UiO : Universitetet i Oslo

Styringsgruppemøte for prosjektene "FAIR@UiO" og "Spesialsamlinger og databaser"

Agenda

- Hvordan henger disse prosjektene sammen?
- Status siden sist for hvert prosjekt
- Veien videre
- Annet
- Evaluering av møtet
- Neste møte

Bakgrunn og begrunnelse

UiOs interne utlysning av forskningsinfrastrukturmidler

Behov for en koordinering på tvers.

Plattform-basert IT-utvikling og anskaffelser - «grunnfjell» (tungvekts-IT), spesialtilpassede løsninger (lettvekts-IT) på toppen

Målbilde / leveranser

"Spesialsamlinger og databaser":

Å etablere en/flere digital(e) plattform(er) for å lagre, katalogisere og tilgjengeliggjøre UiOs spesialsamlinger, faglige arkiv, lokalt utviklede databaser, tekster og/eller audiovisuelt materiale

"FAIR@UiO":

Å etablere en IT-plattform som muliggjør det å gjøre data FAIR på en enkel måte ved UiO

Kost / nytte

- Verdien i data går tapt
- Neste prosjekt får det enklere
- Varighet og kontinuitet
- Krav fra samfunn og finansiører
- Gjenbruk (særlig maskinlæring og AI)

- Først ?? == Innovasjon == Utfordringer

SPESIALSAMLINGER OG DATABASER

Prosjektets leveranser og avgrensning

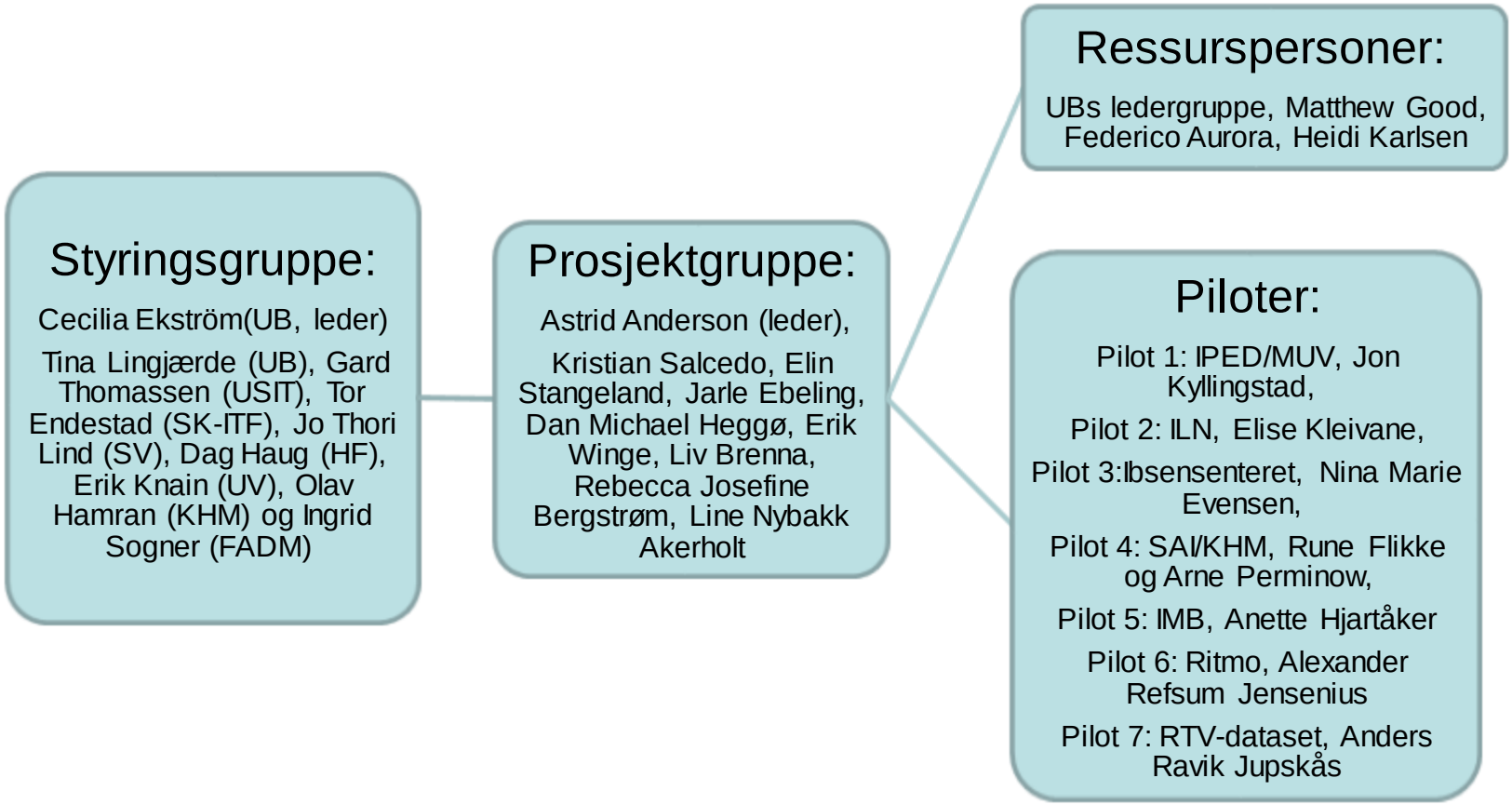
Viktigste leveranser:

- Digital(e) plattform(er)
- Ivaretagelse av behov hos fagmiljøene
- Etablert arbeidsflyt for aktuelle prosjekter
- Hub/node-organisering av infrastrukturarbeidet
- Kompetanse hos UB

