Checksums?!

Een instrument voor betrouwbare digitale langetermijnbewaring

Digitale bestanden zijn kwetsbaar, niet alleen door de snel wijzigende technologie maar ook doordat alle digitale dragers onbetrouwbaar zijn voor langetermijnbewaring als ze niet worden gekoppeld aan o.a. goede back-up- en controleprocedures. Zonder de nodige voorzorgen kunnen digitale gegevens zelfs al op korte termijn verloren gaan of onbedoeld wijzigen. Dit fenomeen noemt men bitrot. De oorzaak hiervan ligt vaak bij de mechanische slijtage van de drager, of in een wijziging van de chemische samenstelling ervan. Daarom is een identieke kopie als back-up steeds noodzakelijk. Ook fouten bij het kopiëren van bestanden kunnen echter gegevensverlies tot gevolg hebben, bv. bij het maken van een back-up.

Een checksum stelt je in staat om dergelijke fouten of informatieverlies op te sporen. Het vertelt je bij de verslechtering van de drager wanneer je het oorspronkelijke bestand moet vervangen door de back-up, en stelt je in staat te verifiëren of de back-up wel een identieke kopie is van het origineel. Iedereen die digitale bestanden duurzaam wil archiveren, zou zonder uitzondering dergelijke checksums moeten aanmaken en ze vervolgens regelmatig moeten controleren.

Het principe van een checksum of controlegetal is erg eenvoudig: op een reeks letters of cijfers wordt met behulp van een algoritme een berekening uitgevoerd, met een nieuwe, kortere tekenreeks als resultaat. Door die berekening achteraf opnieuw uit te voeren en te vergelijken met de vorige uitkomst, kan worden gecontroleerd of de tekenreeks nog correct is. Een bekend voorbeeld is het laatste cijfer van een ISBN-nummer of de eindcijfers van je bankrekeningnummer.

In de informatica wordt deze techniek gebruikt bij datacommunicatie en -opslag. Hierbij wordt een algoritme uitgevoerd op een reeks bits, de verzameling enen en nullen waaruit elk digitaal bestand in essentie bestaat. Wanneer daarvan een bit verandert, levert dit een ander controlegetal op en is het duidelijk dat er iets mis is met het bestand. Zo'n controlegetal kan op elke willekeurige reeks bits worden berekend, dus ook op bijvoorbeeld een digitale afbeelding of tekstbestand.

MD5

Het Message Digest Algorithm 5 (MD5) geeft een checksum van 32 tekens. Ieder teken bestaat uit een cijfer van 0 tot 9 of een letter van a tot f, bijvoorbeeld 5adb6b18a918913e279761a06e5ba73a. Door deze samenstelling zijn 1632 of 2128 verschillende combinaties mogelijk. De kans dat twee bestanden hetzelfde controlegetal opleveren, is extreem klein. Met een MD5-checksum kun je dus een quasi unieke vingerafdruk van elk bestand creëren.

Oorspronkelijk werd MD5 ontworpen als beveiligingsalgoritme, maar intussen blijkt het daarvoor te kwetsbaar te zijn. Als controlemiddel voldoet het echter nog steeds, bijvoorbeeld bij gebruik in een digitaal archief. MD5-checksums worden gecreëerd voor of tijdens de opname van bestanden in het digitaal archief. Op regelmatige tijdstippen en/of bij raadpleging van een bestand, wordt aan de hand van eerder gemaakte checksums gecontroleerd of het bestand nog steeds volledig en ongewijzigd is (en dus niet gecorrumpeerd is).

Dit is belangrijk omdat digitale bestanden vaak in grote hoeveelheden worden bewaard en men onmogelijk ieder individueel bestand visueel kan gaan inspecteren. Bovendien zou een visuele inspectie van alle individuele bestanden in de meeste gevallen nog steeds onvoldoende uitsluitsel geven of de integriteit van de opgeslagen bestanden ongewijzigd is. Wanneer uit een controle van de MD5-checksum blijkt dat de integriteit van een digitaal bestand is gewijzigd, dien je terug te grijpen naar de (niet-gewijzigde) back-up en het gewijzigde bestand te vervangen door een exacte kopie van die back-up.

