# Metadata curation strategy

*Jan Odijk*
*2015-06-29*

- I assume that providers of metadata in the general case cannot improve their metadata. There may be many reasons for that, all completely valid, e.g. in many cases they have no human resources to work on improved metadata, they are often not themselves the sources of the metadata, they work with processes in which CMDI metadata are just a derivative, and the relevant improvements of the original metadata is not possible due to the limited scheme followed for these, or many other reasons.
- Therefore the strategy should be to set up a **metadata curation task force (MCTF)** to improve / extend the metadata **after harvesting.** Such extensions/ improvements should be made **intensionally** (by means of functions / transformation) so that they can be fully automatically re-executed after each new harvesting. New original metadata, or modified original metadata should be automatically reported to the MCTF so that extension / improvements for these can be made or modified. This is basically the strategy proposed in (Odijk, 2014:13-14)
- The improvements / extensions do not modify the original metadata but create a new metadata record based on the original metadata and in accordance with a profile defined by the MCTF. This avoids many problems, inter alia potential copyright problems. This profile will be set up with the purpose of discoverability in mind, and probably defines the facets for search in a faceted search engine (such as the VLO).
- The improvements can in principle be made before the metadata are processed for faceted search, or after that. The experiments I report on below actually operate on the facets and their values already produced in the VLO (on the basis of a design and implementation of the VLO taskforce).
- The MCTF defines a profile with a number of attributes that are deemed relevant for resource discovery , and their possible values, and defines a mapping from the attributes and possible/actual values in the harvested metadata to attributes and values in the MTCF profile. Some attributes will be used as facets (mainly attributes with controlled vocabulary), others will be made available for string search in the attribute's values, but no list of actual values will appear in an interface (esp. if the number of values and their nature is wildly varying)
- I carried out a small experiment for the facet *modality*, reported on already in my VLO paper (Odijk, 2014):
    - the multiple values of the modality attribute have been split up into single values.
    - Each single value is mapped to a combination of 3 attributes: modality, genre and subject.
    - Modality and Genre will have a closed vocabulary as their possible values.
- I started this time on a small experiment for the *resource type* facet.
- In both cases, values are mapped to normalized values , usually of multiple attributes.

- For both, an Excel sheet is included. These experiments must be seen as initial attempts to get to a definition of the MTCF profile mentioned above. I assume that such excel sheets (or the csv files resulting from them) can be easily turned into transformations (e.g. in XSLT).
- It is essential that new facets are added that are important for researchers to be able to discover their resources. In (Odijk, 2014) I argued that facets *for linguistic annotation* and *time period* are currently lacking but are required, at least for linguistic research. For linguistic annotation I made a concrete initial list of possible values, and we experimented with it in the CLARIN-NL portal Data Overview. Lists of required facets must be collected for each humanities discipline. Another interesting and important facet might be *research discipline*, i.e. specifying for which research disciplines a particular data set or piece of software might be relevant (which of course is somewhat subjective, and limited by the limited knowledge and expertise or fantasy of the one who fills this, but nevertheless it is useful. If needed, different facets might be used for different disciplines
- (Odijk, 2014:13) also suggested that small taxonomies for values of attributes are highly desirable. We did small experiments with that in the CLARIN-NL portal
- In order to experiment with different set-ups of the VLO, it is desirable to have full control over one's own version of it. CLARIAH-NL (DB) plans to set up such a VLO. (maybe it can be simply a branch of the EU VLO in the version control system used for that)
- We do propose all our extensions and improvements to the original metadataprovider and suggest them to make these improvements / extensions in the original sources, but we will not be dependent on whether or when they will actually do that
- We suggest that everybody who has to make new metadata follows fuidelines / recommendations / requirements that we derive from our experiences, so that hopefully future metadata will not require post-harvesting curation.

Reference:

Odijk, J. (2014) 'Discovering Resources in CLARIN: Problems and Suggestions for Solutions', internal CLARIN Report, Utrecht University [pdf] [URL]