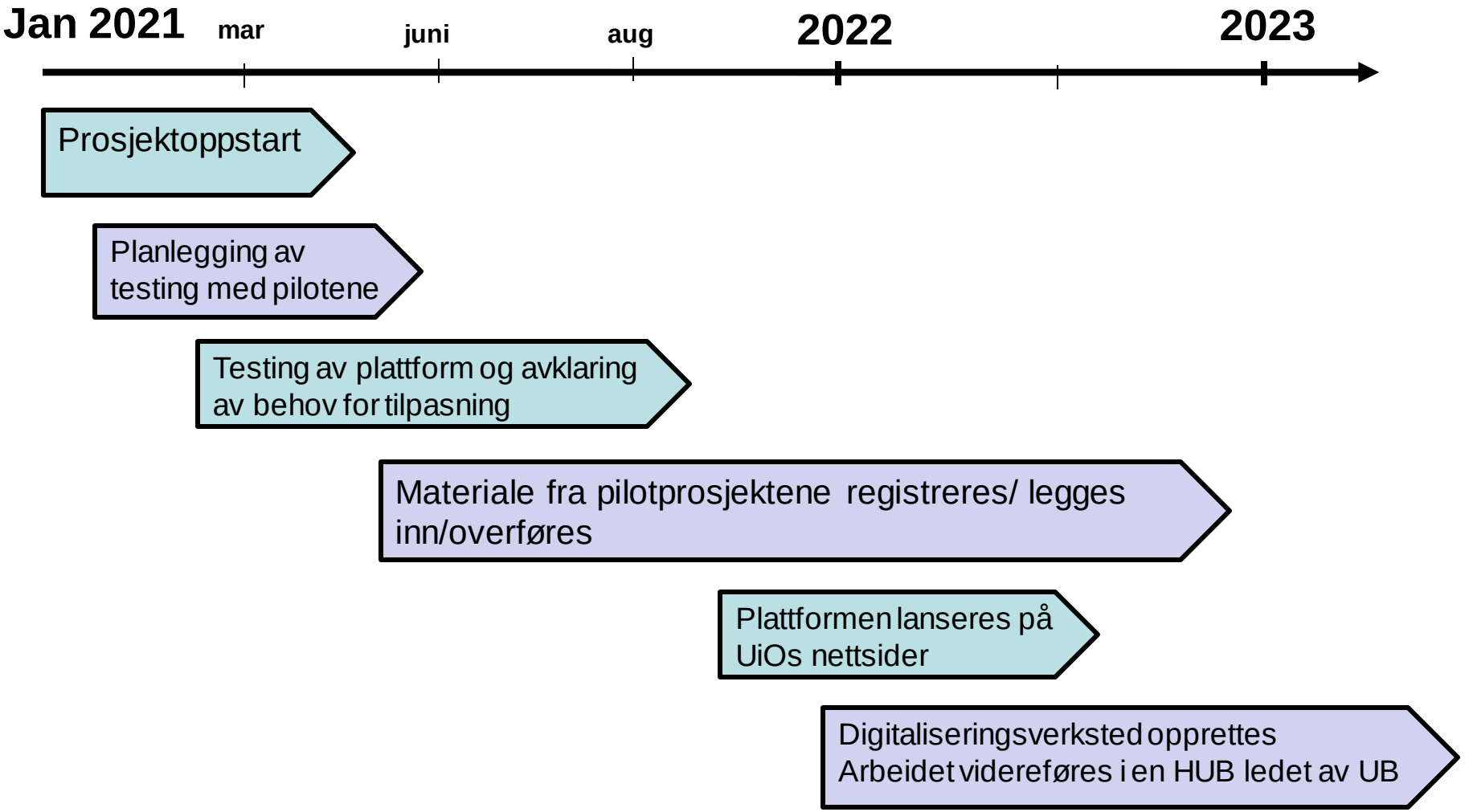
Avgrensning:

- Prosjektet skal ikke legge til rette for fysisk ivaretagelse av spesialsamlinger og arkiver.
- I videreføringen av plattformen skal arbeid med nytt/oppdatert innhold i plattformen gjøres og finansieres prosjektbasert av enhetene som har behovet

Organisering og ansvar



Overordnet tidsplan



Rekrutteringer og ressursbruk

- **Universitetsbiblioteket:** 100 % fast overingeniør finansiert av prosjektmidlene i prosjektperioden er ansatt. Lars Jynge Alvik er historiker og digital arkivar og kommer fra en stilling hos arkivverket i september
- **HumSam-biblioteket:** metadatabibliotekar i full stilling i 2 år, innstilling ligger hos ansettelsesrådet

- **Piloter**:
  - ILN får midler til vit.ass. i 9 måneder og 2 månedsverk til utvikling hos Tekstlab
  - EMMA: Maria Kartveit frikjøpes til å jobbe med bildesamlingene (usikkert hvor mange månedsverk enda)

Status og utfordringer

- Pilotprosjektenes behov og ønsker er avklart i møter med prosjektgruppa
- Prosjektgruppa har hatt kontakt med Marcus om potensielt samarbeid
  - Kristian og Dan Michael har deltatt i workshop om Goobi, et verktøy for arbeidsflyt i digitaliseringsprosjekter som kan bidra til å sikre kvalitet i metadata
- Prosjektgruppa har hatt flere møter med Alvin for å se på muligheter og begrensninger, også sammen med representanter fra pilotprosjektene
- Prosjektgruppa tester i Alvin i april/mai
- Pilotene med behov som ikke kan løses i en felles plattform har hatt individuelle møter med prosjektleder og USIT for å få ivaretatt disse behovene på annet vis.

Det som ikke løses i felles plattform så langt

- **Ritmo**:

Løsning for videovisning lages i vortex i samarbeid med USIT.

- **ILN**:

Fortsetter å samarbeidet med Tekstlab med en databaseløsning, finansiert av prosjektets midler. På sikt kan tekstene fra prosjektet eventuelt også tilgjengeliggjøres i felles plattform.