Checksum tools

Om MD5-checksums te gebruiken zijn een groot

aantal – gratis – programma's beschikbaar. Het principe is steeds hetzelfde en even eenvoudig: het programma creëert checksums van een aantal bestanden. Het resultaat is een klein tekstbestand, dat je samen met de bestanden bewaart. Wanneer je de bestanden wil controleren, vergelijkt het programma de nieuwe checksums met die in het tekstbestand. Wil je zeker zijn dat gegevens door bv. slijtage van de drager niet samen met het bestand verloren gaan, dan kan je het tekstbestandje ook op een andere locatie (bv. een externe harde schijf) opslaan.

Enkele voorbeelden van checksum tools:

- Checksum+ [1]
- Md5sum [2]
- MD5checker [3]
- Checksum Checker [4]
- DROID [5]
- Fixity [6]
- Fsum Frontend [7]
- Hash Functions [8]
- Jacksum [9]
- MD5Summer [10]

Hou er rekening mee dat er regelmatig nieuwe checksum tools verschijnen, en dat de ondersteuning van oudere checksum tools op een gegeven ogenblik mogelijk stopt. De MD5-checksums zelf zijn echter niet afhankelijk van een bepaalde checksum tool.

De keuze voor een bepaalde checksum tool kan bepaald worden door verschillende factoren. Niet elke checksum tool draait onder alle besturingsprogramma's of versies ervan; naargelang je gebruiker bent van Windows, Mac OS X of Linux, of een bepaalde versie ervan, kan het nodig zijn om een andere tool te kiezen. Ook niet alle tools hebben een grafische gebruikersinterface. Tools die enkel met een command line werken kunnen sommige gebruikers afschrikken. Sommige checksum tools bieden ook meer uitgebreide of andere gebruiksmogelijkheden dan andere tools. Met de meeste checksums tools kunnen meestal niet alleen MD5-cheksums worden gemaakt en gecontroleerd, maar ook andere types checksums.

Een uitgebreider overzicht van voorbeelden van checksum tools vind je op http://en.wikipedia.org/wiki/Checksum#Checksum_tools.

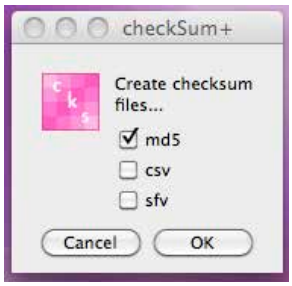
Aan de slag met enkele checksum tools

Ter illustratie demonstreren we hier een drietal mogelijkheden om MD5-checksums te creëren en te controleren. Met het oog op gebruiksvriendelijkheid hebben gekozen voor checksum tools met een grafische gebruikersinterface. We hebben zelf de checksums tools gebruikt op een Apple-computer, maar ze draaien ook onder andere besturingsprogramma's dan Mac OS X. Om de verschillende checksum tools te installeren is het aangewezen om de installatiehandleidingen te raadplegen.

A. CheckSum+: de aanmaak en controle van checksums voor individuele bestanden

Eerste methode

1. Open het programma checkSum+.
2. Ga naar 'File' in menubalk.
3. Kies 'Open...'.
4. Selecteer te controleren bestand.
5. Vink 'md5', 'csv' en/of 'sfv' aan.

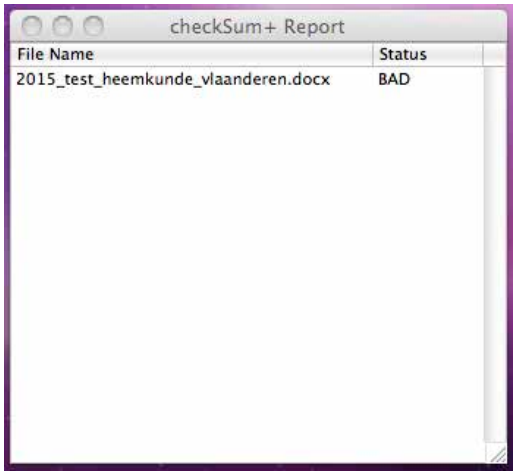


Figuur 1: maak een keuze tussen md5, csv of sfv.

