SAMENVATTING

Context

Grootschalige digitalisering is een belangrijk aandachtspunt in de cultureel erfgoedsector, waar een enorme hoeveelheid origineel materiaal zoals historische kranten, manuscripten, kaarten, enz. wordt omgezet naar digitale *formats*. Nu intussen een enorme hoeveelheid digitale collecties beschikbaar is, kwamen de laatste jaren enkele algemene problemen aan het licht. Zowel de nodige menselijke investering als bepaalde technische uitdagingen bleken zwaar onderschat en bij veel digitaliseringsprojecten blijft de kwaliteitscontrole een uitdaging. Niettemin is kwaliteitscontrole in holistische zin essentieel om de integriteit, de consistentie en de lange termijn bewaring van de geproduceerde bestanden en data, alsook de publieke toegankelijkheid ervan, te garanderen.

Doelstellingen en methode

ADOCHS (2016-2021) was een samenwerking tussen het Centrum Oorlog en Hedendaagse Maatschappij (CegeSoma, de vierde operationele directie van het Belgische Rijksarchief), de Koninklijke Bibliotheek van België (KBR), de Vrije Universiteit Brussel (VUB) en *de Université libre de Bruxelles* (ULB). Het project had tot doel de kwaliteitscontrole voor de digitalisering van het cultureel erfgoed te verbeteren door nieuwe methodologische benaderingen te ontwikkelen en een reeks praktische richtlijnen en bruikbare instrumenten te creëren voor een stapsgewijze aanpak van digitaliseringsprojecten, met als algemeen doel **de kwaliteit te verbeteren van de beelden en metadata** uit digitaliseringsprojecten van erfgoed en documenten. Drie aandachtspunten kwamen naar voren: verbetering van de kwaliteit van de beelden, van de metadata en van de digitaliseringsprocessen.

1/ Wat de beeldkwaliteit in de cultureel-erfgoedsector betreft, werden in dit project drie aspecten van beeldverwerking onderzocht, namelijk DIS (Document Image Segmentation), IQA (Image Quality Assessment) en DDR (Document Damage Recognition). Deze drie aspecten zijn onderling nauw verbonden onderdelen van een digitaliserings- en informatie-extractie workflow. Deze drie thema's werden benaderd vanuit een wiskundig oogpunt door gebruik te maken van 'homogeniteit'. Eén onderzoek (door dr. Tan Lu) conceptualiseerde en modelleerde homogeniteit op zodanige manier, dat relaties tussen elementen kunnen worden gekarakteriseerd en unieke patronen kunnen worden beschreven. Het onderzoek stelde zo een probabilistisch homogeniteitsmodel voor DIS voor, waarbij het begrip teksthomogeniteit werd gedefinieerd vanuit Gestalt-principes. Vervolgens werd een statistisch model geformuleerd dat gebruik maakt van teksthomogeniteit voor zowel tekst- als niet-tekst classificatie in de context van DIS. Dit leidde tot de ontwikkeling van een nieuwe DIS-methode, namelijk **documentsegmentatie met probabilistische homogeniteit** (DSPH), die veelbelovende resultaten liet zien op datasets gebruikt als benchmark.

Het probleem van DDR werd benaderd met **een modellering van zowel globale als lokale homogeniteit**, waarbij de lokale homogeniteit vanuit verschillende invalshoeken werd geëxploiteerd met behulp van grafiekmodellering en wavelet-propagatie. Deze verschillende modellen werden

geïntegreerd in een Bayesiaans raamwerk voor DDR. Testresultaten op echte beeldstalen toonden bemoedigende prestaties van de voorgestelde methode.

Tenslotte hebben we ons gewaagd aan een uniform IQA raamwerk door gebruik te maken van **domeinoverstijgende homogeniteit tussen natuurlijke- en documentaire beelden**. Gebaseerd op 'diepe convolutionele neurale netwerken' (DCNNs) en 'transfer-leren', werden de lessen uit de verwerking van natuurlijke beelden geleidelijk geëxploiteerd, eerst voor de beoordeling van documenten en vervolgens voor een uniforme kwaliteitsbeoordeling. Een verenigd raamwerk gebaseerd op 'generatief adversair leren' werd ook ontwikkeld, waarbij bemoedigende resultaten werden verkregen rond de onbruikbare 'ruis' over twee datasets gebruikt voor benchmark.

2/ Een tweede groot aandachtspunt was het doctoraatsonderzoek van dr. Anne Chardonnens over **archivistische 'authority data' in een Linked Open Data context**. Haar werk was gestructureerd rond twee delen: methoden voor kwaliteitsanalyse en methoden voor het verbeteren van de kwaliteit van (meta)data. Het eerste deel werd uitgevoerd in drie stappen - analyse van het beheer van de metadata/autoriteitsgegevens en hun publicatie, de analyse van de impliciete of expliciete behoeften van de desbetreffende erfgoedinstelling en de analyse van de impliciete of expliciete behoeften van de gebruikers (een semi-geautomatiseerde methode werd toegepast op honderdduizend gebruikersverzoeken, via de de digitale catalogi van KBR en CegeSoma). De wetenschappelijke resultaten werden voorgesteld in het doctoraal proefschrift van Anne Chardonnens (*La gestion des données d'autorité archivistiques dans le cadre du Web de données).*

