Recommendations for the VLO-Facets

Provided by the VLO Working Group:
Thomas Eckart (Uni Leipzig), Peter Fankhauser (IDS), Susanne Haaf (BBAW),
Axel Herold (BBAW), Hanna Hedeland (HZSK), Jörg Knappen (Uni Saarbrücken),
Jens Stegmann (IMS), Thorsten Trippel (Uni Tübingen)

March 2015

I *General information*

Conventions for the Vocabulary

- Generally there should be an English version of all Metadata (though there may be exceptions); all non-English metadata should be indicated by the attribute @xml:lang, defining the language of the respective metadata
- Use existing vocabularies for metadata entries wherever possible.
- Use standardised spellings/versions of names (such as person names, organisation names, etc.).
- In general, each metadata element should be filled by exactly one metadata entry. If there are multiple information/entries for one metadata category, repeat the respective metadata element. Example:
  `<modality>speech, gesture</modality>`
  →
  `<modality>speech</modality><modality>gesture</modality>`
- Different elements should be exclusive concerning their meaning (i.e. there shouldn't be semantic overlaps).
- Prefer explicit information over unspecific element contents (e.g. `<modality>multimodal</modality>` → `<modality>speech</modality><modality>gesture</modality>`)
- Metadata entries which are meant to be integrated in the facetted search should be kept short (one-word-expressions or short phrases).

Terminological Conventions of Facet Definitions

Source Material
The original texts or objects which the resource under consideration is based on.

Resource
The digital object at hand which is described by the metadata record under consideration.

II *Search facets*

1 Facet: Collection

1.1 Definition

The collection to which a resource or tool belongs. A resource or tool can only belong to one collection. A collection is based on a specific (legacy) archive or a certain type of resource within such an archive or centre. The members of a collection share basic metadata vocabularies for the information displayed in the VLO.

1.2 Tooltip

The collection to which the resource or tool belongs

1.3 Recommended Data Categories

None. Element content of `<MdCollectionDisplayName>` is used instead.

1.4 Recommendations on the Vocabulary

A resource has to be assigned *exactly one* collection it belongs to. The usage of more than one `<MdCollectionDisplayName>` element is not valid with respect to the schema. The usage of more than one collection name within one `<MdCollectionDisplayName>` element is not supported.

2 Facet: Continent (deprecated)

2.1 Definition

The continent of origin of the source material of the resource, i.e. not the continent in which the resource was created, but where e.g. original texts were written or speech recordings made.

2.2 Tooltip

The continent of origin of the source material of the resource

→ **This facet is to be eliminated from the VLO**

3 Facet: Country

3.1 Definition
The country of origin of the source material of the resource, i.e. not the country in which the resource was created, but where e.g. original texts were written or speech recordings made.

3.2 Tooltip
The country of origin of the source material of the resource

3.3 Recommended Data Categories
http://www.isocat.org/datcat/DC-2532: location country
→ Nota bene: This data category has a broader definition than the VLO facet “Country”. Decisive for the interpretation by the VLO is the (stricter) facet definition.

3.4 Problematic Data Categories
http://www.isocat.org/datcat/DC-3792: country name
http://www.isocat.org/datcat/DC-2092: country coding
→ Both these data categories are not sufficiently concrete themselves for further interpretation within the VLO. If used, it is necessary that they are embedded in a context (of other data categories) which narrows down the possible readings of these categories to the one given in the facet definition.

3.5 Recommended Vocabulary
Codes or Names according to ISO 3166-1 (Alpha-2, Alpha-3, or Numerical); cf.http://en.wikipedia.org/wiki/ISO_3166-1, http://www.geonames.org/countries/.

4 Facet: Data Provider (deprecated)

4.1 Definition
The indication of the institution providing the metadata on the resource as being a CLARIN centre or not

4.2 Tooltip
The provider of the metadata for this resource

→ **This facet is to be eliminated from the VLO**

5 Facet: Format

5.1 Definition
The mime types of the files in the resource or consumed/produced by the tool.

5.2 Tooltip
The mime types used in the resource or by the tool

5.3 Recommended Data Categories
http://www.isocat.org/datcat/DC-2571: mime type

5.4 Recommended Vocabulary
The *Template* expressions under:
http://www.iana.org/assignments/media-types/media-types.xhtml

5.5 Open Issues
Dieter van Uytvanck may have a list of those mimetypes which are specifically suitable for linguistic resources. Send inquiry; maybe include this list in 5.4.

6 Facet: Genre

6.1 Definition
The conventionalized discourse or text type of the content of the resource, consistently applied within the collection.

6.2 Tooltip
The genre of the content of the resource

6.3 Recommended Data Categories
http://www.isocat.org/datcat/DC-2470: genre
http://www.isocat.org/datcat/DC-3899: subGenre

6.4 Recommended Vocabulary
The usage of a somehow controlled vocabulary is recommended which is

a. homogeneous in itself,
b. sufficiently documented,
c. linked to via some reference within the respective element.

7 Facet: Keywords

7.1 Definition
Keywords containing relevant information on the resource or tool not stated in other VLO metadata facets, consistently applied within the collection.