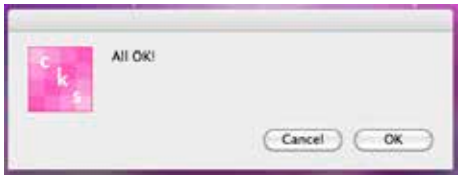
6. Klik 'OK'.
7. Een bestand met dezelfde naam als het origineel bestand maar met extensie .md5 wordt opgeslagen samen met het originele bestand

Figuur 2: bestand met extensie .md5 dat door de software wordt aangemaakt en opgeslagen.

8. Door op dit toegevoegde bestand te klikken, kun je controleren of het bestand integer is: als er iets is veranderd, geeft het programma aan dat de status 'bad' is, als er niks is veranderd, geeft het programma aan 'All OK'.



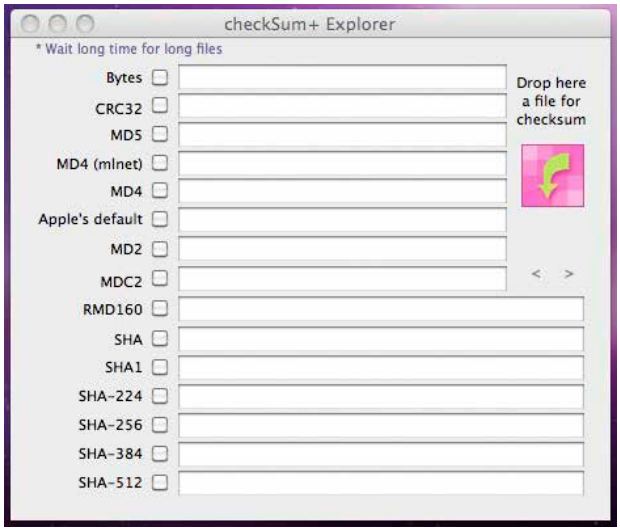
Figuur 3: het controlerapport aangemaakt m.b.v. checkSum+ geeft aan dat de integriteit van het gecontroleerde bestand niet goed is.



Figuur 4: checkSum+ geeft aan dat de integriteit van het gecontroleerde bestand goed is.

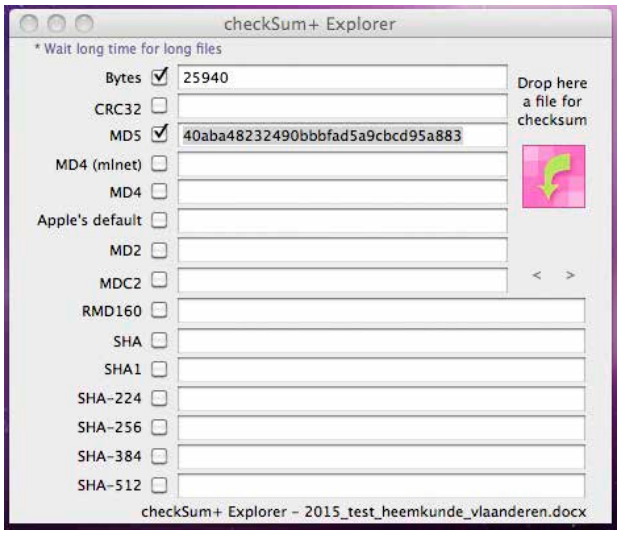
Tweede methode

1. Open het programma checkSum+.
2. Ga naar 'File' in menubalk.
3. Kies 'Open Explorer'.
4. Vink 'Bytes', 'CRC32', 'MD5', 'MD4 (minet)', 'MD4', 'Apple's default', 'MD2', 'MDC2', 'RMD160', 'SHA', 'SHA1', 'SHA-224', 'SHA-256', 'SHA-384' en/of 'SHA-512' aan.



Figuur 5: vink aan welk type checksum je wil creëren.

5. Versleep het te controleren bestand en drop het op het gekleurde icoon linksboven.
6. De gewenste informatie verschijnt nu in het venster.



Figuur 6: weergave van de grootte van het bestand en de gecreëerde MD5-checksum.

