

CLARIN-PL Repository – Preservation Plan (ver. 1.3)

1. Purpose and Scope

The CLARIN-PL DSpace repository undertakes long-term preservation of linguistic and language-related research data and associated metadata to ensure their continued accessibility, usability, and authenticity for the Designated Community. This preservation plan defines the responsibilities, procedures, and measures implemented to safeguard digital content throughout its lifecycle, from ingestion to potential reappraisal and removal.

2. Level of Preservation Responsibility

The repository assumes full responsibility for the long-term preservation of all accepted datasets and their metadata upon the successful completion of the deposition process and signing of the Deposition License Agreement. The depositor transfers custody and preservation responsibility to CLARIN-PL, as specified in the agreement. All deposited datasets are treated equally in terms of preservation level, regardless of depositor or type of data. Persistent Identifiers (PIDs) ensure that all preserved items remain citable and traceable over time, even in the event of withdrawal.

3. Preservation Strategy and Procedures

The CLARIN-PL preservation strategy is based on the Open Archival Information System (OAIS) reference model and implemented through DSpace software adapted to the CLARIN-DSpace profile. Preservation involves two levels: bit-level preservation, ensuring data and metadata integrity, and functional preservation, which includes active measures such as format migration to maintain usability.

a. Format Migration and Emulation

Preferred formats follow the CLARIN Standards and Formats guidelines, emphasizing open, well-documented, verifiable and text-based standards. XML, UTF-8/Unicode, and other non-proprietary formats are prioritized. Proprietary or customized formats must be accompanied by exhaustive documentation to enable future conversions. Migration to newer formats may occur when a format is at risk of obsolescence, with original files retained for authenticity.

b. Metadata Schemas for Preservation

Metadata is stored using CLARIN-recommended schemas (CMDI, Dublin Core, DataCite) and can be converted into self-explanatory, human-readable XML. Preservation metadata includes provenance, checksums, file format, and version history.

c. Verification and Audit

Integrity checks using checksums are performed regularly. Automated weekly reports and manual reviews monitor repository health. Editors verify file formats, metadata completeness, and licensing before publication. All post-ingest modifications are recorded in provenance metadata.

4. Custody Transfer and Documentation

Custody transfer occurs upon formal submission under the Deposition License Agreement. Submission Information Packages (SIPs) are created according to documented criteria. All actions are logged for auditability.

5. Future Migrations and Obsolescence Mitigation

The repository assesses technology trends and format risks. Planned migrations are documented and executed as needed, retaining original and migrated versions. Participation in CLARIN ERIC committees ensures awareness of technological changes and adherence to standards

6. Retention and Reappraisal

CLARIN-PL intends long-term preservation of data and metadata. Digital objects are reappraised every five years or earlier if needed. Outcomes include continued preservation, migration, or justified withdrawal at the request of the depositor.

7. Withdrawal, Deletion, and Impact on PIDs

Withdrawal requests must include the item's PID and justification. Upon approval, data files are removed but administrative metadata and the PID remain. This ensures the scholarly record's integrity. The web interface for the Item landing page is not preserved from this moment.

8. Disaster Recovery and Data Replication

In case of disruption, data will be replicated or transferred to trusted partner sites such as:

- Repository replication AZON <https://zasobynauki.pl>
- Backup kronika.gov.pl
- Backup kmd.pionier.net.pl

TODO: We are currently looking for a suitable partner to provide offshore replication that may be used to mitigate geopolitical risks. It is not implemented at present

9. Governance, Roles, and Review

Preservation is managed by dedicated technical and editorial teams under WUST and CLARIN ERIC best practices. Policies are reviewed annually. Staff participate in international conferences and CLARIN committees.

10. Evidence and Documentation

Deposition License Agreement: <https://clarin-pl.eu/dspace/page/contract>

Standards and Formats: <https://www.clarin.eu/content/standards-and-formats>

Withdrawal Procedure: <https://clarin-pl.eu/dspace/page/item-lifecycle#deleting-item>

Preservation Policy Page: Available under 'About and policies' → 'Preservation policy' on the CLARIN-PL repository website.