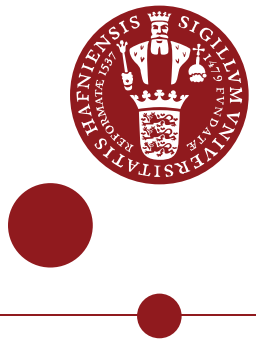
KØBENHAVNS UNIVERSITET



Portfolie 3

Open Data Sciene

Martine Ingemann Jørgensen, JPR328

Natacha Rylander Bech, TGZ940

Stephanie Rose Acampado Soelmark, PZG932

2 december 2019

Martine Ingemann Jørgensen, JPR328
Natacha Rylander Bech, TGZ940
Stephanie Rose Acampado Soelmark, PZG932
02/12/2019

Opgave 1

Beskrivelse af SMK Open

Vores valgte institution er Statens Museum for Kunst (SMK), da denne på nuværende tidspunkt fører et projekt der hedder SMK Open (2016-2020). Dette går ud på "...at stille hele Danmarks kunstsamling til fri afbenyttelse." (SMK Open, u.å.)) ved at digitalisere og tilgængeliggøre museets samling. SMK's formål er at alle danskere skal kunne benytte sig af kunsten og anvende den i "...sit eget liv og bruge på sine egne vilkår" (SMK Open, u.å.). Dermed har SMK tilgængeliggjort alt deres data ved at lade interessenter benytte SMK's egne API (SMK's API (beta-version), u.å.).
API (Application Programming Interface) er en tjeneste der muliggør at software kan tale med andet software. Det er med til at digitale tjenester kan udgive og offentliggøre deres data for andre interessenter på en struktureret måde (SMK's API (beta-version), u.å.).
SMK forestiller sig eksempelvis at deres data kan; tilgås uafhængigt af tid og rum; viderebearbejdes; nærstuderes i detaljer; deles; indsættes i alt fra bøger til forskningsartikler og skoleopgaver; og trykkes på alt fra plakater til sofapuder (SMK Open, u.å.).

Opgave 2

Generering af URL

Vi har genereret vores URL, der søger efter statuetter i SMK's digitale database, som er følgende:
https://api.smk.dk/api/v1/art/search?keys=statuette&rows=1000&encoding=json
Denne har vi indsat i JSON Beautifier, for at kunne beskrive indholdet.

Typer af metadata

Vi har identificeret forskellige slags metadata, som vi har kategoriseret i forhold til deskriptiv, administrativ og strukturel metadata.
De deskriptive metadata, bruges til at kunne identificere og beskrive indholdet på et overordnet plan, uden at gå i dybden med at forklare og vurdere (Digital Bevaring, u.å.).

De administrative metadata repræsenterer informationer om tid/dato, digitalisering og tekniske informationer og rettigheder (Digital Bevaring, u.å.).
De strukturelle metadata bruges til henholdsvist at vise og navigere samt kan det også være information om den interne organisering eller en rækkefølge af dokumenter (Digital Bevaring, u.å.).

Med henblik på førstnævnte, deskriptiv data, har vi eksempelvis observeret benævnelsen af årstal på værket: 'period', materiale: 'material', teknik: 'technique', kunstner: 'creator', fødselsår: 'creator_date_of_birth', dødsår: 'creator_date_of_birth', nationalitet: 'creator_nationality', titel: 'titles', noter: 'frame_notes' og 'content_notes', farver: 'colors', samt beskrivelse: 'content_description'.
Dernæst har vi i relation til administrativ metadata fundet id-numre: 'id', referencenumre: 'object number', digitaliseringsdage: 'created', rettigheder: 'public_domain', 'copyright', afdeling på museet: 'responsibledepartment', ændringsdato: 'modified', udstilling: 'exhibition', placering: 'shelfmark'.
Slutteligt har vi udpeget de strukturelle metadata i vores dataframe, bestående af antal dele: 'parts', kollektion: 'collection', dimensioner: 'dimensions' og hvad kunstværket er en del af 'parts of'.

