Arbeidsnotat
2023:10

Nye kilder til data om forsknings-aktivitet

Notat til NFRs arbeidsgruppe for bibliometri, 2023

Henrik Karlstrøm

NIFU

Arbeidsnotat
2023:10

Nye kilder til data om forsknings-aktivitet

Notat til NFRs arbeidsgruppe for bibliometri, 2023

Henrik Karlstrøm

Arbeidsnotat 2023:10

Utgitt av Nordisk institutt for studier av innovasjon, forskning og utdanning (NIFU)
Adresse Postboks 2815 Tøyen, 0608 Oslo. Besøksadresse: Økernveien 9, 0653 Oslo.

Prosjektnr. 21341-1

Oppdragsgiver Norges forskningsråd
Adresse Drammensveien 288, 0283 Oslo

Fotomontasje NIFU

ISBN 978-82-327-0607-5
ISSN 1894-8200 (online)

Copyright NIFU: CC BY 4.0

www.nifu.no

Forord

Dette arbeidsnotatet omhandler nye kilder til data om forskningsaktivitet. Notatet er utarbeidet som en del av NIFUs rammeavtale med Norges forskningsråd om analyser og utvikling av indikatorer for det nasjonale forsknings- og innovasjonssystemet. Norges forskningsråds arbeidsgruppe for bibliometri har bestilt oversikten.

Oslo, 21.06.2023

Michael Mark
forskningsleder

Innhold

Sammendrag

Dette notatet gir en oversikt over etablerte og nye kilder til informasjon om forskeres forskningsaktivitet. Formålet er å gi en oversikt over status for feltet i lys av en økende interesse for å kunne identifisere og anerkjenne et bredere spekter av aktiviteter knyttet til forskergjerningen.

Notatet diskuterer noen måter å forstå dekning og kvalitet på metadata, og går så gjennom kildene til data om resultater av forskjellige former for forskningsaktivitet: fire etablerte kilder til bibliometriske metadata, fire nye kilder til bibliometriske metadata, fem kilder til data om åpen vitenskap-aktivitet og tre kilder til formidlingsdata.

Hovedkonklusjonen er at det fortsatt kun er bibliometriske metadata som egner seg til generell monitorering av forskningsaktivitet, men at flere nye kilder til slike data kan være aktuelle alternativer til de store etablerte aktørene på feltet. For mer spesifikk bruk, som for eksempel anerkjennelse av en enkeltforskers innsats i en spesifikk faglig og institusjonell kontekst, kan kilder til informasjon om åpen vitenskap-aktivitet eller forskningsformidling være aktuelle. Bruk av slike er i tråd med en bevegelse mot et mer sammensett indikatorsett for anerkjennelse av forskningsaktivitet enn et rent publiserings- og siteringsbasert et.

1 Innledning

Dette notatet gir en oversikt over etablerte og nye kilder til informasjon om forskeres forskningsaktivitet. Der man tidligere i stor grad har fokusert på vitenskapelige publikasjoner i vurderingen av resultatet av forskning har det de siste årene blitt stadig mer interesse for det vi kan kalle forskningsaktiviteter i vid forstand, det vil si hele spekteret av formidling av forskning – fra fagfellevurderte publikasjoner myntet på lesere i samme faglige krets via produksjon av forskningsresultater som bygger opp under disse (forskningsdata, programkode) til formidlingen av forskningsresultater til et videre publikum.

Fremveksten av nye kilder til nye typer data gir anledning til å diskutere eksisterende måter å forstå verdien av forskjellige former for forskningsaktiviteter. En slik bred diskusjon er langt utenfor rammene av et notat som dette, men det er samtidig vanskelig å gi noen god oversikt over kildene og deres styrker og svakheter uten å skjele til evnen til å dekke behovet for forskjellige former for dokumentasjon av aktivitet, enten behovet kommer fra den enkelte forsker, en institusjon eller evaluerende myndighet. Notatet starter derfor med en kort diskusjon av bruksområdene for forskjellige former for aktivitetsdata og en konseptuell modell for å vurdere egnetheten til en kilde basert på det aktuelle bruksområdet.

Dette er ikke en komplett katalog over eksisterende og nye løsninger for innhenting av data om forskningsaktiviteter. Det utvidete bibliometrifelt er i rask utvikling, og med utviklingen av felles standarder for metadata og økt bruk av konsistente, persistente identifikatorer til å koble data dukker nye løsninger stadig opp. Det er heller ment som en oppdatering på status på feltet for et sett med potensielt relevante kilder. Flere av dem er godt etablerte kommersielle produkter og vil i liten grad forandre underliggende struktur. Andre er mindre etablerte, og det kan hefte usikkerhet ved finansiering eller strukturell stabilitet rundt databasen. Samtidig vil disse i større grad følge moderne prinsipper for strukturering av databaser, med fokus på felles standarder for datakobling og interoperabilitet med andre kilder.

1.1 Metadatatypologi

Ulike datakilder egner seg ikke til alle analyseformål. I dette notatet vil jeg tegne opp to dimensjoner ved metadata, generalitet og dekningsgrad (beskrevet under), som vil være bestemmende for hvordan disse kan anvendes, og dermed også kunne fungere som rettesnor for hvorvidt en kilde til slike data kan anvendes eller regnes som autoritativ for formålet. Utover en kort beskrivelse av kilden vil omtalen av hver kilde under søke å plassere kilden langs disse to aksene.

Hvordan en kilde er organisert reflekterer til en viss grad den underliggende filosofien til databaseforvalterne. Grovt forenklet kan man si at det eksisterer en spenning mellom en filosofi om kunnskapsorganisering og en om informasjonsgjenfinning (Hjørland 2021). Der kunnskapsorganisering fokuserer på indeksering, klassifisering og tematisk organisering av dokumenter er informasjonsgjenfinning orientert mot å identifisere så mye data som mulig innenfor rammen av det som kan være relevant og presentere det i en form som muliggjør kobling mot andre data. Et eksempel på det første kan være en base som Norsk vitenskapsindeks, en kilde med klart avgrenset mandat og manuelt kontrollert innhold organisert etter institusjonelle linjer. Det tydeligste eksempelet på en bibliometrisk kilde som ren informasjonsinnhentingsverktøy er Google Scholar, som utelukkende fungerer som en søkemotor for antatt akademisk tekst. De fleste kilder i dette notatet vil befinne seg et sted på spekteret mellom disse to ytterpunktene.

De godt etablerte kildene diskutert her er i stor grad organisert etter kunnskapsorganiseringsprinsipper, men det er en tydelig bevegelse mot å behandle bibliometriske metadata og annen informasjon om forskningsaktivitet som ledd i større analytiske dataflyter med behov for økt interoperabilitet og transparens og reproduserbarhet rundt hvordan kobling av data skjer. For alle som benytter seg av kilder til forskningsaktivitetsmetadata er det viktig å ha et avklart forhold til hvordan disse kildene organiserer sine data, hvordan de utleveres og i hvilken grad de kan inngå i egne analytiske prosesser også i fremtiden, herunder hvordan data er lisensiert og i hvilken grad de er interoperable med andre data av interesse.

1.1.1 Generalitet

En viktig dimensjon i vurderingen av en datakilde er om den egner seg som en generell, global kilde til data, altså i hvilken grad den kan sies å kunne gi informasjon om forskningsaktivitet uavhengig av faglig, institusjonell eller geografisk kontekst.

Data som skal innhentes og brukes i evaluativt øyemed uten at personen eller personene som er gjenstand for monitorering kan motsette seg det må være av generell art, og kilden må kunne demonstrere bred dekning i den relevante faglige, institusjonelle og geografiske konteksten. Motsatt kan data som skal inngå i en

selvpresentasjon være av mer lokalt relevant natur, og i mange tilfeller ikke være underlagt krav om å kunne presentere et sterkt komparativt perspektiv på dataene.

1.1.2 Dekningsgrad

Et avgjørende aspekt ved nytten ved en datakilde er dekningsgraden til en kilde. Dette er ikke bare reflektert i størrelsen på basen generelt, men også i bredden i dekningen i basen, både faglig, geografisk og institusjonelt.

Dekningsgraden til en kilde inngår som en del av vurderingen av selve datakvaliteten på kilden. En database med perfekte metadata om halvparten av de data som regnes som relevante, kan være mindre egnet enn en base med komplett dekning, men 10 % feil i postene i basen. Akkurat hvordan forholdet mellom dekning og datakvalitet skal se ut er et spørsmål som må avklares i hvert konkrete tilfelle, og det er derfor en fordel å ha oversikt over og kunne ta i bruk flere kilder til metadata om forskningsaktivitet i bred forstand.