- **Ibsen-senteret:**

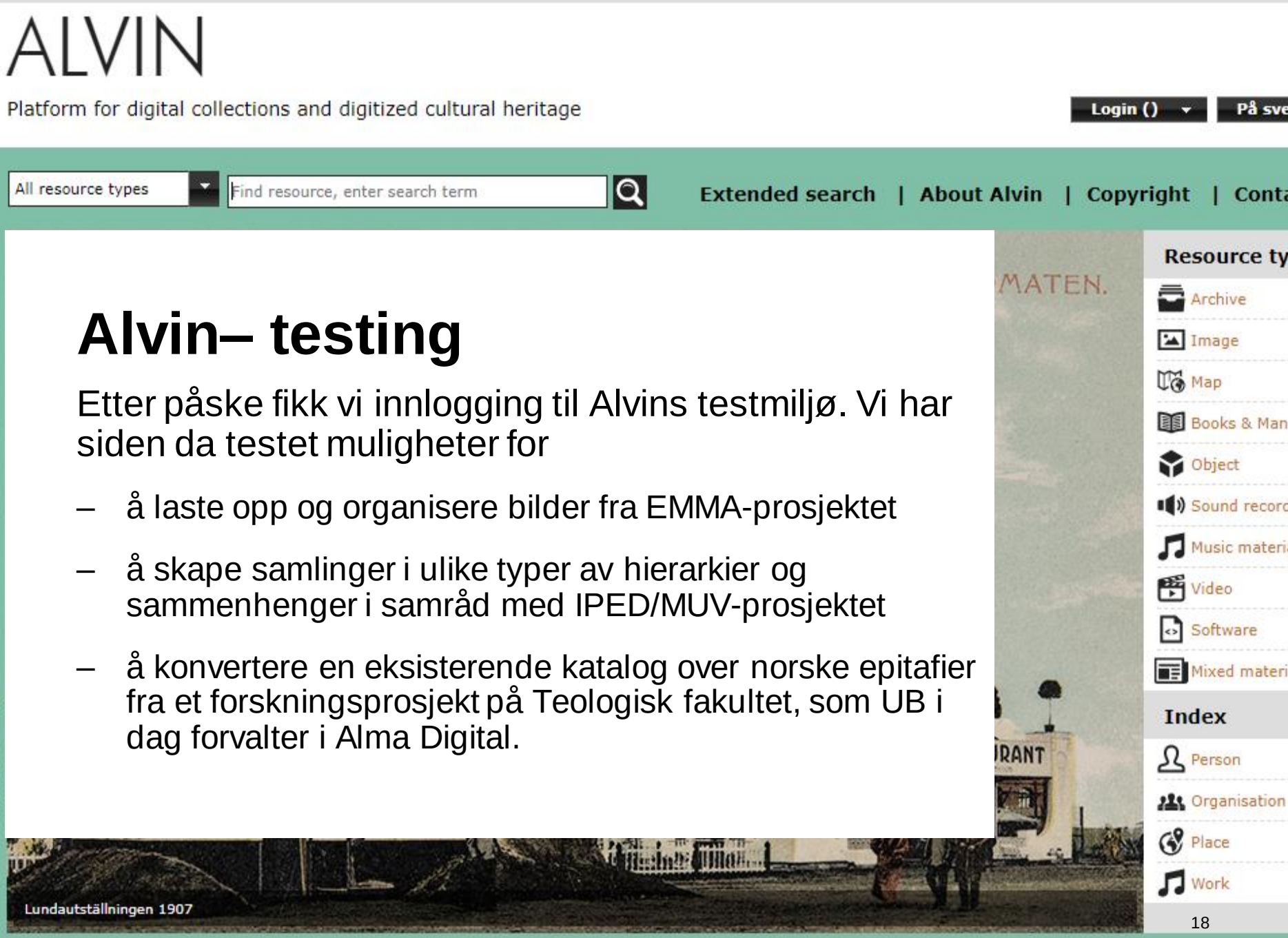
Venter svar på NFR-søknad i desember. USIT og Ibsen-senteret tar opp igjen tråden over sommeren. Forslag om å jobbe med UX-teamet om å se på design av den digitale Ibsen-plattformen.

- **RTV-datasettet:**

Samarbeider med USIT om lagrings- og visningsmulighetene. Tar senere kontakt med prosjektet med andre behov i genren «spesialsamlinger»

- **Nutrifoodcalc:**

Fortsetter samarbeidet med USIT



Alvin– testing

Etter påske fikk vi innlogging til Alvins testmiljø. Vi har siden da testet muligheter for

- å laste opp og organisere bilder fra EMMA-prosjektet
- å skape samlinger i ulike typer av hierarkier og sammenhenger i samråd med IPED/MUV-prosjektet
- å konvertere en eksisterende katalog over norske epitafier fra et forskningsprosjekt på Teologisk fakultet, som UB i dag forvalter i Alma Digital.

Alvin - erfaringer

Musiclab (Ritmo)

Norske epitafier 1537–1700

- Får representert dataene i katalogen bedre enn i Alma Digital. Begrensningen i Alvin ligger først og fremst i at hvert epitafium representeres som én ressurs med ett sett metadata, men egentlig er flere (selve epitafiet har sine metadata, transkripsjonen har sine metadata, fotografiene har sine, osv.).

Fredrik Barths bilder fra feltarbeid

- Gode muligheter for nødvendige metadata, men begrensninger i mulighetene for å vise bilder samlet i Alvin

Testkit, IPED

- Gode muligheter til å uttrykke samlinger/sammenhenger. Ellers samme begrensninger i bildevisning som over.

Ibsen-senteret

- TEI-filer fra Ibsen-senteret kan transformeres til MODS for import, men vil da miste så mye at Alvin i praksis ikke kan erstatte frontend-løsningene Ibsen-senteret har i dag. Kan være den kan fylle et behov for stabil langtidslagring av mediefiler på sikt når IIIF-støtte kommer på plass.

Alvin – eksempler fra andre institusjoner

- University of Yale har brukt en bildegjenkjenningsapp på en samling bilder fra Alvin: http://dh.library.yale.edu/projects/bagge/

- Universitetet i Lunds prosjekt "Witnessing Genocide" baserer seg på Ravensbruck-samlingen i Alvin: https://www.ub.lu.se/hitta/digitala-samlingar/witnessing-genocide

Alvin – begrensninger i dag

- Per i dag er det begrensede APIer for å hente ut data og ingen APIer for å importere og redigere data.
- Metadata registreres etter standarder utviklet i bibliotek-verdenen, primært MODS, og mangler mulighet til å tilføye egne metadatafelt og sette opp samlingsspesifikke skjemaer.
- Tilbyr ikke løsninger for navigasjon i, og presentasjon av, mer komplekse data enn det som kan representeres i MODS som f.eks. TEI-kodete transkriberte tekster, annotert med kommentarer og krysslenker innad i teksten, bibliografier og andre databaser. Kan kanskje beskrives og arkiveres (langtidslagringsperspektivet) i Alvin, men vil trenge eksterne grensesnitt for presentasjon.
- Eneste norske institusjon i Alvin

Ibsen-senterets innspill

1. Av **strategiske grunner** er det veldig lite hensiktsmessig å inngå i ALVIN-konsortiet. Et hovedmål med dette prosjektet er å bygge kompetanse internt ved UiO/UB. Ved å «outsource» til Sverige vil man miste denne muligheten til å styrke DH-kompetansen internt. Valg av mediebase bør være første skritt på veien mot å etablere en infrastruktur for langsiktig drift og vedlikehold av DH-materiale ved UiO, og det oppnår man ikke ved å velge ALVIN.

2. En viktig praktisk begrensning i ALVIN er at den **ikke godtar alle typer formater**, f.eks. to av de aller viktigste og mest utbredte innen bilde- og tekstdata, nemlig RAW-filer og XML-filer. Den tilbyr heller ingen løsning for drift og lagring av databaser. ALVIN vil dermed ikke kunne håndtere en stor del av datasettene som produseres i forbindelse med særlig humanistisk forskning.

