Technical support for reusing the open data set *Radio bulletins from the ANP press agency*

For the open data set *ANP* (http://www.kb.nl/banners-apis-en-meer/dataservices-apis/anp-radiobulletins-digitaal) this document explains and illustrates the

1) conditions for reuse, attribution & licensing,
2) descriptive and structural metadata (mpeg21-didl),
3) images (scans of radio typoscripts in JPG-format) and full-text objects (OCR and ALTO in XML)
4) metadata harvest API based on OAI-PMH,
5) search API based on SRU,

1) Conditions for reuse, attribution & licensing

The Koninklijke Bibliotheek (KB) has released the **metadata** in this set under the CC0 1.0 Universal declaration.

The ANP agency has released the **objects** in this set (JPGs, OCR, ALTO) under the CC-BY-NC-ND 3.0 license

Please use the following format for attribution of the name, source and license

Algemeen Nederlands Persbureau & Koninklijke Bibliotheek.
Source: www.delpher.nl/radiobulletins – CC-BY-NC-ND

2) Metadata

Each typoscript is described by a number of descriptive and structural metadata fields. The descriptive metadata is formatted in Dublin Core (dc), the structural in mpeg21-didl. Using an example record, the table below shows each descriptive metadata field and its meaning.

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

(typoscript no 2 from 01-10-1938)

<dcx:volgnummer>	Typoscript number of the day: per day multiple news bulletins were read on the radio
<dcx:recordIdentifier>	Unique identifier of the typoscript Format: *anp:year:month:day:typoscriptnumber*
<dc:identifier>	Persistent URL to the scan (jpg) of the typoscript Format: http://resolver.kb.nl/resolve?urn=<*dcx:recordIdentifier*>:mpeg21
<dc:language>	Language of the typoscript : Dutch
<dc:rights>	Rights holder of the objects : ANP **But please note**: the ANP has released the objects (JPG, OCR, ALTO) under the CC-BY-NC-ND 3.0 license
<dcx:recordRights>	Rights holder of the metadata: Koninklijke Bibliotheek, Den Haag **But please note**: the KB has waived the copyright on the metadata; making it available under the CC0 1.0 Universal declaration
<dc:title>	Title of the typoscript : Format : *ANP Nieuwsbericht – day-month-year – typoscript number of that day*
<dc:date>	Date of radio broadcast Format : *year-month-day*

The descriptive metadata are embedded as an XML-block at the beginning of each MPEG21-DIDL structural metadata file. In other words: there are no separate XML-files for the descriptive metadata. For every <*dcx:recordIdentifier*> there are persistent URLs to the image, alto and ocr files

- <didl:Resource ref="http://resolver.kb.nl/resolve?urn=<*dcx:recordIdentifier* >:image" mimeType="image/jpeg"/>
- <didl:Resource ref="http://resolver.kb.nl/resolve?urn=<*dcx:recordIdentifier* >:alto" mimeType="text/xml"/>
- <didl:Resource ref="http://resolver.kb.nl/resolve?urn=<*dcx:recordIdentifier* >:ocr" mimeType="text/xml"/>

3) Objects

The metadata of the typoscript with identifier *anp:1938:10:01:2:mpeg21* can be requested via
http://services.kb.nl/mdo/oai?verb=GetRecord&identifier=anp:anp:1938:10:01:2:mpeg21&metadataPrefix=didl

This mpeg21-didl file first has a block of descriptive metadata, according to the table above. It is followed by a block of structural metadata, containing the persistent URLs of the object files (JPG, OCR en ALTO). There are no separate XML-files for the descriptive metadata

The object files associated with the identifier *anp:1938:10:01:2:mpeg21* are obtained as follows

Image files

- *Hi-res JPG of the typoscript*
  http://resolver.kb.nl/resolve?urn=anp:1938:10:01:2:mpeg21:image
  This persistent URL points to a JPG file, in this case to
  http://resources51.kb.nl/anp/data/1938/1938_10/jpeg/anp_1938-10-01_2_access.jpg
  By copy-pasting this non-persistent URL into a request to the KB imaging service, a
- *thumbnail* of the JPG can be obtained (zoom=0.1 → 10% of original size)
  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

Text files

- *OCR/TXT of the typoscript (uncorrected for OCR errors)*
  http://resolver.kb.nl/resolve?urn=anp:1938:10:01:2:mpeg21:ocr
- *ALTO word coordinates* (for highlighting purposes)
  http://resolver.kb.nl/resolve?urn=anp:1938:10:01:2:mpeg21:alto

The metadata and digital objects in the set are appox. 1 TB in size

4) Harvesting metadata using the download API (OAI-PMH)

The ANP dataset has a download API based on the OAI-PMH protocol. OAI-PMH stands for "Open Archives Initiative Protocol for metadata Harvesting". It is typically used for incrementally harvesting entire collections bunch by bunch.

In general it should be avoided to do harvesting via a web browser; especially for large metadata files it's more efficient to use dedicated OAI harvesting software (like this one)