Sammenligning af metadata med Dublin Core

Nedenstående er en oversigt over 15 metadata kerneelementer defineret af Dublin Core, som vi har sammenlignet med vores SMK datasæt og dermed kommet med eksempler på. Følgende 15 kerneelementer er hentet fra Dublin Cores hjemmeside (Dublin Core, 2019):

Contributor – "An entity responsible for making contributions to the resource."

- På trods af at vi har kendskab til at SMK som organisation står for at tilbyde billeder, står dette ikke beskrevet i deres metadata.

Coverage – "The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant."

- Spatial: Vi har fundet lokationen på værket i form at. 'shelfmark : 60875'.
- Temporal: I metadataen kan vi finde både digitaliseringsdag: 'created', ændringsdato: 'modified', start og slutdato på perioden: 'start', 'end', 'period', samt årstal på kunstværket: 'acquisition_date_precision'.
- Jurisdiction: Selve afdelingen som værket hører ind under, er også nævnt som 'responsible_department'.

Creator – "An entity primarily responsible for making the resource."

- Kunstneren er nævnt under produktionsdetaljer om værket som 'creator'.

Date – "A point or period of time associated with an event in the lifecycle of the resource."

- Perioden' som startdato: 'start' og slutdato: 'end' på perioden kan man finde i metadataen under objektet 'production_date'.

Description – "An account of the resource."

- Der findes diverse beskrivelser af værket i metadaten, eksempelvis: noter til indholdet: 'notes', indholdsbeskrivelse: 'content_description', noter til opsætningen: 'frame_notes', dimensioner: 'dimensions', og dokumentation: 'documentation'.

Format – "The file format, physical medium, or dimensions of the resource."

- SMK har inkluderet dimensionerne på deres værker, der indbefatter højde og centimeter under 'dimensions'.

Identifier – "An unambiguous reference to the resource within a given context."

- Her har SMK inkluderet ID numre: 'ID', referencenumre: 'object number', lokation: 'shelfmark' og 'current_location_name'.

Language – "A language of the resource."

- Derudover er forskellige sprogversioner af værkernes titler inkluderet i datasættet under elementet 'titles'.

Publisher – "An entity responsible for making the resource available."

- Vi fandt et element i metadataen under "notes", der står beskrevet som forskellige kataloger og hæfter - hvilket muligvis kan være dem, der er ansvarlige for at gøre statuetterne tilgængelige/dem der har givet SMK disse værker.

Relation – "A related resource."

- Der er metadata omkring relaterede værker i form af en note: 'notes' omkring eksempelvis kataloger, som værket indgår i, derudover står der også information om samlingen.

Rights – "Information about rights held in and over the resource."

- Vi fandt metadata, der angiver rettighederne omkring værket. Herunder er det synliggjort, hvilken form for kreditering værket er registreret med. Eksempelvis ift. copyright kreditering: 'public domæne', og rettigheder: "rights", hvor værdien kan være creative commons.

Source – "A related resource from which the described resource is derived."

- Kilden er beskrevet ved nogle af dem som "source", hvor de har tilhørende værdier som "THL", eller "Eva De La Fuente Pedersen".
- Under notering: 'notes' der står også ved specifikke værker, hvorfra de er erhvervet. Eksempel fra 'notes' er: 'Erhvervet af Herbert Melbye den 24. juni 1942 på auktion hos Winkel & Magnussen. kat. 54 (for DKK 517,50). Bruun Rasmussens bogfortegnelse nr. 35.'

Subject – "The topic of the resource."

- Kollektionen, udstillingen som værket er del af, indenunder 'content_description', får man nogle nøgleord og sætninger omkring værket, 'text' der beskriver værket.

Title – "A name given to the resource."

- Titlen på værket er ligeledes til stedet samt forskellige sproglige versioner af den, under 'title'.

Type – "The nature or genre of the resource."

- Dette er beskrevet under "titles", og herunder "type", hvoraf det kan være "Museum".

