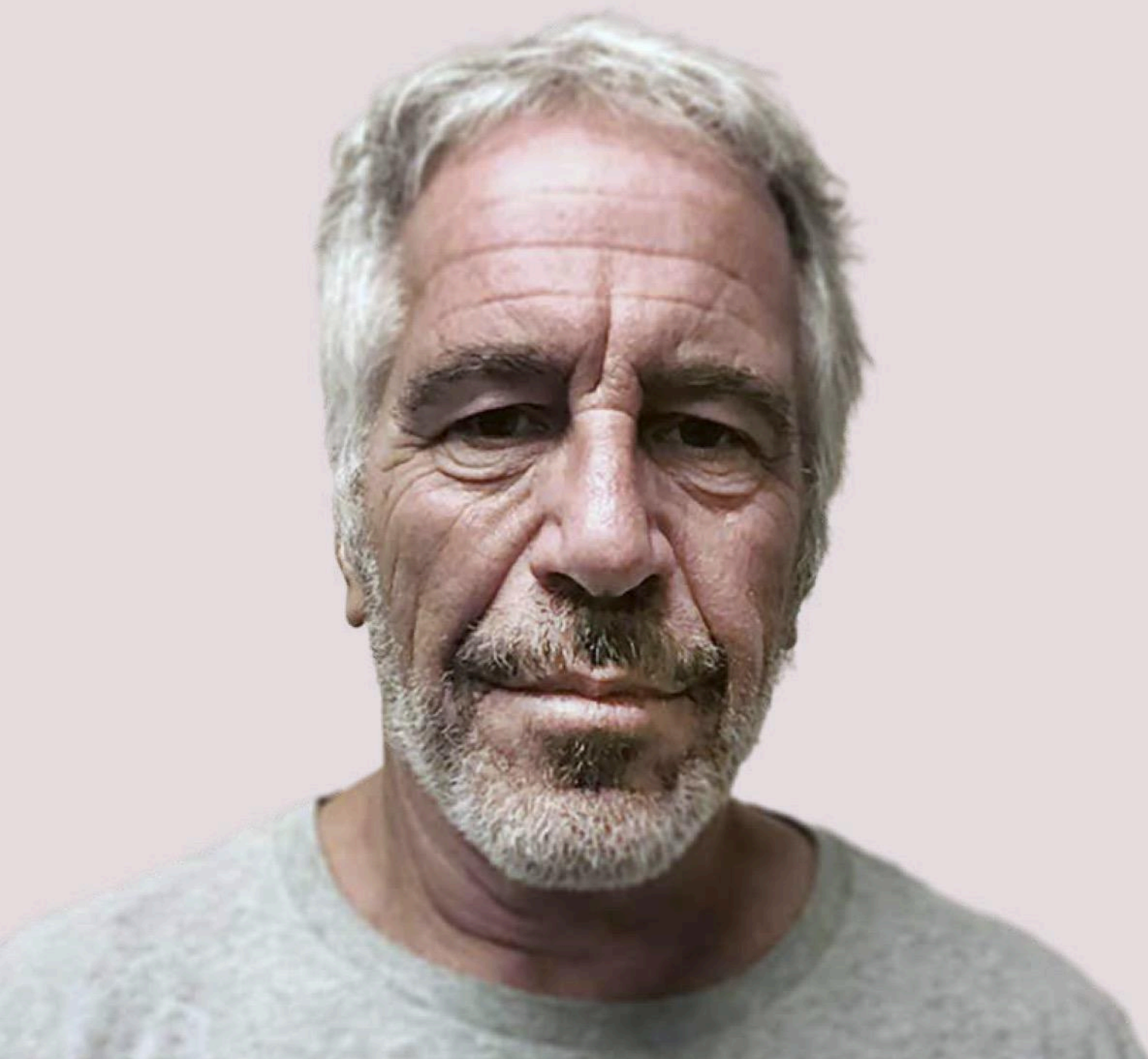


VG

Metoderapport Data-SKUP 2026

EPSTEIN-FILENE

Einar Otto Stangvik, Jarle Grivi Brenna, Elvin Baloku, Anders Haarr, Jari Bakken og Johannes Gorset



Innsendere:

Einar Otto Stangvik, Jarle Grivi Brenna, Elvin Baloku, Anders Haarr, Jari Bakken og Johannes Gorset

Takk til:

Nyhetsredaksjonen i VG.