1.1.3 Evaluativt formål

Formålet med innsamling av metadata vil til en viss grad diktere egnetheten til data man skal bruke. I henhold til anbefalingene fra CoARA[1] og i tråd med rådende forståelse av forsvarlig bruk av bibliometri (Hicks, et al. 2015) stilles det andre krav til kilder til data som ligger til grunn for løpende monitorering av fagmiljøer enn data som skal understøtte en enkeltpersons søknad til en stilling eller til opprykk. Begge deler er en form for evaluering, men de har forskjellig opphavspunkt og formål. En person som innhenter data om egen aktivitet på eget initiativ, må kunne sies å samtykke til en evaluativ vurdering av eget arbeid i langt større grad enn en prosess med ekstern evaluering.

Formålet med den evaluative situasjonen, enten det er bred monitorering av større institusjoner eller fagmiljøer eller individuell promotering, vil være styrende for hva slags datakilde man kan nyttiggjøre seg. Data som viser bred formidlingsaktivitet eller vektige bidrag til åpen vitenskap kan være akseptable for en kandidat å selv hente frem og få anerkjennelse for, uten at de dermed er passende for en finansiør eller myndighetsaktør å bruke til å sammenstille personer eller forskningsmiljøer.

[1] https://coara.eu/agreement

2 Kilder til bibliometriske data

I dette kapittelet tar jeg for meg etablerte og nye kilder til klassiske bibliometriske analyseoppgaver. Første delkapittel diskuterer fire godt etablerte kilder til bibliografiske data og siteringsdata og gir en indikasjon på deres dekningsgrad og eventuelle kjente feilkilder i basene. Andre delkapittel går gjennom fire nye kilder til slike data, som alle søker å enten utvide dekningen eller presentere nye typer metadata. Disse kildene kan enten potensielt erstatte etablerte kilder eller fungere som nyttige supplement til dem.

2.1 Etablerte bibliometriske kilder – referanseverdier

Meningen med dette notatet er å presentere nye kilder til bibliometriske data, men det kan være hensiktsmessig å se på noen av de mer etablerte kildene for å ha et sammenligningsgrunnlag. Her presenterer jeg kort de fire basene Web of Science, Scopus, Dimensions og Crossref for å kunne vise til noen referanseverdier for kvalitet og dekningsgrad.

2.1.1 Web of science

Web of Science, som eies av Clarivate, er den eldste og mest etablerte av de bibliometriske databasene. De dekker vitenskapelig publisering i et selektivt utvalg tidsskrifter, bøker og konferansebidragsamlinger. Deres filosofi om en kuratert, selektiv oversikt over kjernen av vitenskapelig publisering har opphav i grunnleggeren Eugene Garfields arbeid på seksti- og søttitallet med å identifisere en kjerne av vitenskapelige tidsskrift som sto for de viktigste bidragene til den vitenskapelige litteraturen (Garfield 1972). I takt med at databaser med langt større dekningsgrad har tatt opp konkurransen med WoS har de likevel utvidet dekningen av kjernebasen med supplerende baser som dekker flere fagfelt, et større geografisk område og flere tidsskrift som regnes som potensielt viktige om ikke i kjernen av kunnskapsproduksjon. I tråd med selektivitetskriteriet er det kun rundt 10 % av tidsskriftene som vurderes som blir indeksert i Core Collection (Clarivate 2020).

WoS er lukket lisensiert, og bruk av basen må skje i tråd med lisensavtalen som er inngått. Flere norske institusjoner har i dag lisens på WoS og den tilhørende analyseprogramvaren InCites, i stor grad gjennom den nasjonale konsortieavtalen for siteringsdatabaser. Den norske versjonen av WoS som ligger under Nasjonal infrastruktur for bibliometri gir anledning til å tilgjengeliggjøre publiseringsdata på aggregert nivå, men kun metadata på mest granulerte nivå med tilgangskontroll.

WoS indekserer totalt rundt 155 millioner dokumenter i ~34 000 tidsskrift, konferanseserier, bøker og datasett (Birkle, et al. 2020), med rundt halvparten av disse tilgjengelig i NIB. Dekningsgraden varierer med fagfelt, dokumenttype, geografi, tidsperiode og kildetype, men generelt er en tommelfingerregel at dekningsgraden minker jo lenger unna tidsmessig, geografisk eller tematisk man er nåtidig, engelskspråklig forskning innen naturvitenskap eller teknologi.

WoS baserer seg på en blanding av automatisk indeksering og manuell røkting av databasen. Fordi dette arbeidet har pågått over lang tid er det få feil og mangler i WoS, og metadatakvaliteten er høy også bakover i tid. WoS har enkelte problemer med nøyaktigheten på metadata for forhåndspubliserte artikler (Online First), samt mangler i indekseringen av alle referanser i enkelte artikler og korrekt navngivning på forfattere (feil i hhv. 1,9 % og 1,65 % av tilfellene, ifølge (Franceschini, Maisano og Mastrogiaco 2016)).

2.1.2 Scopus

Scopus, eid av Elsevier, kom som en konkurrent til WoS i 2004. Som kommersielt produkt er de det nest største per i dag. Scopus har hatt som mål å dekke mer av vitenskapelig publisering enn WoS, og baserer seg i større grad enn WoS på automatisk indeksering av publikasjoner. Scopus bygger på en rekke mindre indekseringsbaser eid av Elsevier, som EMBASE og GEOBASE. De kobler også basen i tettere grad til andre produkter de tilbyr til vitenskapelige institusjoner, blant annet referansehåndteringsverktøyet Mendeley og CRIS-systemet Pure. Samtidig har den elementer av selektivitet, med et eget fagfelleråd som legger føringer for hvilke tidsskrift som skal indekseres og ikke. Fagrådet aksepterer for indeksering rundt halvparten av tidsskriftene de vurderer (Elsevier 2017).

Scopus er lukket lisensiert, og bruk av basen må skje i tråd med lisensavtalen som er inngått. Den generelle lisensen til Scopus tillater deling av data på publikasjonsnivå i forskningsøyemed, men ikke til generelle analyseformål. Flere norske institusjoner har i dag lisens på Scopus og den tilhørende analyseprogramvaren SciVal, i stor grad gjennom den nasjonale konsortieavtalen for siteringsdatabaser.

I omfang er Scopus omtrent 10 % større enn WoS i absolutte tall, og omtrent 20 % større for nyere årganger (Visser, van Eck og Waltmann 2021). Scopus har flere

dokumenter enn WoS i alle publikasjonstyper og fagdisipliner. Akkurat som WoS er Scopus dominert av engelskspråklig forskning, med rundt 90 % av dokumentene i basen på engelsk.

Kvaliteten på metadata i Scopus er høy, med få og relativt ufarlige feil. Der WoS har enkelte tilfeller av underindeksering av referanser har Scopus relativt flere duplikater av publikasjoner (van Eck og Waltmann 2019), og har som WoS problemer med korrekt navngivning for forfattere og institusjoner utenfor Vesten (Selivanova, Kosyakov og Guskov 2019).

2.1.3 Dimensions

Dimensions er den nyeste av de «tradisjonelle» siteringsdatabasene. Den ble lansert av selskapet Digital Science (et selskap som igjen eies av Holtzbrinck Publishing Group, som også eier SpringerNature) i 2018. I tillegg til vitenskapelige publikasjoner inneholder basen også metadata om patenter, kliniske studier, datasett, finansieringskilder (NFRs prosjektbank er integrert i Dimensions), altmetric.com-data og politikkdokumenter (Hook, Porter og Herzog 2018). Dimensions er i stor grad basert på data fra Crossref (med noe data hentet fra PubMed), utvidet, beriket og koblet til andre datakilder ved bruk av maskinelle prosedyrer. Digital Science eier også altmetric.com-basen for sporing av ikke-akademiske referanser til forskning (se ellers 3.3.1 om altmetrikk), og denne er tett integrert i Dimensions. Dimensions gjør ikke redaksjonelle vurderinger av innholdet i basen på samme måte som Web of Science eller Scopus.

Dimensions fins i en begrenset åpen versjon, mens den fulle versjonen av databasen er lukket lisensiert. Bruk av basen til for eksempel større dataeksport må skje i tråd med lisensavtalen som er inngått, men det er mulig å få tilgang til basen i sin helhet til bruk i forskning, og Digital Science tillater også tilgjengeliggjøring av data på publikasjonsnivå i forskningsøyemed, men ikke til generelle analyseformål.

Dimensions har omtrent dobbelt så mange dokumenter som Scopus, mange av dem i kategorier som ikke dekkes av de andre basene i samme grad, som patenter, prosjektbeskrivelser, politikkdokumenter og kliniske studier (Bode, et al. 2019). Det er stor grad av overlapp mellom basene, og Dimensions inneholder 95 % - 97 % av publikasjonene som er å finne i Web of Science og Scopus (Visser, van Eck og Waltmann 2021), med sammenlignbare siteringstall (Thelwell 2018). Den ekstra mengden publikasjoner i basen i forhold til de to andre store basene fordeler seg relativt jevnt på artikler, bokkapitler og konferansebidrag.