Statistiske beregninger

I dataen er værdier man kan anvende til statiske beregninger. Man kan blandt andet regne på årstal herunder digitaliseringsdato og produktionsdato. Med disse kan man eksempelvis lave en statistik over antal værker produceret i et bestemt år, eller en oversigt over værker digitaliseret på bestemte datoer. En anden værdi er periode, denne kan man eksempelvis anvende til at undersøge hyppigheden af perioder på SMK. Derudover kan man lave statistik over teknikker og materialer, eksempelvis hvor hyppigt et bestemt materiale eller en bestemt teknik er anvendt. Kunstnere i datasættet er også relevante at anvende, her kan man undersøge, hvor mange værker en bestemt kunstner har udstillet på SMK, eller hvilke materialer en bestemt kunstner anvender sig af hyppigst.

Eksempler på relevante spørgsmål til analyse af omtalte datasæt

- I hvilket årstal producerede man flest statuetter? (i vores sample bestående af 1000 værker)
    - Hvad er frekvensdistribueringen af statuetterne inden for hvert årstal?

- Hvor mange unikke kunstnere har skabt disse statuetter og med hvilket materialer?

- Hvordan er fordelingen af kunstnere der stadig lever og ikke lever?

- Hvordan er fordelingen af kunstnernes nationaliteter i vores datasæt? og hvilken nationalitet er hyppigst?

- Hvilken periode er den mest hyppigste inden for skulpturering?

Opgave 3

Importering, beskrivelse og rensning af datasæt

Vores datasæt består af 537 rækker og 49 kolonner før vi har renset det. Herunder er der adskillige kolonner, hvis værdier betegnes som 'NaN' not a number, da der mangler data - eksempelvis ved de kolonner, der starter med **'**image_'. Efter rensningen af dataframen har vi nu 537 rækker og 26 kolonner. Det vil sige at der er lige så mange statuetter som før, men med (49-26) = 23 færre kolonner (elementer/typer data). Hovedparten af de kolonner vi har fjernet, er dem, hvis værdier står som 'NaN', såsom billeder eller billedinformationer. Vi vil gerne fokusere på kolonner, der er mere konkrete og kan være behjælpelige i statistiske analyser, eksempelvis; kunstner, produktionsår, nationalitet, materialer.

I vores rensede datasæt er datatyperne 'objects', 'floats' og 'booleans'. Der 537 rækker, samt 26 kolonner. Vi har fravalgt visse kolonner da disse ikke er relevante for os, eksempler er**;** 'iiif_manifest', 'object_history_note', 'image_native' og 'image_thumbnail'.

Se bilag 1 for tabeller over rensning af datasættet.

Opgave 4

Beregning af udvalgt data

Vi kan lave en optælling af årstallene for at undersøge, hvornår der blev fremstillet flest statuetter. Samme metode kan benyttes med henblik på typer, hvor vi kunne observere at kategorier såsom statuetter, tegninger, skulpturer indgik i vores dataframe. Vi har specificeret statuette som nøgleord i vores søgning, hvilket ikke udelukker andre former for værker.

Endvidere har vi foretaget samme slags optælling af materialer og teknikker, således vi kunne skabe et overblik over de hyppigste anvendte.
Vi har udtrukket alle 'NaN' værdierne for at danne overblik over manglende værdier og udregnet summen for hver kolonne. Således kan vi se, at der forekommer adskillige celler i dataframen, hvor angivelse af tilhørende værdi mangler.
Vi har udregnet hvor mange af værkerne, der står som 'public domain', og fundet frem til at det er 245 ud af 537 værker. Således er der copyright rettigheder på hovedparten af værkerne i SMKs database.

Opgave 5

Producering af dataframe over udvalgt data

Vi har valgt at fokusere på acquisition_date_precision, da det viser årstallet for hvornår værket antageligvis blev overleveret til en kulturinstitution og/eller arkiveret
Først har vi lavet en optælling af disse årstal og fremstillet dem i procent. Dernæst har vi lavet en dataframe bestående af årstal i en kolonne og procentvise optællinger i en anden.

