



ONGERUBRICEERD
MT DAD

Justitiële Informatiedienst

Burg. Raveslootsingel 2
7607 GK Almelo
Postbus 337
7600 AH Almelo
www.justid.nl

Contactpersoon
Digitalisering en
Archiveringsdienst

T 088 99 890000

Datum
27 februari 2017

memo

Fixity checks in het CDD+

Om te voldoen aan de eisen voor certificering en om de betrouwbaarheid van het CDD+ te verhogen, wordt het AMT verzocht om fixity checks in te laten voeren op de opslag van het CDD+. Daarnaast wordt het AMT verzocht om deze functionaliteit later uit te breiden met fixity checks op de ingest en de access, voor een verdere verhoging van de betrouwbaarheid.

Achtergrond

Opgeslagen digitale bestanden zijn niet onkwetsbaar. Als gevolg van beschadiging van de hardware of fouten die ontstaan bij het kopiëren of verplaatsen van bestanden kan de inhoud van een opgeslagen bestand wijzigen. Daarnaast kan *bitrot*¹ optreden.

Van documenten die in het CDD+ worden opgeslagen moet de integriteit en de bruikbaarheid worden gegarandeerd. Daarom is het belangrijk dat fouten die ontstaan op tijd worden gedetecteerd, zodat dergelijke fouten kunnen worden hersteld door een back-up terug te plaatsen.

Het CDD+ heeft enige bescherming tegen hardwarefouten (waaronder *bitrot*) doordat het is opgeslagen in een RAID systeem. In een dergelijk systeem wordt de data verdeeld over meerdere harde schijven, en wordt er aan elk blok data een controlegetal toegevoegd, een zogenaamde *parity*. Als een van de harde schijven een beschadiging oploopt, dan klopt de parity niet meer, en wordt de fout herkend. Met behulp van de parity kan de data worden hersteld.

Dit systeem helpt echter niet tegen fouten die ontstaan doordat een bestand verkeerd wordt weggeschreven bij kopiëren of verplaatsen ervan. Daarvoor is een andere methode nodig namelijk gebruik maken van een controlegetal dat wordt berekend over het bestand als geheel, een zogenaamde *hash*. Ook hier geldt dat als het controlegetal is gewijzigd, dit aangeeft dat het bestand is gewijzigd. In dat geval moet er een back-up worden teruggeplaatst. Het controleren van een dergelijke hash wordt een *fixity check* genoemd. Fixity is een benaming voor het "onveranderlijk" zijn van een bestand.

¹ Zie: <https://nl.wikipedia.org/wiki/Bitrot>

Standaarden

Het gebruik van fixity metadata wordt beschreven in het OAIS model² en PREMIS³. Fixity checks zijn een noodzakelijke voorwaarde voor het verkrijgen van certificering. Zo kent de ISO-16363 de volgende eis⁴:

Datum

27 februari 2017

The repository shall actively monitor the integrity of AIPs.

Vervolgens worden de volgende voorbeelden genoemd van hoe een organisatie kan aantonen dat er aan de eis voldaan is:

Fixity information (e.g., checksums) for each ingested digital object/AIP; logs of fixity checks; documentation of how AIPs and Fixity information are kept separate; documentation of how AIPs and accession registers are kept separate.

De Data Seal of Approval heeft een soortgelijke eis, namelijk "Data integrity and authenticity", met de volgende uitleg⁵:

For this Requirement, responses on data integrity should include evidence related to the following:

- *Description of checks to verify that a digital object has not been altered or corrupted (i.e., fixity checks).*

² Reference Model for an Open Archival Information System (OAIS), Magenta Book, pagina 4-30:

"Fixity Information provides the Data integrity checks or validation/verification keys used to ensure that the particular Content Information object has not been altered in an undocumented manner. Fixity Information includes special encoding and error detection schemes that are specific to instances of Content Objects. Fixity Information does not include the integrity preserving mechanisms provided by the OAIS underlying services, error protection supplied by the media and device drivers used by Archival Storage."

³ Data Dictionary for Preservation Metadata: PREMIS version 3.0, pagina 258:

"In the digital domain, hash algorithms that create a message digest can be used to implement a fixity check for an object. If the message digest created by an algorithm at one point is identical to the message digest created by the same algorithm at a later point, this indicates the object did not change during the interim. In fact, it is common to create and test two or more message digests using different algorithms to be certain that an object is fixed."

<https://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf>

⁴ ISO 16363: Space data and information transfer systems — Audit and certification of trustworthy digital repositories, pagina 4-21.

⁵ DSA-WDS Partnership Working Group, Catalogue of Common Requirements (v2.2), pagina 11.

Fixity checks worden onder andere ingezet door de Britse National Archives⁶, Het Nationaal Archief⁷ en Beeld en Geluid⁸. Als DAD haar klanten maximale zekerheid wil geven dat de integriteit van de documenten behouden blijft, dan is het noodzaak om ook in het CDD+ fixity checks in te zetten

Datum

27 februari 2017

Verdere toepassingen

De standaarden schrijven het gebruik van fixity checks voor om te controleren dat de bestanden tijdens de opslag niet veranderd zijn. Echter, bestanden kunnen ook tijdens transport beschadigd raken. Een fixity check kan worden gebruikt om te bepalen of een verzonden bestand correct is aangekomen. De hash wordt dan meegestuurd met het bestand. Bij ontvangst van het bestand wordt de meegestuurde hash gecontroleerd, om te zien of hij nog steeds klopt. Is dit niet het geval, dan wordt ervan uitgegaan dat het bestand beschadigd is geraakt.