Kontaktperson:

Johannes Gorset, johannes@vg.no, 977 62 202

Redaksjon:

VG

Publisert:

30. januar – 28. februar 2026

Kort fortalt

Fredag 30. januar 2026 offentliggjorde Department of Justice (DOJ) nye datasett i Epstein-filene. DOJ opplyste at publiseringen omfattet mer enn tre millioner sider, over 180 000 bilder og mer enn 2 000 videoer. VGs første spørsmål var:

Hvilke nordmenn kunne vi finne i de nye filene, og hva var deres tilknytning til Jeffrey Epstein?

Redaksjonell utvikling i VG kastet seg rundt, og innen tirsdag kveld kunne journalistene boltre seg i en verktøykasse som gjorde et enormt og multimodalt dokumentslipp søkbart for hele redaksjonen.

Dokumentene vi ikke kunne laste ned

DOJ publiserte materialet som datasett. Det så ved første øyekast ut som et vanlig dokumentslipp, med mulighet for å laste ned alle filene enten manuelt eller i samlede arkiver. Men så enkelt ble det ikke. De samlede arkivene var i praksis umulige å laste ned: Nedlastingene gikk i sneglefart og stoppet av seg selv, og innen kort tid ble de fjernet helt.

Alternativet var å laste ned alle filene individuelt. Det kunne selv ikke VGs redaksjon gjøre. Hvis vi satte ti journalister på saken og hver journalist lastet ned ti filer i minuttet fra hvert sitt utgangspunkt, ville det tatt over 20 døgn i strekk å laste ned over tre millioner filer med perfekt koordinasjon. Det var åpenbart ikke en farbar vei

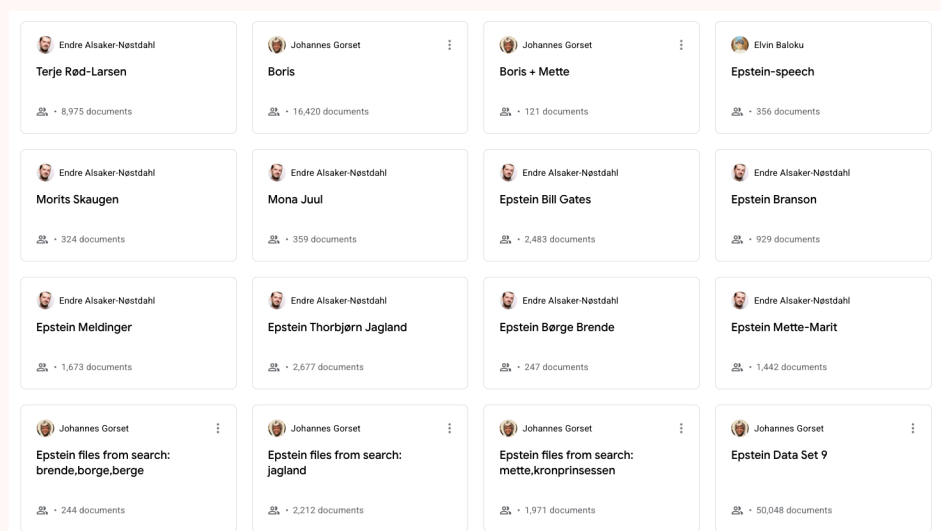
av rent nyhetsmessige hensyn, og ikke minst fordi filene etter hvert ble mer sladdet eller rett og slett forsvant.

Den første oppgaven ble derfor å sikre materialet, så vi ikke ville miste viktige bevis, og så vi kunne søke og prosessere det på bedre måter enn DOJs eget søkefelt muliggjorde.