7. Selecteer de gewenste informatie in één van de velden.
8. Kies 'Copy'
9. Open een rekenblad en plak de gekopieerde informatie in het document; maak links van het veld een nieuw veld aan waarin je de naam plakt van het bestand waarop de checksum betrekking heeft.
10. Indien je de integriteit van het bestand nadien wil controleren, herhaal je de procedure en vergelijk je de bekomen checksum met de opgeslagen checksum. Als deze niet meer exact dezelfde is, is er iets gewijzigd in het bestand.

Bestandsnaam
2015_test_heemkunde_vlaanderen.docx

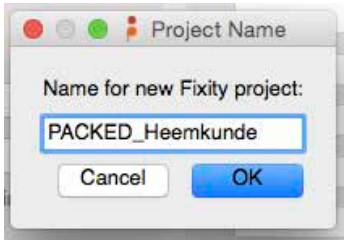
MD5-checksum
40aba48232490bbbfad5a9cbcd95a883

Je kunt deze procedure ook toepassen op ZIP-bestanden van folders waarin verschillende bestanden zijn opgenomen. Het nadeel is dan echter dat er maar één checksum wordt aangemaakt en geen checksum voor ieder individueel bestand dat onderdeel is van het ZIP-bestand. Met het oog op langetermijnbewaring is het ook beter om een SIP (Submission Information Package) aan te maken (zie verder). De bestanden worden dan samen met

metadata samengevoegd in één digitaal pakket. Bij de aanmaak van deze SIP worden er checksums per bestand aangemaakt die samen met de bestanden worden bewaard.

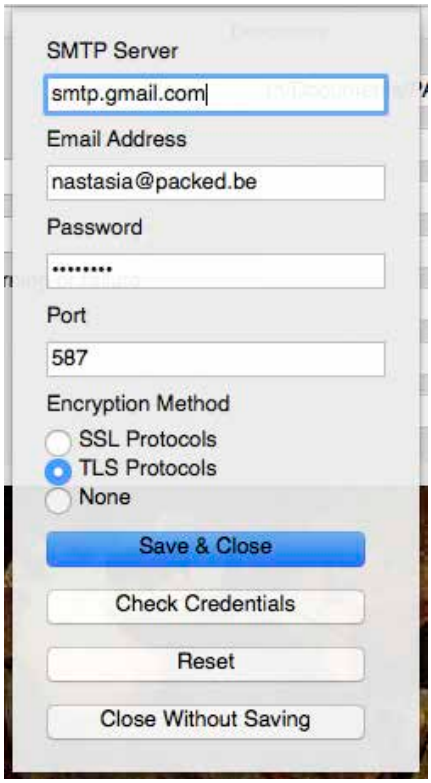
B. Fixity: de aanmaak en controle van checksums voor bestandsmappen

1. Open het programma Fixity.
2. Ga naar 'File' in de menubalk.
3. Kies 'New Project'.



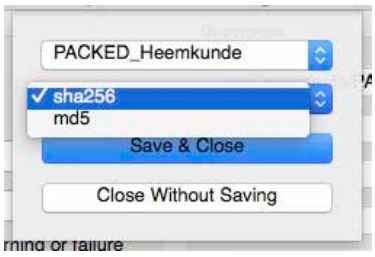
Figuur 7: maak de naam aan voor een nieuw project.

4. Selecteer de bestandsmap die je wil scannen
5. Indien je wil dat Fixity het rapport naar je mailt, vul je e-mailadres dan aan. Ga naar 'Preferences E-mail settings' om de gegevens van je e-mailadres in te geven.



Figuur 8: geef aan hoe de resultaten naar het e-mailadres moeten worden gestuurd.

6. Kies bij 'Preferences - Select Checksum Algorithm' volgens welk algoritme je de checksums wil aanmaken. Je kan kiezen tussen MD5 of SHA256.



Figuur 9: kies het algoritme voor de creatie van de checksum.

7. Je kan ook kiezen hoe regelmatig en op welk moment je wil dat Fixity de documenten controleert.

Figuur 10: geef aan hoe vaak dat je wil dat de integriteit van de bestanden wordt gecontroleerd.