Aanbeveling

Zodra een nieuw bestand wordt opgenomen in het CDD+ wordt er al wel een hash uitgerekend. Er wordt echter nooit gecontroleerd of de hash gewijzigd is. Het voorstel is daarom op deze controle op twee momenten toe te voegen:

1. Bij elke migratie van de bestanden, zoals het verplaatsen naar een ander medium, moet de fixity van elk verplaatst bestand worden gecontroleerd.
2. De fixity van elk bestand (originelen en conversies) wordt periodiek gecontroleerd.
Door het uitvoeren van deze controle wordt ook in kaart gebracht hoe vaak zich fouten voordoen. Op basis van die gegevens kan de frequentie later worden bijgesteld.

⁶ Website National Archives:

"The National Archives undertakes a parsimonious approach to digital preservation, focusing on preserving the original files. This is achieved by making sure they are virus free, undertaking fixity checks to ensure the integrity of the files over time and by "knowing what we've got" (identifying all the file formats in our collections). We achieve file format identification by using the tools described in the section below."

<http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/our-role/>

⁷ *Preservation policy Nationaal Archief, versie 1.1, november 2015*, pagina 18:

"Een ander integriteitscontrole is de fixity check. De checksum voor elke content file vergelijken we met de originele checksum zoals gespecificeerd in de metadata. Vóór de ingest vindt deze controle plaats na elk transport (ftp, kopiëren enzovoort). Na de ingest gebeurt deze controle periodiek."

http://www.nationaalarchief.nl/sites/default/files/docs/preservation_policy_nationaal_archief_juli_2016.pdf

⁸ *Digitale Preservering Beeld en Geluid*, pagina 26:

Bij elke bewerking of verplaatsing van een versie van het object voor (hernieuwde) opslag, kopiëring, migratie of voor uitlevering aan gebruikers, wordt met een fixity check teruggerepen op de checksum, zodat zeker is het object 'qua bits' identiek is aan de content die in een eerdere fase werd overgebracht.

http://publications.beeldengeluid.nl/pub/387/digitale_preservering_Beeld_en_Geluid.pdf

Afhankelijk van de benodigde rekencapaciteit kan ervoor worden gekozen om hiernaast ook nog een controle uit te voeren bij het opvragen van een bestand, zodat ook op dat moment zeker is dat het bestand dat wordt uitgeleverd, correct is.

Datum
27 februari 2017

Verder wordt voorgesteld om optionele controles mogelijk te maken bij het deponeren en opvragen van een bestand:

- Het wordt mogelijk gemaakt voor de klant om bij het deponeren van een bestand de hash⁹ van dat bestand mee te geven. Is deze hash meegegeven, dan wordt deze bij ontvangst in het CDD+ gecontroleerd voordat het bestand de verdere ingestprocedure doorloopt. Is de hash veranderd, dan wordt hier melding van gemaakt, en krijgt de klant (of het klantsysteem) de mogelijkheid om het bestand opnieuw te deponeren.
- Bij het opvragen van een bestand wordt de hash van het bestand meegezonden. De klant (of het klantsysteem) kan aan de hand van de hash controleren of het bestand goed is overgekomen.

In beide gevallen is het dus aan de klant om te bepalen of hij gebruik wil maken van deze controles.

Het toevoegen van de controles moet worden uitgevoerd door de ontwikkelaars van Justid. Hiervoor moet een wijzigingsverzoek worden ingediend.

De controles kunnen worden geautomatiseerd, hetzij door de controle in het CDD+ toe te voegen, hetzij door externe software te gebruiken. Elke controle dient te worden gelogd, zodat er bewezen kan worden dat de fixity is gecontroleerd. Wordt er een verschil in de hash geconstateerd, dan moet de controlesoftware ook een melding versturen aan beheer, zodat er actie kan worden ondernomen om de fout te herstellen. Dit betekent ook dat beheer alleen in actie hoeft te komen als er een fout wordt gemeld. Ervan uitgaande dat het aantal fouten zeer beperkt is, is de impact van de eigenlijke controles op de beheeruren zeer klein¹⁰.

Merk op dat deze fixity checks uitsluitend bedoeld zijn om fouten op te sporen die optreden bij normaal gebruik door klanten en Justid beheer. De fixity check is dus geen betrouwbaar middel om dergelijke wijzigingen aan te passen en tegen te gaan.

⁹ Deze hash kan dezelfde hash zijn zoals die bij de opslag wordt gebruikt, maar er kan (in overleg met de gebruikers) worden gekozen om bij de ingest en access een andere hash te gebruiken, die beter aansluit bij de wensen van de gebruikers.

¹⁰ Naast de tijd die functioneel beheer kwijt is aan deze acties naar aanleiding van de controles, zal er ook tijd nodig zijn voor het beheer van de nieuwe functionaliteit.