3. Det er **ikke mulig å hente ut materiale via API-er** i ALVIN. Det er vilje til å utvikle dette, men vi vet ikke når det kommer, og det er en funksjonalitet som er viktig å ha på plass helt fra starten av.

4. ALVIN er en **relasjonell database med hovedfokus på digitalisert svensk kulturmateriale**. Det betyr at forskningsdataene fra UiOs HumSam-miljøer vil bli koblet inn i denne konteksten, noe som ikke vil være relevant i de aller fleste tilfeller. Merverdien i å inngå i en relasjonell database blir dermed sterkt redusert.

5. ALVIN er et veletablert konsortium som allerede har **mange underinstanser** fra ulike arenaer. Dette vil høyst sannsynlig gjøre det vanskelig for UiO – med en helt annen profil enn de fleste andre etablerte partnere – å nå gjennom med egne behov.

6. ALVINs brukergrensesnitt er **tospråklig svensk og engelsk, men er ikke konsekvent**. Velger man svensk visning, får man en blanding av svensk og engelsk språk i systeminfo og innhold. Dette vil bli ytterligere komplisert og uryddig ved å legge til norsk som språkvalg.

Alvin - muligheter

- Gode organiseringsmuligheter for samlinger og arkiver
- Kjente standarder for registrering som det finnes god kompetanse i ved UB
- Samarbeidsarena som legger til rette for kompetanseheving
- Tilgang til bred kompetanse på digitaliseringsprosjekter, og hjelp med import/konvertering av eksisterende samlinger
- Samarbeid om utvikling - Alvin vil prioritere UiOs behov i prosjektperioden
- Sikker drift over tid
- Lave kostnader frigjør midler til utvikling lokalt
- Tilgang til Alvins mangfoldige samlinger fra Sverige

Alvin - anbefaling

Prosjektgruppa foreslår at

- prosjektet tester Alvin videre gjennom et begrenset medlemskap i prosjektperioden og velger om vi skal bli fulle medlemmer i konsortiet først ved prosjektets slutt (årsskiftet 2022/23). Prosjektgruppa jobber tett med Alvin om UiOs behov i prosjektperioden for å få reell innsikt i hva man kan få til i samarbeid og hva gevinstene av konsortium-deltagelse vil være
- vi utforsker alternativer til Alvin parallelt
- det settes av ressurser til å lage portaler/innganger til enkelt-samlinger i Alvin fra UiOs nettsider. Selv om prosjektet skal jobbe med back end, trenger pilotprosjektene å se hva som er mulig å få til av frontend-løsninger for å kunne vurdere løsningen (bildevisning, samlingspresentasjoner mm.)
- vurderingsgrunnlaget utvides til å omfatte også andre prosjekter UB jobber med for UiOs fagmiljøer (kan legges fram ved neste styringsgruppemøte)

Beslutningspunkter

- Fortsette med testing i Alvin gjennom et begrenset medlemskap i prosjektperioden
- Utforske alternative/parallelle løsninger
- Inkludere testing av front end-muligheter i prosjektet

Til neste møte:

- Vurderingsgrunnlag videre


FAIR@UIO

Prosjektets leveranser og avgrensning

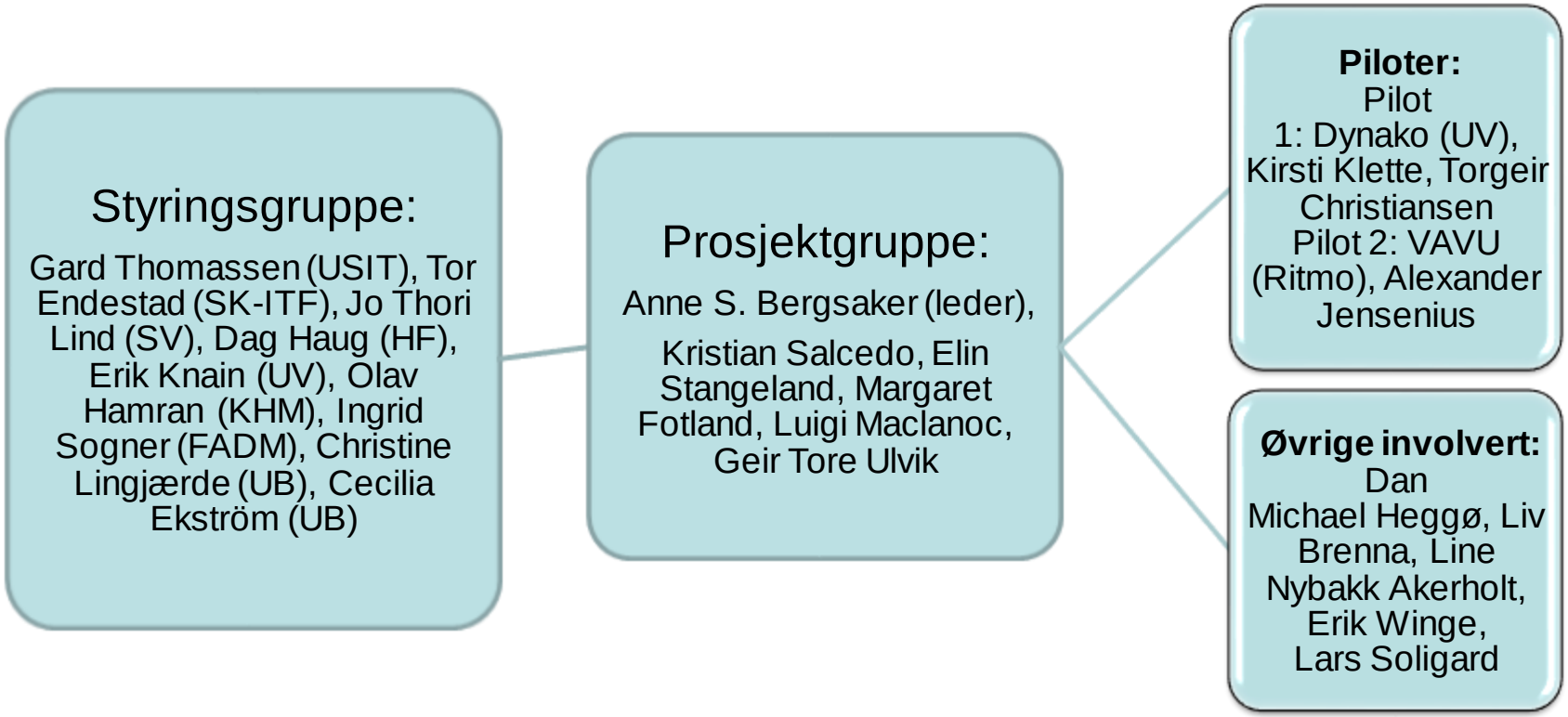
Viktigste leveranser:

- Etablere en minimumsstandard for metadata
- Hente ut metadata fra kjente kilder
- Pilotere dette på data fra PSI (SV), Dynako (UV) og VAVU (HF)
- Utvikle en ”oppdager-løsning” for å finne data
- Eksport av metadata

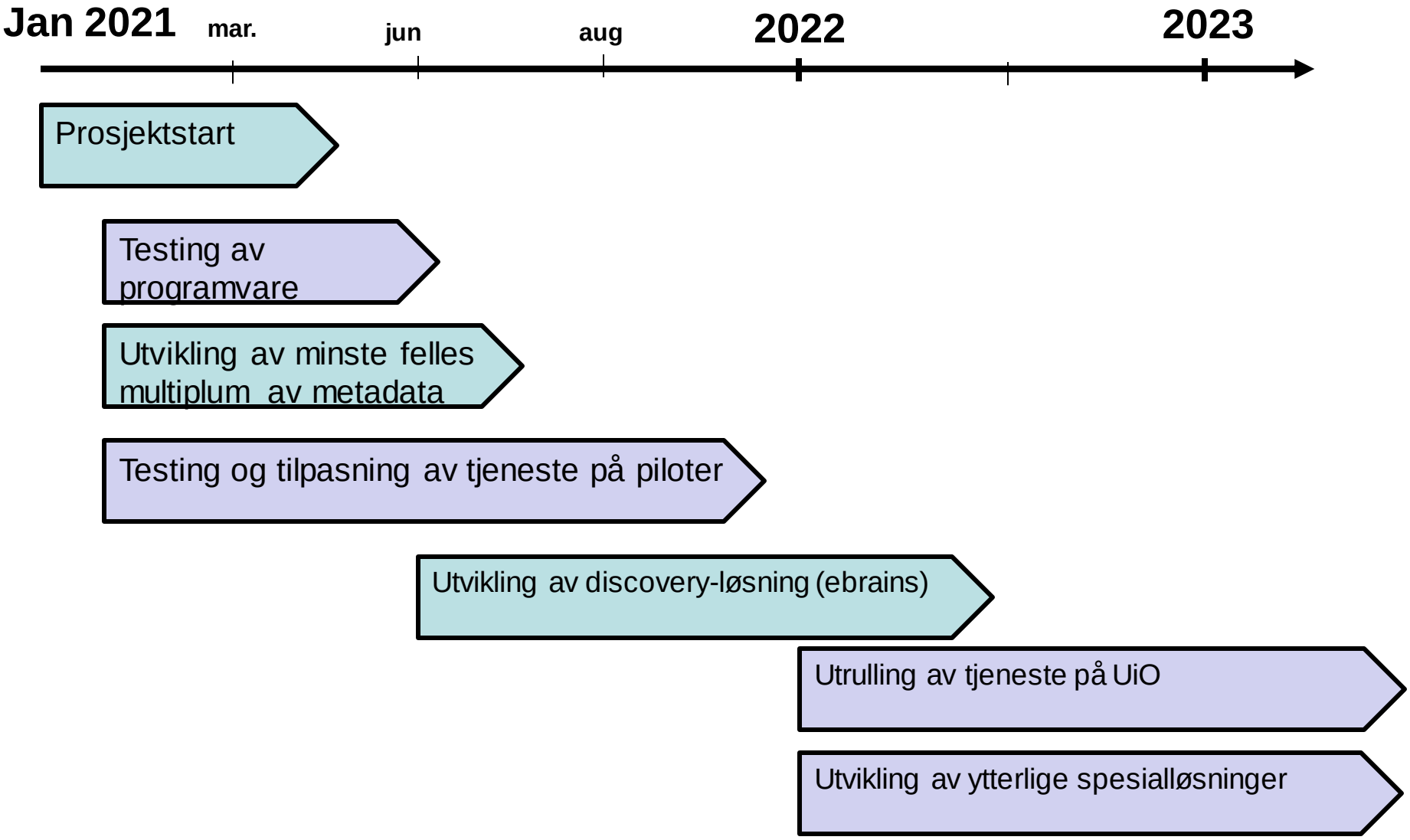
Avgrensninger:

- Skal ikke utvikle eller fungere som arkiv
- Skal ikke lære opp forskere i bruk av tjenesten
- Skal ikke selv ”tagge” data med metadata, annet enn via automatikk

Organisering og ansvar



Overordnet tidsplan



Ressursbruk

Midler:
2021: 7 MNOK (6MNOK SK-ITF, 1 MNOK FI-utvalget
2022: 4,8MNOK (SK-ITF, budsjettert)