7.2 Tooltip
Keywords describing the resource or tool

7.3 Recommended Data Categories
http://www.isocat.org/datcat/DC-5436: metadata tag

7.4 Recommended Vocabulary
The usage of a somehow controlled vocabulary is recommended which is

a. homogeneous in itself,
b. sufficiently documented,
c. linked to via some reference within the respective element.

8 Facet: Language

8.1 Definition
The object language relevant for the resource or tool, i.e. the language of the source material of a resource, the object language of a language description, or the language supported by a linguistic tool.

8.2 Tooltip
The object language relevant for the resource or tool

8.3 Recommended Data Categories
http://www.isocat.org/rest/dc/2482: languageID

http://www.isocat.org/rest/dc/2484: languageName
→ N.B.: If possible, the languageName-category should only be used in combination with a language ID corresponding to ISO 639-3 (see: Recommendations on the Vocabulary).

http://www.isocat.org/rest/dc/5361: langUsage with http://www.isocat.org/rest/dc/5358: language
→ N.B.: The langUsage and language categories are modeled analogous to the corresponding TEI Header elements. They should only be used in combination with one another.

8.4 Recommendations on the Vocabulary

- Language Codes according to ISO 639-3 (Languages) or ISO 639-5 (Language Families)
- “und” (undeterminded) for resources which cannot be assigned one exact language (e.g. language independent tools)
- If a tool is language independent, please use the ISO 639-3 'und' for 'undetermined'.

9 Facet: Availability

9.1 Definition

A rough description of the conditions under which the resource or tool can be used.

9.2 Tooltip

The usage conditions for the resource or tool

9.3 Recommended Data Categories

http://www.isocat.org/datcat/DC-2453: i.e. availability, such as “free; free for academic use; restricted use; request required; user licence required; registration required; unknown”
http://www.isocat.org/datcat/DC-6846: i.e. rights, meaning “Any rights information for this resource.” (hence: very broad definition)

9.4 Recommended Vocabulary

free
free for academic use
restricted
upon-request

9.5 Open Issues

- Test implementation available under: http://aspra11.informatik.uni-leipzig.de:8080/vlo/search?0
- This facet should be accompanied by a display facet “License”.
- A mapping from different license types to the corresponding availability values has to be prepared.

10 Facet: Lifecycle Status

10.1 Definition

The status in the life cycle of the resource or tool.

10.2 Tooltip
The status in the life cycle of the resource or tool

10.3 Recommended Data Categories (not discussed yet!)
http://www.isocat.org/datcat/DC-3818: life cycle status, such as: “.”

10.4 Recommended Vocabulary
As far as possible use the vocabulary proposed in DC-3818, i.e.:
planned
development
released
production
withdrawn
retired
superseded
archived

11 Facet: Modality

11.1 Definition
The channel by which the signs in the content of the resource were transmitted or the modality for which a tool is intended, e.g. to regonize speech, gestures or entities of a text.

11.2 Tooltip
The modality of the content of the resource or intended for the tool

11.3 Recommended Data Categories
http://www.isocat.org/datcat/DC-2490: modalities

11.4 Recommended Vocabulary
(based on the examples under http://www.isocat.org/datcat/DC-2490 and the data within the VLO’s modality facet)

actOut / performance / demonstration		e.g. a cooking show
dance		
eyeGaze		
facialExpressions		
gestures	**(languageLike)**	e.g. pantomimes, emblems, sign language
	(coSpeech)	e.g. iconics, beats, cohesives, deictics, metaphorics
	(enblematics)	
image	**(painting / drawing)**	
	(photo)	
	(other)	

music	**(instrumental)**	
	(singing)	
speech	e.g. an interview, storytelling	
writing	**(print)**	e.g. a book
	(handwritten)	e.g. a letter

unspecified
other e.g. drum signals, whistling

multimodal: In case that several modalities are involved, do not provide more than one term per element. Instead make use of the respective element several times. If that is not possible, the unspecified term "multimodal" may be used.

11.5 Open Issues

- Pass the vocabulary list on to the Discipline-specific working group 6 (Speech and Other Modalities)

12 Facet: National Project

12.1 Definition

The national CLARIN project providing the resource or tool.

12.2 Tooltip

The national CLARIN project providing the resource or tool

12.3 Recommended Data Categories

/c:CMD/c:Header/c:MdCollectionDisplayName/text()

12.4 Recommended Vocabulary

CELR
CLARIN-AT
CLARIN-D
CLARIN-DK-UCPH
CLARIN-EU
CLARIN-NL
CLARIN-PL
CLARINO
SWE-CLARIN
Other

12.5 Open Issues

- CLARIN-DK anfragen, wie sie heißen wollen (CLARIN-DK oder CLARIN-DK-UCPH)

- a description of the resources harvested as well as the criteria for harvesting would be helpful

13 Facet: Organisation

13.1 Definition
The name of the organisation currently responsible for the resource or tool, i.e. to be contacted with any questions or requests regarding the metadata or access to the resource/tool.

