Rapport

Datum: 9 april 2017
Van: Matthias Vandermaesen
Betreft: Verslag dienstreis LDCX, Stanford University

Verslag congres, lezing, workshop, ...

Dienst/instelling:	Vlaamse Kunstcollectie VZW
Deelnemer:	Matthias Vandermaesen

Naam van de activiteit:	LDCX: Library Developer Conference X
Inrichter:	Stanford University
Adres:	Stanford University, Stanford, California 94305
Plaats:	Stanford Univ.
Datum:	26.03.2018 - 28.03.2018

Was uzelf of uw instelling spreker/mede-inrichter:

Ik was geen aangekondigde spreker, niettemin kreeg ik een half uur tijd om onze VKC-datahub toe te lichten en een half uur om de ervaringen met catmandu te duiden. Elke occassionele tussenkomst werd door het aanwezig publiek beoordeeld. De VKC-datahub kreeg werd zeer positief onthaald.

Evaluatie van congres, lezing, workshop, …
Inhoudelijk:

LDCX is een unconference. Dit is een evenement waarbij er vooraf geen afgelijnd programma is vastgelegd. Van de deelnemers wordt verwacht dat zij zelf thema's en presentaties brengen en dus de inhoud actief bepalen.

Het doelpubliek bestaat uit peers die actief betrokken zijn in de ontwikkeling van digitale technologie voor GLAM - galleries, libraries, archives, museums - ruimte. LDCX legt wel vooraf een aantal high-level thema's op om de tussenkomsten toch enigszins af te bakenen. Dit jaar waren de afgebakende thema's: artificiële intelligentie en machine learning, data-forward architectures en API's, ETL, data pipelining, serverless system design, web archivering, distributed digital preservation, moderne front-end integratie en IIIF.

Om inhoudelijk een heldere lijn te behouden, opteerden de organisatoren - Stanford Digital Library Systems and Services (DLSS) - om de conferentie op te delen in 5 parallele tracks:

- Data Architecture Track
- Artificial Intelligence & Machine Learning Track
- Digital Preservation Track
- Front End Integration track

- Web archivering track

De Vlaamse Kunstcollectie werkte de voorbije jaren voornamelijk op metadata via de projecten Datahub en Persistente Identificatoren in samenwerking met PACKED vzw. Er werd gekozen om de Data Architecture track bij te wonen. Enerzijds om actief een bijdrage in de discussies te leveren, anderzijds om de meest recente evoluties in de Angelsaksische wereld eigen te kunnen maken.

Maandagvoormiddag (26/03)

Na een (praktische) introductie door de organisatoren werden de deelnemers uitgenodigd om voorstellen voor plenaire sessies in te dienen (session pitches). Nadien volgende een aantal lightning talks. Dit zijn 7 minuten durende presentaties die een specifiek onderwerp belichten:

- Fedora: An Update (Andrew Woods, DuraSpace)
- Sword protocol (Neil Jefferies, Bodleian Libraries, University of Ofxord)
- Adding Tests to Universal Viewer (Jack Reed, Digital Library Systems and Services, Stanford)
- Changing the Interview Process - A case study (Jeremy Friesen, Digital Library Technologies, University of Notre Dame)
- The Other Development (Steven Ng, Temple University Libraries)
- LD4L Questioning Authority (QA) Server (E. Lynette Rayle, Cornell University)
- Query-time deduplication, Leveraging Solr to integrate records from multiple sources (Kerrick Rogers, Penn University Libraries)
- LAKEsuperior, A New, Old Fedora implementation (Stefano Cossu, The ArtUniversity of Chicago)
- TACOs are good. (Christina M. Harlow, Digital Library Systems and Services, Stanford)
- The Intersection of ETL, Machine Learning, Collections as Data and Decentralized Data (Matt Zumwalt, MediaShelf)

Maandagnamiddag (26.03.2018)

Plenaire sessies welke werden opgevat als plenaire discussies, en niet zozeer formele presentaties. Deze duurden telkens 60 minuten.

Plenaire sessie: Metadata Aggregation Systems (Randy Stern)

Aggregatie van metadata uit diverse bronnen kan op de meest diverse manieren worden geïmplementeerd. Welke use cases zijn er? Welke tools gebruikt men? Hoe benadert men een gemeenschap van stakeholders? Bestaande technologieën? Metadata kwaliteit doorheen de tijd? Domein-specifieke ontologieën? Use cases die in de sessies aan bod kwamen: Harvard Libraries, Cincinnati University Libraries, Oxford University Libraries, Europeana, British Library, **Vlaamse Kunstcollectie**.