Vi satte AI på problemet: «Lag et skript som leser oppføringssidene, finner filadressene og laster ned hver PDF.» Innen 15 minutter hadde vi den første løsningen. Den fungerte først ikke, men ved nærmere undersøkelse viste det seg at aldersverifiseringen på oppføringssiden var nødvendig for enhver nedlasting (informasjonskapselen `justiceGovAgeVerified=true`). Flere HTTP-hoder måtte dessuten ligne trafikken fra en vanlig nettleser: blant annet `User-Agent`, `X-Requested-With`, `Referer` og `Origin`. Etter ytterligere 15 minutter var nedlastingene i gang.

På grunn av den voldsomme trafikken ga DOJs løsning også tidvis midlertidige feil. Koden ble derfor forbedret ved å lagre feilede nedlastinger og prøve dem på nytt etter gradvis lengre ventetid, i stedet for å gi opp filen. Etter en stund innså vi at én nedlasting om gangen ville gå for tregt, og bygget om skriptet til å laste ned med ti parallelle tråder. Selv da tok det hele natten å laste ned filene vi hadde identifisert i datasett 9, 10 og 11.

Mens vi ventet på at de komplette datasettene ble lastet ned, utviklet vi egne skript som brukte DOJs enkle søkeløsning til å finne og laste ned filer med ord som «mette», «mm», «princess», «jagland» og feilstavinger som «jabland», samt andre referanser vi kunne gjette oss til eller lese oss til at ble brukt om nordmenn i filene. Uttrekkene ble fortløpende lastet opp til Google Pinpoint og ble dermed søkbare og «chatbare» med AI. Pinpoint tillot maksimalt 200 000 dokumenter per samling, og en profesjonell konto hadde 100 GB samlet lagring. Vi laget derfor flere mindre samlinger for bestemte personer og temaer.



FIGUR 1: Noen av Pinpoint-samlingene som ble opprettet for ulike personer og temaer.

Lørdag morgen var datasett 9, 10 og 11 sikret på vår egen lagring. I dagene etter supplerte vi materialet med større kildearkiver som inneholdt filer DOJ senere hadde fjernet.

Slik fikk vi det mest mulig komplette datasettet

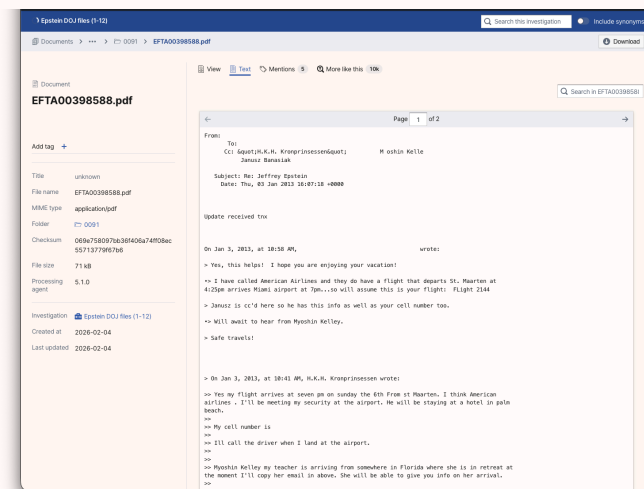
Materialet kom til oss gjennom flere veier: vår egen nedlasting, arkiver fra kilder og filer som tidligere hadde ligget på DOJ's nettsider, men senere var fjernet. Det samme dokumentet kunne derfor finnes i flere arkiver. Filene hadde ofte samme originalnavn, men innholdet kunne være ulikt sladdet.

En nyere versjon var ikke nødvendigvis mer komplett. Ofte var det omvendt. Noen ganger var en eldre fil mindre sladdet enn den som lå ute senere. Vanlig deduplisering, der man bare beholder siste fil eller én tilfeldig kopi, kunne derfor ha fjernet den mest informative versjonen.

Vi beholdt originalfilnavnene og brukte dem til å gruppere filer som skulle være samme dokument. Deretter automatiserte vi en visuell sammenligning av PDF-ene i hver gruppe. Slik kunne vi beholde den versjonen som hadde minst synlig sladding blant kopiene vi faktisk hadde funnet, uten at journalistene måtte åpne alle dublettene manuelt.

