P3 SMK af Camilla Ejby Pedersen og Sara Liv Broe

1. **Beskriv (med referencer) hvorfor institutionen har en API tilgang til deres datasæt:**

* API står for ”Application Programming Interface” og er linket mellem ens forespørgsel og det svar man får tilbage. SMK bruger altså bl.a. API til at brugerne kan gå ind og finde kunsten og hente den ned på computeren. Uden API’en ville det ikke være muligt for brugerne at sende en request på f.eks. ”Picasso” og få svar tilbage med de værker som museet har liggende (<https://www.youtube.com/watch?v=s7wmiS2mSXY>).

**Hvad er det blevet brugt til?**

* Det er bl.a. blevet brugt af Codrin-Mihail, som er en canadisk designer. Han har taget nogle 3D-scanninger og leget rundt med dem. Givet dem et andet udtryk og på den måde visualiseret de idéer og tanker han havde (<https://www.behance.net/gallery/81900161/METAMORPHOSE-FUTURISTIC-SCULPTURES?fbclid=IwAR0B-pLFVi7wV_6wdQPfV87ggH09AeIpm3uaaSzl5SaKuDOfwi8JpsqUJhs>).

**Hvad tænker de at det kunne bruges til?**

* Man kan finde data på alle digitaliserede kunstværker og fotografier (<https://www.smk.dk/article/smk-api/>).
* Når man besøger SMK, så ser man kun 0,7% af samlingen. Dette vil SMK gøre op med. De vil sørge for at folk kan se kunsten, lege med den og forme den, som de har lyst til. Der skal ikke være en begrænsning på kunst (<https://www.smk.dk/article/smk-open/>).

1. **Sammensæt en URL der søger efter kulturinstitutionens elementer (fx. genstande, kunstværker, personer, avisartikler, osv.) i APIen (gerne 100+ stk) med output i JSON.**

* ﻿[https://api.smk.dk/api/v1/art/search/?keys={}&offset={}&rows={}](https://api.smk.dk/api/v1/art/search/?keys=%7b%7d&offset=%7b%7d&rows=%7b%7d)
* Keys er søgeordet f.eks. Picasso
* Offset er hvor langt inde i listen man starten
* Rows er hvor mange resultater man vil have returneret

Så vi kan f.eks. skrive Picasso i ”keys”, 0 i ”offset” og 200 i ”rows”. Da vi blot skal have over 100, så tænker vi at 200 er et passende antal til formålet.

***Load denne URL i en JSON beautifier og beskriv indholdet:***

https://api.smk.dk/api/v1/art/search/?keys=Picasso&offset=0&rows=200

**Hvilke typer metadata er der for hvert element?**

* Metadataen beskriver kunstværket, såsom f.eks. højde og bredde hvis det er et maleri, samt hvilket materiale der er brugt såsom ”lærred” og ”olie”. Endvidere står der hvilken udstilling som maleriet hører under eller har hørt under. Der står hvem der kreereret værket, samt værkets navn og kunstnerens nationalitet og hvilket årstal at kunstværket er fra. Endvidere har hvert element sit eget ID, hvilket giver god mening hvis man hurtigt skal kunne finde det pågældende element - eller kunstværk.

**Sammenlign metadata med Dublin Core. Er der overensstemmelse?**

* Der er 15 kerneelementer i Dublin Core og det er med udgangspunkt i disse at vi sammenligner metadataen.
  + Contributor - ”an entity responsible for making contributions to the resource.
    - Ingen “contributor”
  + **Coverage** – “the spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.”
    - Ingen “coverage”
  + **Creator** – “An entity primarily responsible for making the resource”
    - ”Creator” stemmer godt overens med DC. Her får vi navnet på personen der har lavet værket, i denne forbindelse, Picasso. Der står hvornår han blev født og hvornår han døde, samt hans nationalitet.
  + **Date** – “A point or period of time associated with an event in the lifecycle of the resource.”
    - “Date” står ikke alene og er ikke kun relateret til kunstværket. ”Date” relaterer sig også til udstillinger og kunstneren.
  + **Description** – “An account of the resource.”
    - “Description” i forbindelse med vores elementer, beskriver f.eks. om der er et vandmærke – og beskriver således noget omkring værket.
  + **Format** – “The file format, physical medium, or dimensions of the resource.”
    - Intet der hedder format i denne forstand. Vi har ”dimensions” som bl.a. beskriver højde og bredde.
  + **Identifier** – “An unambiguous reference to the resource within a given context.”
    - Ingen identifer
  + **Language** – “A language of the resource.”
    - Language beskriver sproget og der står også om værket er oversat til andre sprog end det oprindelige.
  + **Publisher** – “An entity responsible for making the resource available.”
    - Ingen “publisher”.
  + **Relation** – “A related resource.”
    - Ingen “relation”
  + **Rights** – “Information about rights held in and over the resource.”
    - Under “rights” står der hvem der har rettighederne til værket. Dog er dette open-source og alle billeder, kunstværker mv. er dermed ”public domain”.
  + **Source** – “A related resource from which the described resource is derived.”
    - Ingen “source”
  + **Subject** – “The topic of the resource.”
    - Ingen subject
  + **Title** – “A name given to the resource”
    - Værkets navn eller navne
  + **Type** – “The nature or genre of the resource.”
    - Vi har bade “type” som relaterer sig til værket, men også til kunstneren. Der står f.eks. ved Picasso ”type: artist”, eller ”type: museum”. Der står dog også ”type: image”. Så det virker til at ”type” bliver brugt i en bredere sammenhæng i vores datasæt.