Conclusies:

- Instellingen maken gebruik van gelijkaardige standaarden (OAI-PMH, MODS, EAD, Dublin Core, MARC,..) en technologieën (Solr, Apache Camel, Amazon SQS, Apache Spark, etc.) omdat ze voor gelijkaardige uitdagingen staan wanneer men metadata wil aggregeren (uniform datamodel,

rechten, gecontroleerd beheer, automatisering,...)

- De uiteindelijke implementatie van aggregatieplatforms per instelling verschilt dan weer vrij sterk. De businesscases waarvoor men aggregatie implementeert, verschillen danig. Dat leidt tot verschillen in functionele vereisten en dus verschillende technische oplossingen die eerst en vooral gericht zijn op het aansluiten bij de eigen businesscase.

Plenaire sessie: Measuring Metadata Quality (Péter Kiraly)

Hoe meten we de kwaliteit van metadata? Welke criteria kunnen we hanteren om kwaliteit te meten? Kernuitdaging: Welke aspecten van metadata kunnen we kwantificeren en welke aspecten zijn zo subjectief/interpretatief dat ze niet te vatten zijn in een mathematisch model?

De kwaliteit van metadata is een topic in de verschillende Europese en Amerikaanse projecten.

Europeana Data Quality Committee [1], ADOCHS [2], Conquaire [3], DLF Assessment Interest Group Metadata Working Group [4] en UNT Libraries [5]. Er zijn verschillende succesvolle lopende en afgelopen doctoraatsonderzoeken (in Spanje, België en Duitsland) [6].Een community based bibliografie wordt onderhouden op [7]. Er zijn ook enkele inspanningen om een generiek schema te ontwikkelen voor gemeenschappelijke bibliotheekformaten zoals MARC21 [9] en PICA [10]; verder ook tools om data tegen scheam's te valideren [11] [12]. Enkele tools om te overwegen: Avram JSON schema, command line validation met Catmandu, hoe MARC21 checken, tools voor linked data quality assurance (ShEx [13], SHACL [14], Luzzu [15]).

[1]: https://pro.europeana.eu/project/data-quality-committee
[2]: http://adochs.be/
[3]: http://conquaire.uni-bielefeld.de/
[4]: http://dlfmetadataassessment.github.io/
[5]: https://vphill.com/journal/
[6]: Seth van Hooland, *Metadata Quality in the Cultural Heritage Sector: Stakes, Problems and Solutions* (2009); Sascha Tönnies, *Quality Control using Semantic Technologies in Digital Libraries* (2013); Nikos Palavitsinis, *Metadata Quality Issues in Learning Repositories* (2014). [7]:
https://www.zotero.org/groups/metadata_assessment/items
[8]: http://format.gbv.de/schema/avram/specification
[9]: https://pkiraly.github.io/2018/01/28/marc21-in-json/
[10]: https://metacpan.org/pod/PICA::Schema
[11]: https://github.com/pkiraly/metadata-qa-marc
[12]: https://metacpan.org/pod/Catmandu::Validator::PICA
[13]: https://www.w3.org/community/shex/
[14]: https://www.w3.org/TR/shacl/
[15]: https://eis-bonn.github.io/Luzzu/team.html

Conclusies:

- Kwantificeren van kwaliteitsparameters is een uitdaging.
    - FAIR: Findable, Accessible, Interoperable, Reusable. De eerste twee aspecten van metdata kunnen objectief worden gekwantificeerd, de laatste twee kunnen dat niet omdat ze te contextafhankelijk zijn en dus onderhevig aan menselijk interpretatie.
- Authorities (i.e. gestandaardiseerde thesauri) vormen een uitdaging:
    - Bij het normaliseren van lokale authoriteiten (i.e. lokale thesauri gehanteerd in lokale instellingen) naar gestandaardiseerde, algemene authoriteiten (i.e. AAT, ICONCLASS,...) loop je het risico dat waardevolle lokale betekenis verloren gaat.
    - "Slechte" metadata bevat ook een betekenisvol verhaal.
    - Het beste is om de beide naast elkaar te behouden waar mogelijk.
    - Linked Data kan voldoende context voorzien om betekenis af te leiden uit metadata.
    - Het "schonen" van metadata is een "forever" job zonder de juiste toolset.
- Hoe scoor je velden in een datamodel op basis van "volledigheid" en "descriptiveness"?
- Linked data: wat is de kwaliteit van persistente URL's? Werken ze wel correct?

