

# Using GitBook for user-friendly documentation of a vocabulary

**Simon Musgrave**

Language Data Commons of Australia / University of Queensland

# LDaCA vocabulary - background

- Based on the metadata schema developed by Open Language Archives Community (OLAC) c20 years ago
  - Part of broader Open Archives Community (therefore ultimately DC base)
  - Input also from project Electronic Metadata for Endangered Language Documentation
- These origins mean some gaps in the schema
  - E.g. the possibilities for LinguisticGenre include Formulaic and Ludic but not Informational
- We have added classes, properties and terms as needed
  - First preference: schema.org as source
  - Second preference: Other linkable sites
  - Then: 'home-grown'

# Versions of the vocabulary

- Initially, we made the vocabulary available in two formats:
  - a. json-ld
  - b. Documentation automatically generated from a.
- Both available at:  
<https://github.com/Language-Research-Technology/language-data-commons-vocabs/>

# The formal (machine-readable) version

```
399 {
400   "@id": "https://purl.archive.org/language-data-commons/terms#Annotation",
401   "@type": "rdfs:Class",
402   "name": "Annotation",
403   "sameAs": {
404     "@id": "http://www.language-archives.org/REC/type-20020628.html#annotation"
405   },
406   "rdfs:label": "Annotation",
407   "rdfs:comment": "The resource includes material which adds information to some other linguistic record.",
408   "rdfs:subClassOf": {
409     "@id": "http://schema.org/CreativeWork"
410   }
411 },
412 {
413   "@id": "https://purl.archive.org/language-data-commons/terms#Code",
414   "@type": "DefinedTerm",
415   "name": "Coded",
416   "description": "The resource contains an analysis or annotations represented by a code (such as the International Phonetic Alphabet).",
417   "inDefinedTermSet": {
418     "@id": "https://purl.archive.org/language-data-commons/terms#CommunicationModeTerms"
419   }
420 },
```

# Documentation of the formal version

## Class: Annotation [↗](#)

---

The resource includes material which adds information to some other linguistic record.

### Subclass of: [↗](#)

[ <http://schema.org/CreativeWork> ] |

### Properties [↗](#)

[ [annotationType](#) ] | [ [linguisticGenre](#) ] |

### Same as: [↗](#)

[ [annotation](#) ] |

:

[Top of page](#)

## Defined Term: Coded [↗](#)

---

The resource contains an analysis or annotations represented by a code (such as the International Phonetic Alphabet).

[Top of page](#)



Australian Research Data Commons



# Is this enough?

- Using the vocabulary has the obvious benefits for:
  - a. Data managers
  - b. Developers
  - c. ....
- But does it have benefits for users?
- Benefit should be that data should be (more) FAIR
- People need to understand the vocabulary for at least two purposes:
  - a. To find data efficiently
  - b. To describe their data so that purpose (a) can be achieved by others

# Vocabularies for language data - a sad story

- As mentioned, the story goes back at least 20 years
- Various schemas have been proposed:
  - OLAC
  - General Ontology for Language Description (GOLD)
  - IMDI / CMDI
- Take-up has not been good
- Specific example: Leipzig Glossing Rules
  - Suggestion to standardise abbreviations for common grammatical categories
  - c200 items included in list
  - General usage accepts only a small subset

# Providing a resource to improve uptake

- LDaCA uses terms from the vocabulary in records displayed in our data portal
  - First problem is relevant: do users understand these terms (or how they are being used specifically)?
- LDaCA provides advice on good practice to people collecting language data
  - Second problem is relevant
  - Obviously, we recommend using our vocabulary
  - Do data collectors understand these terms and how to apply them?
- To try to meet these needs, we are creating another resource documenting the vocabulary using GitBook:  
<https://ldaca.gitbook.io/metadata-for-language-data/>



# Why GitBook?

- Presentation style and layout are familiar (for most of our users)
- Editing interface and procedures are straightforward
- No need to develop everything from scratch
- Delivery platform is reasonably stable
  - But export to pdf or Markdown is easy if needed
- Now some examples of the additional content we are providing:
  - Explanation
  - Usage examples
  - References to literature

# More text - further explanation

## M Metadata for Language Data

Metadata for Language Data


- Introduction

Classes

- CollectionEvent
- CollectionProtocol
- PersonSnapshot
- DataLicense
- DataDepositLicense
- DataReuseLicense
- PrimaryResource
- DerivedResource
- Annotation**

Properties

- access
- annotationOf
- annotationType
- collectionEventType
- collectionProtocolType
- channels
- derivationOf

Powered By  GitBook

## Annotation

The resource includes material which adds information to some other linguistic record.