Dimensions har som Scopus og formodentlig Web of Science hatt problemer med datakvaliteten i starten av databasens levetid. De har etter påpekninger av problemer med klassifiseringen på publikasjonsnivå (Bornmann 2018) gjort

endringer i metodene for disse, blant annet ved å gå bort fra å basere seg på et klassifiseringssystem laget for annen bruk og over til et egenutviklet system tilpasset dokumenttypene i basen (Herzog og Lunn 2018). Det har også vært spørsmål knyttet til variabel dekning av siteringsrelasjoner internt i basen, med en større andel ufullstendig indekserte referanselister enn for de to eldre basene (Stahlschmidt og Stephen 2020). Det er uklart om dette fortsatt er et problem, eller om Dimensions som Scopus har tatt grep for å forbedre datakvaliteten etter at den første runden med ekspansjon av basen er ferdig. Det er grunn til å tro at basen over tid vil forbedre kvaliteten også for eksisterende poster, i takt med at Digital Science kan vie mer ressurser til manuell kvalitetskontroll.

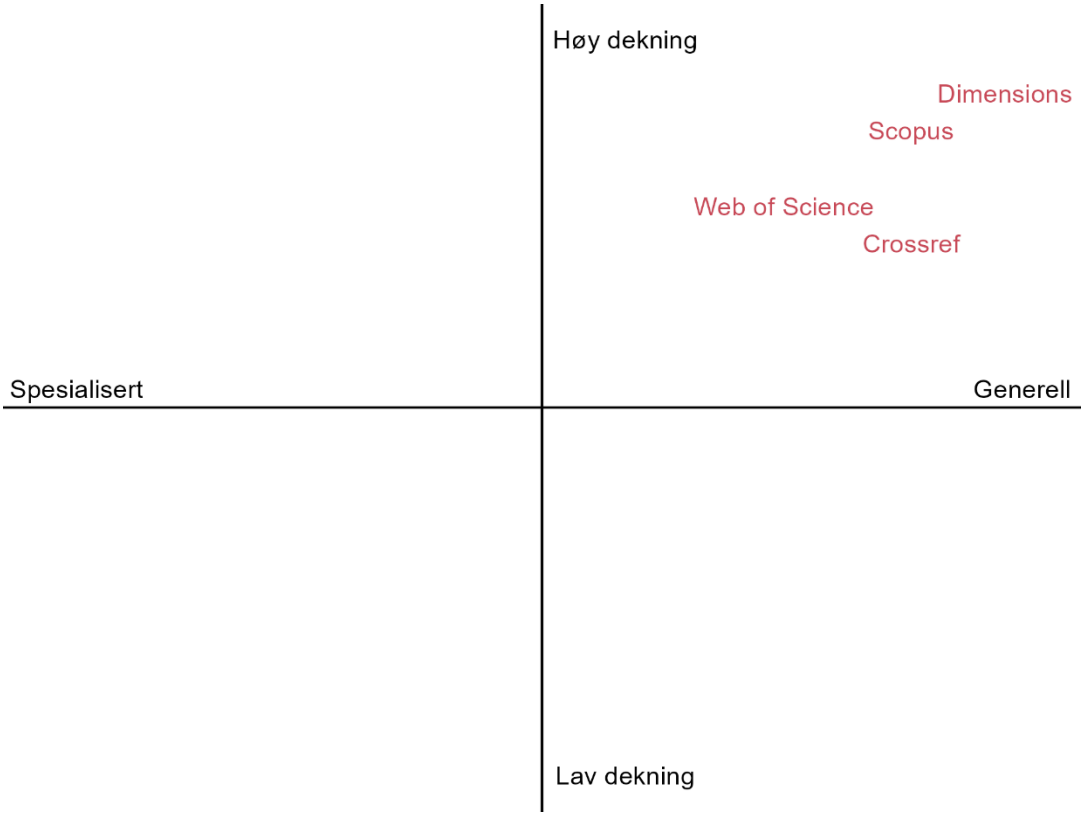
2.1.4 Crossref

Crossref er både en siteringsdatabase og en registreringstjeneste for persistente identifikatorer for forskningsresultater. Gjennom prosessen med å registrere en publikasjon med DOI gir forlag og tidsskriftutgivere fra seg metadata om publikasjonene som blir gjort åpent tilgjengelige. Metadata fra Crossref er åpent og gratis tilgjengelige.

Dekningen og kvaliteten på data i Crossref er avhengig av kvaliteten på det som blir registrert av forlag. Det er først nylig at alle de store forlagene har gått med på å tilgjengeliggjøre mer enn det absolutte minimum av metadata, så det er fortsatt spørsmål rundt omfanget av og kvaliteten på metadata. Kvaliteten har blitt forbedret for nyere årganger, spesielt for tidsskriftartikler (van Eck og Waltmann 2022), men Chudlarský og Dvořák (2020) finner fortsatt skjev dekning av fagfelt (mellom 78 % og 25 % dekning, avhengig av fagfelt) når siteringsrelasjoner funnet i Web of Science har blitt forsøkt gjenskapt.

2.1.5 Kvalitativ vurdering

Figur 1 gir en grafisk fremstilling av hvordan de fire etablerte kildene til bibliometriske data fordeler seg langs de to dimensjonene generalitet og dekningsgrad. Alle fire kilder vil egne seg godt som grunnlag for generell monitorering av vitenskapelig publiseringsaktivitet. Web of Science, Scopus og Crossref klassifiserer publikasjoner etter fagfelt på tidsskriftnivå, hvilket medfører noe merarbeid med å korrekt klassifisere enkelte publikasjoner som havner i svært generelle kategorier, mens Dimensions klassifiserer på publikasjonsnivå. Dette gjør kildene velegnet til komparative analyser og bruk av siteringsindikatorer som er sensitive til forskjeller i siteringspraksis på tvers av fag.



Figur 1. Anslag på generalitet og dekningsgrad for etablerte bibliometriske kilder

2.2 Nye kilder til bibliometriske data

Det har de siste årene dukket opp mange nye kilder til bibliometriske data, som til en viss grad er basert på en annen underliggende filosofi for dataforvaltning og -presentasjon. Noen av disse diskuteres under.

2.2.1 OpenAlex

OpenAlex er arvtakeren til Microsofts Academic Graph (MAG), selskapets forsøk på å lage en base for akademiske søk som kunne konkurrere med Google Scholar. MAG baserte seg i likhet med Scholar på høsting av metadata ved bruk av søkeroboter heller enn å høste fra faste lister over akademiske tidsskrift og lignende (Wang, et al. 2020). En slik prosedyre gjør basene større og i stand til å identifisere svært mye forskning som publiseres utenfor det tradisjonelle forlagsøkosystemet og ikke fanges opp av konkurrentene, men de fanger også opp mange dokumenter som mindre entydig kan kalles forskning.

Etter at MAG ble lagt ned i 2021 ble basen overtatt av et uavhengig team som har sikret finansiering til å bygge ut produktet med mer metadata, blant annet klassifisering av publikasjoner på dokumentnivå og direkte kobling til det

internasjonale institusjonsregisteret Research Organization Registry, og flere kilder. Metadata fra OpenAlex er åpent og gratis tilgjengelige.

OpenAlex har omtrent tre ganger så mange dokumenter som WoS og Scopus, og 50 % flere enn Dimensions, og er den største samlingen dokumenter produsert i forskningsøyemed som er tilgjengelig for bruk i bibliometriske analyser (Priem, Piwowar og Orr 2022). En av årsakene til dette er den språkagnostiske tilnærmingen til dataindeksering, som gjør at basen har langt bedre dekning av ikke-engelsk forskning enn konkurrentene.

Fordi OpenAlex er såpass ny er det få studier av datakvaliteten ennå. Visser et al. (2021) fant at MAG hadde høyere grad av manglende siteringsrelasjoner mellom dokumenter enn WoS og Scopus for årgangene 2018 og 2019, men det er usikkert om dette har blitt adressert i den nye utgaven av basen. En studie av forfatternavn i OpenAlex finner at også denne basen har problemer med korrekt tildeling av forfatternavn (Zhang, Lu og Yang 2022).

2.2.2 OpenCitations

OpenCitations er både en database med siteringsrelasjoner mellom dokumenter hentet fra flere åpne kilder og en organisasjon som jobber med infrastruktur for åpne metadata for forskningspublikasjoner. Basen vedlikeholdes av Research Centre for Open Scholarly Metadata ved Universitetet i Bologna, med finansiering fra OpenAIRE og Global Sustainability Coalition for Open Science Services (SCOSS). OpenCitations er ment som en komplementær tjeneste til de basene de lenker sammen, og basen består derfor av en kjernemodul med bibliografiske metadata som dekker alle dokumentene i basen og en gruppe modulære baser med siteringsdata hentet fra forskjellige tjenester. I skrivende stund består siteringsbasene av COCI, med Crossref-data (Heibi, Peroni og Shotton 2019), DOCI med data fra DataCite, POCI med PubMed-data og et folkeforskningsgenerert datasett ved navn CROCI (Heibi, Peroni og Shotton 2019). Metadata fra OpenCitations er åpent og gratis tilgjengelig under en Creative Commons CC0-lisens.