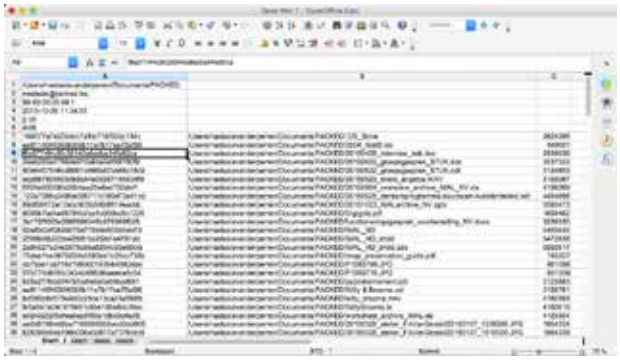
8. Ga naar 'File' en kies 'Save project'
9. Het project is opgeslagen. Kies 'Run now' om het scannen te beginnen.

Figuur 11: de waarschuwing om Fixity niet te sluiten terwijl het de opgeslagen bestanden scant.

10. Sluit Fixtiy niet tijdens het proces
11. Wanneer Fixity klaar is, krijg je een e-mail met het rapport.

Fixity slaat de rapporten op op je harde schijf in twee TSV-bestanden[11]: een TVS-bestand waarin aangeduid wordt hoeveel nieuwe, bevestigde of gewijzigde bestanden er zijn; en een TSV-bestand met een lijst van alle checksums. Het eerste TSV-bestand vind je in de Fixity map onder 'reports', het tweede bestand vind je in de 'history'-map.

Figuur 12: een TSV-bestand met een overzicht van de bestanden waarvan de integriteit in orde is.



Figuur 13: een CSV-bestand met in de eerste kolom de MD5-checksums en in de tweede de naam en locatie van de corresponderende bestanden.

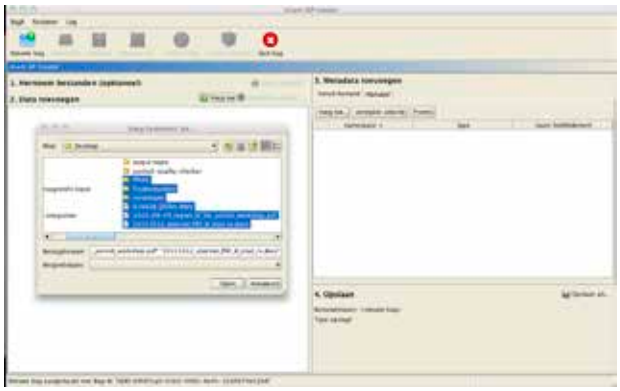
3. Ugent SIP Creator:[12] bestanden opslaan in een SIP, samen met checksum

De SIP Creator is een gebruiksvriendelijke software die toelaat om digitale bestanden en een uitgebreide bijhorende set metadata te verpakken in eenzelfde digitaal pakket dat is bedoeld voor opname in een digitaal archief. Deze tool is enerzijds gebaseerd op BagIt, wat toelaat om bestanden te verpakken, veilig te transporteren over een netwerk en op te slaan op harde schijven. Anderzijds is het gebaseerd op XML en een aantal open metadatastandaarden (DC, EAD, MARC21, METS) om een zo rijk mogelijke set metadata te kunnen bewaren naast de data. Bij het verpakken van een set digitale bestanden wordt per bestand ook een MD5-checksum gecreëerd die samen met het bestand wordt opgeslagen in het pakket.

1. Open de UGent SIP Creator software.
2. Kies 'New Bag' door op icoon linksbovenaan te klikken.

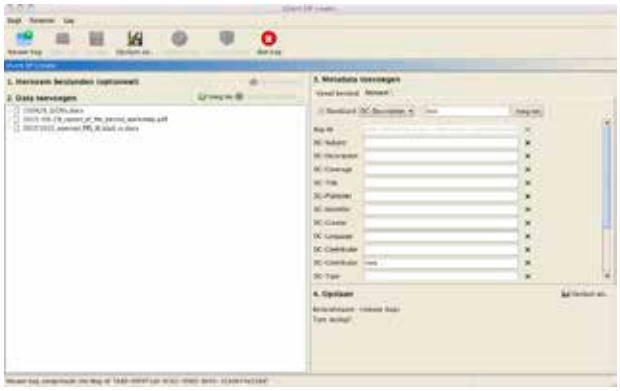
Figuur 14: geef de identifier voor de bag die je wil aanmaken.