I uken etter publiseringen fikk vi hele det rensede korpuset inn i OpenAleph, en åpen kildekode-plattform for gravejournalister som muliggjør lagring, indeksering, søk og analyse av store mengder strukturerte og ustrukturerte data. Da kunne redaksjonen

søke på tvers av datasett, samle dokumenter i undersøkelser og gå fra et funn tilbake til originalfilen. Pinpoint-uttrekkene, som muliggjorde umiddelbare søk i timene etter dokumentslippet, ble derfor gradvis erstattet av denne mer robuste løsningen.



FIGUR 2: Et dokument fra det sammenstilte materialet i OpenAleph, med søkbar tekst, dokumentdata og kobling tilbake til originalfilen.

Fra 1,1 millioner bilder til Terje Rød-Larsen i sofaen

Bildene var det største metodiske problemet og det tydeligste gjennombruddet. Store deler av materialet var skannede PDF-er eller PDF-er som i praksis bare inneholdt fotografier. Tradisjonelt fulltekstsøk kunne ikke fortelle hvem som var avbildet, om to bilder viste samme rom, eller om en bestemt gjenstand gikk igjen.

Det ville vært for trekt og kostbart å sende alle dokumentene direkte til de tyngste bilde- og ansiktsmodellene. Vi bygget derfor en grov-til-fin arbeidsflyt:

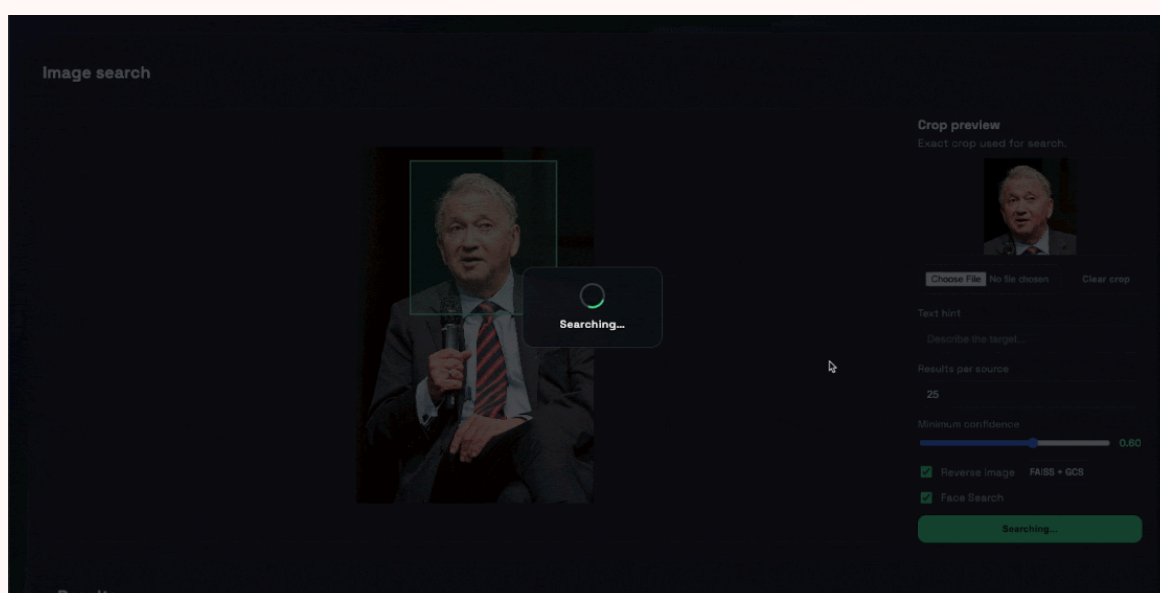
1. Koden analyserte innebygde bilder på minst 200×200 piksler og reduserte hvert bilde til 100×100 piksler for rask analyse.
2. PDF-en gikk videre dersom minst ett bilde hadde gjennomsnittlig metning over 3 og standardavvik over 10 på en skala fra 0 til 255, ingen metningsverdi dominerte mer enn 80 prosent av pikslene og minst 5 prosent av pikslene lignet hudtoner.
3. Vi hentet ut kandidatene og laget vektorrepresentasjoner som gjorde ansikts-, likhets- og objektsøk mulig. Journalistene kontrollerte alltid treffene mot originalbildene.

