Technische uitleg dataset ANP

Dit document geeft uitleg over en voorbeelden van de

- Beschrijvende en structurele metadata (MPEG21-DIDL),
- Beeldobjecten (typoscripten in JPG)
- Full-text objecten (OCR en ALTO in XML)
- Metadata harvest-API op basis van OAI-PMH,
- Zoek-API op basis van SRU,

voor de set ANP (http://www.kb.nl/banners-apis-en-meer/dataservices-apis/anp) die de Koninklijke Bibliotheek als open data aanbiedt.

Om snel aan de slag te kunnen met ANP bieden we een Snelstart ANP (pdf, in het Engels) aan. Ook kan het nuttig zijn de Snelstart SGD (pdf) en de Snelstart EDBO (pdf) te raadplegen om a.d.h.v. andere open KB-sets binnen de KB-infrastructuur meer inzicht te krijgen hoe SRU-queries opgebouwd moeten worden.

Voorwaarden hergebruik & licenties

Het ANP heeft de **objecten** in deze set (JPGs, OCR's, ALTO's) beschikbaar gesteld onder een CC-BY-NC 3.0 licentie

Gebruik bij naams-, bron- en licentievermeldlng het volgende format:

Algemeen Nederlands Persbureau (ANP) & Koninklijke Bibliotheek (KB). Bron: <persistente link naar JPG> – CC-BY-NC

voorbeeld persistente link naar de JPG http://resolver.kb.nl/resolve?urn=anp:1938:10:01:2:mpeg21:image

Commercieel hergebruik van de objecten dient vooraf schriftelijk te worden goedgekeurd door ANP. U kunt daarvoor contact opnemen met business@anp.nl.

De Koninklijke Bibliotheek heeft afstand gedaan van het auteursrecht op de **metadata**. Deze bestanden zijn derhalve beschikbaar onder de CC0 1.0 Universal verklaring.

Metadata

Elk typoscript is van een aantal beschrijvende metagegevens voorzien. Voor deze metadata wordt de standaard de Dublin Core (dcx) formatering gebruikt. De volgende beschrijvende metadata zijn opgenomen.

Voorbeeldrecord
http://services.kb.nl/mdo/oai?verb=GetRecord&identifier=anp:anp:1938:10:01:2:mpeg21&metadataPrefix=didl

(bericht nummer 2 van 01-10-1938)

<dcx:recordIdentifier>	Unieke identifier van het typoscript Formaat: *anp:jaar:maand:dag:berichtnummer*
<dc:identifier>	Persistente URL naar de scan (jpg) van het typoscript Formaat: http://resolver.kb.nl/resolve?urn=<*dcx:recordIdentifier*>:mpeg21
<dc:language>	Taal van het typoscript : Nederlands
<dc:rights>	Rechthebbende van de objecten : ANP Het ANP heeft de objecten (JPGs, OCR's, ALTO's) beschikbaar gesteld onder een CC-BY-NC 3.0 licentie
<dcx:recordRights>	Rechthebbende van de metadata : Koninklijke Bibliotheek, Den Haag De Koninklijke Bibliotheek heeft afstand gedaan van dit auteursrecht, de metadata is dus beschikbaar onder de CC0 1.0 Universal verklaring.
<dc:title>	Titel van het typoscript : Formaat : *ANP Nieuwsbericht – dag-maand-jaar – berichtnummer van die dag*
<dc:date>	Datum van uitzending Formaat : *jaar-maand-dag*
<dcx:volgnummer>	Berichtnummer op een bepaalde dag : per dag werden meerdere bulletins voorgelezen
De beschrijvende metadata zijn als XML-blok aan het begin van de structurele metadatabestanden (MPEG21-DIDL) opgenomen, er is dus geen apart XML-bestand voor de beschrijvende metadata beschikbaar	

Objecten

De metadata van een typoscript met identifier *anp:1938:10:01:2:mpeg21* is op te vragen via http://services.kb.nl/mdo/oai?verb=GetRecord&identifier=anp:anp:1938:10:01:2:mpeg21&metadataPrefix=didl

Dit mpeg21-didl bestand bevat eerst een blok beschrijvende metadata over het typoscript (zie tabel hierboven), daarna volgt de structurele metadata van de pagina. Hierin staan de persistente URLs van de content-bestanden (JPG, OCR en ALTO). Er is dus geen apart bestand voor de beschrijvende metadata beschikbaar.

De bijbehorende content-bestanden worden als volgt uit de identifier verkregen

Afbeeldingen

- *Hi-res JPG van het typoscript*
  http://resolver.kb.nl/resolve?urn=anp:1938:10:01:2:mpeg21 of
  http://resolver.kb.nl/resolve?urn=anp:1938:10:01:2:mpeg21:image
  Deze permanente URL verwijst naar een JPG bestand. In bovenstaand geval naar:
  http://resources51.kb.nl/anp/data/1938/1938_10/jpeg/anp_1938-10-01_2_access.jpg
  Door deze niet-persistente URL in een aanroep van de KB-imageserver te plakken, kan een
- *Thumb van de JPG* gemaakt worden (zoom=0.1 → 10% oorspronkelijke grootte)
  http://imageviewer.kb.nl/ImagingService/imagingService?id=http://resources51.kb.nl/anp/data/1938/1938_10/jpeg/anp_1938-10-01_2_access.jpg&zoom=0.1&useresolver=false

Teksten

- *OCR/TXT van typoscript (ongecorrigeerd)*
  http://resolver.kb.nl/resolve?urn=anp:1938:10:01:2:mpeg21:ocr
- *ALTO woordcoördinaten van typoscript*
  http://resolver.kb.nl/resolve?urn=anp:1938:10:01:2:mpeg21:alto

De metadata en digitale objecten in de set beslaan ongeveer 1 TB.