Som nevnt over er OpenCitations strukturert annerledes enn de andre, mer monolittiske siteringsdatabasene. Med separate datasett for hver kilde vil de forskjellige settene delvis overlappe, men også reflektere bedre fordelene og begrensningene i de underliggende datakildene. Det kan gi mer mening å forstå OpenCitations som et potensielt supplement til analyser som benytter seg av en av underlagskildene heller enn en kilde til bibliometriske data i seg selv.

2.2.3 Semantic Scholar

Semantic Scholar, eid av Allen Institute for Artificial Intelligence, tilbyr en bibliografisk database beriket med semantisk informert metadata om innholdet i publikasjoner, samt en base strukturert for maskinell tekstanalyse. Semantic Scholar fins i en begrenset åpen versjon, mens den fulle versjonen av databasen er lukket lisensiert. Bruk av basen til for eksempel større dataeksport må skje i tråd med lisensavtalen som er inngått, men det er mulig å få tilgang til basen i sin helhet til bruk i forskning. Semantic Scholar består av to moduler, en relasjonsdatabase som er på størrelse med Dimension og et korpus for kvantitativ tekstanalyse og tekst- og datauthenting (Wade 2022).

Det er gjort noen studier av kvaliteten på metadata i Semantic Scholar. Disse viser omtrent full dekning av sentrale publikasjoner i tekniske fag (Hannouse 2021), men bildet i humaniora og samfunnsfag er mer uklart. En gjenfinningsstudie for referanser finner at mellom 90 % og 99 % av referanser i basen lar seg gjenfinne, avhengig av fagfelt, men at rundt 4 % av publikasjonene som ble testet er duplikater eller versjoner av andre publikasjoner i basen (Hiltabrand 2022).

Semantic Scholar retter seg i stor grad mot analyser av innhold i og semantiske relasjoner mellom publikasjoner. Det gjør at basen i mindre grad egner seg til monitorering, men den kan være en verdifull for mer spissede analyser som søker å avdekke relasjoner mellom dokumenter utover siteringssignalet.

2.2.4 Scite

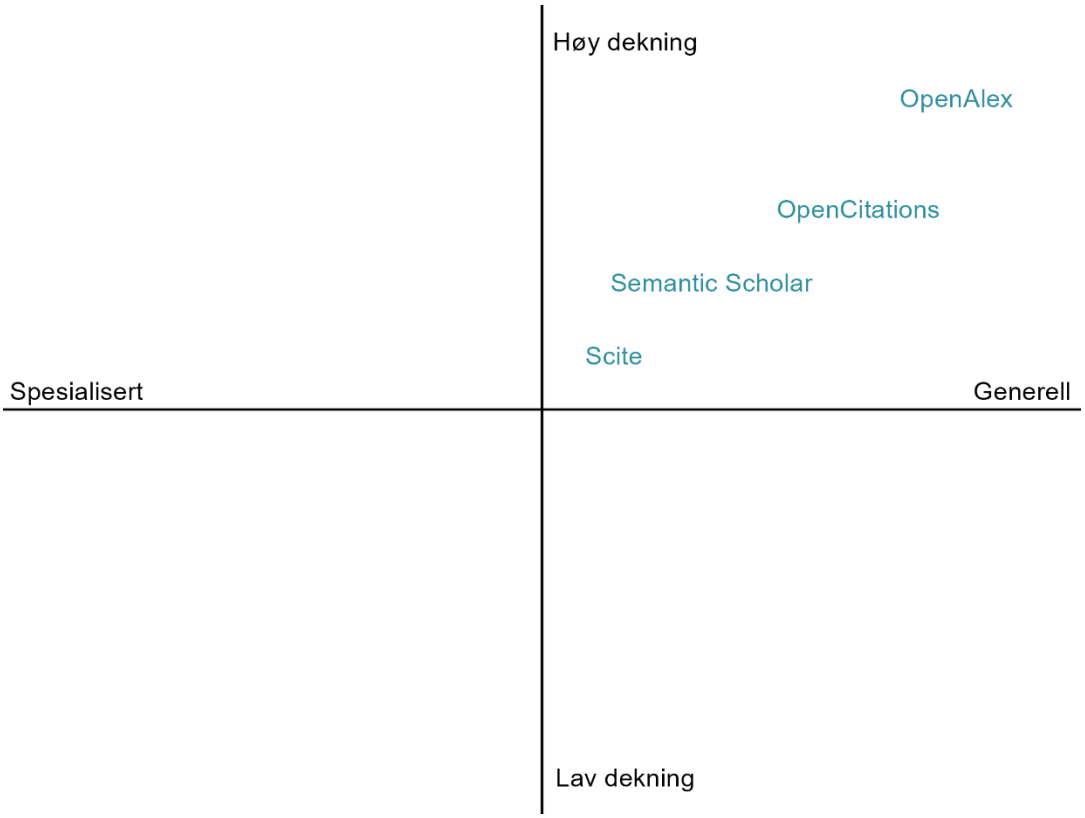
Scite er en tjeneste som leverer kontekstsensitive data om siteringsrelasjoner, basert på det semantiske innholdet i setningen en referanse opptrer i. Til dette bruker de en egenutviklet maskinlæringsalgoritme for klassifisering av tekst som ser på hvor i et dokument en referanse opptrer og holdningen til referansen slik det uttrykkes i setningen referansen opptrer i. Scite er lukket lisensiert, og bruk av basen må skje i tråd med lisensavtalen som er inngått.

Ved lansering hadde Scite i underkant av en milliard siteringer klassifisert etter hvor i artikkelen den opptrer, og hvorvidt referansen var støttende, kritisk eller nøytral (Nicholson, et al. 2021). I skrivende stund er antall siteringer mer enn doblet til 1,9 mrd., og antall publikasjoner dekket er 180 millioner.

Det er ikke gjort noen studier på kvaliteten på siteringskoblingene som fins i Scite, men studier av den underliggende maskinlæringsmodellen SciBERT viser at den i snitt feilklassifiserer tekst i rundt 15 % av tilfellene (Beltagy, Lo og Cohan 2019). Det er uklart om Scite har gjort forbedringer i sin implementering av modellen. Samtidig har verktøyet allerede blitt tatt i bruk i studier av uenighet i forskning (Rosali og Simboli 2023) og som supplerende metadata i tradisjonelle bibliometriske studier (Yeung, Cushing og Lee 2022, Suzan, et al. 2021).

2.2.5 Kvalitativ vurdering

Figur 2 gir en grafisk fremstilling av hvordan de fire nyere kildene til bibliometriske data fordeler seg langs de to dimensjonene generalitet og dekningsgrad. OpenAlex og OpenCitations vil egne seg godt som grunnlag for generell monitorering av vitenskapelig publiseringsaktivitet, mens bruk av Semantic Scholar og Scite bør begrenses til situasjoner hvor det er ønskelig med kvantitative vurderinger av semantisk innhold i tekst.



Figur 2: Anslag på generalitet og dekningsgrad for nye bibliometriske kilder

3 Kilder til data om annen forskningsaktivitet

Forrige kapittel diskuterte et sett med ressurser orientert mot analyser av tradisjonelle forskningsaktivitet i form av forskningspublikasjoner. Selv om basene har noe forskjellig filosofi rundt indeksering og kuratering og enkelte produkter baserer seg på å tilby ekstra metadata utover standardpakken er de relativt homogene i utforming og nytteområder, og kan brukes til mange av de samme analysene.

I dette kapittelet skal vi se på noen ressurser for å identifisere et bredere sett av indikatorer på forskningsaktivitet som går utover publisering av fagfellevurdert forskning. Dette kan være formidlingsaktivitet, sporing av innflytelse av publikasjoner utover referanser i forskningslitteraturen eller indikatorer på arbeidsinnsats knyttet til åpen vitenskap, som det er økt fokus på å dokumentere og evt. belønne.

Kildene til data som kan være indikatorer på slik aktivitet er langt mer heterogene enn kildene til bibliometriske data. Det er også i mange tilfeller større uklarhet rundt hva man måler, og ikke minst spørsmål om dekningsgraden av disse kildene. Det betyr at få av disse ressursene vil kunne egne seg som kilde til en løpende monitorering av aktiviteten slik for eksempel de bibliometriske kildene kan egne seg til. Samtidig kan de ha høy verdi som en legitimerende dokumentasjon på kvalifikasjon og egnethet på det individuelle nivå, selv om lav generalitet og dekning gjør sammenligning vanskelig.

3.1 Åpen vitenskap

Det er en generell bevegelse mot mer dokumentasjon av alle momenter som inngår i produksjonen av en forskningspublikasjon, utover den beskrivelsen av sluttresultatet som en publikasjon utgjør. Åpen vitenskap innbefatter det som trengs for å verifisere og validere prosessen fram mot endelig publikasjon, herunder dokumentasjon av forhåndsregistrering av prosjekter og protokoller for datainnsamling og tilgjengeliggjøring av underliggende forskningsdata og analysekode. Dette er prosesser som er nødvendige for å sikre kvalitet og legitimitet for

forskningsresultater, men som kan være krevende for forskere å bruke tid og ressurser på, spesielt i situasjoner hvor man får lite anerkjennelse for arbeidet.