De kerneelementer som vi ikke har fundet, betyder ikke nødvendigvis at de ikke findes i vores datasæt, men det kan være at de hedder noget andet. Vi har taget vores URL, sat den i en browser og åbnet den. Herefter har vi søgt efter kerneelementerne. Vi så det f.eks. med ”format” og ”dimension”. Vi har ikke noget der hedder ”format” i vores data, men vi har ”dimensions” som dækker over det samme.

**Er der nogle tal der kan bruges til statistiske beregninger (fx. årstal)?**

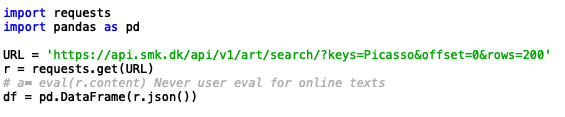
* Ja, der står både produktionsåret og hvornår kunstneren er født og hvornår han døde. Der står årstal på udstillingerne og der står mål på de forskellige værker. Man vil derfor f.eks. kunne finde ud af hvor gammel Picasso var da han var mest produktiv (i forhold til de samlinger som SMK har liggende).

**Hvilke slags spørgsmål kunne besvares gennem analyse af dette datasæt? (gerne med eksempler):**

* Hvad var Picassos foretrukne materiale (eller hvad brugte han mest)?
* Hvor mange malerier har Picasso malet?
* I hvilket år var Picasso mest produktiv?
* Osv.

1. **Ved hjælp af Pandas og request, importere nu jeres valgte datasæt.**

***Beskriv og rens datasættet.***



Vi prøvede først at importere datasættet ved at skrive denne kode, men vi fik meddelsen: ValueError: arrays must all be same length

Vi har fundet frem til, at det er på grund af, at JSON datasættet indeholder items, hvor hvert item er et værk. Dog indeholder det også bl.a. rows of offset. Vi tror så, at det der er sket er, at den har talt rows som en række, offset som en række og items som en række, i stedet for kun at hente listen med items. Derfor tilføjede vi noget mere til koden:

A close up of a piece of paper

Description automatically generated

Vi har her specificeret, at det er items som skal hentes. For at sikre os at det virkede valgte vi så at bede om at få at vide hvor mange items der var.

**Hvilke data og datatyper er der?**

* Vi gik ind på siden: <https://api.smk.dk/api/v1/docs> og klikkede på ”art search” hvorefter vi søgte på Picasso i ”keywords”. Under ”responses” klikkede vi på model, hvorefter vi kunne se alle de forskellige datatyper der var.
  + String f.eks. ved ”id” og ”current location name”
  + Boolean f.eks. ved frame (har værket en ramme eller ej)
* Mht. hvilke data der, så er det beskrevet, overordnet, højere oppe.

Dette er overraskende, at der kun er er ”strings” og ”booleans” da vi også havde regnet med, at der ville være ”floats”.

**Hvor mange kolonner og rækker?**

* Ved at kalde ”head”, kan vi se at vi har 49 kolonner. Vi har 200 rækker, da vi selv har sat den til dette, men hvis vi f.eks. sætter den til 2000 i stedet for, så får vi 234 rækker tilbage.

**Er der kolonner I ikke skal bruge?**

* Da vi skal sidde og analysere datasættet, så har vi ikke brug for kolonnerne med billeder i. Vi skal kunne beregne på den data vi har udvalgt og vi har derfor valgt at fjerne stort set alle kolonner på nær ”production date”. Vi har sat ”ID” som vores indeks, da det giver mening at kunne søge direkte på nummeret, fremfor at benytte det autogeneret indeks fra vores dataframe.

1. **Lav nu udtræk og beregn på det data I har udvalgt. Hvilke informationer kan det give os om de udvalgte elementer?**

* Det kan f.eks. give os information om i hvilket år at Picasso var mest produktiv og i hvilket år at han var mindst produktiv.

1. **Brug value\_counts() (**[**https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.value\_counts.html**](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.value_counts.html)**) funktionen til at producere en dataframe over det udvalgte data, optælling og procent.**

* Se koden.

1. **Visualisér det udvalgte data med procent i en graf**.

Antal afsluttede værker i procent

A screenshot of a cell phone

Description automatically generatedÅr