COLLECTIONS AS DATA

Digitale collecties van cultureel-erfgoedinstellingen, zoals bibliotheken, archieven en musea, worden steeds vaker gebruikt voor digitaal onderzoek in de geesteswetenschappen. Traditionele manieren om toegang te verlenen tot dergelijke collecties, bijvoorbeeld via digitale bibliotheekplatformen zijn niet altijd ideaal voor onderzoekers die datasets rond specifieke onderzoeksvragen willen opbouwen. Het *Collections as Data*-initiatief, afkomstig uit de Verenigde Staten, stimuleert nauwere interdisciplinaire samenwerking tussen cultureel-erfgoedexperten, *digital humanities*-onderzoekers en *data scientists* om samen na te denken over hoe toegang tot digitale collecties verleend kan worden. Zo kan analyse vergemakkelijkt worden met behulp van digitale tools en methoden. Wat is *Collections as Data* precies? Kan het interessant zijn voor cultureel-erfgoedinstellingen in België?

TEKST **Sally Chambers (KBR & GhentCDH) en Frédéric Lemmers (KBR)**

COLLECTIONS AS DATA: WAT IS DAT PRECIES?

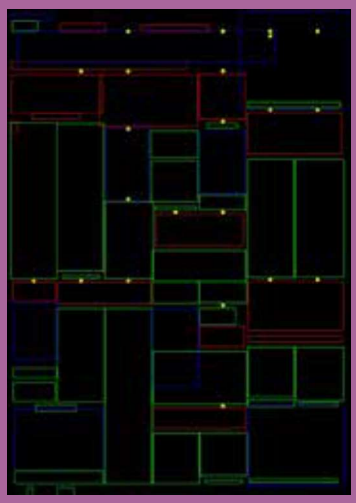
Collections as Data biedt een nieuwe denkwijze aan over hoe toegang tot digitale (gedigitaliseerde en born-digital) collecties in cultureel-erfgoedinstellingen aangeboden kan worden.

De eerste fase van het initiatief, *Always Already Computational*: *Collections as Data* (2016-2018), focuste op het uitwisselen van ervaringen en het delen van kennis over potentiële benaderingen om toegang aan te bieden tot het onderliggende data-niveau van digitale collecties.

In de huidige, tweede fase *Collections as Data: Part to Whole* (2019-2021), ligt de focus op de implementatie en het gebruik van *Collections as Data* door een aantal collaboratieve casestudy's. Een belangrijk onderdeel van deze benadering is dat de interdisciplinaire samenwerking tussen cultureel-erfgoedexperten en *digital humanities*-onderzoekers gelijkwaardig moet zijn. Elke *Collections as Data*-casestudy wordt gezamenlijk geleid.

Tot nu toe heeft de implementatie van *Collections as Data* grotendeels plaatsgevonden in de Verenigde Staten, maar het gebeurt langzamerhand ook in Europa. Er zijn al een aantal Europese voorbeelden, onder andere de Data Foundry van de nationale bibliotheek van Schotland, het Open Data Platform van de nationale bibliotheek van Luxemburg en de datasets van het KB Lab van de Koninklijke Bibliotheek van Nederland. Verder is het gebruik van collaboratieve notebooks zoals de GLAM Workbench van Tim Sherratt of de GLAM Notebooks van de labs van de Biblioteca Virtual Miguel de Cervantes, een andere manier om digitale collecties te benaderen op computationele wijze. (*GLAM staat voor galleries, libraries, archives en musea, n.v.d.r.*) Maar, waar staan we eigenlijk met *Collections as Data* in België?

Automatische layout-analyse van de historische kranten van KBR, door IDLab, UGent.



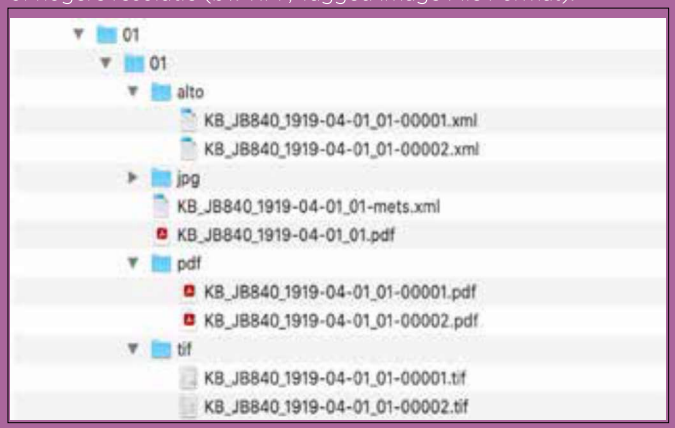
Automatische layout-analyse van de historische kranten van KBR, door IDLab, UGent.