Opgave 6

Visualisering af data

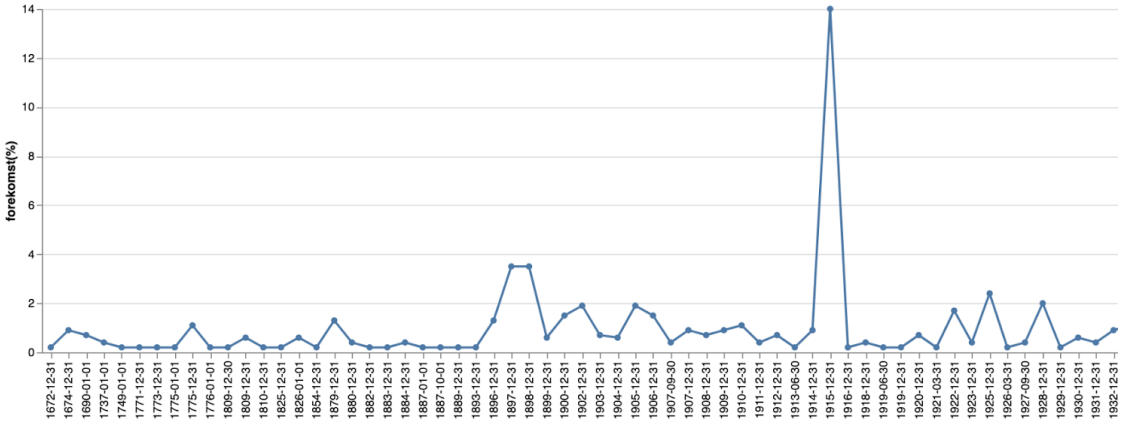
Fra altair biblioteket har vi hentet funktionen, der kan visualisere vores dataframe. Vi har indsat dataframen og angivet år og forekomst som x og y i grafen.

Nedenstående er et udsnit af vores endelige graf. Vi har beskåret denne til fordel for det visuelle, da den fulde graf er mindre overskuelig. Den fulde graf kan findes i bilag 2.

Martine Ingemann Jørgensen, JPR328
Natacha Rylander Bech, TGZ940
Stephanie Rose Acampado Soelmark, PZG932

02/12/2019

[67]:



Litteraturliste

Digitalbevaring.dk (u.å.). Metadata - Hvad er metadata, og hvorfor er de vigtige for digital bevaring?. Lokaliseret d. 2. december 2019 på: https://digitalbevaring.dk/viden/metadata/

Dublin Core Metadata Initiative (2019). DCMI Metadata Terms. Lokaliseret d. 29. november på: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

SMK (u.å.). SMK's API (beta-version). Lokaliseret d. 28. november 2019 på: https://www.smk.dk/article/smk-api/

SMK (u.å.). SMK Open. Lokaliseret d. 28. november 2019 på: https://www.smk.dk/article/smk-open/

Martine Ingemann Jørgensen, JPR328
Natacha Rylander Bech, TGZ940
Stephanie Rose Acampado Soelmark, PZG932

02/12/2019

Bilag

Alle bilag er lavet i JupiterLab.

