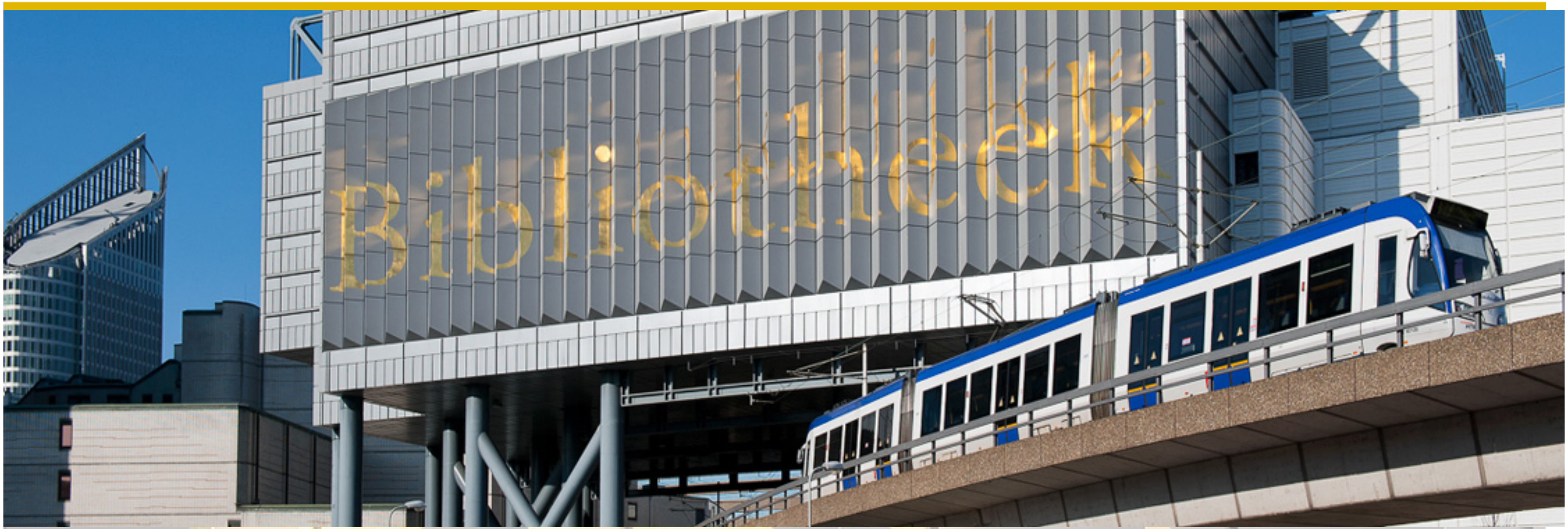
KB Koninklijke Bibliotheek
National Library of the Netherlands



Introduction to the KB APIs

Juliette Lonij, 27 Oct 2016

www.kb.nl

Why APIs?