3. Klik op '+ voeg toe' rechts van '2. Data toevoegen'.
4. Selecteer het bestand of map die je wil verpakken in het pakket.



Figuur 15: selecteer de bestanden en/of bestandsmappen die je wil opnemen in de bag.

5. Voeg metadata toe door onder '3. Metadata toevoegen' te klikken op 'Vanuit bestand' of 'Manueel'.



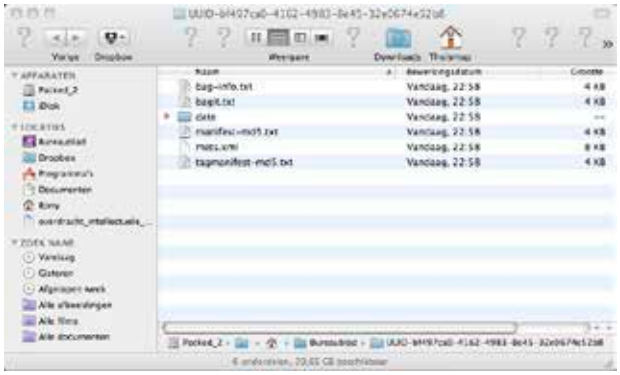
Figuur 16: voeg de gewenste metadata aan de bag toe door ze manueel of als bestaand bestand toe te voegen.

6. Klik nadien op ‘Opslaan als..’ rechts van 4. ‘Opslaan’.
7. Controleer en/of wijzig naast ‘Bestand’ de naam van het digitaal pakket.
8. Kies naast ‘Type opslag’ het type bestandformaat dat je verkiest voor het digitaal pakket: map, zip, tar, tar.gz of tar.bz3.
9. Kies naast ‘Type checksum’ het type checksum dat je wil gebruiken: MD5, SHA1, SHA256 of SHA512.

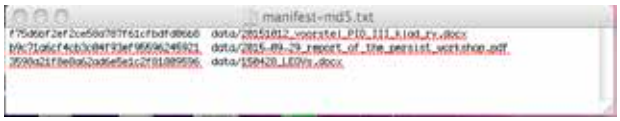
Figuur 17: kies het type checksum dat je wil gebruiken.

10. Klik op ‘Voltooien’.

Als je het digitale pakket opent, zie je:



Figuur 18: de inhoud van de gecreëerde bag.



Figuur 19: de MD5-checksums.

Rony Vissers
PACKED vzw

m.m.v. Nastasia Vanderperren
PACKED vzw

Henk Vanstappen

Dit artikel is de eerste aflevering in een artikelenreeks die PACKED vzw, expertisecentrum voor digitaal erfgoed, levert aan ‘Bladwijzer. Wegwijs met Heemkunde Vlaanderen’, het methodologisch tijdschrift van Heemkunde Vlaanderen. Indien u vragen hebt over de inhoud van dit artikel of over de uitdagingen van digitaal erfgoed in het algemeen, zijn ze steeds welkom op info@packed.be.

1 Voor meer informatie en download, zie: http://www.julifos.com/soft/checksum/
2 Voor meer informatie en download, zie: http://www.md5summer.org/
3 Voor meer informatie en download, zie: http://getmd5checker.com/
4 Voor meer informatie en download, zie: http://checksumchecker.sourceforge.net/
5 Voor meer informatie en download, zie: http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/
6 Voor meer informatie en download, zie: https://www.avpreserve.com/tools/fixity/
7 Voor meer informatie en download, zie: http://fsumfe.sourceforge.net/
8 Voor meer informatie en download, zie: http://www.fileformat.info/tool/hash.htm
9 Voor meer informatie en download, zie: http://www.jonelo.de/java/jacksum/
10 Voor meer informatie en download, zie: http://www.md5summer.org/
11 Tsv staat voor tab-separated value. Het is een gestructureerd tekstbestand waarin de waarden gescheiden worden door tabs. TSV-gegevens kunnen in een rekenblad- of een databaseprogramma worden ingelezen en vervolgens als tabel worden gepresenteerd.
12 Voor meer informatie en download, zie: http://projectcest.be/wiki/UGent_SIP_Creator