SWISSUbase

LaRS Data Curation

Principles for Linguistic Data

FORS
explore.understand.share.

UNIL | Université de Lausanne

Universität
Zürich UZH

Version:	1.0	Persistent link:	https://resources.swissubase.ch/linguistics_data_curation_en/
Last modified:	01.03.2024	License:	CC BY
Introduced:	01.03.2024	Publisher:	Language Repository Switzerland (LaRS) swissubase@ub.uzh.ch

1. STUDY: LEVEL OF DATA CURATION

On study level, LaRS performs data curation level B: a) Checking all metadata entered on SWISSUbase. C.f. Section 2. Dataset: Levels of Data Curation (B: Basic Curation, a) Checking all metadata entered on SWISSUbase).

2. DATASET: LEVELS OF DATA CURATION

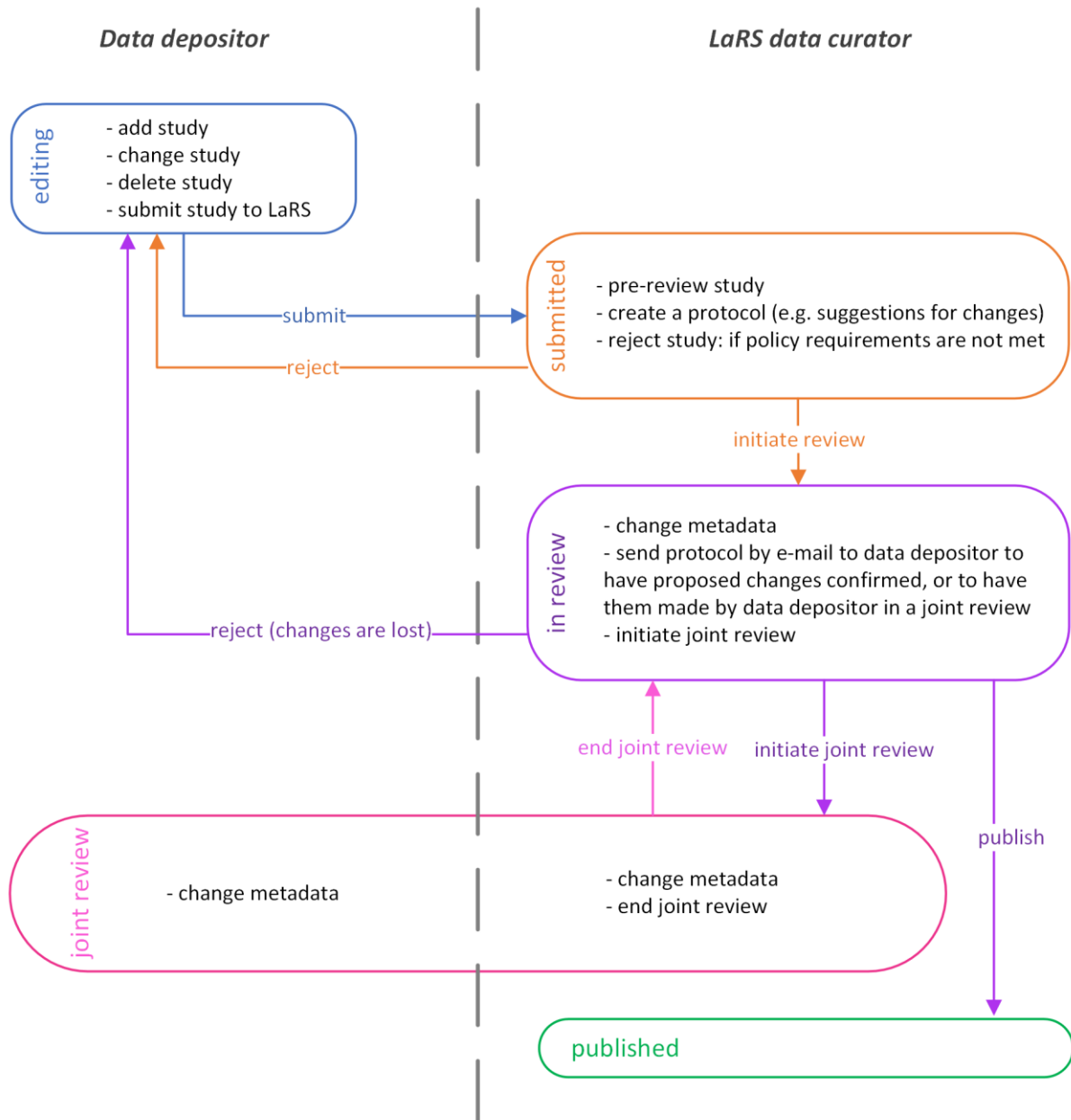The levels of data curation are based on the levels proposed by CoreTrustSeal for its certification process: https://doi.org/10.5281/zenodo.7051096

LaRS performs data curation levels A to C for every submitted dataset.

