Overview of integrity and authenticity

Purpose of this document

This document provides an overview of the implementation of measures for ensuring integrity and authenticity on an operational level in the current digital repository system DM1.5.

Integrity

The concept of ‘integrity’ as defined in the Preservation plan 2019-2022 ensures that objects and collections are complete and that changes to the data take place in a controlled and documented manner.
This document uses the subdivisions of integrity as introduced in the Preservation plan. These are as follows:

- Bit integrity
- Version integrity
- IP integrity
- Information integrity
- Collection integrity

Bit integrity

Storing the checksum

Producers are requested to provide a checksum along with the data. The ingest specification details in which flows a checksum is delivered. Not all producers delivers checksums in which case alternative methods are used to verify bit integrity after transfer. As an alternative method verification based on file size may be used or extraction of files from a package file.
If a checksum is delivered by the producer verification is done by calculation of the checksum. If the checksum is not in SHA512-format, a new checksum is created.

Use of the checksum during ingest

The checksum is used during the ingest process to verify if delivery was successful and all files have been completely delivered. The checksum is also used to detect 0-byte files. This process is described in the ingest specification. Checksum calculation is also done for certain transfer actions during ingest, as also described in said specification. During ingest a checksum is also created for the AIP. This checksum is stored as a text file with .sha-512-extension on the Silent Cube in the same folder as the publication.

After the ingest phase a process is started to verify that files are safely stored on the Silent Cubes. This involves an ETL[1] process on a Pentaho server that verifies whether the SHA512 checksum on the Silent Cube matches the checksum in MDS. If this is the case, the status of the publication in the MDS database is set to status 7 which means the publication is completely stored and verified.

Checksum verification

Bit integrity during storage is ensured by checksum verification. In DM1.5 the checksum is verified in the Silent Cube storage system. A report is sent monthly containing audit information based on the scheduled checksum verification process. This report ensures that files have not been inadvertently changed on the storage.

Version integrity

In DM1.5 it is not possible to relate versions of publications. Different versions are stored as separate publications without a link between them. In MDO versions can be linked based on metadata elements such as owner-id (of the publisher) and publication id and a unique key (recordIdentifier). In the process the old NBN is stored as 'invalid' and the new NBN becomes the valid NBN.
Publications that are duplicates are skipped in the ingest process. The ingest specification details how the mechanism for detecting duplicates works.

IP integrity

IP integrity is described in the preservation plan as the completeness of the *Information Packages*. This is applicable to *Submission Information Packages* (SIP), *Archival Information Packages* (AIP) and *Dissemination Information Packages* (DIP).

During ingest the SIP is verified for completeness at different stages. Completeness is checked after transfer by the producer completeness. Before ingest into the repository the SIP is verified once more. These checks are described in the SIP specification. When the SIP is transformed to an AIP a further check is done to ensure the transformation will succeed. This is an implicit completeness check. The DIP is not checked for completeness since DIP and AIP are identical. The only difference is that storage locations are replaced by resolver links and file names are restored to their original filenames during transformation from AIP to DIP.

Information integrity

Information integrity is defined in the preservation plan as a measure that takes into account all the information needed to understand the publication, both now and in the future. In terms of the

1 ETL = extract, transform, load

OAIS model this information is considered *Representation Information*. This *Representation Information* is stored and related to the publication so it is clear at all times what information is crucial for understanding the publication.
At the moment no Representation Information is stored as such for that purpose.

Some metadata may also be considered *Representation Information*. This information is stored in the AIP schema but not identified as such. An example would be the structural metadata of the AIP or the qualification of a file as a master or a supplemental file.

Collection integrity

The concept of collection integrity is defined in the preservation plan as a measure for determining completeness of the collection. This means that all AIPs that would be expected as part of a collection have been received and stored completely. Collection integrity as defined in this way is currently not implemented and no systematic checks are done to ensure this aspect of integrity. The ETL process for checking integrity doesn't take into account collection integrity. There is no end-to-end monitoring that checks whether all material that has been delivered also has been stored and can be accessed.

Authenticity

Authenticity is defined in the preservation plan as consisting of three aspects. One of these aspects is provenance of the object. This is stored in the metadata of the AIP. A description of the AIP is available in the KB datamodel. There it is defined which fields contain provenance information.

Authenticity is also determined based on the history of the object. This consists of all the actions performed on an object in the whole lifecycle from delivery to access. Mainly during the ingest process history is stored as event metadata in the AIP manifest. Which information is stored for a flow is defined in the specification of that flow.