I dette underkapitlet presenterer jeg fire typer kilder til data om åpen vitenskap-aktivitet, og vurderer i hvilken grad de kan egne seg som dokumentasjon av åpen forskningsaktivitet.

3.1.1 OpenAIRE

OpenAIRE er en europeisk infrastruktur for åpen vitenskap som dekker alle stegene i forskningsprosessen. De tilbyr egne tjenester for monitorering av forskningsresultater, inkludert manuskripter før fagfellevurdering, forskningsdata, programkode og andre supplerende materialer, samt en oversikt over vitenskapelige publikasjoner i åpne og lukkete kanaler. Monitoreringstjenestene er basert på OpenAIRE Graph, en modulær database som høster data fra åpne kilder og kobler dem til det europeiske organisasjonsregisteret, finansiørdata fra EUs forskningsprogrammer samt en del nasjonale forskningsfinansiører og enkeltforskere gjennom ORCID. Data fra OpenAIRE er åpent lisensiert og tilgjengelige uten kostnad i rå form, og under en lukket lisensavtale for tilpassete produkter.

I størrelse er OpenAIRE på linje med Scopus i antall publikasjoner, men med langt mer metadata om forskningsdata, programkode og finansierte forskningsprosjekter. Dette skyldes at mye av publikasjonsbasen har samme kilde som OpenAlex, Microsoft Academic Graph. Det er vanskelig å anslå den totale dekningsgraden for publikasjoner, og også omfanget av feil i basen, men det er indikasjoner på at koblingen mellom publikasjoner og prosjekter fra de inkluderte 26 finansiørene er bedre i OpenAIRE enn i EUs egen prosjektdatabase (Mugabushaka, et al. 2021).

Strukturen på basen gjør den i utgangspunktet velegnet for monitoreringsformål. De forskjellige dataenhetene er koblet mot globale standarder for persistente identifikatorer, og metadataskjemaet er det samme som er i bruk i det nye nasjonale vitenarkivet, som gjør sammenstilling med nasjonale datakilder relativt enkelt. Samtidig viser tidligere studier at avhengigheten av datadeling fra nasjonale kilder gjør at dekningen utover det som rapporteres til finansiører kan være mangelfull (Abad-Garcia, González-Teruel og González-Llinares 2018), selv om dette kan ha bedret seg i årene siden.

3.1.2 DataCite

DataCite er en tjeneste for registrering og sporing av metadata om forskningsresultater som datasett og andre supplerende materialer til forskningspublikasjoner. DataCite er mest å forstå som en registreringstjeneste som utstyrer datasett med persistente identifikatorer for gjenbruk i andre databasetjenester, på samme

måte som Crossref er det for dokumenter. Det er stor variasjon i hvordan metadata er registrert for postene i basen, og studier viser at det kan være problematisk å konsolidere data fra flere institusjoner til bruk i for eksempel indikatorproduksjon (Dudek, Mongeon og Bergmans 2019). Dette betyr at tjenesten per dags dato ikke egner seg til bred monitorering, selv om den kan være en viktig kilde til data for promoteringsformål.

3.1.3 Kodearkiver

Produksjon av analyse- og programkode og forskningsdata i forbindelse med forskningsprosjekter blir stadig vanligere, og det stilles også oftere krav om dokumentasjon av kode for databehandling og -analyse som underbygger publikasjoner. Slik dokumentasjon kan utgjøre en betydelig andel av innsatsen knyttet til en publikasjon, og dermed eksemplifisere bidraget til enkeltforskere involvert i prosessen. Det er også nødvendig for å sikre reproduserbarhet av forskningsresultater, og er slik en viktig del av begrunnelsen for bevegelsen mot åpen vitenskap i seg selv.

Det er mange arkiver hvor forskere kan opprette prosjektbaserte kodedepoter for deling av programkode og forskningsdata, og disse kan potensielt utgjøre viktige kilder til informasjon om slik aktivitet. Arkiver som Zenodo, Figshare og GitHub tilbyr alle både individuelle og institusjonelle løsninger for arkivering av programkode og forskningsdata, og gir også programmatisk tilgang til både metadata og innhold til de depoter hvor innholdet er åpent tilgjengelige. Dette letter arbeid med å identifisere slik aktivitet.

Selv om det i teorien er lett å hente ut metadata for kode og forskningsdata gjenstår det fortsatt utfordringer med å standardisere koblingen mellom forskningspublikasjon og supplerende materialer, samt en utvetydig kobling av bidragsyter til et kodedepot og medforfatter på en publikasjon. Standarder for persistente identifikatorer for integrasjon mellom et personregister som ORCID og brukerkontoer til slike tjenester er fortsatt under utvikling, og det vil ta tid før slik integrasjon kan garantere for gjenfinnbarhet og dekningsgrad i en slik grad at data om bidrag til programkode og forskningsdata kan brukes i monitoreringsøyemed. Samtidig bør slike bidrag kunne anerkjennes i en mer individuelle vurderinger av bidrag.

3.1.4 Forhåndstrykkarkiver

Bruken av nettarkiver for tilgjengeliggjøring av manuskripter før de er fagfellevurdert har økt kraftig de siste årene. Det gir forskere en mulighet til å offentliggjøre og spre publikasjoner når de er i nesten ferdig tilstand uten å måtte vente på den

tidvis sendrektige fagfellevurderingsprosessen før de kan få anerkjennelse for sine bidrag. Det er også en måte for forskere å gjøre hevd på funn med en pålitelig tidfesting av ideer, og fungerer i noen tilfeller som den faktiske publiseringsløsningen for vitenskapelige konferanser.

Arkiver for forhåndstrykk som arXiv, bioarXiv, SSRN og socarXiv er svært vanlig i enkelte fag, spesielt innen naturfag og informatikk, men blir også vanligere i samfunnsvitenskapene. Samtidig varierer dekningsgraden betydelig på tvers av fag, og legitimiteten til det som blir lastet opp i arkivene henger tett sammen med fagfeltets tradisjoner for å lese og reagere på det som blir lastet opp. I hvilken grad produksjon av forhåndstrykk gir grunnlag for personlig anerkjennelse er altså til en viss grad fagavhengig.

Arkivene er også i liten grad satt opp for høsting for bibliometriske formål. Selv om metadata for alle poster lar seg hente ut er det ingen standardløsninger for feltdefinisjoner og interoperabilitet på tvers av plattformer, og hvert arkiv holder seg med sine egne identitets- og institusjonsregistre. Det må derfor påregnes en betydelig innsats for å koble metadata fra arkivene med andre bibliometriske data, noe som reduserer muligheten til å bruke disse som kilder til kontinuerlig monitorering.