Zie: https://drive.google.com/open?id=1LwvrnFMG5ZzG04prMcOOhjYSc8xe6PqccMWTkGDxCvI

Theme track: Data planning

In de laatste 90 minuten werd er in besloten kring besproken hoe de Data Architecture Track voor dinsdag zou worden ingevuld. Welke topics komen aan bod? Welke use cases willlen we bespreken? Wat willen we leren? Wie wil iets presenteren? Etc.

Het resultaat werd vastgelegd in onderstaand document:

https://docs.google.com/document/d/1FH3BbQWjq7pVoszUOgjjVARFfDBu1HeZ5wjvvU8GBKg/edit?usp=sharing

Dinsdagvoormidag 27.03.2018

Plenaire sessie: Collections as data (Hannah Frost & Kate Lynch)

Discussie thema's:

- Wat is de rol van repositories (contentproviders) in het ondersteunen van de "computational analysis" van de inhoud van digitale collecties?
- Welke types onderzoeks/computer-gegenereerde datasets maken we aan wanneer we digitale collecties online toegankelijk maken?
- Welke technologieën zijn noodzakelijk om bruikbare data toegankelijk te maken voor gebruikers?

De discussie tussen de deelnemers aan deze sessie wisselde tussen enerzijds beschikbare tools en technieken (types API's, diverse use cases) en anderzijds de vraag rond IP en licenties. Ook hier was de discussie vooral gericht op de uitwisseling van ervaringen met een vooral technologische insteek.

Zie: https://drive.google.com/open?id=13vbh2B-ZQ-QykS9JEyT57SmsRecPWN8jg53BaR0v8s4

Dinsdagvoormiddag + namiddag (27.03.2018)

De rest van de dag werd voornamelijk gespendeerd in de Data Architecture track. Deze werd onderverdeeld in drie grote blokken:

Theme track: Group Retro / Fail4Data (60 minuten)

In deze sessie werden lessons learned uit gefaalde projecten aangekaart in de groep:

- **Concrete use cases horen de ontwikkeling van toepassingen vooruit te drijven, in plaats van omgekeerd.** Bekijk eerst of er ook effectief vraag is naar data / bekijk eerst de vooropgestelde toepassing!
- **Persistentie =/= Data presentatie**. Het formaat waarin je data bewaart in een databank (XML, Linked Data,...) is niet hetzelfde formaat dat ook effectief een meewaarde creëert voor de eindgebruiker.
- **Ken je data!** Zorg dat je een inzicht hebt in de inhoud en de structuur van de dataset. Zo kan je makkelijker prioriteiten bepalen. Als je je data kent, wordt het eenvoudiger om een systeem rond je data te ontwerpen.
- **Ken je klanten.** Zij bepalen immers welke data relevant is, en het formaat waarin data ook effectief een meerwaarde kunnen brengen.
- **Schoon je data op terwijl je je klanten bedient**
- **Perfectionisme is ook een type van falen!**

Zie: https://docs.google.com/document/d/1Cp84rZCkV8I6UXpeIGaN1_pmvCNgBBBHQIfw2DCSbT8

Theme track: Project Deep Dives (3 x 30 minuten)

In dit onderdeel werden 3 projecten gepresenteerd:

- Europeana Data Quality Assessment (Peter Kiraly)
- Catmandu (Matthias Vandermaesen)
- Figgy (Trey Pendragon)

Zie: https://docs.google.com/document/d/1QccAg5AjNcAZSLjaBJITuwRHNA_8gcN5YQuD6Ljcxkk

Theme track: Cloud/Serverless (45 minuten)

In deze korte discussie werd nagegaan hoe cloud technologie (AWS) dan wel serverless systemen kunnen worden gehanteerd als onderbouw voor digital repositories. Daarbij werd niet alleen de technologische aanpak belicht, maar ook de enkele vraagstukken inzake copyright.