13.2 Tooltip
The organisation currently responsible for the resource or tool

13.3 Recommended Data Categories
www.isocat.org/datcat/DC-2459: organization

13.4 Problematic Data Categories
www.isocat.org/datcat/DC-2979: Organisation
→ Underspecified, i.e. this DC doesn't specify the purpose of the organisation named here in the context of the resource.

13.5 Eliminated/Ignored Data Categories
www.isocat.org/datcat/DC-6134:publisher (from TEI)
http://purl.org/dc/terms/publisher: publisher (Dublin Core)
→ Both underspecified, i.e. this DCs don't specify the entity named here (organisation, person, etc.) now the purpose of the entity in the context of the resource.

13.6 Recommendations on the Vocabulary
(1) There should be an english reading provided for each institution name which will be represented within the VLO. The english name should be marked as such by usage of a language-defining attribute and an ISO 639-3 language code (e.g. `@xml:lang="eng"`).
(2) There should only be one variant of an institution name used within all metadata records provided by this respective institution.

14 Facet: Project

14.1 Definition
The name of the projects originally involved in the creation of the resource or tool. These projects may no longer exist and are usually not the ones to be contacted regarding the resource/tool.

14.2 Tooltip
The project within which the resource was created

14.3 Recommended Data Categories
http://www.isocat.org/datcat/DC-2536: project name
http://www.isocat.org/datcat/DC-2537: project title

14.4 Recommendations on the Vocabulary
open

15 Facet: Resource Type

15.1 Definition
The type of the resource or tool (e.g. corpus, lexicon, grammar, tool...).

15.2 Tooltip
The type of the resource or tool (e.g. corpus, lexicon, grammar, tool...)

15.3 Recommended Data Categories
http://www.isocat.org/datcat/DC-3806: resource class
http://www.isocat.org/datcat/DC-5424: type → based on the TEI Header definition of type

15.4 Problematic Data Categories
http://purl.org/dc/terms/type
http://purl.org/dc/elements/1.1/type
→ This DC is underspecified since “genre” is included in the definition as well. If used, it is necessary that it is embedded in a context (of other data categories) disambiguating the possible readings of this category.

15.5 Recommended Vocabulary
(based on the examples under http://www.isocat.org/datcat/DC-3806 and the data within the VLO’s Resource Type facet)

audioRecording
collection i.e. a collection which cannot be used as a corpus in a linguistic sense
corpus
database
dataset:experimentalData
dataset:fieldworkMaterial
dataset:surveyData
dataset:testData
image i.e. a digital image

lexicalResource
physicalObject e.g. book, picture, photo, stone with inscription, ...
transcribedText
annotatedText
tool
toolChain
videoRecording
webApplication
webService

unspecified
other

Excluded DC3806-vocabulary
grammar → rather belongs to facet genre
lexicon → use lexicalResource instead
resourceBundle → use collection instead
teachingMaterial → rather belongs to facet genre

16 Facet: Subject

16.1 Definition
The subject or topic of the content of the resource, consistently applied within the collection.

16.2 Tooltip
The subject or topic of the content of the resource

16.3 Recommended Data Categories
http://purl.org/dc/terms/subject: subject
http://purl.org/dc/elements/1.1/subject: subject

http://www.isocat.org/datcat/DC-6147: domain of use

http://www.isocat.org/datcat/DC-5316: classification code
→ This DC is designed analogous to the TEI-Header element classCode. It is thus underspecified for the subject facet. Hence, when used the classification scheme has to be determined and the usage of this element according to the definition of the subject facet should be specified by the context somehow.

16.4 Recommended Vocabulary
The usage of a somehow controlled vocabulary is recommended which is

a. homogeneous in itself,

b. preferably sufficiently documented.

17 Facet: Temporal Coverage

17.1 Definition

The temporal coverage of the source material of the resource, i.e. not the time within which the resource was created, but when e.g. original texts were written or speech recordings made.

17.2 Tooltip

The temporal coverage of the source material of the resource

17.3 Possible Data Categories

(1) http://www.isocat.org/datcat/DC-3664: Time coverage

(2) http://www.isocat.org/datcat/DC-3654: Start range
→ This data category has to be used together with DC-3655

(3) http://www.isocat.org/datcat/DC-3655: End range
→ This data category has to be used together with DC-3654

17.4 Problematic Data Categories

(1) http://www.isocat.org/datcat/DC-4343: interval
→ Underspecified ("a : a space of time between events or states"): could be filled with values like "200 years", "a decade", etc.

(2) http://www.isocat.org/datcat/DC-5742: End time
→ incomplete; corresponding start time category missing

(3) http://www.isocat.org/datcat/DC-2502: time coverage
→ Definition ("The time period that the content of a resource is about.") does not suit the focus of this facet (cf. 17.1)

17.5 Proposed vocabulary

In case of (1): Open Date Range format www.ukoln.ac.uk/metadata/dcmi/date-dccd-odrf

In case of (2) and (3): W3C DateTime
http://www.w3.org/TR/NOTE-datetime