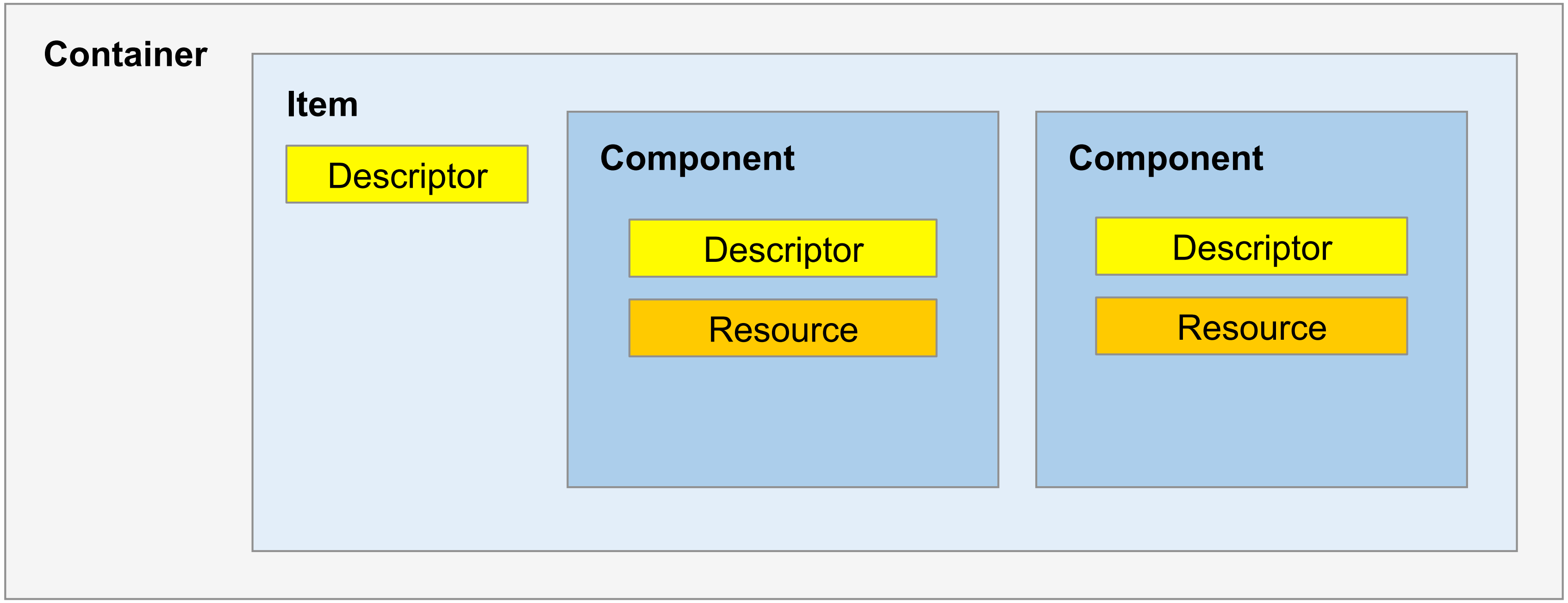
- KB websites
  - Browse, search, view, download via Graphical User Interface (GUI)
- KB APIs
  - Search, download underlying data via Application Programming Interface (API)

Introduction to the KB APIs

- Types of files and metadata associated with digital objects
- Overview of KB data access infrastructure
- Getting files through the resolver using persistent identifiers
- Querying a collection with the search API
- Collecting metadata with the harvest API

What data?

- Structural metadata in MPEG21 DIDL format
- Descriptive metadata in the Dublin Core (DC) vocabulary, with some KB-specific extensions (DCX)
- ALTO files with content and layout information
- OCR files containing only the textual content of the item
- One or more images of the object, usually in JPEG format