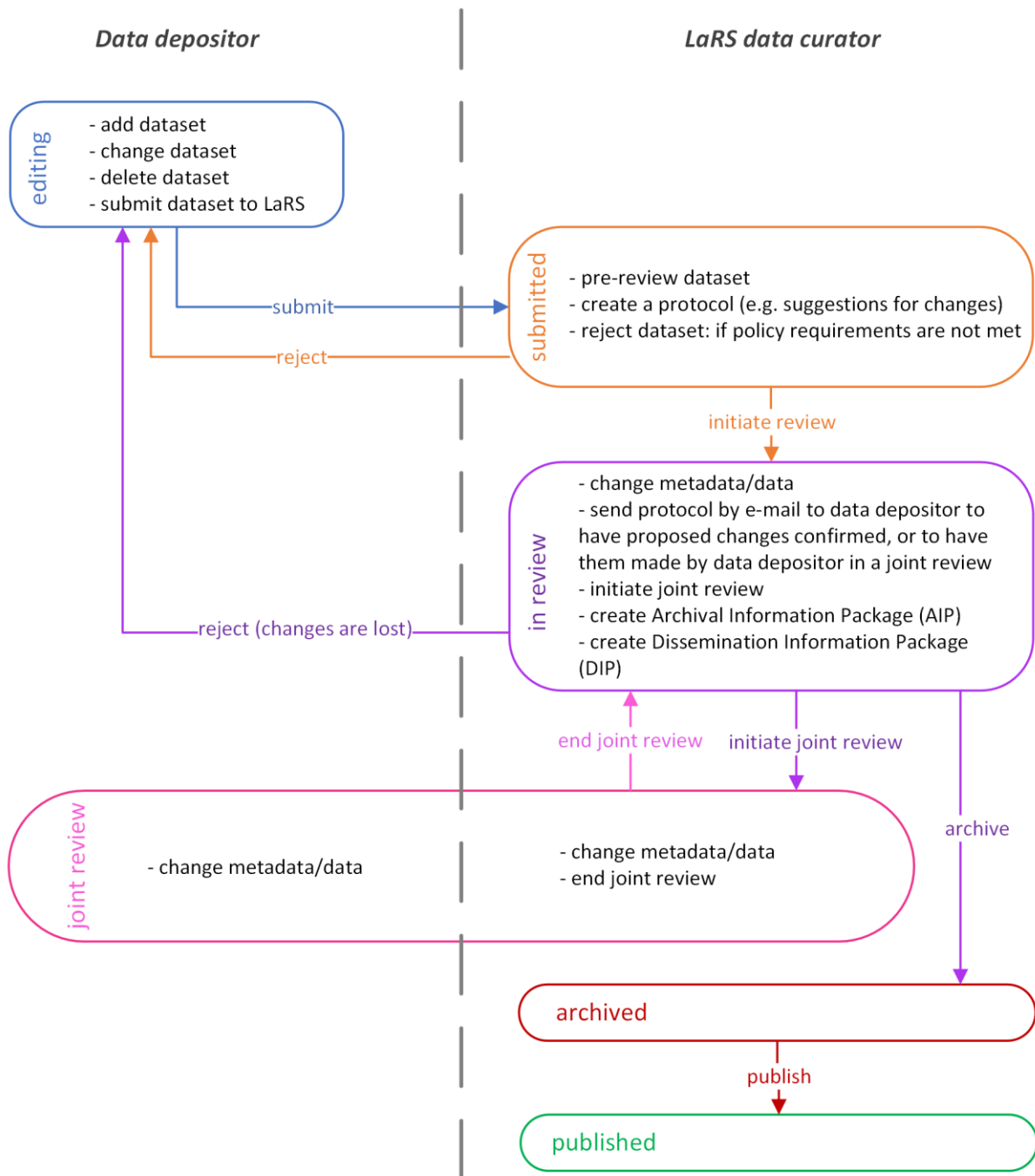| Data curation level	LaRS data curation
A: Content distributed as deposited	As a matter of principle, LaRS data curators check each submitted study and dataset. Only in very rare cases can content be distributed as deposited (c.f. levels B and C). However, adhering to the ISO OAIS model for archiving data, SWISSUbase stores the Submission Information Package (SIP) submitted by the data depositor. The SIP remains unchanged. The SIP is copied to an Archival Information Package (AIP) based on which all further data curation activities are executed. After submitting the data to SWISSUbase, the SIP is no longer available to the data depositor.