Het tweede deel van haar doctoraatsonderzoek, gericht op de ontwikkeling van nieuwe methodologieën om de kwaliteit van (meta)data te verbeteren, had tot doel het potentieel te verkennen van **een semi-gecentraliseerd beheer van autoriteitsdata met behulp van de vrije en open-source Wikibase software**. De (meta)gegevens van CegeSoma werden opgeslagen via een Wikibase-beheersysteem, wat een reeks taken met zich meebracht zoals de standaardisatie en verwerking van een reeks zeer heterogene gegevens (opgeslagen in verschillende nominatieve lijsten/databanken of in het verzamelingsbeheersysteem "Pallas"), het creëren van scripts voor het koppelen van de records, een koppeling en integratie van persoonsentiteiten met externe datasets en van plaatsentiteiten met de formele lijst van het Rijksarchief, de ontwikkeling van een datamodel en de implementatie daarvan in een Wikibase omgeving, de aanpassing van een programma voor de massale data import vanuit csv-bestanden, en tot slot ook het beschrijven van de installatie en de configuratie, het maken van SPARQL queries en het updaten en kopiëren van de Wikibase omgeving. De resultaten van dit vernieuwende experiment werden ook geanalyseerd in het bovenvermelde doctoraatsproefschrift.

3/ Een laatste en derde *deliverable* is de **Gids voor Kwaliteitscontrole**[1] geschreven door Chloé Brault met de steun van Anne Chardonnens. In 2017 voerde Nicolas Roland (KBR) al een eerste studie uit van de gespecialiseerde literatuur, internationale normen, gidsen voor goede praktijken en algemene monografieën over dit onderwerp. Zijn werk werd verder gezet door Chloé Brault, die in oktober 2020

[1] https://www.cegesoma.be/nl/publication/gids-digitalisering-kwaliteit-cegesoma

door CegeSoma werd aangeworven. Zij bracht de meest frequente kwaliteitsproblemen in kaart en voerde aanvullend onderzoek uit, waaronder interviews met de digitaliseringsteams van het Algemeen Rijksarchief en het KBR. Mede op basis van de schat aan onderzoek dat door het projectteam (en in het bijzonder door dr. Anne Chardonnens en dr. Tan Lu) is verricht, publiceerde zij in september 2021 de ADOCHS-gids voor 'best practices' in het Frans, Nederlands en Engels.

De kwaliteitsgids biedt een *state-of-the-art* benadering van kwaliteitsprocessen, maar is ook praktijkgericht en spitst zich toe op concrete methoden die kunnen worden gebruikt door heel uiteenlopende collectiebeherende instellingen. Achtereenvolgens komen aan bod: een kort overzicht van de context van digitalisering en de bijhorende uitdagingen; een verfijning van het begrip 'kwaliteit' volgens de internationale norm ISO-9001 in overeenstemming met de specifieke realiteit van het digitaliseringsproces; de kwaliteit van beelden en hun metadata; de digitaliseringsomgeving en goed atelierbeheer; en tot slot essentiële 'samenvattende fiches' als praktisch geheugensteuntje om de kwaliteit van een digitaliseringsproject te garanderen. De gids werd op grote schaal verspreid onder de talrijke institutionele netwerken van de betrokken instellingen.

Besluiten en aanbevelingen

Dit onderzoek was een unieke gelegenheid om wiskundige modellering, *machine learning* en taalverwerkingsmethoden te combineren met meer traditionele methoden zoals interviews. Het project heeft als een concrete impuls gediend in de beide federale wetenschappelijke partnerinstellingen, om reflecties en experimenten te initiëren met betrekking tot processen van collectiedigitalisering. De testen en analyses die zijn uitgevoerd op basis van de casestudies van het CegeSoma/Rijksarchief en de KBR-collecties, maken het mogelijk nieuwe manieren te bedenken om aan de behoeften en verwachtingen van gebruikers te voldoen. Zij tonen ook de relevantie en het belang aan van **document- en beeldverwerkingstechnieken in de context van de digitalisering van het cultureel erfgoed,** waarbij menselijke kennis en expertise kunnen worden gekoppeld aan computeralgoritmen om de werkprocessen van digitalisering te ondersteunen en de exploitatie van de digitale data te verbeteren.

Naast het enthousiasme dat de studiedagen in het kader van dit project[2] teweeg brachten, is een van de grootste successen van ADOCHS waarschijnlijk dat de projectdynamiek na afronding van het project intussen wordt voortgezet door de lancering van **het Wikibase-Verzetsproject**[3] binnen het Rijksarchief en de ontwikkeling van een **Data Science Lab**[4] bij de KBR. Wij hebben goede hoop dat deze beide initiatieven, die concreet door de verschillende leden van het ADOCHS-team worden geleid, zullen voortbouwen op de lessen van het project en ook voortwerken op de vragen en problemen die nog moeten worden aangepakt, zoals de uitdagingen van de samenwerking met ICT, de onderlinge relatie tussen metadata en beelden, de automatisering van de kwaliteitsverbetering en de integratie van de resultaten binnen volledige erfgoedcollecties.

[2] See http://adochs.be/linking/ and http://adochs.be/idpchs/
[3] https://www.cegesoma.be/nl/project/wikibase-verzet
[4] http://adochs.be/wp-content/uploads/2021/09/AnnDooms_FredericLemmers_KBRDataScienceLab.pdf

Aangezien het project heeft aangetoond hoe belangrijk het is methoden en modellen aan te passen en te ontwikkelen die inspelen op de specifieke kenmerken van het desbetreffende cultureel erfgoed, koesteren we vooral ook de hoop dat dit project een van de eerste zal zijn in een lange reeks, waarbij technologie wordt gebruikt om de ruimere toegang tot collecties te verbeteren.

Trefwoorden
Datakwaliteit; Digitalisering; Semantisch Web; Linked Open Data; Beeldverwerking