Metadata harvesten via de download-API (OAI-PMH)

De ANP-dataset heeft een download-API o.b.v. het OAI-PMH-protocol. Hiermee kun je zowel de metadata harvesten m.b.v. een OAI-harvester

De naam van de set binnen de KB OAI-repository is “anp” (met kleine letters)

1. De OAI-PMH-syntax is hoofdlettergevoelig!
2. Beschrijving KB OAI-repository : http://services.kb.nl/mdo/oai?verb=Identify
3. Voor het harvesten van de eerste 400 records metadata:
   http://services.kb.nl/mdo/oai?verb=ListRecords&set=anp&metadataPrefix=didl (let op: anp dus met kleine letters!)
4. *ResumptionToken:* met de laatste XML-tag in deze respons (de ResumptionToken) kan de volgende batch van 400 records opgevraagd worden, bijvoorbeeld:

http://services.kb.nl/mdo/oai?verb=ListRecords&resumptionToken=anp!2008-09-24T09:09:13.816Z!!didl!0 (metadataPrefix=didl hoeft dan niet meer in de query string aangegeven te worden.) Op die manier kan iteratief de hele set metadata overgehaald worden

5. *Identifiers:* met http://services.kb.nl/mdo/oai?verb=ListIdentifiers&set=anp&metadataPrefix=didl kan een lijst van de identifiers worden opgevraagd. Met de ResumptionToken onderaan deze respons kan de volgende batch identifiers worden opgevraagd.

Dataset doorzoeken via de zoek-API (SRU)

De ANP-dataset heeft ook een zoek-API o.b.v. het SRU-protocol. Hiermee kun je de set doorzoeken. . SRU is gebaseerd op CQL

De naam van de set binnen SRU is “ANP” (deze keer met hoofdletters!)

- De SRU-syntax is hoofdlettergevoelig!
- Informatie over de KB metadata and full-text index: http://jsru.kb.nl/sru/sru?operation=explain
- *Eenvoudige zoekvraag*: http://jsru.kb.nl/sru/ANP?&operation=searchRetrieve&version=1.1&query=hond&recordSchema=didl geeft de beschrijvingen terug van de eerste 20 typoscripten waar ‘hond’ in de full-text (ocr) voor komt.
- Als we die full-text (OCR) manifest in de respons willen zien, voegen we “&x-fields=content” aan de request toe en veranderen het recordSchema naar dc http://jsru.kb.nl/sru/ANP?version=1.1&operation=searchRetrieve&recordSchema=dc&query=hond&x-fields=content
- Als we i.p.v. de hele full-text alleen de abstract willen zien, laten we “&x-fields=content” weg: http://jsru.kb.nl/sru/ANP?version=1.1&operation=searchRetrieve&recordSchema=dc&query=hond
- *Maximum aantal records* : Standaard worden 20 resultaten teruggegeven. Als je er maximaal 50 wil opvragen, geef je maximumRecords=50 als argument mee: http://jsru.kb.nl/sru/ANP?&operation=searchRetrieve&version=1.1&query=hond&maximumRecords=50&recordSchema=didl
- De aanroep http://jsru.kb.nl/sru/ANP?&operation=searchRetrieve&version=1.1&query=hond&maximumRecords=1&recordSchema=didl geeft dus maar 1 hond-typoscript terug
- *Beginrecord*: Als je records 4 t/m 10 uit de hond-resultaatset wilt tonen, gebruik je de startRecords-parameter, samen met de maximumRecords parameter als volgt: http://jsru.kb.nl/sru/ANP?&operation=searchRetrieve&version=1.1&query=hond&startRecord=4&maximumRecords=7&recordSchema=didl
- *Samengestelde zoekvraag*: Als we willen zoeken op ‘man en vrouw’ gaat dit als volgt:

http://jsru.kb.nl/sru/ANP?&operation=searchRetrieve&version=1.1&query="man%20en%20vrouw"&recordSchema=dc&x-fields=content

- Zoek naar alle typoscripten voorgelezen tussen 1 en 4 januari 1960, en laat de OCR manifest zien in de respons. Let op de URL-encoding! http://jsru.kb.nl/sru/sru?version=1.2&maximumRecords=250&operation=searchRetrieve&startRecord=1&recordSchema=dc&x-collection=ANP&query=%28date%20within%20%2201-01-1960%2004-01-1960%22%29&x-fields=content (dit geeft 232 records terug)
- Zie ook de Snelstart ANP (pdf, in het Engels)
- Ook kan het nuttig zijn de Snelstart SGD (pdf) en de Snelstart EDBO (pdf) te raadplegen om a.d.h.v. andere open KB-sets binnen de KB-infrastructuur meer inzicht te krijgen hoe SRU-queries opgebouwd moeten worden.

Vragen of opmerkingen over de ANP-set?

Mail naar dataservices{at}kb.nl

Laatste update : 19-12-2012