Resultatet var omtrent **1,1 millioner uttrukne og søkbare bilder**. Omtrent **70 000** bilder klassifisert som menneskebilder ble samlet for seg selv. Alle ble

tilgjengeliggjort i PhotoPrism, en åpen kildekode-plattform for organisering av bilder.

Parallelt bygget vi søk som dekket de ulike journalistiske behovene vi hadde i perioden:

- **Ansiktssøk:** Journalisten lastet opp et referansebilde, markerte ansiktet og søkte etter samme person i materialet.
- **Bilde-til-bilde-søk:** Et helt bilde eller et utsnitt ble brukt til å finne visuelt lignende bilder.
- **Tekst-til-bilde-søk:** Journalisten kunne skrive et motiv eller en gjenstand, for eksempel «yellow sofa».
- **Kombinert søk:** Et bilde kunne kombineres med en tekstlig presisering for å spisse resultatet.



FIGUR 3: *Bildeverktøyet under ansiktssøk etter Terje Rød-Larsen.*

Det tekstbaserte søket brukte Vertex AI Multimodal Embeddings til å lage numeriske representasjoner av både bilder og søketekst, og Vertex AI Vector Search til å finne de nærmeste treffene blant alle de 1,1 millioner bildene som var lagret i Google Cloud Storage. Det lokale søket brukte FAISS, et bibliotek for raske likhetssøk mellom slike vektorer, til å søke i delmengden på omtrent 70 000 menneskebilder. Verktøyet slo sammen og rangerte treffene fra begge indeksene i én resultatliste.

Det første eksempelet på at verktøyene fungerte, fikk vi ved å laste opp et referansebilde av Terje Rød-Larsen. Umiddelbart fant vi Rød-Larsen i en sofa mens han viste frem et bilde av en lettkledd kvinne på telefonen sin. Deretter måtte stedet dokumenteres. Vi lastet opp utsnitt med kunst og møbler for å finne flere bilder fra

rommet og kryssrefererte dem med verifiserte bilder fra Epsteins leilighet i Paris. Slik kunne vi plassere Rød-Larsen i leiligheten.

Metoden kunne også gå motsatt vei. Mange bilder viste Epsteins boliger uten at alle personene var kjent. Et utsnitt av et rom, et møbel eller et bilde på veggen kunne fungere som et visuelt fingeravtrykk. Vi kunne søke etter samme rom og finne andre personer fotografert der. Slik fant vi blant annet DOJs sladdede versjon av Aftenpostens bilde av Mette-Marit i sofaen med Epsteins assistent.

Saken «Til bords med Epstein» ble bygget på bilder funnet gjennom det samme interne bildeverktøyet. Flere av motivene ble kontrollert mot politiets åstedsbilder fra Epsteins hjem i New York.

I arbeidet med Zorro Ranch begynte vi med tekstsøk som «ranch», «Zorro», «cowboy» og «horses». Da ett bilde var stedfestet, brukte vi omvendt bildesøk til å finne bilder fra samme sted og trolig samme dag. Klær, rom og andre detaljer bandt bildene sammen. Vi brukte også salgsannonser for ranchen som ekstern bildekilde i det omvendte søket.

De skjulte videoene

En av de viktigste oppdagelsene kom ikke fra en modell, men fra en merkelig PDF. Tidvis viste PDF-er bare teksten «No Images Produced». Samtidig hadde DOJ opplyst at publiseringen inneholdt tusenvis av videoer. Vi prøvde derfor å endre filendelsen i nettadressen fra .pdf til blant annet .mp4, .mov og .mkv. Plutselig svarte serveren med en video.



FIGUR 4: En av PDFene som skjulte en video.

Vi utviklet umiddelbart et skript som prøvde samtlige filendelser som videoer og lyd kunne være lagret i, og kjørte det mot alle PDF-ene som viste «No Images Produced». I løpet av natten til tirsdag hentet vi på den måten ut **2 828 video- og lydfiler**.