Zie: https://docs.google.com/document/d/1FQZRg-C7NL5b4jlueEg-TZ6_-jch2s_xS9YY0psZWBc/edit -heading=h.dgt1vwcozrgg

Dataflow (45 minuten)

In deze sessie werd er diep gedoken in Islandora-CLAW. Islandora is een respositorysysteem gebouwd met Drupal en Fedora Commons. Islandora CLAW is de next gen versie van Islandora gebouwd met Drupal 8. In deze sessie werd belicht hoe Islandora CLAW data ingest afhandelt en welke technische componenten hier onderliggend voor worden gebruikt.

Woensdagvoormiddag (28.03.2018)

Plenaire sessie: Accessing Authorities in your Application (Lynette Rale)

De discussie opende met een voorstelling van de LD4L (Linked Data for Libraries) Authority Lookup Service (http://lookup.ld4l.org/authorities). Deze service laat toe om een aantal gedistribueerde authorities (LOC Subject, Geonames, DBPedia, Agrovoc, OCLC Fast) te benaderen vanuit een centraal platform. Het achterliggende idee is het duurzaam bestendigen van de beschikbaarheid van de authority service.

Integratie tussen authorities en tools die door collectiebeheerders worden gebruikt (registratiesystemen!) is absoluut noodzakelijk. Hoewel een dergelijke integratie een aantal technische uitdagingen met zich mee brengt:

- Caching van de opgevraagde thesaurus om beschikbaarheid te garanderen.
- Het up-to-date houden van de authority langs de zijde van het registratiesysteem
- Verschillende formats die door de diverse authority services worden aangeboden via hun API's.

Er wordt gekeken naar integratie van Linked Data in LD4L QA service, eventueel Linked Data Fragments, maar dergelijke integratie is lange termijn muziek.

De discussie ging dan dieper in op de pijnpunten van het toegankelijk maken van authorities in een gebruikersinterface. Het beschikbaar maken van termen via een autocomplete formulierveld is een must, maar niet altijd even eenvoudig te implementeren.

Ook het indexeren van termen voor zoekopdrachten blijkt niet altijd even eenvoudig of eenduidig te kunnen gebeuren. Ook al worden de "correcte" termen door de collectiebeheerders toegewezen aan een record, dan nog is niet gegarandeerd dat daarmee een relevante zoekresultaat zal geven voor een eindgebruiker. Veel hangt af van de context waarin een zoekopdracht wordt gelanceerd: full text search, browsing, ... maar ook: hoe de eindgebruiker de term precies interpreteert. Veel termen kennen immers meerdere (dubbele) betekenissen die niet altijd expliciet zijn gemaakt.

Zie: https://docs.google.com/document/d/1oBh5EMxLXTpCTyWFz0rzpEZa-eMat59n794LXQE3bkQ

Theme track: Group Design Session

De groep werd onderverdeeld in 3 kleinere groepen. Doel was om via een whiteboard oefening een ruw ontwerp te maken van een forward data architecture. De use case: "Stel, je hebt enkele millioenen records, met bijhorend enkele miljoenen digitale objecten (audio, beeld, video), hoe kan je die zo snel mogelijk on line krijgen?"

Het doel van de opdracht was om na te gaan in welke mate de drie ontwerpen gelijkenissen met zich mee

dragen. Samengevat:

- Er wordt uit gegaan van het idee dat alle objecten en records (metadata) reed permanent zijn geïdentificeerd (ARK ID: https://en.wikipedia.org/wiki/Archival_Resource_Key). Tevens wordt er van uit gegaan dat er een gestandaardiseerd formaat wordt gehanteerd.
- Metadata wordt via een data pipeline met een message queue in een indexer service gestopt.
- De indexer service transformeert de binnenkomende records naar het datamodel dat de toepassing onderliggend aan de zoektoegang hanteert (bv. Project Blacklight)
- De indexer service upload de getransformeerde records in een Apache Solr Index.
- De digitale objecten gaan door een "characterization" service (bv. JHOVE) om na te gaan of het om audio, video of beeldmateriaal gaat. Afhankelijk van het formaat wordt het materiaal naar een andere component gestuurd.
- Voor beeldmateriaal betekent dit dat er eerst "derivatives" of afgeleiden gemaakt worden (i.e. TIFF naar JPG). Eventueel wordt beeldmateriaal ook nog eens on-the-fly OCR'ed (tekstextractie) via Tesseract OCR (https://github.com/tesseract-ocr/tesseract)
- De beelden worden ter beschikking gesteld via een LORIS IIIF compliant server (https://github.com/loris-imageserver)
- De URL's worden via een message queue doorgegeven aan de Indexer service waardoor ze mee in de metadata asynchroon worden verwerkt.
- Via Project Blacklight wordt het geheel toegankelijk gemaakt (doorzoekbaar via een gefacetteerde zoekinterface)

Feedback & Wrap Up

LDCX eindigde met een een Feedback / Wrap up waarbij plenair ruimte werd gelaten voor een Plus/Delta (feedback) ronde. Die feedback wordt verwerkt in functie van de volgende editie van LDCX.