Bilag 1

Rensning af datasættet

Før rens af datasættet

	id	created	modified	acquisition_date_precision	responsible_department	content_description	frame_notes	materials	object_names	part_of	...	image_iiif_id
0	1180035377_object	2019-08-07T05:10:34Z	2019-08-12T09:33:32Z	1898-12-31	Samling og Forskning (KAS)	[Erstattet med KAS2229]	[Baglædning: false, Mikroklimaramme: false]	[{'material': 'gips'}]	[{'name': 'statuette'}]	[KAS2229, ORIG3096]	...	NaN
1	1180070230_object	2019-08-07T07:24:30Z	2019-08-08T08:29:02Z	1976-12-31	Samling og Forskning (KMS)	[Puys er et fiskerleje i nærheden af Dieppe.]	[Baglædning: false, Mikroklimaramme: false]	[{'material': 'bronze'}]	[{'name': 'skulptur'}]	NaN	...	NaN
2	1180032239_object	2019-08-07T05:01:38Z	2019-08-12T09:32:26Z	1915-12-31	Samling og Forskning (KAS)	NaN	[Baglædning: false, Mikroklimaramme: false]	[{'material': 'gips'}]	[{'name': 'statuette'}]	[ORIG3101]	...	NaN
3	1180081678_object	2019-08-07T08:11:07Z	2019-08-12T09:50:58Z	1915-12-31	Samling og Forskning (KAS)	NaN	[Baglædning: false, Mikroklimaramme: false]	[{'material': 'gips'}]	[{'name': 'statuette'}]	[ORIG3097]	...	NaN
4	1180014421_object	2019-08-07T04:07:35Z	2019-08-08T08:33:53Z	1925-12-31	Samling og Forskning (KAS)	NaN	[Baglædning: false, Mikroklimaramme: false]	[{'material': 'gips'}]	[{'name': 'statuette'}]	[ORIG3104]	...	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...
532	1180058626_object	2019-08-07T06:37:21Z	2019-08-08T07:06:23Z	1968-12-30	Samling og Forskning (KKS)	NaN	[Baglædning: false, Mikroklimaramme: false]	NaN	[{'name': 'tegning'}]	NaN	...	NaN
533	1180017629_object	2019-08-07T04:17:22Z	2019-08-08T08:33:58Z	1970-12-31	Samling og Forskning (KKS)	NaN	[Baglædning: false, Mikroklimaramme: false]	NaN	[{'name': 'tegning'}]	NaN	...	NaN
534	1180060426_object	2019-08-07T06:44:35Z	2019-08-08T11:22:07Z	1968-12-30	Samling og Forskning (KKS)	NaN	[Baglædning: false, Mikroklimaramme: false]	NaN	[{'name': 'tegning'}]	NaN	...	NaN
535	1180004533_object	2019-08-07T03:37:36Z	2019-08-12T09:22:53Z	1690-01-01	Samling og Forskning (KMS)	NaN	[Baglædning: false, Mikroklimaramme: false]	[{'material': 'alabast'}]	[{'name': 'friskulptur'}]	NaN	...	NaN
536	1180022529_object	2019-08-07T04:31:43Z	2019-08-12T09:29:04Z	1672-12-31	Samling og Forskning (KMS)	NaN	[Baglædning: true, Mikroklimaramme: false]	[{'material': 'lærred'}, {'material': 'olie'}]	[{'name': 'maleri'}]	[KMS3076]	...	https://iip.smk.dk/iiif/jp2/KMS3075.tif.recons...

537 rows × 49 columns

Efter rens af datasættet

```
df.head()
```

	id	created	modified	acquisition_date_precision	responsible_department	content_description	frame_notes	materials	object_names	part_of	...	object_number	public_domain
0	1180035377_object	2019-08-07T05:10:34Z	2019-08-12T09:33:32Z	1898-12-31	Samling og Forskning (KAS)	[Erstattet med KAS2229]	[Baglædning: false, Mikroklimaramme: false]	[{'material': 'gips'}]	[{'name': 'statuette'}]	[KAS2229, ORIG3096]	...	KAS384	False
1	1180070230_object	2019-08-07T07:24:30Z	2019-08-08T08:29:02Z	1976-12-31	Samling og Forskning (KMS)	[Puys er et fiskerleje i nærheden af Dieppe.]	[Baglædning: false, Mikroklimaramme: false]	[{'material': 'bronze'}]	[{'name': 'skulptur'}]	NaN	...	KMShm3	True
2	1180032239_object	2019-08-07T05:01:38Z	2019-08-12T09:32:26Z	1915-12-31	Samling og Forskning (KAS)	NaN	[Baglædning: false, Mikroklimaramme: false]	[{'material': 'gips'}]	[{'name': 'statuette'}]	[ORIG3101]	...	KAS1910	False
3	1180081678_object	2019-08-07T08:11:07Z	2019-08-12T09:50:58Z	1915-12-31	Samling og Forskning (KAS)	NaN	[Baglædning: false, Mikroklimaramme: false]	[{'material': 'gips'}]	[{'name': 'statuette'}]	[ORIG3097]	...	KAS1906	True
4	1180014421_object	2019-08-07T04:07:35Z	2019-08-08T08:33:53Z	1925-12-31	Samling og Forskning (KAS)	NaN	[Baglædning: false, Mikroklimaramme: false]	[{'material': 'gips'}]	[{'name': 'statuette'}]	[ORIG3104]	...	KAS2049	False

5 rows × 26 columns

Bilag 2

Visualisering af den udvalgte data med procent i en graf.