AI-modellene vi hadde tilgjengelig på det tidspunktet, var ikke raske eller presise nok til å søke og oppsummere hele videomengden sikkert. Vi gikk derfor manuelt

gjennom 1 642 videoer og merket dem som interessante eller uinteressante. Denne gjennomgangen fant blant annet videoen der Epstein løper etter unge kvinner rundt et kjøkkenbord på den private øya.

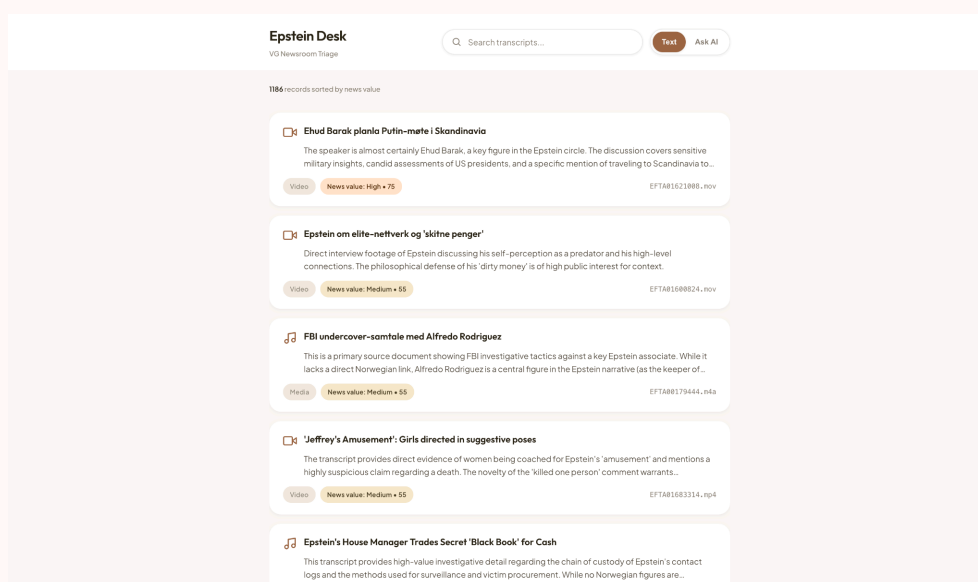
De 1 186 lydfilene var for lange til å lytte gjennom manuelt. Vi bygget derfor en egen arbeidsflyt for tale:

1. Systemet forsøkte å hente ut lydsporet fra hver fil.
2. Lyden ble transkribert med ElevenLabs Scribe.
3. Transkripsjonene ble klassifisert etter innhold.
4. Sang og støy ble skilt fra materialet med meningsfull tale.
5. Resten fikk maskinelt genererte sammendrag, nøkkelord og navngitte entiteter.

Det interne verktøyet inneholdt **1 186 filer med brukbare transkripsjoner**.

Journalistene kunne søke med nøkkelord eller stille spørsmål i naturlig språk med en språkmodell, sortere på innholdstype og maskinelt vurdert nyhetsverdi, lese transkripsjonen og hoppe direkte til riktig sted i originalopptaket.

Det øverste treffet i en nyhetsrangert visning var et nesten tre og en halv time langt middagsopptak med Jeffrey Epstein, Israels tidligere statsminister og forsvarsminister Ehud Barak og USAs tidligere finansminister og Harvard-rektor Larry Summers. Samtalen handlet blant annet om internasjonal politikk og planer om et konsultentselskap der Terje Rød-Larsen skulle delta.



FIGUR 5: Verktøyet for å behandle transkripsjonene.

Da maskinene var ferdige, begynte jobben for menneskene

Tekstsøkene fant store mengder dokumenter med navn som Mette-Marit, Torbjørn Jagland og Rød-Larsen. De viktigste publiseringene krevde likevel et omfattende manuelt arbeid.

Det tydeligste eksempelet var rekonstruksjonen av nær 400 meldinger mellom Mette-Marit og Epstein. Fjorten journalister arbeidet i flere dager med et stort regneark for å dobbeltsjekke og sammenstille alle meldingene. Hver rad inneholdt:

- Tidspunkt
- Avsender
- Mottaker
- Meldingstekst
- Dokumentnavn

Den ferdige kronologien inneholdt alle verifiserte, unike meldinger redaksjonen fant, med ett bevisst unntak: Enkelte meldinger med medisinske opplysninger og uten annen relevans for saken ble holdt tilbake av personvern hensyn.