B: Basic curation – e.g. brief checking, addition of basic metadata or documentation	The following data curation actions are executed:
	
	a) Checking all metadata entered on SWISSUbase:
	  – Correction of typographical errors without further consultation with the data depositor
	  – Correction of wrong or unclear metadata in consultation with the data depositor
	  – Addition of new metadata or information in consultation with the data depositor
	b) Spot-checking documentation:

	– Checking if sufficient documentation has been submitted – Screening of documentation for plausibility/comprehensibility, and suggesting to data depositor to add additional documentation or to re-work documentation c) Spot-checking data: – Downloading each dataset – Spot-checking random files for readability (e.g. opening an audio file with a media player) d) Checking the selected licenses/contracts
C: Enhanced curation – e.g. conversion to new formats during ingest, changing the dataset structure (folders)	As a matter of principle, LaRS data curators respect the decision of the data depositor as to how the data are structured and only interfere if deemed necessary. The following data curation actions are executed: a) Checking the structure of the data collection (e.g. for unnecessarily nested folders in the ZIP file) and suggesting to change the structure in consultation with the data depositor b) Checking the file collection for hidden files and suggesting their removal in consultation with the data depositor c) Checking submitted file formats for FAIRness/adequacy and suggesting data format conversions in consultation with the data depositor d) Checking submitted file formats for FAIRness/adequacy and adding converted data formats to the Archival Information Package (for archiving only) and/or to the Dissemination Information Package (for publication); depending on the case this will be done in consultation with the data depositor e) Performing sample checks with regard to potential violations of ethical or data protection regulations (e.g. lack of anonymization)
D. Data-level curation – as in C above, but with additional editing of the deposited data	LaRS does not currently perform level D data curation. The content of a data collection remains unchanged (e.g. the content of an XML or a CSV file).

3. DATA CURATION WORKFLOW

3.1 Study: Data curation workflow

Data depositor | **LaRS data curator**

editing
- add study
- change study
- delete study
- submit study to LaRS

submit →

← reject

submitted
- pre-review study
- create a protocol (e.g. suggestions for changes)
- reject study: if policy requirements are not met

initiate review

in review
- change metadata
- send protocol by e-mail to data depositor to have proposed changes confirmed, or to have them made by data depositor in a joint review
- initiate joint review

← reject (changes are lost)

end joint review | initiate joint review

publish

joint review
- change metadata

- change metadata
- end joint review

published

3.2 Dataset: Data curation workflow

Data depositor

LaRS data curator

editing
- add dataset
- change dataset
- delete dataset
- submit dataset to LaRS

submit →

← reject

submitted
- pre-review dataset
- create a protocol (e.g. suggestions for changes)
- reject dataset: if policy requirements are not met

↓ initiate review

in review
- change metadata/data
- send protocol by e-mail to data depositor to have proposed changes confirmed, or to have them made by data depositor in a joint review
- initiate joint review
- create Archival Information Package (AIP)
- create Dissemination Information Package (DIP)

← reject (changes are lost)

end joint review ↑   ↓ initiate joint review

archive ↓

joint review
- change metadata/data   - change metadata/data
                           - end joint review

archived

↓ publish

published

3.3 Remarks

*Status **editing:***

– The data depositor can add, change, or delete the dataset

*Trigger for the status change from **editing** to **submitted**:*

– The data depositor must submit the dataset

*Status **submitted**:*

– The data curator must create a protocol that contains proposed changes, changes made and other remarks on the data curation
– The data curator pre-reviews the dataset:
  o The data curator can reject a dataset if the requirements stipulated in the LaRS policy are not met; the data depositor receives a corresponding message; the status of the dataset is changed from **submitted** to **editing**; the "reject" function is generally only used if a dataset cannot be published at all; the data curation process is complete and the dataset will likely have to be deleted
  o The data curator accepts the dataset for the proper review process

*Trigger for the status change from **submitted** to **in review**:*

– The data curator must initiate the review process

*Status **in review**:*

– The data curator can review and change the data/metadata
– The data curator can send suggestions (via the protocol) to the data depositor
  o The data depositor needs to confirm the changes proposed in the protocol
– The data curator can decide whether changes need to be made by the data depositor directly, or in a joint review session (e.g. via Zoom)

*Trigger for the status change from **in review** to **in joint review**:*

– The data curator must initiate the joint review process
– The data curator usually sends the protocol with suggestions to the data depositor

*Status **in joint review**:*

– The data depositor can change the data/metadata
– The data curator can change the data/metadata

*Trigger for the status change from **in joint review** to **in review**:*

– The data curator must end the joint review process

*Status **in review***:

- The data curator can continue the review process
- The data curator can make archiving decisions such as the creation of an Archival Information Package (which will be archived) and a Dissemination Information Package (which will be published)

*Trigger for the status change from **in review** to **archived**, and to **published**:*

- The data curator must initiate the archiving and publishing process

Status **published**:

- The data depositor receives an automated notification about the successful publication of the dataset (including a link to the dataset and the DOI)