> *Collection as Data* zou een game-changer kunnen zijn voor de samenwerking tussen cultureel-erfgoedinstellingen en digital humanities-onderzoekers

DATA-KBR-BE: COLLECTIES ALS DATA BIJ KBR

Geïnspireerd door de *Collections as Data*-beweging, financierde het Federaal Wetenschapsbeleid (BELSPO) *DATA-KBR-BE*, een project van 24 maanden (2020-2022). Dat gebeurde in het kader van het onderzoeksprogramma BRAIN-be (Belgian Research Action through Interdisciplinary Networks). Het project is een interdisciplinaire samenwerking tussen cultureel-erfgoedexperten, *digital humanities*-onderzoekers en *data scientists* om toegang tot de gedigitaliseerde en born-digital collecties van KBR aan te bieden op data-niveau voor *digital humanities*-onderzoek. In eerste instantie ligt de focus op BelgicaPress, de gedigitaliseerde historische krantencollectie van KBR. Op de langere termijn zou er naar andere digitale collecties gekeken kunnen worden, zoals gedigitaliseerde collecties (bv. BelgicaPeriodicals) en born-digital collecties, bv. de gearchiveerde websites van het *PROMISE*-project (2017-2019) en een archief van sociale media van het *BESOCIAL*-project (2020-2022).

Toegang tot digitale collecties op data-niveau houdt in dat je toegang krijgt tot de onderliggende databestanden, bv. XML-bestanden in ALTO-formaat (Analysed Layout and Text Object) met de volledige tekst- en opmaakinformatie van de automatische tekstherkenning (OCR, Optical Character Recognition) of de structurele metadata over de gedigitaliseerde krantenpagina's, bv. in METS-formaat (Metadata Encoding and Transmission Standard). Verder kunnen het pdf's zijn van de gescande krantenpagina's of de afbeeldingen in een lagere (bv. JPEG, Joint Photographic Experts Group) of hogere resolutie (bv. TIFF, Tagged Image File Format).



Een voorbeeld van de onderliggende data van een gedigitaliseerde historische krant bij KBR.

DIGITAL HUMANITIES-CASESTUDY'S

Onderzoeksteams bij het Ghent Centre for Digital Humanities (GhentCDH) en het Antwerp Centre for Digital Humanities and Literary Criticism (ACDC) werken nauw samen met de digitaliserings-, collecties- en ICT-experten van KBR om drie interdisciplinaire onderzoeksscenario's te ontwerpen. Deze onderzoeksscenario's zijn bedoeld als eerste *digital humanities*-casestudy's om het wetenschappelijke potentieel van *Collections as Data* aan te tonen. Deze casestudy's worden uitgevoerd in nauwe samenwerking met het KBR Digital Research Lab. De interdisciplinaire onderzoeksscenario's die geselecteerd werden voor het project zijn:

- **Collective Action Belgium**, geleid door GhentCDH, concentreert zich op de sociale geschiedenis in het interbellum en tijdens de Tweede Wereldoorlog en wil de dynamiek van stakingen, demonstraties en andere vormen van collectieve actie in België traceren, zoals gerapporteerd in Belgische kranten;
- **Het feuilleton in België**, geleid door ACDC, richt zich op literaire studies in de periode 1830-1930 en heeft tot doel de publicatie van literatuur in Belgische kranten tijdens de eerste eeuw van de Belgische natiestaat in kaart te brengen;
- **De geschiedenis van de Belgische journalistiek**, geleid door CAMILLE (Centrum voor Archieven over Media en Informatie, ULB-KBR), focust op de mediageschiedenis van 1886 tot nu en wil de geschiedenis van de Belgische journalistiek traceren door de lens van kritische discoursen over journalistiek in Belgische kranten.

Bovendien worden relevante thematische datasets uit BelgicaPress geëxtraheerd op basis van de expertise van *data scientists* bij het Internet Technology and Data Science Lab (IDLab). Hiervoor worden artikelen uit historische kranten onderworpen aan een automatische layout-analyse en een semi-automatische extractie en classificatie.

Een belangrijke onderdeel van het project is om te begrijpen hoe *Collections as Data* op een duurzame manier bij KBR geïmplementeerd kan worden.

COLLECTIES ALS DATA IN BELGIË STIMULEREN?

Collections as Data staat momenteel nog in de kinderschoenen in België. *DATA-KBR-BE* zal de implementatie van *Collections as Data* bij KBR op gang brengen, maar het heeft ook de visie om andere instellingen in Vlaanderen, België en daarbuiten te inspireren om te experimenteren met *Collections as Data*. Zijn jullie er klaar voor? ▮▮