This class is a subclass of [schema:CreativeWork](#).

This class will have an [annotationType](#) as a property.

This class is the same as OLAC [annotation](#).

The OLAC page linked above has the following comment:  
'A linguistic annotation is defined as structured linguistic information that is explicitly aligned to some spatial and/or temporal extent of some other linguistic record.'  
This allows for the possibility that an annotation can be aligned with the entire extent of a linguistic record, but it is more common that annotations are linked to portions of the linguistic record. In principle, this could mean each annotation (each piece of information aligned to part of a linguistic record) being stored separately. But common practice is to store all annotations of a particular type together, or even to store several types of annotation in a single document or file, treating them as distinct tiers. Various software tools (e.g. [ELAN](#)) have implemented this structure for annotations.

This schema treats any information added to a [PrimaryResource](#) as annotation. [Transcription](#) can be considered the paradigm case. A recording of spoken language is a [PrimaryResource](#); a [Transcription](#) is a written representation of aspects of the original record and the parts of the [Transcription](#) are aligned with parts of the recording by, for example, use of time codes or assignment of utterances to speakers.

An example of several types of annotation being brought to together is aligned interlinear text (also [interlinear glossing](#)), a common medium for the presentation of language material.

← Previous [DerivedResource](#) Next [Properties](#) →



# More text - examples of usage


## M Metadata for Language Data

Metadata for Language Data >

Classes >

Properties ▾

- access
- annotationOf
- annotationType
- collectionEventType
- collectionProtocolType
- channels
- derivationOf
- doi
- geoJSON
- hasAnnotation
- hasDerivation
- indexableText
- linguisticGenre
- communicationMode
- hasCollectionProtocol
- isDeidentified
- undefined
- annotator

Powered By  GitBook

## subjectLanguage

The language(s) that this annotation resource is about, a language which the content of the resource describes or discusses.

**Values expected to be one of these types:**

<http://schema.org/Language>

Recommended values:

- for Australian languages: Uses standard names and codes from <https://collection.aiatsis.gov.au/austlang>
- for other languages: use codes from [Glottolog](#)

This property and the property [inLanguage](#) make up a pair. [inLanguage](#) describes the language in which a resource is written, while [subjectLanguage](#) describes the language or languages which a resource is about.

For example, the following work is about the Italian language (as used in Australia) and is written in English:

Caruso, Marinella. *Italian language attrition in Australia: The verb system*. Franco Angeli, 2010.

The [subjectLanguage](#) value for this work would therefore be **ital1282**, the Glottocode for Italian, and the [inLanguage](#) value would be **stan1293**, the Glottocode for Standard English.

**Used on these types:**

[Annotation](#)

← Previous sponsor    Next transcriber →

# References to literature

M Metadata for Language Data


- researchParticipant
- researcher
- responder
- signer
- singer
- speaker
- sponsor
- subjectLanguage
- transcriber
- translator

Defined Term Sets

- AccessTypes
- AnnotationTypeTerms
- AuthorizationWorkflow
- CollectionEventTypeTerms
- CollectionProtocolTypeTerms
- LinguisticGenreTerms
- CommunicationModeTerms
- WrittenLanguageTypeTerms

Defined Terms

- accessControlList

Powered By  GitBook

## WhistledLanguage

The resource contains data for which the medium of interaction was whistling.

This term is an expected value for the following property:  
[communicationMode](#)

This term is part of the set:  
[CommunicationModeTerms](#)

Meyer, J; Gautheron, B. (2006) "Whistled Speech and Whistled Language". In Brown, K. (editor-in-chief). *The Encyclopedia of Language and Linguistics*. Oxford: Pergamon. pp. 573-576.  
<https://doi.org/10.1016/B0-08-044854-2/00034-1>

← Previous **WrittenLanguage** Next Relationships →

Last modified 1mo ago

# Conclusion

- Gaining maximum benefit from using a vocabulary means making the vocabulary accessible to users of our products and services
- To achieve this, we are aiming for a rich documentation of the vocabulary including:
  - Detailed explanations
  - Examples of usage
  - References to further literature
- We are using GitBook to deliver this material:
  - Easy to create and edit content
  - Good production values and familiar format for users



# Thank you

For more information,  
please contact:

Name

[s.musgrave@uq.edu.au](mailto:s.musgrave@uq.edu.au)

[ldaca@uq.edu.au](mailto:ldaca@uq.edu.au)