constraints>
        ]]></r:Content>
    </pr:Instructions>
</pr:Used>
```

Not mandatory

Used for user documentation

Extended semantics This element is "recommended"

# CMV


Consortium of European Social Science Data Archives
cessda
MV Metadata Validator
Configuration
Validation Gate
BASIC
Profile
BY_PREDEFINED
BY_URL
BY_UPLOAD
jar:file:/cmv-server/application.jar!/BOOT-INF/lib/cmv-core-0.3.1.jar!/demo-documents/ddi-v25/cdc25_profile.xml
Clear
Documents
BY_PREDEFINED
BY_URL
BY_UPLOAD
Choose Files
No file chosen
cdc_example_document.xml
Validate
Reports
Document
cdc_example_document.xml
Constraint Violations
'/codeBook/stdyDscr/citation/titlStmt/titl/@xml:lang' is mandatory
'./@xml:lang' is mandatory in /codeBook/stdyDscr/citation/titlStmt/parTitl (lineNumber: 21)
'./@xml:lang' is mandatory in /codeBook/stdyDscr/citation/titlStmt/parTitl (lineNumber: 22)
'/codeBook/stdyDscr/citation/distStmt/distrbtr' is mandatory
'/codeBook/stdyDscr/citation/distStmt/distrbtr/@xml:lang' is mandatory

# Published DDI profiles

Available to use in CESSDA Metadata Validator
https://zenodo.org/record/4050124

- CDC 2.5 Profile
- CDC 1.2.2 Profile

Released at end of 2020

- CDC 3.2 Profile

Released 2021

- EQB 2.5 Profile
- EQB 3.2 Profile


cessda metadata-office@cessda.eu

# Lessons learnt

- Limited semantics available natively in profile e.g. “MandatoryIfParentPresent” constraint necessary
- Review process – move from email/BaseCamp to BitBucket issue
- Different stakeholder types have different priorities
- Have a single design authority – profile development by consensus often leads to lots of changes.

# CMM & DDI profiles in 2021

Preparatory work: for new CMM version (to be released in 2022)

Updates: DDI profiles and UML model (in 2022) based on feedback of CESSDA Service Providers

metadata-office@cessda.eu



Illustration: pixabay

# Questions & Answers




cessda

metadata-office@cessda.eu

Illustration: pixabay

# Further Reading


Bell, Darren. (2020, September 28). CESSDA Data Catalogue - DDI Codebook Profile (Version 1.0.1). Zenodo. http://doi.org/10.5281/zenodo.4050124

Borschewski, Kerrin. 2020. CESSDA Metadata Office – status quo, future developments & special focus on CMM [Webinar report]. doi: http://dx.doi.org/10.5281/zenodo.4072183

Borschewski, Kerrin, André Förster, Tanja Friedrich, Wolfgang Zenk-Möltgen, Patrícia Miranda, Pedro Moura Ferreira, Jelena Banovic, Aleksandra Bradic-Martinovic, Larisa Malic, Henri Ala-Lahti, Taina Jääskeläinen, Katja Moilanen, Sunniva Hagen, Morten Jakobsen, Silje Storviken, Anne Marie Try Laundal, Katrine Utaaker Segadal, Lorna Balkan, Suzanne Barbalet, Jeannine Beeken, Darren Bell, and Sharon Bolton. 2019. CMM CESSDA Metadata Model v1.0. doi: http://dx.doi.org/10.5281/zenodo.3543756

Borschewski, Kerrin, Esra Akdeniz, Darren Bell, Alexander Mühlbauer, Jeannine Beeken, Sharon Bolton, and Taina Jääskeläinen. 2020. "The CESSDA Metadata Office: Status quo, future developments & special focus on CMM [Webinar]." 06.05.2020. doi: 10.5281/zenodo.4072184

Förster, André, Kerrin Borschewski, Sharon Bolton, and Taina Jääskeläinen. 2020 (Forthcoming). "The matter of meta in research data management: Introducing the CESSDA Metadata Office Project." IASSIST Quarterly (44, 3). doi: http://dx.doi.org/10.29173/iq970

*metadata-office@cessda.eu*

# Thank you


- Darren Bell – UKDS, dbell@essex.ac.uk
- Kerrin Borschewski – GESIS, kerrin.borschewski@gesis.org