Container

Item

Descriptor

Component

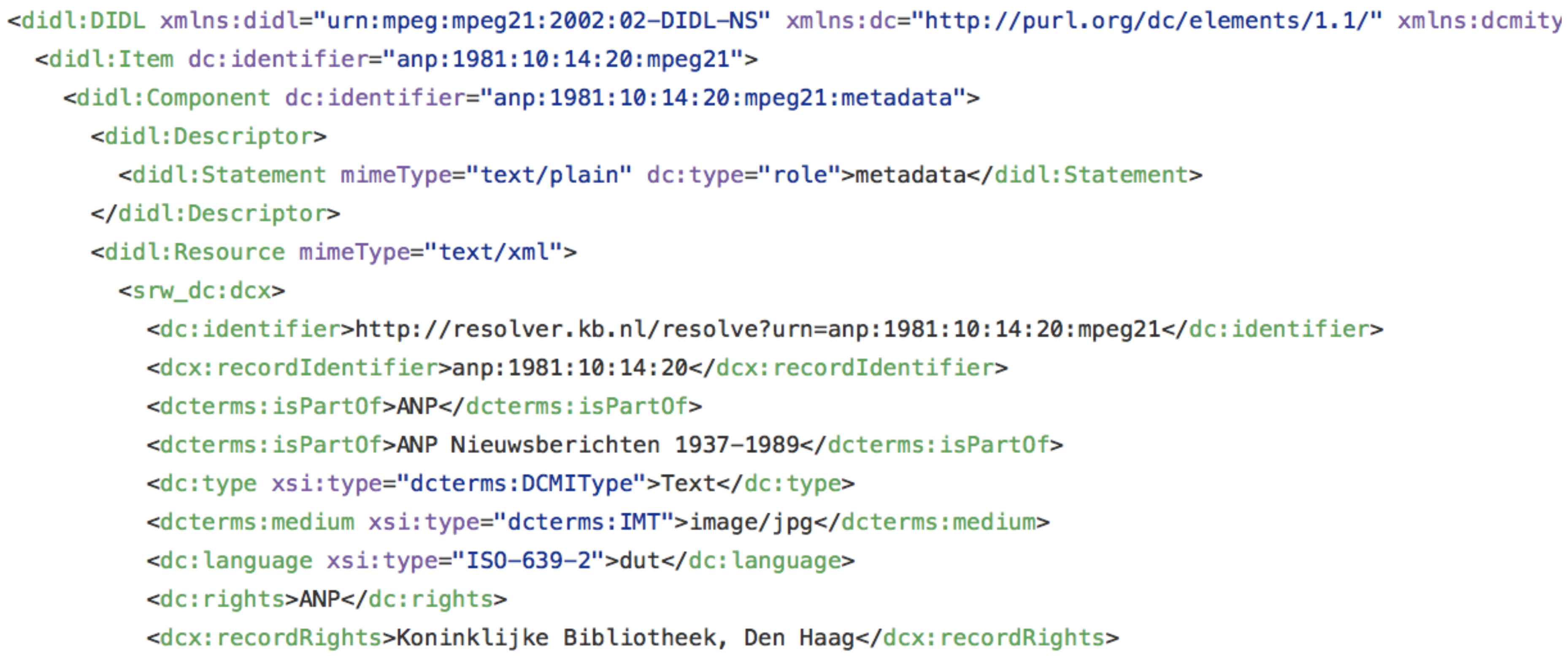
Descriptor

Resource

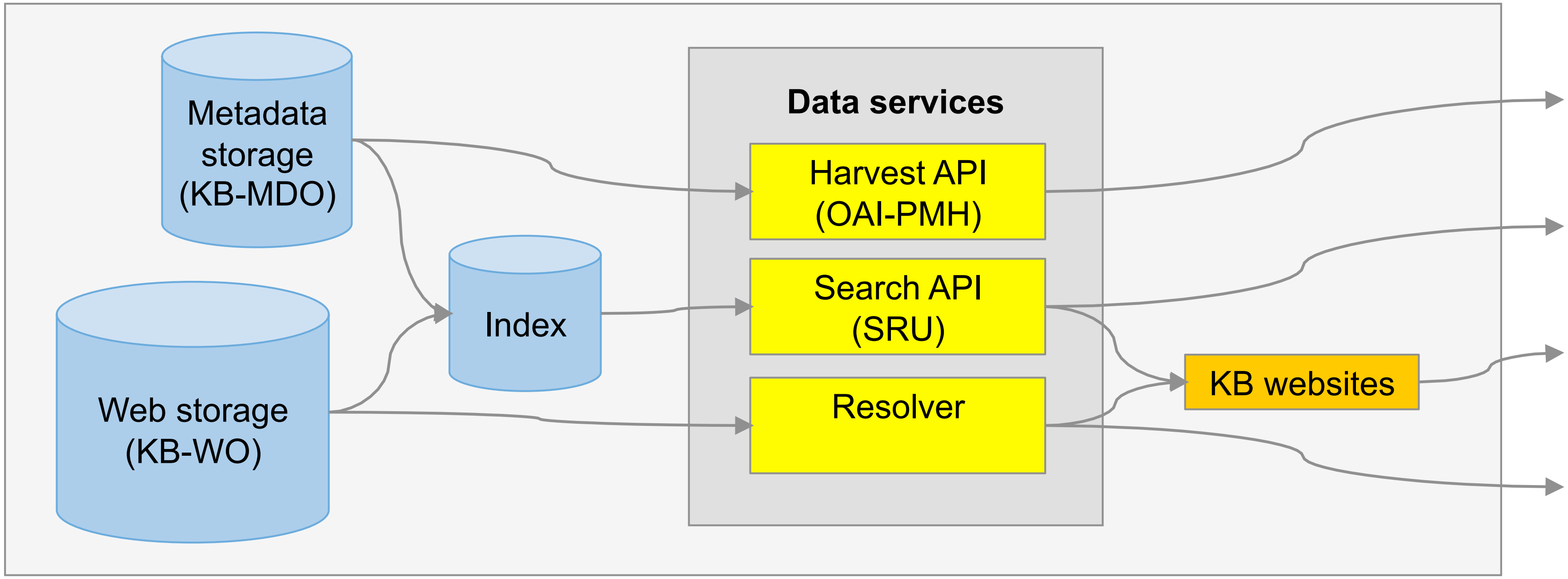
Component

Descriptor

Resource



```
<didl:DIDL xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS" xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcmity
  <didl:Item dc:identifier="anp:1981:10:14:20:mpeg21">
    <didl:Component dc:identifier="anp:1981:10:14:20:mpeg21:metadata">
      <didl:Descriptor>
        <didl:Statement mimeType="text/plain" dc:type="role">metadata</didl:Statement>
      </didl:Descriptor>
      <didl:Resource mimeType="text/xml">
        <srw_dc:dcx>
          <dc:identifier>http://resolver.kb.nl/resolve?urn=anp:1981:10:14:20:mpeg21</dc:identifier>
          <dcx:recordIdentifier>anp:1981:10:14:20</dcx:recordIdentifier>
          <dcterms:isPartOf>ANP</dcterms:isPartOf>
          <dcterms:isPartOf>ANP Nieuwsberichten 1937-1989</dcterms:isPartOf>
          <dc:type xsi:type="dcterms:DCMIType">Text</dc:type>
          <dcterms:medium xsi:type="dcterms:IMT">image/jpg</dcterms:medium>
          <dc:language xsi:type="ISO-639-2">dut</dc:language>
          <dc:rights>ANP</dc:rights>
          <dcx:recordRights>Koninklijke Bibliotheek, Den Haag</dcx:recordRights>
```


Metadata storage (KB-MDO)
Web storage (KB-WO)
Index
Data services
Harvest API (OAI-PMH)
Search API (SRU)
Resolver
KB websites

An example data set: ANP Radio Bulletins

- About 1.5 million digitized typoscripts from radio news broadcasts between 1937 and 1984
- Available through the Delpher website as Radiobulletins collection
- KB offers the data under (semi) open licenses:
    - CC0-license for the metadata
    - CC-BY-NC-ND-licenses for images and full-text objects

Getting files through the resolver

- Persistent identifier in the form of a resolver link:
  http://resolver.kb.nl/resolve?urn=anp:1973:10:18:44:mpeg21
- Image can be retrieved by adding the **:image** suffix:
  http://resolver.kb.nl/resolve?urn=anp:1973:10:18:44:mpeg21:image
- OCR can be retrieved by adding the **:ocr** suffix:
  http://resolver.kb.nl/resolve?urn=anp:1973:10:18:44:mpeg21:ocr
- ALTO file can be retrieved using the **:alto** suffix:
  http://resolver.kb.nl/resolve?urn=anp:1973:10:18:44:mpeg21:alto

Exercise 1: Getting files through the resolver

- Find the full text of this tutorial on GitHub: https://github.com/jlonij/intro-kb-apis
- Navigate to the Delpher website with your browser and search the collection of radio bulletins for a typoscript of your choice. Find the unique identifier associated with this typoscript.
- Open a new tab or window in your browser and retrieve the image, OCR and ALTO files associated with this typoscript. Inspect the results to get an idea of the content and structure of the data that was returned.

The metadata harvest API

- Based on the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) protocol
- Harvesting metadata of a single record or an entire set
- Administrative information to keep copy up to date

Example OAI-PMH request

- http://services.kb.nl/mdo/oai?verb=GetRecord &identifier=anp:anp:1981:10:14:20:mpeg21&metadataPrefix=didl
- Base URL followed by query string with a number of parameters:
  - Parameter **verb** with value **GetRecord** to specify the operation to be performed
  - Parameter **identifier** with the unique object identifier **anp:anp: 1981:10:14:20:mpeg21**
  - Parameter **metadataPrefix** to select the required metadata format **didl**

Exercise 2: Getting started with OAI-PMH

- Open a new window or tab in your browser and navigate to the harvest API base URL. Now construct the parameters for retrieving the DIDL structural metadata record for the typoscript of your choice.
- Inspect the data returned. In particular, identify the different blocks or components that are included in the response, giving an overview of all data associated with this particular typoscript and the ways in which to obtain them.