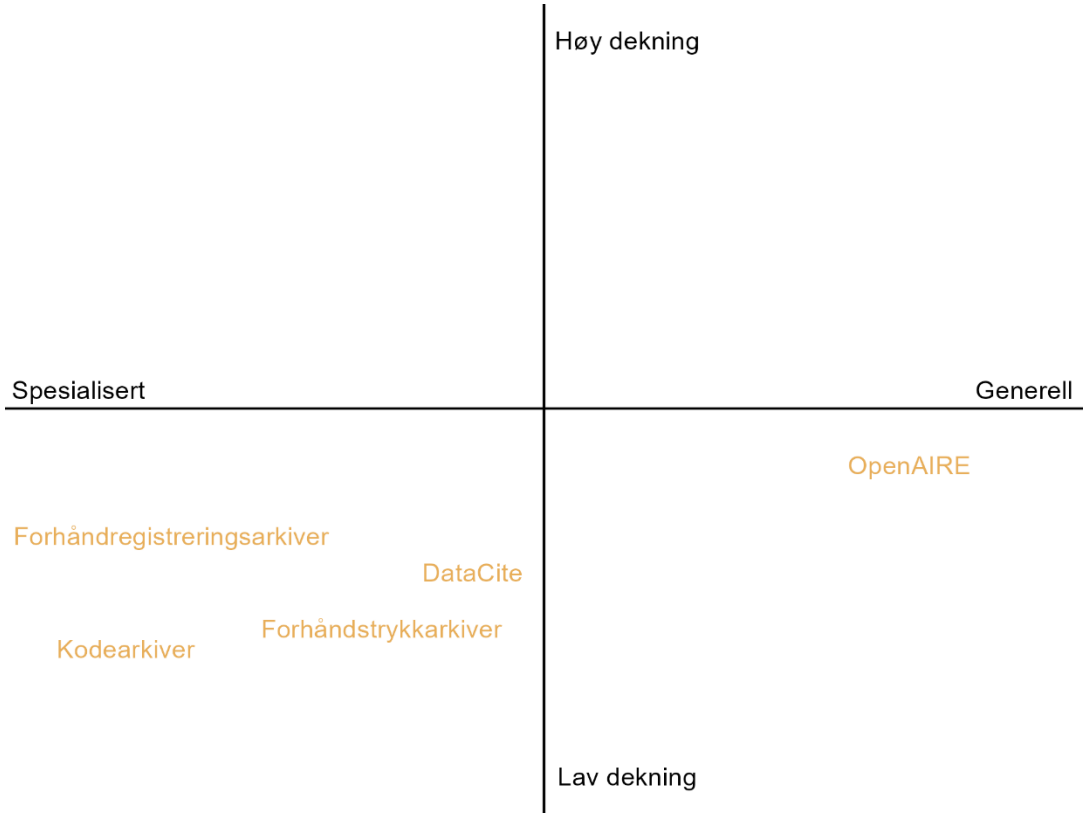
3.1.5 Forhåndsregistreringsarkiver

Et ledd i åpen vitenskap er dokumentasjon av selve forskningsprosessen, og i mange fag innebærer dette et behov for å klargjøre rammene for et forskningsprosjekt før prosjektet igangsettes, for eksempel for å tydeliggjøre analyseopplegg eller eksperimentelle parametere før innsamling og analyse av data begynner. Forhåndsregistrering av analyseopplegg skal hindre tilpassing av dataanalysen etter at data foreligger, noe som kan være et problem for enkelte former for forskning, og også for systematiske gjennomganger av eksisterende forskning. Det finnes derfor flere registre over forhåndsregistrerte analyser, ofte tilpasset behovene til faget registeret betjener.

Å produsere analyseprotokoller er tid- og innsatskrevende. Slike protokoller har form av små publikasjoner i seg selv, og har tydelig fordeling av forfatterskap og ansvarsområder. De egner seg derfor i utgangspunktet godt som underlag for informasjon om bidrag til et forskningsprosjekt som strekker seg utover produksjon av et endelig manuskript til en vitenskapelig publikasjon. Samtidig befinner forhåndsregistreringer seg på samme stadium som kodearkiver nevnt over – det er en bevegelse mot å kunne koble bidrag til forhåndsregistreringer til individuelle registre som ORCID (for eksempel krever tjenesten Open Science Foundation Preregistrations en slik kobling), men dekningen av slike koblinger er langt fra komplett.

3.1.6 Kvalitativ vurdering

Figur 3 gir en grafisk fremstilling av hvordan de kildene til data om åpen vitenskap-aktivitet fordeler seg langs de to dimensjonene generalitet og dekningsgrad. Generelt er dette kilder som både er spesialiserte i innretning og med usikker dekningsgrad også innenfor de fagene eller områdene hvor de kan være aktuelle å benyttes innen. Resultater hentet herfra bør kunne brukes i karrierevurderingsøyemed, men ikke til monitorering. OpenAIRE er som en generell database i utgangspunktet en kandidat for dette, men det er fortsatt noe usikkerhet knyttet til dekningsgrad og datakvalitet på denne kilden. OpenAIRE er også i rask utvikling og har sterk institusjonell støtte fra EU-apparatet, så denne kilden bør følges med på for fremtidig nytte.



Figur 3: Anslag på generalitet og dekningsgrad for kilder til data om åpen vitenskap

3.2 Formidling

Å kunne identifisere den delen av arbeidsinnsatsen til norske forskere som går til å formidle resultatene av egen forskning har lenge vært et mål i norsk forskningsforvaltning. Dette er formidling som går utover den rapporteringen av forskning som skjer gjennom vitenskapelig publisering, og som gjerne er rettet mot et

bredere og mindre spesialisert publikum. Innsatsen ligger i arbeidet med å tilrettelegge innsikter fra ens forskningsarbeid for et slikt publikum.

Tidligere forsøk på å opprette formidlingsindikatorer på linje med forskningsog undervisningsindikatorer har strandet på problemer knyttet til å definere hva formidling er og hvordan det kan måles. Samtidig har formidlingsdelen av forskeroppdraget fått økt oppmerksomhet i senere år. Det er ønskelig å kunne vurdere om kildene vi har til data om forskeres formidlingsaktivitet kan innlemmes i en infrastruktur for sporing av slik aktivitet, eller om kildene har store nok problemer med dekning og kvalitetssikring til at formidlingsinformasjon må henvises til vurdering på individnivå i særskilte tilfeller.

I dette delkapitlet diskuterer jeg tre kilder til formidlingsdata og diskuterer deres bruksområder og begrensninger.

3.2.1 Atekst

Atekst er et mediearkiv eid av Retriever som dekker nordiske og internasjonale nyhetskilder. Basen har omtrent hundre millioner medieartikler med muligheter for søk og programmatisk tilgang. Atekst er lukket lisensiert, og bruk av basen må skje i tråd med lisensavtalen som er inngått.

Det er ingen mulighet for å si noe om dekningen av norsk forskning i Atekst, men det er rimelig å anta at den er temmelig komplett for formålet å spore medietilstedeværelsen for norske forskere og forskningsinstitusjoner. Å oppnå funksjonelle treff i søkene krever mye innsats med kvalitetssikring av søkeresultater, spesielt for navnekobling, før disse kan inngå i en dataflyt for identifisering av formidlingsinnsatsen for forskere. Mange forskningsinstitusjoner har institusjonelle avtaler med Retriever om datatilgang, og bruker ressurser på å tilpasse søk etter de rette kombinasjoner av institusjons- og individnavn.

Gitt vanskene med å standardisere data fra Atekst for sammenstilling med andre data om forskere og deres aktivitet er det vanskelig å se for seg at det kan inngå i en fast dataflyt for registrering og rapportering av formidlingsdata. Atekst kan likevel være en nyttig kilde for punktrapportering og -evaluering som ikke inngår i en generell monitorering, så lenge det gjøres et grundig arbeid med å klargjøre data for slike formål.

3.2.2 Cristin utenfor NVI

Cristin blir hovedsakelig brukt som kilde til publiseringsdata, men er også den viktigste kilden til informasjon om forskeres formidlingsaktiviteter. 86 % av postene i basen er ikke NVI-publikasjoner, fordelt på 64 kategorier med «lisensiatavhandling» (49 poster) som den minste og «vitenskapelig foredrag» som den største

(387 340 poster). Disse postene er i mange tilfeller den eneste informasjonen som kan spores om mye av formidlingsaktiviteten til norske forskere, og gi utdypende informasjon om aktiviteter som i mange tilfeller vektlegges tungt i spørsmål om egnethet og produktivitet.

Det er stor usikkerhet knyttet til kvaliteten på postene som registreres utenfor det formaliserte NVI-systemet. Institusjonene har ingen forpliktelse til å kontrollere at opplysningene stemmer, og det er også frivillig å registrere formidlingsaktivitet der. Det er grunn til å tro at det er betydelig skjevfordeling i registreringsaktiviteten, og dermed også i mengden data som er å finne på forskeres aktivitet. Det er også uklart om kategoriene for registrering av resultater er dekkende for formidlingsaktiviteten til norske forskere. Dette gjør det vanskeligere å kunne basere seg på Cristin som en autoritativ kilde til formidlingsdata i monitoreringsøyemed.

Samtidig er det ingen tvil om at Cristin kan være en viktig kilde til formidlingsdata. Med en struktur som allerede er tilpasset tilordning av resultater til personer og institusjoner som allerede er definert i systemet slipper man mange av problemene som gjør seg gjeldende for en del av de andre kildene diskutert i dette notatet. Det er grunn til å være forsiktig med å bruke Cristin til monitoreringsformål på individnivå på grunn av faren for mangler i registreringen. Forskjellige institusjoner påskjønner registrering av formidlingsaktivitet forskjellig, og det er også ujevn fordeling mellom andelen registrerte NVI-publikasjoner og andelen formidlingsposter på tvers av institusjoner. Likevel kan Cristin sannsynligvis gi et godt bilde av formidlingsaktiviteten for større grupper forskere.

