Translation of CIDOC CRM and XML integration

FORTH
E. Tzortzakakis
European French Translation Initiative
R. Krummeich, B. David, A. Guillem, M. Van Ruymbeke

Related to

- Issue 564 : Tools and formats relevant for the translation work (technical means of the exchange)
- Issue 528 : Guidelines and Protocols for Translating CIDOC CRM

57th SIG Meeting, 9-13 October 2023, issue
Tuesday October 10th, 2023
CIDOC CRM SIG issues

528 & 564 short introduction to the XML translation interchange format

Issue 528: Guidelines and Protocols for Translating CIDOC CRM


Issue 564: Tools and formats relevant for the translation work (technical means of the exchange)

XML file that is sufficient to represent the complete contents of each stable CIDOC-CRM Stable version. Images are handled as external links.

Goals:

1. Ability to provide online the most up-to-date translation for each new CIDOC-CRM stable version
XML interchange format

2. Minimise the effort for translation of a newer version based on a previous one (may require some alignment effort)
XML schema Key-notes

- Images are included as links e.g. https://cidoc-crm.org/html/version_images_full_text/images_7.1.2/image_04.png

- The sources of the images for each version are available: https://cidoc-crm.org/figures

- Classes and properties declarations follow the exact same format as with the XML provided for declarations in the Encodings column of each stable version https://cidoc-crm.org/versions-of-the-cidoc-crm

- All other model documentation sections (apart from the classes and properties declarations) are compliant with the following schema where each section is split in title/contents/footnotes parts
section elements include child elements that require translation (title, contents, footnotes/note).

All attributes of section elements are just used for the representation of the XML in a format resembling the official CIDOC-CRM documentation. @siblingSortOrder determines the relative order of the element among the elements with the same parentDescriptiveId. It is set to be optional as it may be preferable to be removed in order to facilitate comparison.

```xml
<!ELEMENT section ((title*, contents*, footnotes?))>
<!ATTLIST section
descriptiveId CDATA #REQUIRED
parentDescriptiveId CDATA #REQUIRED
siblingSortOrder CDATA #IMPLIED
isHeaderSection (True | False) #REQUIRED
isAutoText (True | False) #REQUIRED
>
Envisioned workflow

1/Create and download an XML translation template by specifying
   ○ a Stable Cidoc-CRM version of interest e.g. 7.1.2
   ○ a language code tag for the translation e.g. el
   ○ a language name for the translation e.g. Ελληνικά
   ○ Optionally: upload the translated template file of a former version
Envisioned workflow

- Fill in the missing translation placeholders of the downloaded XML Manual; semi-automated; automated

```xml
<contents xml:lang="el"/>
</section>

<section descriptiveId="[auto] table of contents" parentDescriptiveId="">
<title xml:lang="en">Table of Contents</title>
<title xml:lang="el">Πίνακας περιεχόμενων</title>
</section>

<section descriptiveId="introduction" parentDescriptiveId="" isAutoTex="">
<title xml:lang="en">Introduction</title>
<title xml:lang="el">Εισαγωγή</title>
<contents xml:lang="en">This document is the formal definit</contents>
<contents xml:lang="el">Αυτό το έγγραφο είναι ο επίσημος όρ</contents>
</section>