OAI-PMH verbs

- **GetRecord** will retrieve a single metadata record from the repository
- **Identify** will retrieve general information about a repository
- **ListSets** will retrieve the names of the data sets in the repository
- **ListMetadataFormats** will retrieve the metadata formats available
- **ListRecords** will retrieve all metadata records from a specified set
- **ListIdentifiers** will retrieve only the identifiers from a specified set

Harvesting an entire set

- ListIdentifiers request to harvest first 400 identifiers for the set ANP: http://services.kb.nl/mdo/oai?verb=ListIdentifiers&set=anp&metadataPrefix=didl
- Resumption token can be inserted with subsequent requests: http://services.kb.nl/mdo/oai?verb=ListIdentifiers&resumptionToken=anp!2008-09-24T09:09:16.332Z!!didl!2317275
- Resulting list of identifiers can now be used to retrieve the full metadata record for each item by issuing GetRecord requests for each identifier

Exercise 3: Harvesting a set with OAI-PMH

- Go back to the browser tab or window containing the GetRecord request from the previous exercise. Adjust the query string to get some general information about the KB-MDO repository, using verbs such as Identify, ListSets and ListMetadataFormats.

- Now enter the URL to harvest the first 400 identifiers from the ANP set. Find the resumption token at the end of the response and use it to request the next 400 identifiers. Repeat this step a few of times to get an idea of how a piece of software could automatically harvest the identifiers for an entire set by means of this mechanism.

The search API

- If you do not wish to retrieve an entire set, but are more interested in getting results for a particular search query
- Based on the SRU (Search and Retrieval via URL) standard protocol
- Uses CQL (Contextual Query Language) as its query language

Example SRU request

- http://jsru.kb.nl/sru/sru?operation=searchRetrieve&x-collection=ANP&query=nobelprijs
- Base URL followed by query string with a number of parameters:
  - Parameter **operation** with value **searchRetrieve** to specify the operation
  - Parameter **x-collection** with value **ANP** to select the collection to search
  - Parameter **query** to enter the search query **nobelprijs**

Other SRU parameters

- With the **operation** parameter set to **explain**, some general information about the service can be obtained
- Adjust the number of results returned with the **maximumRecords** parameter
- Use the **startRecord** parameter if you want to start viewing the result set from a particular record onwards
- If you want to add particular bits of information to the display, you can use the **x-fields** parameter

Exercise 4: Querying the index with SRU

- Open a new window or tab in your browser and navigate to the search API base URL. Now construct the parameters for querying the ANP collection with a single keyword of your choice. Find the total number of results for your query.
- Use the maximumRecords, startRecord and x-fields parameters to expand and navigate through the results.

CQL query syntax

- Keyword string: query="nobelprijs literatuur”
- OR- or AND-queries: query=nobelprijs AND literatuur
- Wildcards: query=nobelpr*
- Restrict to a particular field: query=date=01-01-1960
- Restrict to period in time: query=date within "01-01-1960 01-01-1961”

URL encoding

- Certain characters, such as spaces, may not appear as part of an URL
- Replace them with code: a space should be replaced by %20 or +, for example, and a double quotation mark with %22
- Modern browsers do this automatically
- More information and encoding service at W3Schools URL Encoding Reference: http://www.w3schools.com/tags/ref_urlencode.asp

Faceting with SRU

- http://jsru.kb.nl/sru/sru?operation=searchRetrieve &x-collection=ANP&query=nobelprijs&maximumRecords=0 &x-facetprefix=1&x-facetname=periode &x-facets=indexes:ANPfacets:periode
- The **maximumRecords** parameter has been set to **0**, so that only the facetted results are shown
- The **x-facetprefix** parameter can take values from 0 to 3, resulting in different temporal resolutions of the facet

Exercise 5: Advanced SRU options

- Experiment with the CQL query language to form more elaborate queries. Search for a string consisting of multiple keywords, for example, and filter the results by selecting a particular date range.

- Note the URL encoding that your browser applies to the query (or, if your browser does not offer this functionality, the URL encoding that is returned by the URL Encoding Reference).