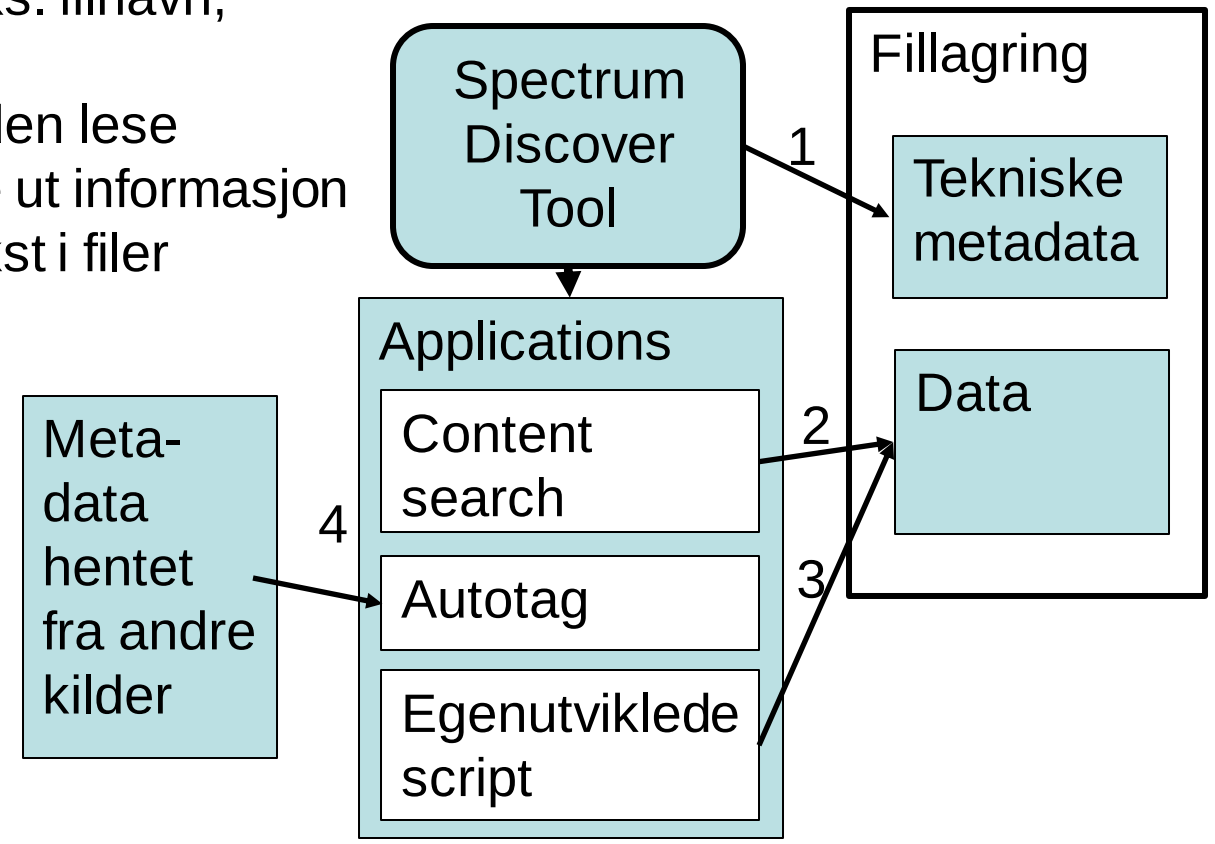
Bruk:
4,3 MNOK på 3 år lisenser Discover (10PiB)
Frikjøp USIT 2 MNOK pr år
Frikjøp UB 1,5 MNOK i 2021, og 5-800KNOK i 2022
Pilotene 1 MNOK

Status

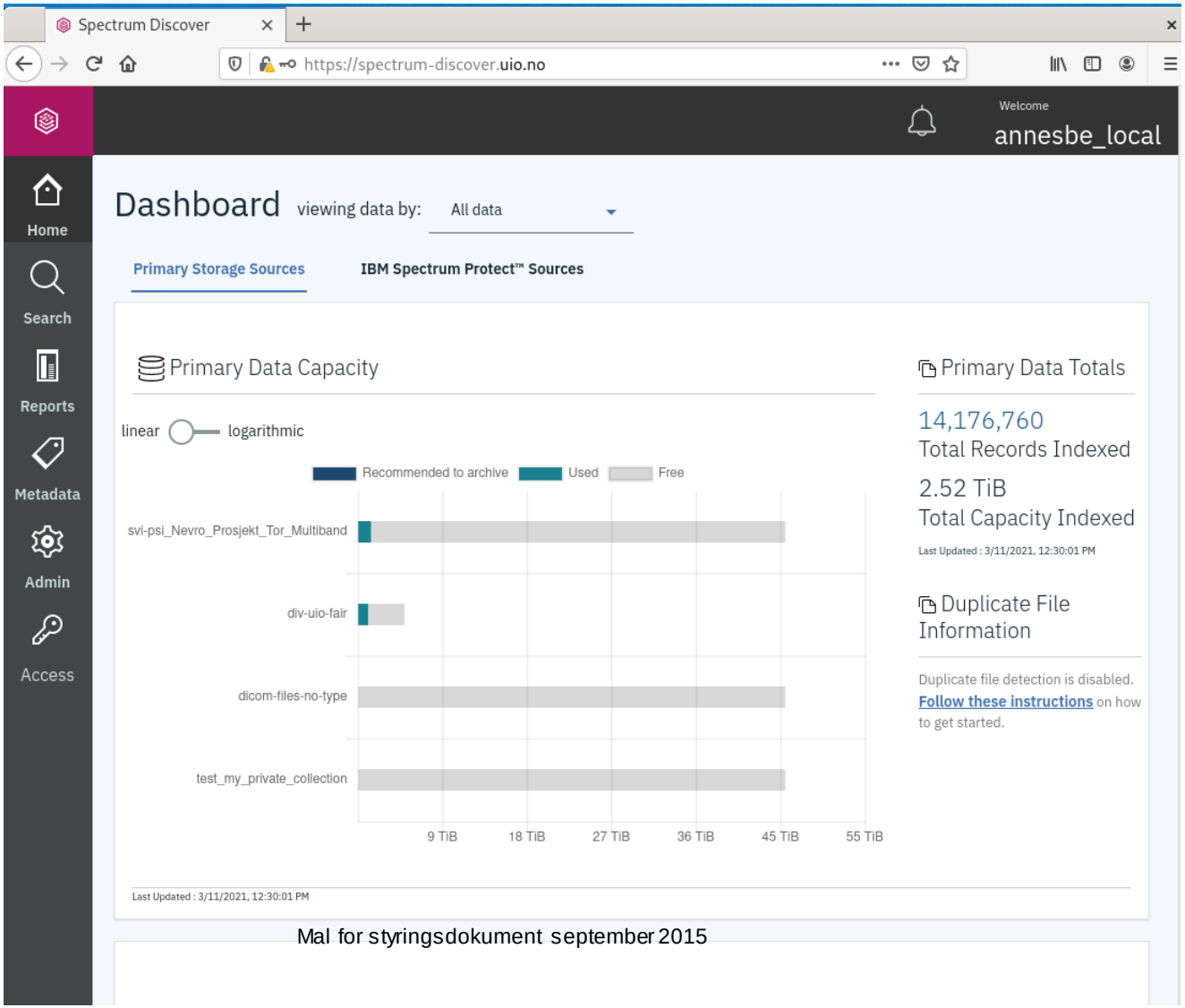
- Programvare er nå installert, men vi venter på en oppdatering
- Møter med IBM annenhver uke for å følge opp utviklingen av teknisk side
- Avklart med piloter hva de ønsker
- Workshop om behov og ønsker for metadata for ulike aktører og brukere
- Workshop om metadata på datasett/prosjekt-nivå
- Utvikling av tjenester for økt uthenting av metadata
- Presentasjon av overordnede metadata er påbegynt