1. The name of the ANP set within the KB OAI-repository is "anp" (so in lower case)
2. The OAI-PMH syntax is case sensitive.
3. For most efficient data transfer harvesting activities should be performed outside Dutch office hours (9 am -17pm).
4. Description of the KB OAI-repository : http://services.kb.nl/mdo/oai?verb=Identify
5. You can browse the records in the anp set http://services.kb.nl/mdo/browse?set=anp
6. Harvesting the first 400 mpeg21-didl records: http://services.kb.nl/mdo/oai?verb=ListRecords&set=anp&metadataPrefix=didl ("anp" in lower case!). By default 400 records are harvested by one request
7. *ResumptionToken:* using the last XML-tag in this response (the so-called ResumptionToken) the following batch of 400 records can be downloaded, e.g. http://services.kb.nl/mdo/oai?verb=ListRecords&resumptionToken=anp!2008-09-24T09:09:13.816Z!!didl!0 (*metadataPrefix=didl* can be omitted from the query string). Iteratively the whole set of metadata can be transferred in this way.
8. *Identifiers:* with http://services.kb.nl/mdo/oai?verb=ListIdentifiers&set=anp&metadataPrefix=didl a list of identifiers can be requested. By iteratively using the ResumptionToken at the end of the response the full set of identifiers can be obtained.
9. With the "from" parameter you can add a datestamp (format *YYYY-MM-DDTHH:MM:SS.msZ*) to selectively harvest records which were added to the collection since a particular date. Also an "until" parameter may be provided, for example: http://services.kb.nl/mdo/oai?verb=ListIdentifiers&set=anp&metadataPrefix=didl&from=2008-10-03T11:11:**44.576Z**&until=2008-10-03T11:**11:45.000Z**
10. An individual record can be obtained by using a GetRecord command: http://services.kb.nl/mdo/oai?verb=GetRecord&identifier=anp:anp:1937:10:01:2:mpeg21&metadataPrefix=didl

5) Searching the dataset using the search API (SRU)

- The ANP dataset has a search API based on the SRU protocol. SRU stands for "Search and retrieval via URLs". More information on this standard can be found at http://www.loc.gov/standards/sru/
- JSRU is a Java implementation of the SRU protocol at the KB. SRU is based on CQL
- Top level information about the KB metadata and full-text index: http://jsru.kb.nl/sru/sru?version=1.2&operation=explain
- Within the KB data infrastructure the name of the ANP search index (*x-collection*) is "ANP" (with capitals, as the SRU-syntax is case sensitive)

Simple queries

- *Simple query in the ANP index for* "hond" = Dutch for "dog"): http://jsru.kb.nl/sru?&operation=searchRetrieve&version=1.2&x-collection=ANP&query=hond&recordSchema=didl returns the descriptions of the first 20 typoscripts with 'hond' occurring in the full-text (ocr).
  - The field <srw:numberOfRecords> shows that are in total 903 such records
  - The recordShema 'didl' used here manifests the persistent URLs to the image, alto and ocr files. With "recordSchema=dcx" less information is retrieved
- If we want to see the full-text (OCR) manifested in the response, we add "&x-fields=content" to the request and change the recordSchema to 'dc' http://jsru.kb.nl/sru?version=1.2&operation=searchRetrieve&recordSchema=dc&x-collection=ANP&query=hond&x-fields=content
- If we want to see the *volgnummer* and *recordRights* manifested in the response as well, we add "&x-fields=content,volgnummer,recordRights" http://jsru.kb.nl/sru?version=1.2&operation=searchRetrieve&recordSchema=dc&x-collection=ANP&query=hond&x-fields=content,volgnummer,recordRights
- By removing "&x-fields=content" from the request, the abstract instead of the full-text is displayed http://jsru.kb.nl/sru?version=1.2&operation=searchRetrieve&recordSchema=dc&query=hond&x-collection=ANP

Number of records returned

- *Maximum number of records* : By default 20 results are given back. If for instance you want to request 53 records, you need to add maximumRecords=53 to the request: http://jsru.kb.nl/sru?&operation=searchRetrieve&version=1.2&query=hond&maximumRecords=53&recordSchema=didl&x-collection=ANP
- Thus the request http://jsru.kb.nl/sru?&operation=searchRetrieve&version=1.2&query=hond&maximumRecords=1&recordSchema=didl&x-collection=ANP only returns 1 typoscript description with 'hond' in the full-text

- *Start record*: If you want to request records 4 to 10 from the “hond” result set, you combine the startRecords and maximumRecords parameters: http://jsru.kb.nl/sru?&operation=searchRetrieve&version=1.2&query=hond&startRecord=4&maximumRecords=7&recordSchema=dc&x-collection=ANP

Combined queries

- *Combined query*: If we want to look for ‘man en vrouw’ (meaning “man and wife”) in the full-text, this can be done via the request http://jsru.kb.nl/sru?&operation=searchRetrieve&version=1.2&query="man%20en%20vrouw"&recordSchema=dc&x-fields=content&x-collection=ANP. In total there are 153 of these typoscripts

- If we want to search all typoscripts that were read on the radio between 1st en 4th January 1960, and show the OCR manifestly, the request looks like this 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
  Please pay attention to the encoding of the CQL-statement *date within "01-01-1960 04-01-1960"*

Facets

- The ANP-index supports a *periode* facet as well http://jsru.kb.nl/sru/sru?query=hond&version=1.2&startRecord=1&operation=searchRetrieve&maximumRecords=10&x-collection=ANP&x-fields=volgnummer&**x-facetprefix=1&x-facetname=periode&x-facets=indexes:ANPfacets:periode**
  The ‘facetprefix’ can take values 0,1,2 or 3, resulting in different temporal resolutions of the facet (0=decade, 1= year, 2= month, 3=day)

- The *period* facet can be filtered as well. Say you want facets with a resolutions of a month (facetprefix=2), but you are only interested in the decade 1950-1959, the query would be http://jsru.kb.nl/sru/sru?query=hond&version=1.2&startRecord=1&operation=searchRetrieve&maximumRecords=10&x-collection=ANP&x-fields=volgnummer&x-facetprefix=2&x-facetname=periode&x-facets=indexes:ANPfacets:periode**&x-filter=%28periode%20exact%20%220/1950-1959/%22%29**
  Please pay attention to the encoding of the CQL-statement *(periode exact "0/1950-1959/")*

Questions or remarks about the ANPset / APIs?

Please contact us via dataservices{at}kb.nl

Last update : 21-11-2014