<section descriptiveId="objectives of the cidoc crm" parentDescriptiveId="">
<title xml:lang="en">Objectives of the CIDOC CRM</title>
<title xml:lang="el"/>
<contents xml:lang="en">The primary role of the CIDOC CRM i</contents>
</section>
```
Envisioned workflow

3/ Upload back the XML file with the translations and download a generated HTML page that provides side-by-side comparison of the English version vs the Translated version.

Introduction

This document is the formal definition of the CIDOC Conceptual Reference Model ("CIDOC CRM"), a formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information and similar information from other domains, as further detailed below. The CRM is the culmination of more than two decades of standards development work by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). Work on the CRM itself began in 1996 under the auspices of the ICOM-CIDOC Documentation Standards Working Group. Since 2000, development of the CRM has been officially delegated by ICOM-CIDOC to the CIDOC CRM Special Interest Group (SIG). The SIG, in turn, collaborates with the ISO working group ISO/TC46/SC4/WG9 to bring the CRM to the form and status of an International Standard. This set of collaborations has resulted in the production of ISO21127:2004 and ISO21127:2014, the ISO standard editions of the CIDOC CRM. This collaboration will be continued in order to support the the next update of the ISO standard edition. The present document belongs to the series of evolving versions of the formal definition of the CRM, which serve the ISO working group as community draft for the standard. Eventual minor differences, in semantics and notation, of the ISO standard text from the present, community CIDOC CRM version, which the ISO working group requires and implements, will be harmonized in the subsequent versions of the present, community CIDOC CRM formal definition document.

Objectives of the CIDOC CRM

The primary role of the CIDOC CRM is to enable the exchange and integration of information from heterogeneous sources for the reconstruction and interpretation of the past at a human scale, based on all kinds of material evidence, including texts, audio-visual material and oral tradition. It starts from, but is not limited to, the needs of museum documentation and research based on museum holdings. It aims at providing the semantic definitions and clarifications needed to transform disparate, localised information into interoperable, comparable representations of cultural and historical phenomena.
Useful Links

GitLab issues:


Readme:


FORTH translations Website:

Problem addressed and Objectives

- Automatic integration of the markdown files of the Fr-Fr translation into the XML file for publication on CIDOC CRM website
- Facilitation of the navigation in the CIDOC CRM documentation (versions, languages etc)
- Integration of the translation information (metadata & paratext)
Integration Workflow for XML

- Integration for HTML publication

We have previously set up a Continuous Integration (YAML) workflow to build the HTML webpage of the French translation from markdown working files.

- **New**: translated text and metadata integration for XML

We define some script to do similar tasks to convert markdown to XML.

  - Step 1/ extract text from .md files and convert in XML (issue #289)

  - Step 2/ extract paratext & metadata from issues’ documentation (status, issue #285, title, labels, translation notes etc.)
Workflow for XML schema

Consolidated workflow (gitlab CI)

HTML french translation

.md files

.xml interchange format files

build

deploy

.zip

“XML workflow”

E54 Dimension

Class Name: E54 Dimension

SubClass Of: E1. Entité CRM

Scope Note:

Dimensions

This class comprises quantifiable properties that can be measured by some calibrated means and can be approximated by values, i.e., by points or regions in a mathematical or conceptual space, such as natural or real numbers, RGB values etc.

An instance of E54 Dimension represents the empirical or theoretically derived quantity including the precision tolerances resulting from the particular method or calculation.

CRM SIG translated HTML edition


E53_Lieu

Sous-classe de E1.Entité CRM

Super-classe de

Note d’application:

Cette classe permet de définir des étendues spatiales, notamment à la surface de la terre, au sens purement physique, indépendamment des phénomènes temporels et matériels.

Les implémentations de type E53_Lieu déterminent en général la position d’objets immobiliers tel qu’un bâtiment, une ville, une montagne, une rivière, ou un repère géodésique dédié. Un lieu peut être défini par la combinaison d’une localisation dans un cadre.

Il est parfois avantageux que les lieux soient mieux identifiés par des coordonnées globales ou des systèmes de référence absolus, mais des références relatives sont souvent plus pertinentes dans le contexte d’une documentation.
Starting point: .md file with markdown syntax

Step 1: transformation flow from md files to XML

- Identification of sed REGEX to collect text and create XML tags
  - for classes and properties
  - for texts
- Integration of the transformation in the consolidated workflow (gitlab CI)
  - Shell script with sed operations
  - Zip file building
- Dialog with FORTH routine for CRM SIG HTML semi-automated integration
Step 1: `sed` REGEX substitutions

Examples with CIDOC Classes

<table>
<thead>
<tr>
<th>Labels</th>
<th>Scope note</th>
</tr>
</thead>
<tbody>
<tr>
<td># E53_Lieu</td>
<td></td>
</tr>
</tbody>
</table>

markdown syntax

```markdown
s/^| Note\s\d\['\']application\[ \]*: *|\[ \]*\(.*\) *|/\t<scopeNote xml:lang="fr">\2</className>;/
```

sed rules

```bash
s/^#/ /g;
```

```
\$\#\s_/_ \$/g;
```
Step 1: REGEX operations in the XML workflow

Example of sed instructions (.sed) for properties items operated with bash script command (.sh) in XML workflow
Step 2 : Translation information & metadata

For each item or subitem, the transformation to XML may benefit information on translation process.

A new tag or attribute “status” may support information using Gitlab API:

- “iid” issue number (#163 for P20, see infra) to get access to translation notes
- “state” of the issue (open or close) respective to the state of the translation adoption flow
- “title” of issue to set values to proper item CIDOC label
- “labels” relative to the translated version for example (CIDOC CRM version, translation steps etc.)
Step 2: `curl` with GitLab API (json)

```
curl --header "PRIVATE-TOKEN: abc-123-def-456"
```

```json
{
  "id": 10307,
  "iid": 285,
  "project_id": 380,
  "title": "Construction version XML",
  "description": "Peut-on inclure l'état des tickets ? pour savoir si l'entité, la propriété est en coulisses",
  "state": "opened",
  "created_at": "2023-05-19T14:21:45.464Z",
  "updated_at": "2023-05-26T18:07:27.655Z",
  "closed_at": null,
  "closed_by": null,
  "labels": [
    "Techniques en coulisses"
  ],
  "milestone": null,
  "assignees": []
  ...
}
Step 2: Information scrapped from GitLab

Access to part of the “whole” story of the translation

- issue id, state, id class (or property) name (inversename), gitlab labels

```
curl -s --header "PRIVATE-TOKEN:$TOKEN_READ_API" "$REPO_URL_API?per_page=100"
| jq '.[]
| "(.iid):(.state):(.title):(.labels)"
| grep [EP][[:digit:]]
| sed ‘s/…/…’
```

- **translation notes:**
  - Gitlab issue URL

- **state:**
  - adopted (closed)
  - in progress (opened)

- **labels:**
  - steps of translation (3 - Traduit …, 5 - validé etc.)
  - type of item (entité, propriété, texte…)
  - actual CIDOC version under translation (latest has to be considered)
Conclusions

Main results

● Semi-automated workflow to push adopted french translation on different platform, notably CIDOC CRM website [Work in Progress]
● Tracking the information from the French initiative translation platform
● Cooperative process that benefits both partners

Further work

● Building translation metadata and further informations (translation notes e.g.) as attribute or tag instances of XML interchange format

-> to be discussed/evaluated

● Towards a fully automated workflow between fr-fr translation initiative WG to FORTH website