1 Million Dutch Newspaper Images available for Researchers: the KBK-1M Dataset

Martijn Kleppe
National Library of the Netherlands (KB)
Desmond Elliott
University of Amsterdam (UvA)

1,6 million images * Dutch National Newspapers * 1922-1994 *
Photographs, drawings, cartoons & weather forecasts * Download as zip per year * Available for researchers

Problem

- Visualisation of news has exploded
- **But**: Methods to collect & analyse large amount of visual data are labour-intensive
- Delpher.nl allows researchers to find 'Images with illustration'
- **But**: Researchers need to find, select and download each image manually

Solution

- KBK-1M dataset
- Containing **1.603.395 captioned images**
- Extracted from Dutch digitised newspapers stored in the Dutch National Library (KB) Newspaper archive
- Period 1922-1994.

Delpher

Creating the Dataset

We created a Harvester that…:

- …used existing Metadata**:** 1) location of each image, 2) caption, 3) the article on that page.
- … was built with Python programming language
- … prepared and extracted the images & captions using KB-internal RESTful APIs.

Set-up of the dataset

Zipfiles that correspond with each year and contain:

- JPEGs of each image
- JSON files containing information about the image
- 1 Excel file containing all information of each individual JSON file.

De Burmese minister van buitenlandse zaken vertoeft te Djakarta met enige leden van zijn staf; hij werd op Kemajoran verwelkomd door Indonesische autoriteiten en de Burmese ambassadeur. (Ipphos) j

Listing 1: JSON Metadata Format


```
{
    "caption": "\n\t\n\tDe Burmese minister van
        buitenlandse zaken vertoeft te Djakarta
        met enige leden van zijn staf; hij werd
        op Kemajoran verwelkomd door
        Indonesische autoriteiten en de Burmese
        ambassadeur. (Ipphos) j\n",
    "paper_title": "De nieuwsgier",
    "page": 1,
    "content_block_url": "http://imageviewer.kb.
        nl/ImagingService/imagingService?id=ddd
        :010474896:mpeg21:p001:image&colour=
        fefe56&s=0&x=2463&y=4545&w=1589&h=790",
    "text_block": "ddd:010474896:mpeg21:a0018",
    "content_block": "P1_CB00003",
    "date": "1951/01/27 00:00:00",
    "image_name": "1951/DDD:ddd:010474896:mpeg21/
        p001-P1_CB00003.jpg",
    "jp2_url": "http://imageviewer.kb.nl/
        ImagingService/imagingService?id=ddd
        :010474896:mpeg21:p001:image",
    "alto_url": "http://resolver.kb.nl/resolve?
        urn=ddd:010474896:mpeg21:a0018:ocr"
}
```


Application Dataset

- **Humanities**: 1) (photo) graphic style changes, 2) the representation of people, societal issues and concepts and 3) build new tools for exploring photographic reuse via image-similarity based search
- **Computer Vision & Natural Language Processing** : 1) Automatic Image Captioning and multimodal ranking using machine learning techniques & 2) Data-to-text Generation

Future Work

- Seperate photographs, drawings, cartoons & weather forecasts
- Include images of recently released newspapers
- Similarlike datasets of advertisements & family notices

Obtaining the Dataset

The dataset is a**vailable for scientific or scholarly purposes.** Requests for use and download of the dataset can be addressed to dataservices@kb.nl
More info at lab.kbresearch.nl/get/downloads

Current Situation:
Find & select each *individual* image *manually*

KBK-1M Dataset:
Download *all* images, captions & links to authentic lay-out *at once*

Koninklijke Bibliotheek
National Library of the Netherlands