Overzicht integriteit en authenticiteit

Doel van het document

In dit document wordt er een overzicht gegeven op welke manier de integriteit en authenticiteit in DM1.5 geïmplementeerd wordt.

Integriteit

Het begrip integriteit is gedefinieerd in het Preservingsplan 2019-2022 als de garantie dat objecten en collecties volledig zijn en dat wijzigingen gecontroleerd en gedocumenteerd verlopen.
In het document worden er meerdere niveau's van integriteit onderscheiden. De indeling van het Preserveringsplan wordt hier overgenomen:

- Bit integriteit
- Versie integriteit
- IP integriteit
- Informatie integriteit
- Collectie integriteit

Bit integriteit

Opslaan van checksum

De checksum wordt gevraagd aan de uitgever om mee te leveren. In de specificaties van ingest wordt beschreven bij welke stromen de checksum door de uitgever worden aangeleverd. Niet alle uitgevers sturen checksums mee met de aanlevering, maar daarom worden er alternatieve methoden gebruikt om de overdracht te controleren op bit-integriteit. De alternatieven zijn bijvoorbeeld het controleren van de bestandsgrootte voor en na overdracht, of het uitpakken van gecomprimeerde bestanden.
De checksum, zoals aangeleverd door de uitgever, wordt gecontroleerd. Als deze checksum geen SHA512 is, wordt er een nieuwe checksum gemaakt.

Gebruik van checksum tijdens ingest

De checksum wordt tijdens het ingestproces gebruikt om na te kijken of de aanlevering goed verlopen is en de bestanden volledig zijn aangeleverd. Daarnaast wordt de checksum ook gebruikt om 0-byte bestanden op te sporen. Dit wordt in de ingestspecificatie beschreven. Daarnaast wordt de checksum berekend bij bepaalde verplaatsacties bij de ingest. Dit wordt ook beschreven in de ingestspecificatie. Daarnaast wordt er tijdens ingest ook een checksum

gemaakt van de AIP en opgeslagen op de Silent Cube in dezelfde map als de publicatie met als extensie sha-512.checksum.

Na ingest wordt er een controleproces uitgevoerd om te controleren of de bestanden goed opgeslagen zijn op de Silent Cubes. Daarom draait er een ETL[1] proces op de Pentaho server die controleert of de SHA512 checksum van de Silent Cube overeenkomt met de checksum in MDS. Dan wordt de status in de MDS database van de publicatie aangepast (status 7) en dit betekent dat de publicatie goed zijn geïngest en gecontroleerd.

Controleren van checksum

De bit integriteit wordt gecontroleerd door de checksum te controleren. In het DM1.5 wordt de checksum gecontroleerd in het opslagsysteem (Silent cube). Er wordt maandelijks een rapport gestuurd over de audit op basis van de checksum om aan te tonen dat de bestanden ongewijzigd zijn.

Versie integriteit

Het is niet mogelijk om versies vast te leggen in DM1.5. De verschillende versies worden als aparte en niet-gelinkte publicaties geïngest. De versies worden wel samengebracht in MDO, waar op basis van een aantal elementen (owner-id van de uitgever en publication-id) een sleutel (recordIdentifier) wordt gemaakt die gebruikt wordt om de versies te identificeren. Het oude NBN wordt opgeslagen als invalid en de nieuwe NBN wordt de valid NBN.

Dubbele publicaties worden niet geïngest en ze worden tegengehouden tijdens ingest. Er staat in de ingestspecificaties beschreven hoe dubbelcontrole geïmplementeerd is.

IP integriteit

IP integriteit wordt beschreven in het preserveringsplan als de volledigheid van de *Information Packages*. Het gaat hier over zowel de *Submission Information Package* (SIP), *Archival Information Package* (AIP) als *Dissemination Information Package* (DIP).

Bij ingest wordt de SIP op verschillende momenten gecontroleerd. Bij de aanlevering door de uitgever wordt de volledigheid gecontroleerd. Voor het ingest in het DM wordt de SIP nog een keer gecontroleerd. De controles staan beschreven in de SIP specificaties. Bij de omzetting van SIP naar AIP wordt er gekeken of de AIP kan omgezet worden. Dit is een impliciete volledigheidscontrole. De DIP wordt niet op volledigheid gecontroleerd op zich, omdat de AIP en DIP inhoudelijk gelijk zijn. Enkel de opslag locaties zijn aangepast naar resolver links en bestanden krijgen terug de originele bestandsnaam bij de omzetting naar DIP.

1 ETL = extract, transform, load

Informatie integriteit

In het preserveringsplan wordt informatie integriteit beschreven als alle informatie die nodig is om de publicatie te begrijpen, ook in de toekomst. In OAIS termen gaat het dan over *Representation Information*. Deze *Representation Information* wordt ook gerelateerd aan de publicatie zodat het duidelijk is, welke informatie belangrijk is voor het begrijpen van een publicatie.
Op dit moment wordt er geen Representation Information bewaard of opgeslagen met dat doel.

Daarnaast wordt sommige metadata ook gezien als *Representation Information*. Deze informatie wordt nu in het AIP schema vastgelegd, maar wordt niet als zodanig aangeduid. Een voorbeeld is de structuurinformatie vastgelegd in een AIP of het verschil tussen verschillende aangeleverde bestanden (bijvoorbeeld het verschil tussen het hoofdbestand en onderdelen van het hoofdbestand).

Collectie integriteit

In het preserveringsplan wordt collectie integriteit omschreven als het volledig zijn van de collectie. Dit betekent dat een collectie alle verwachte AIPs bevat en dat de volledige AIP is opgeslagen. Deze collectie integriteit wordt niet gebruikt op deze manier en wordt ook niet gecontroleerd op deze manier. Er is een ETL proces waar de integriteit wordt gecontroleerd, maar in deze controle wordt geen volledigheidscontrole uitgevoerd. Er is ook geen volledigheidscontrole in de zin van alles wat aangeleverd is, daadwerkelijk ook bewaard wordt en beschikbaar gesteld kan worden. Het opvolgen of alle geleverde bestanden ook daadwerkelijk beschikbaar gesteld worden, gebeurt niet.

Authenticiteit

In het preserveringsplan wordt authenticiteit opgedeeld in een aantal onderdelen. Eén daarvan is herkomst van een object. Hier wordt de herkomst in de metadata van het AIP vastgelegd. De beschrijving van het AIP is opgenomen in het KB metadatamodel. Hierin staat aangegeven welke metadata de herkomst vastlegt.

Daarnaast wordt de authenticiteit aangetoond door de geschiedenis van een object. Hiermee wordt bedoeld welke acties uitgevoerd zijn op een object vanaf aanlevering tot beschikbaar stellen. Voornamelijk tijdens het ingestproces wordt er event metadata vastgelegd in het AIP manifest. Welke informatie vastgelegd wordt, is beschreven in de specificatie van elke stroom.