Spectrum Discover Tool

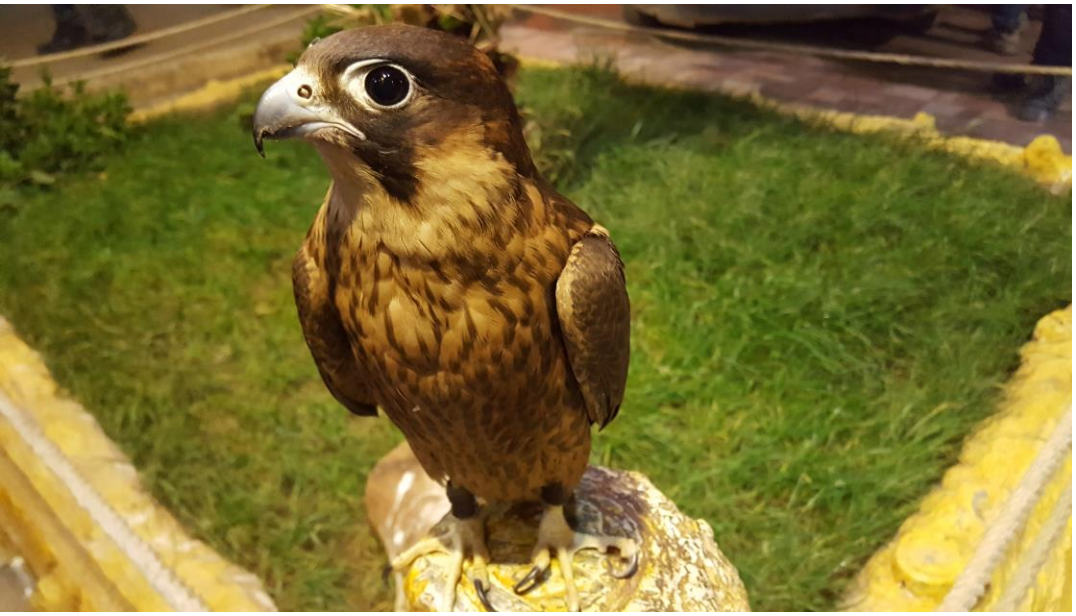
1. Spectrum henter automatisk tekniske metadata, f.eks. filnavn, størrelse, eier, etc
2. På forespørsel kan den lese gjennom data og hente ut informasjon basert på tilgjengeli tekst i filer
3. Ytterligere innhold og mer skjult metadata kan hentes ut med ytterligere skript som vi skriver
4. Metadata som hentes fra andre kilder kan hektes på data, vha «autotagging»



