|  |  |  |
| --- | --- | --- |
| **Title** | Connecting CLARIN and C2CAMP: a first exploration |  |
| **Version** | 1 |
| **Author(s)** | Dieter Van Uytvanck |
| **Date** | 2018-03-08 |
| **Status** | Draft |
| **Distribution** | Internal |
| **ID** | CE-2018-1179 |

# Introduction

C2CAMP is a recently started initiative with the aim of establishing a domain of interoperable digital objects, inspired by the efforts from several working and interest groups of the Research Data Alliance:

* Data Fabric, <https://rd-alliance.org/group/data-fabric-ig.html>
* Data Type Registries, <https://rd-alliance.org/groups/data-type-registries-wg.html>

The base idea of C2CAMP is to come up with a simple yet widely accepted system for describing digital objects. Such a digital object is an online available file[[1]](#footnote-1) which:

1. Is described by metadata
2. Is stored in a trustworthy data repository
3. Has a persistent identifier
4. Has a small set of embedded metadata (“kernel information[[2]](#footnote-2)”) included into the persistent identifier, with the aim to make processing of the data easier, both for machines and human beings.

The first three requirements are already fulfilled for many CLARIN repositories[[3]](#footnote-3). The fourth one would require some additional work. Therefore, it is important to illustrate the added-value of including kernel information in persistent identifiers.

# Added value

CLARIN’s objectives include providing language analysis tools to process existing data sets. The language resource switchboard [LRS] is a module that can suggest and direct a researcher towards suitable processing tools for a specific file. To perform this matching, it needs information on the file format (mime type), the content language of the file and the task to be done – as illustrated in Figure 1. While the LRS is connected to multiple possible input sources (Figure 2), the possibility of directly processing files is currently limited in that mime type and content language are not straightforward to be derived automatically[[4]](#footnote-4).



*Figure 1: schematic overview of the Language Resource Switchboard.*

**

*Figure 2: current (solid lines) and future (dotted lines) connections between applications and the Language Resource Switchboard*

When the required meta-information would be available inside the digital object (Figure 3), it would be easier to match data and tools.



*Figure 3: schematic overview of the connection between a C2CAMP-compliant digital object and existing metadata and data within the CLARIN infrastructure*

This would already be convenient inside the CLARIN ecosystem, allowing the processing of individual files without the need of performing a deep metadata inspection. However, for files outside the CLARIN ecosystem (not being described with CMDI metadata), the advantage would be even bigger: as long as the digital object contains the necessary information, the switchboard would be able to deal with it.

# References

[LRS] <https://office.clarin.eu/v/CE-2016-0881-CLARINPLUS-D2_5.pdf>

# Recommend reading material

Common Patterns in Revolutionary Infrastructures and Data by Peter Wittenburg, George Strawn;

<http://hdl.handle.net/11304/34c5c223-adf9-49fa-9afe-968a7a932970>

C2CAMP presentation by Larry Lanom;

<https://www.rd-alliance.org/sites/default/files/CENDI-15.Nov_.17-Lannom-Final-2.pdf>

1. In principle it can also be a (recursive) collection, which ultimately leads to files. See <https://rd-alliance.org/groups/research-data-collections-wg.html> for details. For the sake of simplicity, we stick here to the notion of a single file. [↑](#footnote-ref-1)
2. See <https://rd-alliance.org/group/pid-kernel-information-wg/wiki/pid-kernel-information-guiding-principles> [↑](#footnote-ref-2)
3. There are discussions about requirement (3), issuing persistent identifiers to individual files (bitstreams). According to the current practice and B-centre requirements, each centre can choose whether or not to publish direct links to bitstreams and whether or not to issue handles to those bitstreams. This discussion falls outside the scope of this document but is obviously of importance for the potential adoption of C2CAMP by the CLARIN community. [↑](#footnote-ref-3)
4. There is some automatic mime type and even language detection included in the LRS for text files. However, this is a best-effort detection and not applicable to more complex cases like sound files. [↑](#footnote-ref-4)