De volledige verslagen van alle tracks in LDCX zijn hier terug te vinden: http://bit.ly/ldcx2018
Het volledige programma is hier terug te vinden: http://ldcx2018.sched.com/

Organisatorisch:

LDCX werd zeer goed en professioneel georganiseerd.

- De timing van het programme werd rigoureus gevolgd en liep nooit uit.
- Er werd voor voldoende catering voorzien: ontbijt/lunch.
- De sessies werden georganiseerd met een facilitator (moderator), een notekeeper (secretaris) en een gatekeeper (iemand die lette op vakjargon)
- Er was meer dan voldoende ruimte om zelf ook aan de discussies deel te nemen. Dit werd zelfs aangemoedigd.
- Op maandag- en dinsdagavond waren er uitgebreide netwerkmomenten (pizza-avond / meer formele receptie)
- De organisatoren zorgden ervoor dat de focus voldoende bewaakt werd om tot optimale inhoudelijke resultaten te komen.

Eigen presentatie (indien van toepassing):

Catmandu Deep Dive (30 minuten)

In deze korte presentatie werd voor een publiek van 30 library technologists (Amerikaanse Universiteitsbibliotheken) een hands-on demonstratie gegeven van Catmandu (http://librecat.org) Deze tool wordt door Vlaamse Kunstcollectie intensief gebruikt om museale data te bewerken en uit te wisselen. De presentatie bestond uit twee onderdelen: een demo van de functionaliteit van Catmandu, en een demo van de integratie van Catmandu in het Datahub/Arthub platform welke de Vlaamse Kunstcollectie operationeel onderhoudt.

Resultaten

Diego Pino Navarro, Release manager van Islandora (http://islandora.ca/) en developer bij de New York Metropolitan Library Council (https://metro.org/) toonde interesse in Catmandu en de software-componenten, gebouwd tijdens het Datahub project, die op Github staan gepubliceerd: http://github.com/thedatahub.

Stefano Cossu, Senior Application Developer bij de The Art Institute of Chicago (http://www.artic.edu/) en Islandora contributor, toonde eveneens interesse in Catmandu en de software componenten, gebouwd tijdens het Datahub project, die op Github staan gepubliceerd: http://github.com/thedatahub.

De verworven kennis draagt op volgende wijzen bij tot de verdere uitbouw en verfijnining van het VKC-ecosysteem:

- Het gericht kiezen van nieuwe technologieën die hun waarde bewezen hebben en aansluiten bij de noden en behoeften van collectiebeherende instellingen.
- Het gericht kiezen van formaten & standaarden die hun waarde bewezen hebben en aansluiten bij de noden en behoeften van collectiebeherende instellingen.
- De aansluiting van de technische onderbouw op een conceptueel model van internationale, in de praktijk bewezen, zienswijzen en visies.
- De adoptie van werkwijzen die in internationale context gangbaar zijn om een digitaal architectuur voor de ontsluiting van museale collecties te ontwikkelen, en het herschalen van dergelijke werkwijzen naar Vlaamse leest.
- De bevestiging dat de door VKC reeds gehanteerde werkwijzen, technologieën en standaarden aansluiten bij de gangbare, internationale praktijken.

De kennis die tijdens deze conferentie werd verzameld - en vervat zit in dit verslag en de rapporten waarnaar werd verwezen in dit verslag - zal verder worden gedeeld met de Vlaamse erfgoedsector.

Datum:
13 april 2018

Handtekening:

Bijgevoegd: programma congres, workshop, lezing, ...
De volledige verslagen van alle tracks in LDCX zijn hier terug te vinden: http://bit.ly/ldcx2018
Het volledige programma is hier terug te vinden: http://ldcx2018.sched.com/