3.2.3 Institusjonelle arkiver

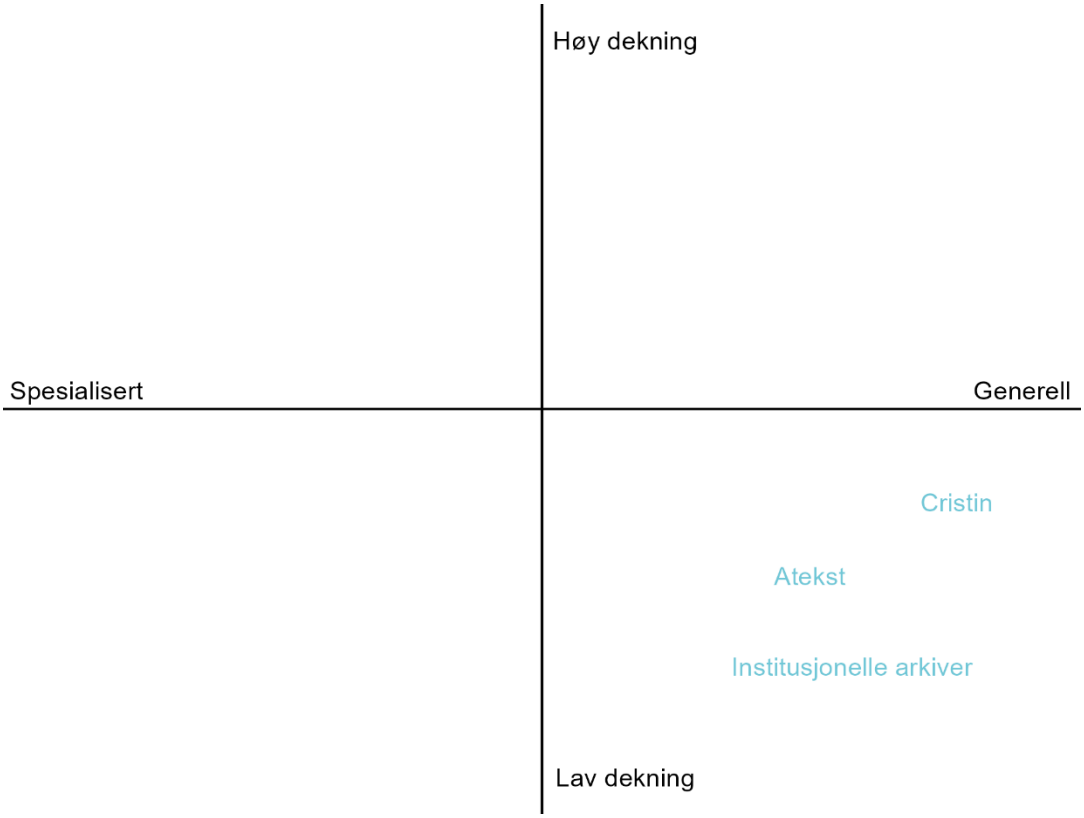
De fleste forskningsinstitusjoner i Norge holder seg med institusjonelle arkiver for lagring og tilgjengeliggjøring av faglig arbeid produsert ved institusjonen. Disse arkivene inneholder mange dokumenter av faglig og vitenskapelig art som ikke nødvendigvis ender opp som fagfellevurderte publikasjoner, for eksempel rapporter, utredninger, tekniske notater og lignende. Ved utdanningsinstitusjonene benyttes arkivene også til å publisere avhandlinger og masteroppgaver. Disse dokumentene utgjør potensielt en kilde til informasjon om forskningsaktiviteter som ikke nødvendigvis blir registrert i Cristin/NVI.

De norske institusjonelle arkivene benytter seg stort seg av den nasjonale arkivinfrastrukturen Brage, som er satt opp for enkel høsting av metadata og forvaltes av Sikt. Der det er relevant er også poster i arkivene koblet til sin respektive Cristinpost, slik at det er lett å koble relevante data. Samtidig er det stor variasjon i hvilken grad institusjonene benytter seg av arkivene til å tilgjengeliggjøre relevant arbeid, og det er også mye som ikke kan kalles faglig eller vitenskapelig som

eventuelt må filtreres bort før data kan brukes til å berike for eksempel Cristin-statistikk.

3.2.4 Kvalitativ vurdering

Figur 4 gir en grafisk fremstilling av hvordan kildene til data om formidlingsaktivitet fordeler seg langs de to dimensjonene generalitet og dekningsgrad. Merk at denne fremstillingen tar for seg registrert formidlingsaktivitet i Cristin, og ikke vitenskapelig publisering, hvor dekningsgraden på norsk forskning er svært god. Alle tre kilder er generelle i karakter med tanke på fagfordeling, men det er vanskelig å se for seg at de er dekkende for all aktivitet på denne fronten. Til monitorering av formidling er hver enkelt kilde uegnet, men alle tre kilder bør kunne være grunnlag for overordnete analyser av formidlingsaktiviteten til norske forskere.



Figur 4: Anslag på generalitet og dekningsgrad for formidlingskilder

3.3 Innflytelse

I tillegg til å registrere og anerkjenne produksjon og formidling av forskningsresultater er det mulig å måle forskjellige former for innflytelse denne produksjonen har. Historisk har dette blitt gjort gjennom å måle innflytelsen til en vitenskapelig

publikasjon på andre publikasjoner gjennom siteringsanalyser, men det er økende interesse for å innhente data om indikasjoner på former for innflytelse som har gjenklang hos et bredere publikum enn det akademiske.

3.3.1 Altmetrikk

Altmetrikk er en sekkebetegnelse for metrikk om vitenskapelige publikasjoner som går utover siteringer i den akademiske litteraturen. Altmetrikk søker å spore akademisk innflytelse i den generelle medieoffentligheten eller i andre kanaler som henviser til forskningslitteratur. Å kunne identifisere slikt gjennomslag kan være viktig for forskere, spesielt de som opererer i fag som har stort overlapp med den offentlige sfære, som samfunnsvitenskapene eller de samfunsrettede helsefagene.

Verktøy som PlumX, Mendeley og altmetric.com sporer referanser til forskningspublikasjoner i flere kanaler, både tradisjonelle medier og i nettfora hvor forskning ofte diskuteres, som Twitter og Facebook. De har også med noen sentrale politikk- og myndighetskilder, samt noe sporing av kilder som ofte omtaler akademisk litteratur, som Wikipedia og sider satt opp spesifikt for å diskutere akademiske publikasjoner. Alle tjenester for å spore altmetrics er lukket lisensierte.

Studier av disse databasen viser en betydelig dominans av sosiale medier (spesielt Twitter) som kilde (Torres-Salinas, Robinson-Garcia og Arroyo-Machada 2022), og også en klar slagside mot engelskspråklige kilder (Ortega 2020). Til gjengjeld gjør bruken av persistente identifikatorer og direkte koblinger mot bibliometriske databaser det relativt enkelt å koble altmetrikkdata på eksisterende metadata om forskningspublikasjoner.

3.3.2 Overton

Overton er en tjeneste for å spore referanser til forskning i politikk- og myndighetsdokumenter. Selv om denne typen referanser er et undersett av metrikken som tilbys av altmetricstjenestene nevnt over er basen både mer omfattende og mer spesialisert, med over ti millioner referanser til forskningslitteratur i basen, mot altmetric.com sine tre millioner. Overton er heller ikke like avhengig av direkte referanser til dokumenter med en DOI for å kunne registrere henvisninger til forskning. Metadata fra Overton er lukket lisensiert, og bruk av tjenesten må skje i tråd med lisensavtalen som er inngått.

Studier av dekningsgraden til Overton finner at basen har god dekning for fagfelt som typisk forventes å henvises til i offentlige og halvoffentlige publikasjoner, som for eksempel økonomi, statsvitenskap, helseforskning og miljøvitenskap (Szomszor og Adie 2022). Samtidig er det vanskelig å kontrollere for eventuelle

skjevheter i dekningen med henblikk på nasjonal og tematisk dekning. Metadata fra Overton er heller ikke koblet opp mot internasjonale standarder for interoperable identifikatorer, så det må regnes med arbeid med å koble for eksempel institusjonsnavn mot et eksisterende institusjonsregister.

3.3.3 Kvalitativ vurdering

Figur 5 gir en grafisk fremstilling av hvordan kildene til data om innflytelse fordeler seg langs de to dimensjonene generalitet og dekningsgrad. Her er det både generelle kilder innen altmetrikk som har møtt kritikk for skjev dekning og svært spesialiserte baser som Overton, som er den mest omfattende i sitt slag, men bare har nytte i spesifikke settinger.

Figur 5: Anslag på generalitet og dekningsgrad for kilder til data om innflytelse

4 Oppsummering

I dette notatet har jeg diskutert noen momenter til vurdering av kilder til data om forskningsaktiviteter. På det konseptuelle plan kan datakilder vurderes langs flere dimensjoner, hvorav de to viktigste er generalitet og dekningsgrad. Plasseringen av kilder langs disse to aksene dikterer hvilke formål kilden egner seg til. Figur 6 plasserer alle kildene diskutert i notatet på samme skjema, mest for å illustrere spredningen og diversiteten i kilder til data om forskningsaktivitet i dag.

Figur 6: Anslag på generalitet og dekningsgrad for alle kilder samlet

Det er nyttig å skille mellom bibliometriske analyser av forskningspublisering og analyser av annen relatert aktivitet. Når det kommer til bibliometriske analyser har det etter hvert dukket opp mange kilder som leverer data som i all hovedsak

er ganske likt innrettet, men hvor det er aspekter som underliggende databasefilosofi og innretning av metadata som skiller basene. For de eldste og mest etablerte databasene som Web of Science og Scopus har vi et godt etablert bilde av dekning og feilkilder i basen, mens dette i noe større grad er mer usikkert for kilder som Dimensions og OpenAlex. Til gjengjeld har de nyere produktene den fordel at de er bygd opp etter moderne prinsipper for utlevering og kobling av data, som i større grad er interoperable og kan brukes sammen med andre datakilder med mindre forarbeid enn basene som holder seg med egne, internt definerte identifikatorer på alle deler av basen.