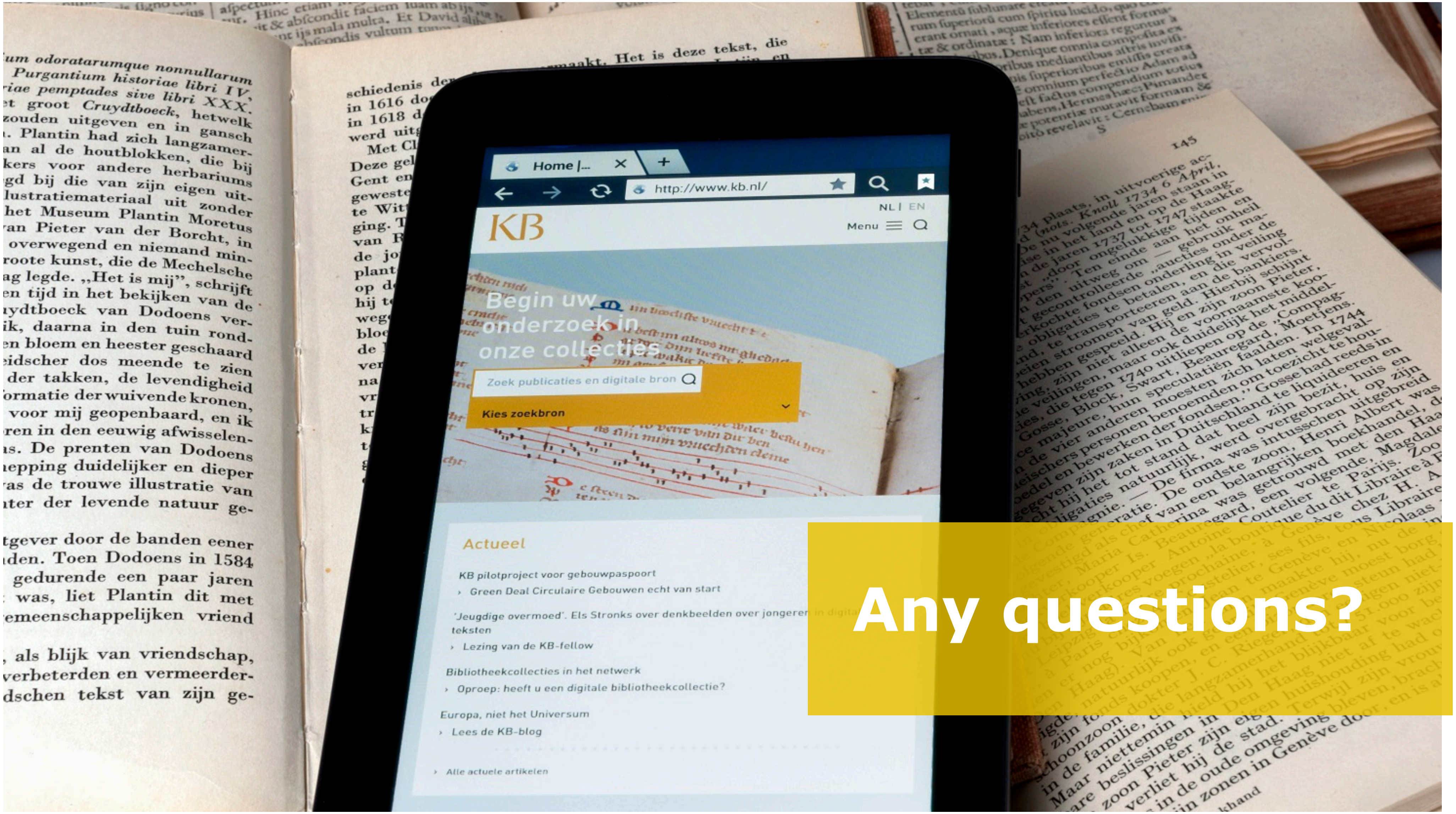
- Create a faceted view of a result set of your choice by using the x-facets, x-facetname and x-facetprefix parameters.

Other data sets

- Data available may vary somewhat across collections, but the use of the APIs will be the same
- Data form other collections will be more complex if objects with a hierarchical structure, such as a newspaper issue comprising multiple pages and articles, are involved
- Detailed technical instructions for a number of other sets available at the Data Services page of the KB website: https://www.kb.nl/en/resources-research-guides/data-services-apis

Exercise 6: Moving on to other sets

- Visit the Data Services page of the KB website at https://www.kb.nl/en/resources-research-guides/data-services-apis and take a look at the open datasets that are available.
- Open the technical instructions for a dataset of your choice and try it out with the search and harvest APIs.


Home |...
http://www.kb.nl/
NL | EN
KB
Menu
Begin uw onderzoek in onze collecties
Zoek publicaties en digitale bron
Kies zoekbron
Actueel
KB pilotproject voor gebouwpaspoort
Green Deal Circulaire Gebouwen echt van start
'Jeugdige overmoed'. Els Stronks over denkbeelden over jongeren in digitale teksten
Lezing van de KB-fellow
Bibliotheekcollecties in het netwerk
Oproep: heeft u een digitale bibliotheekcollectie?
Europa, niet het Universum
Lees de KB-blog
Alle actuele artikelen

Any questions?