IBM Spectrum Discover Tool - demo

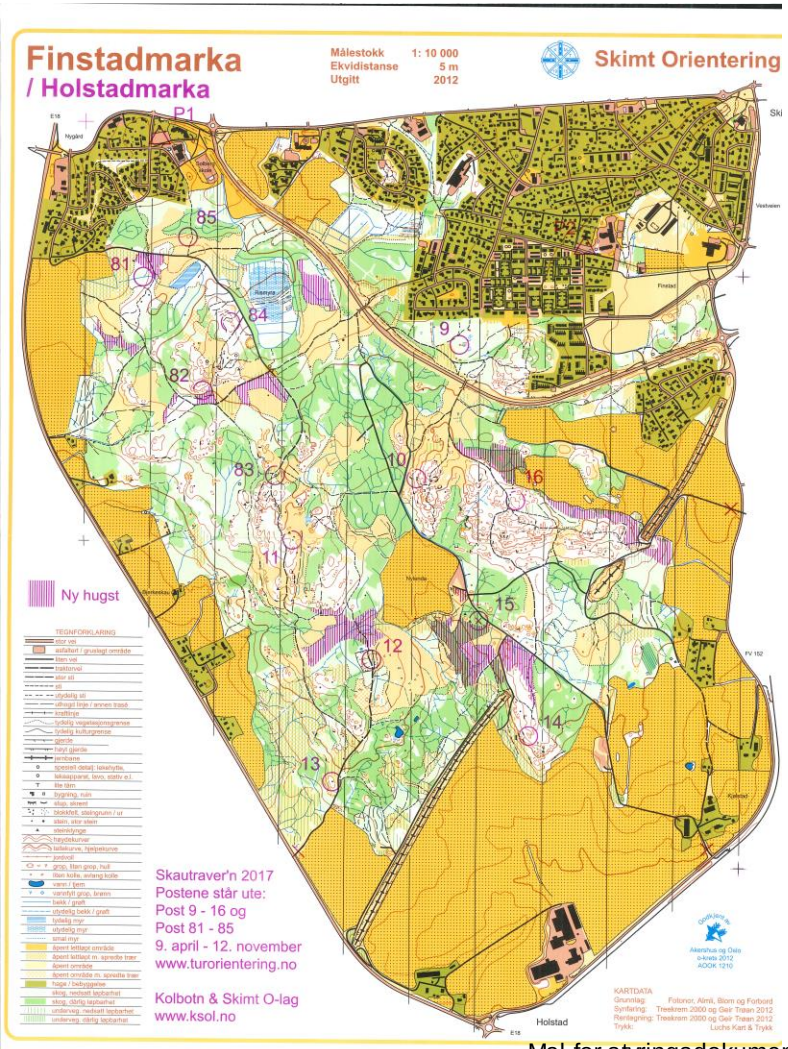


Eksempel på bilde med metadata




```
"Component 1": "Y component: Quantization table 0, Sampling factors 2 horiz/2 vert",
"Component 2": "Cb component: Quantization table 1, Sampling factors 1 horiz/1 vert",
"Component 3": "Cr component: Quantization table 1, Sampling factors 1 horiz/1 vert",
"Compression Type": "Baseline",
"Content-Type": "image/jpeg",
"Creation-Date": "2016-11-10T20:33:22",
"Data Precision": "8 bits",
"Exif IFD0:Date/Time": "2016:11:10 19:33:22",
"Exif IFD0:Image Height": "2988 pixels",
"Exif IFD0:Image Width": "5312 pixels",
"Exif IFD0:Make": "samsung",
"Exif IFD0:Model": "SM-G920F",
"Exif IFD0:Orientation": "Top, left side (Horizontal / normal)",
"Exif IFD0:Resolution Unit": "Inch",
"Exif IFD0:Software": "G920FXXS4DPI4",
"Exif IFD0:X Resolution": "72 dots per inch",
"Exif IFD0:Y Resolution": "72 dots per inch",
"Exif IFD0:YCbCr Positioning": "Center of pixel array",
"Exif SubIFD:Aperture Value": "f/1.9",
"Exif SubIFD:Brightness Value": "-1.57",
"Exif SubIFD:Color Space": "sRGB",
"Exif SubIFD:Date/Time Digitized": "2016:11:10 19:33:22",
"Exif SubIFD:Date/Time Original": "2016:11:10 19:33:22",
"Exif SubIFD:Exif Image Height": "2988 pixels",
"Exif SubIFD:Exif Image Width": "5312 pixels",
"Exif SubIFD:Exif Version": "2.20",
"Exif SubIFD:Exposure Bias Value": "0 EV",
"Exif SubIFD:Exposure Mode": "Auto exposure",
"Exif SubIFD:Exposure Program": "Program normal",
"Exif SubIFD:Exposure Time": "0.05 sec",
"Exif SubIFD:F-Number": "f/1.9",
"Exif SubIFD:Flash": "Flash did not fire",
"Exif SubIFD:FlashPix Version": "1.00",
"Exif SubIFD:Focal Length": "4.3 mm",
"Exif SubIFD:Focal Length 35": "28 mm",
"Exif SubIFD:ISO Speed Ratings": "640",
"Exif SubIFD:Makernote": "[98 values]",
"Exif SubIFD:Max Aperture Value": "f/1.9",
"Exif SubIFD:Metering Mode": "Spot",
"Exif SubIFD:Scene Capture Type": "Standard",
"Exif SubIFD:Shutter Speed Value": "1/19 sec",
"Exif SubIFD:Unique Image ID": "A16LLIC08SM A16LLIL02GM",
"Exif SubIFD:User Comment": "",
"Exif SubIFD:White Balance Mode": "Auto white balance",
"Exif Thumbnail:Compression": "JPEG (old-style)",
"Exif Thumbnail:Image Height": "288 pixels",
"Exif Thumbnail:Image Width": "512 pixels",
"Exif Thumbnail:Orientation": "Top, left side (Horizontal / normal)",
"Exif Thumbnail:Resolution Unit": "Inch",
"Exif Thumbnail:Thumbnail Length": "15898 bytes",
"Exif Thumbnail:Thumbnail Offset": "928 bytes",
"Exif Thumbnail:X Resolution": "72 dots per inch",
"Exif Thumbnail:Y Resolution": "72 dots per inch",
"File Modified Date": "Thu Apr 29 12:43:55 +02:00 2021",
"File Name": "apache-tika-1329376857939179720.tmp",
"File Size": "3980926 bytes",
"Image Height": "2988 pixels",
"Image Width": "5312 pixels",
"Last-Modified": "2016-11-10T20:33:22",
"Last-Save-Date": "2016-11-10T20:33:22",
"Number of Components": "3",
```

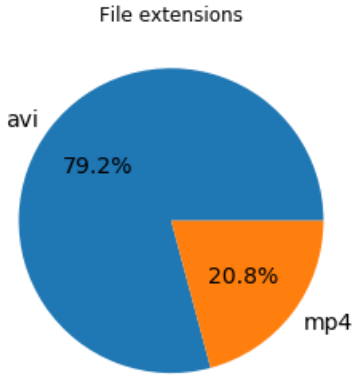
Eksempel på bilde med metadata

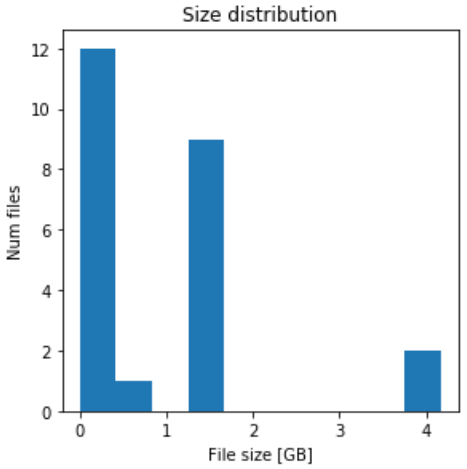


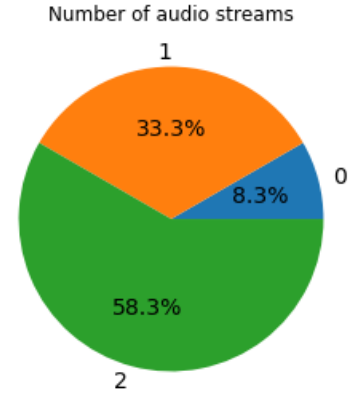

```
"Component 1": "Y component: Quantization table 0, Sampling factors 2 horiz/2 vert",
"Component 2": "Cb component: Quantization table 1, Sampling factors 1 horiz/1 vert",
"Component 3": "Cr component: Quantization table 1, Sampling factors 1 horiz/1 vert",
"Compression Type": "Baseline",
"Content-Type": "image/jpeg",
"Data Precision": "8 bits",
"File Modified Date": "Thu Apr 29 12:44:59 +02:00 2021",
"File Name": "apache-tika-2003137717656400459.tmp",
"File Size": "5609937 bytes",
"Image Height": "9921 pixels",
"Image Width": "7016 pixels",
"Number of Components": "3",
"Number of Tables": "4 Huffman tables",
"Resolution Units": "inch",
"Thumbnail Height Pixels": "0",
"Thumbnail Width Pixels": "0",
"X Resolution": "600 dots",
"X-Parsed-By": [
  "org.apache.tika.parser.DefaultParser",
  "org.apache.tika.parser.jpeg.JpegParser"
],
"Y Resolution": "600 dots",
"language": "",
"resourceName": "b'P1L03FB03_D1_FolloFinstad_#DYNAKO2021.jpeg'",
"tiff:BitsPerSample": "8",
"tiff:ImageLength": "9921",
"tiff:ImageWidth": "7016"
```

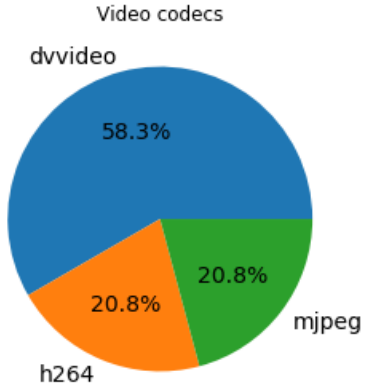
Foreløpig presentasjon av generelle metadata

- Bruker Jupyter notebooks
- Demo









Utfordringer

- Finne ut hvor i arbeidsflyten verktøyet passer inn
- Finne gevinster som gjør at forskere vil bruke dette
- Koble opp mot eksisterende tjenester
  - TSD
  - EduCloud
  - Forskpro
  - Vortex/UiO-nettsider

Veien videre

- Oppdatere til nyeste versjon av programvare
- Begynne arbeidet med en søkeløsning for å kunne søke i metadata
- Avklare med piloter mer detaljer rundt ønsket metadata og hvordan disse bør gjøres tilgjengelige og fremstilles innad i et prosjekt, for økt oversikt