Det er viktig at det er samsvar mellom hva kilden kan og ikke kan levere og formålet med datainnhentingen. Flere av databasene som leverer data som kan supplere tradisjonelle bibliometriske analyser har fortsatt ikke nådd en grad av dekning og kvalitet hvor de kan brukes til løpende kartlegginger av status for feltet. De egner seg heller som kilde til informasjon i punktanalyser av spesifikke miljøer hvor slike data kan ha spesiell betydning eller i prosesser hvor en person søker å sannsynliggjøre at de har bidrag til forskningen som ikke fanges opp av klassiske bibliometriske kilder.

Bibliometriske data har hatt lang modningstid, og har tidligere hatt mange av de samme problemene med validitet som plager nye metadata om forskningsaktivitet. Det er grunn til å tro at utviklingen mot et sett med standard indikatorer for ting som bidrag til formidling, åpen vitenskap og innflytelse utover den akademiske litteraturen vil fortsette. Enten vil dagens kilder til data om forskningsaktivitet utvikle seg i retning av å fange opp mer slik aktivitet og klassifisere den etter slike standarder, eller så vil nye kilder som gjør det dukke opp. Det er viktig å være klar over denne bevegelsen, ha et avklart forhold til hvordan slike kilder skal brukes i dag, og også være beredt til å inkludere slike kilder i bredere analyser etter hvert som de modnes.

Referanser

Abad-Garcia, M.-F., A. González-Teruel, and J. González-Llinares. 2018. "Effectiveness of OpenAIRE, BASE, Recolecta, and Google Scholar at finding spanish articles in repositories." *Journal of the Association for Information Science and Technology.* doi:10.1002/asi.23975.

Beltagy, I., K. Lo, and A. Cohan. 2019. "SciBERT: A Pretrained Language Model for Scientific Text." *arXiv.* doi:10.48550/arXiv.1903.10676.

Birkle, C., D. A. Pendlebury, J. Schnell, and J. Adams. 2020. "Web of Science as a data source for research on scientific and scholarly activity." *Quantitative Science Studies* 363-376.

Bode, C., C. Herzog, D. Hook, and R. McGrath. 2019. *A guide to the Dimensions data approach: A collaborative approach to creating a modern infrastructure for data describing research: where we are and where we want to take it.* Technical, Digital Science.

Bornmann, L. 2018. "Field classification of publications in dimensions: A first case study testing its reliability and validity." *Scientometrics.*

Chudlarský, T., and J. Dvořák. 2020. "Can Crossref Citations Replace Web of Science for Research Evaluation? The Share of Open Citations." *Journal of Data and Information Science* 35-42. doi:10.2478/jdis-2020-0037.

Clarivate. 2020. *Web of Science: Summary of Coverage.* Clarivate.

Dudek, J., P. Mongeon, and J. Bergmans. 2019. "DataCite as a Potential Source for Open Data Indicators." *Proceedings of the International Society for Scientometrics and Informetrics.* ISSI. 2037-2042.

Elsevier. 2017. *Scopus: Content Coverage Guide.* Elsevier.

Franceschini, F., D. Maisano, and L. Mastrogiaco. 2016. "Empirical analysis and classification of database errors in Scopus and Web of Science." *Journal of Informetrics* 933-953.

Garfield, E. 1972. "Citation analysis as a tool in journal evaluation." *Science* 471-479.

Hannouse, A. 2021. "Searching relevant papers for software engineering secondary studies: Semantic Scholar coverage and identification role." *IET Software.* doi:10.1049/sfw2.12011.

Heibi, I, S. Peroni, and D. Shotton. 2019. "Crowdsourcing open citations with CROCI -- An analysis of the current status of open citations, and a proposal." *Proceedings of the 17th International Conference on Scientometrics and Informetrics.* ISSI.

Heibi, I, S. Peroni, and D. Shotton. 2019. "Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations." *Scientometrics* 1213-1228. doi:10.1007/s11192-019-03217-6.

Herzog, C., and B. K. Lunn. 2018. "Response to the letter 'Field classification of publications in Dimensions: a first case study testing its reliability and validity'." *Scientometrics* 641-645.

Hicks, D., P. Wouters, L. Waltman, S. de Rijcke, and I. Rafols. 2015. "Bibliometrics: The Leiden Manifesto for research metrics." *Nature.* doi:10.1038/520429a.

Hiltabrand, R. 2022. "Assessing Near Duplicity and Document Linking Fidelity of the Semantic Scholar Open Research Corpus."

Hjørland, B. 2021. "Information Retrieval and Knowledge Organization:: A Perspective from the Philosophy of Science." *Information.* doi:0.3390/info12030135.

Hook, D. W., S. J. Porter, and C. Herzog. 2018. "Dimensions: building context for search and evaluation." *Frontiers in Research Metrics and Analytics.*

Mugabushaka, A.-M., M. Baglioni, A. Bardi, and P. Manghi. 2021. *Scholarly outputs of EU Research Funding Programs: Understanding differences between datasets of publications reported by grant holders and OpenAIRE Research Graph in H2020.* arXiv. doi:10.48550/arXiv.2109.10638.

Nicholson, J. M., M. Mordaunt, P. Lopez, A. Uppala, D. Rosati, N. P. Rodrigues, P. Grabitz, and S. C. Rife. 2021. "scite: A smart citation index that displays the context of citations and classifies their intent using deep learning." *Quantitative Science Studies* 882-898.

Ortega, J. L. 2020. "Blogs and news sources coverage in altmetrics data providers: a comparative analysis by country, language, and subject." *Scientometrics* 555-572. doi:10.1007/s11192-019-03299-2.

Priem, J., H. Piwowar, and R. Orr. 2022. "OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts." *Proceedings of the 26th International Conference on Science, Technology and Innovation Indicators.* STI.

Rosali, D., and B. Simboli. 2023. "Discovering substantive disagreement with review articles?" *arXiv.* doi:10.48550/arXiv.2303.04219.

Selivanova, I. V., D. V. Kosyakov, and A. E. Guskov. 2019. "The Impact of Errors in the Scopus Database on the Research Assessment." *Scientific and Technical Information Processing* 204-212.

Stahlschmidt, S., and D. Stephen. 2020. *Comparison of Web of Science, Scopus and Dimensions database.* German Centre for Higher Education Research and Science Studies (DZHW).

Suzan, V., B. B. Kanat, H. Yavuzer, and A. Doventas. 2021. "Themes and trends for osteoporosis: the bibliometric and altmetric approach." *Archives of Osteoporosis.* doi:10.1007/s11657-021-00983-w.

Szomszor, M., and E. Adie. 2022. "Overton: A bibliometric database of policy document citations." *Quantitative Science Studies* 624-650. doi:10.1162/qss_a_00204.

Thelwell. 2018. "Dimensions: A competitor to Scopus and the Web of Science?" *Journal of Informetrics* 430-435.

Torres-Salinas, D., N. Robinson-Garcia, and W. Arroyo-Machada. 2022. "Coverage and distribution of altmetric mentions in Spain: a cross-country comparison in 22 research fields." *Profesional de la información.* doi:10.3145/epi.2022.mar.20.

van Eck, N. J., and L. Waltmann. 2019. "Accuracy of citation data in Web of Science and Scopus." *Proceedings of the 16th International Conference of the International Society for Scientometrics and Informetrics.* 1087-1092.

van Eck, N. J., and L. Waltmann. 2022. *Crossref as a source of open bibliographic metadata.* MetaArXiv.

Visser, M., N. J. van Eck, and L. Waltmann. 2021. "Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic." *Quantitative Science Studies* 20-41.

Wade, A. 2022. "The Semantic Scholar Academic Graph (S2AG)." *Proceedings of the World Wide Web Conference.* New York: Association for Computing Machinery.

Wang, K., Z. Shen, C. Huang, C.-H. Wu, Y. Dong, and A. Kanakia. 2020. "Microsoft Academic Graph: When experts are not enough." *Quantitative Science Studies* 396-413. doi:10.1162/qss_a_00021.

Yeung, A. W. K., C. A. Cushing, and A. L. F. Lee. 2022. "A bibliometric evaluation of the impact of theories of consciousness in academia and on social media." *Consciousness and Cognition.* doi:10.1016/j.concog.2022.103296.

Zhang, L., W. Lu, and J. Yang. 2022. "LAGOS-AND: A large gold standard dataset for scholarly author name disambiguation." *Journal of the Association for Information Science and Technology.* doi:10.1002/asi.24720.

Figuroversikt

Nordisk institutt for studier av
innovasjon, forskning og utdanning

Nordic institute for Studies in
Innovation, Research and Education

www.nifu.no