Metoden viser arbeidsdelingen vi brukte gjennom hele prosjektet:

- Kode og modeller fant, sorterte og rangerte kandidater.
- Journalister leste originalmaterialet, fjernet dubletter, bygget kronologi og vurderte relevans.
- Publiserte opplysninger ble kontrollert mot dokumentet, bildet eller opptaket de kom fra.

Fire døgn fra råfiler til selvbetjening

For å få mest mulig fart og ikke gå i beina på hverandre, organiserte vi arbeidet i parallelle spor:

- Johannes og Einar fokuserte på nedlasting og deduplisering.
- Jarle fokuserte på tekstsøk.
- Einar og Anders fokuserte på bildesøk.
- Elvin fokuserte på talesøk.
- Jari fokuserte på søk i verdipapirer og kvitteringer.

Før verktøyene fikk egne grensesnitt, gjorde vi søk sammen med eller på vegne av journalistene. Vi opprettet et levende arbeidsdokument som forklarte hva hvert

verktøy kunne brukes til, og hvor det fantes. Verktøyene ble demonstrert på morgenmøter og i én-til-én-arbeid med journalister som fulgte konkrete spor. Slik kunne vi oppdage nye behov og gjøre verktøyene enklere å bruke.

Innen tirsdag kveld 3. februar – omtrent fire døgn etter DOJ-publiseringen – kunne journalistene selv søke i tekst, bilder og tale uten kode eller SQL. Fra 30. januar til og med 28. februar publiserte VG 254 artikler i Epstein-komplekset. 91 forskjellige personer hadde byline.

Funn, presentasjon og følger

Verktøykassa ga funn i alle materialtypene: [nær 400 sammenstilte meldinger mellom Epstein og Kronprinsesse Mette-Marit](#), [Rød-Larsen i Epsteins Paris-leilighet](#), [bildeserier fra flere boliger og Zorro Ranch](#), [middagsopptaket med Epstein, Barak og Summers](#) og [kjøkkenvideoen fra øya](#).

Prosjektet fikk stor rekkevidde. De 254 VG-artiklene som ble publisert fra 30. januar til og med 28. februar 2026 med Epstein i tittel, tagg eller URL, hadde til sammen **74 414 368 sidevisninger** frem til 18. august. Den mest leste var «Kronprinsessen delte private detaljer med Epstein», med **1 927 495 sidevisninger**.

Flere av verktøyfunnene ble publisert først av VG i Norge. Dette kan dokumenteres særlig tydelig i to tilfeller: Da VG fant Thorbjørn Jaglands e-post om «extraordinary girls», skrev Dagbladet i sin oppfølging at VG hadde omtalt saken først. Da VG senere avdekket lydopptaket hvor Ehud Barak diskuterte et konsulentfirma med «denne Terje», fulgte iNyheter opp knappe to timer senere og krediterte VG for funnet.

Søk i norske medier viser også at VG var først med bildet av Terje Rød-Larsen med en mobiltelefon i Epsteins Paris-leilighet og bildefortellingen «Til bords med Epstein». Vi fant heller ingen tidligere norsk publisering av videoen hvor Epstein løper etter to kvinner på kjøkkenet, eller av den konkrete bildeserien fra Zorro Ranch. Materialet var offentlig tilgjengelig, så dette var ikke eksklusiv tilgang. VGs fortrinn var at verktøyene gjorde det mulig å finne, verifisere og publisere enkeltfunnene før andre redaksjoner.

Den dokumenterbare hovedprestasjonen er likevel den metodiske: På fire døgn gikk VG fra et ustabilt utslipp av enkeltfiler til selvbetjente søk i tekst, bilder og tale. Verktøyene gjorde at en stor del av redaksjonen kunne arbeide parallelt og førte til flere konkrete, publiserte funn som ellers ville krevd svært omfattende manuell gjennomgang.