

Teknisk Udvalgs Grundpapir

Version 1, den 24.9.2007

Grundpapiret har til formål overordnet at klarlægge Teknisk Udvalgs tekniske/praktiske forståelse af opgaven, således som den pt. er beskrevet i kommissorium og det generelle princippapir. Herved kan samspillet med Fagligt Udvalg fremmes, ligesom kommende, fokuserede konsulentundersøgelser kan basere sig på en sammenhængende beskrivelse af det totale system.

Grundpapiret skaber således et overblik over det samlede systemlandskab og opregner kortfattet for de enkelte delområder: dansk status, internationale aspekter, udviklingsbehov og udfordringer, behov for konsulentundersøgelse. Grundpapiret behandler i den sammenhæng følgende emner:

1. Systemlandskabet
2. Høst til nationale formål
3. Udvekslingsformat og vokabularer
4. Validering og deduplicering
5. Hjælpedatabaser og autoritetsdata
6. National analyse og beregning (bibliometrisk kvalitetsindikator)
7. National forskningsformidling (og international eksponering)
8. Universiteternes lokale forskningsdatabaser
9. Opsummering

Indledning

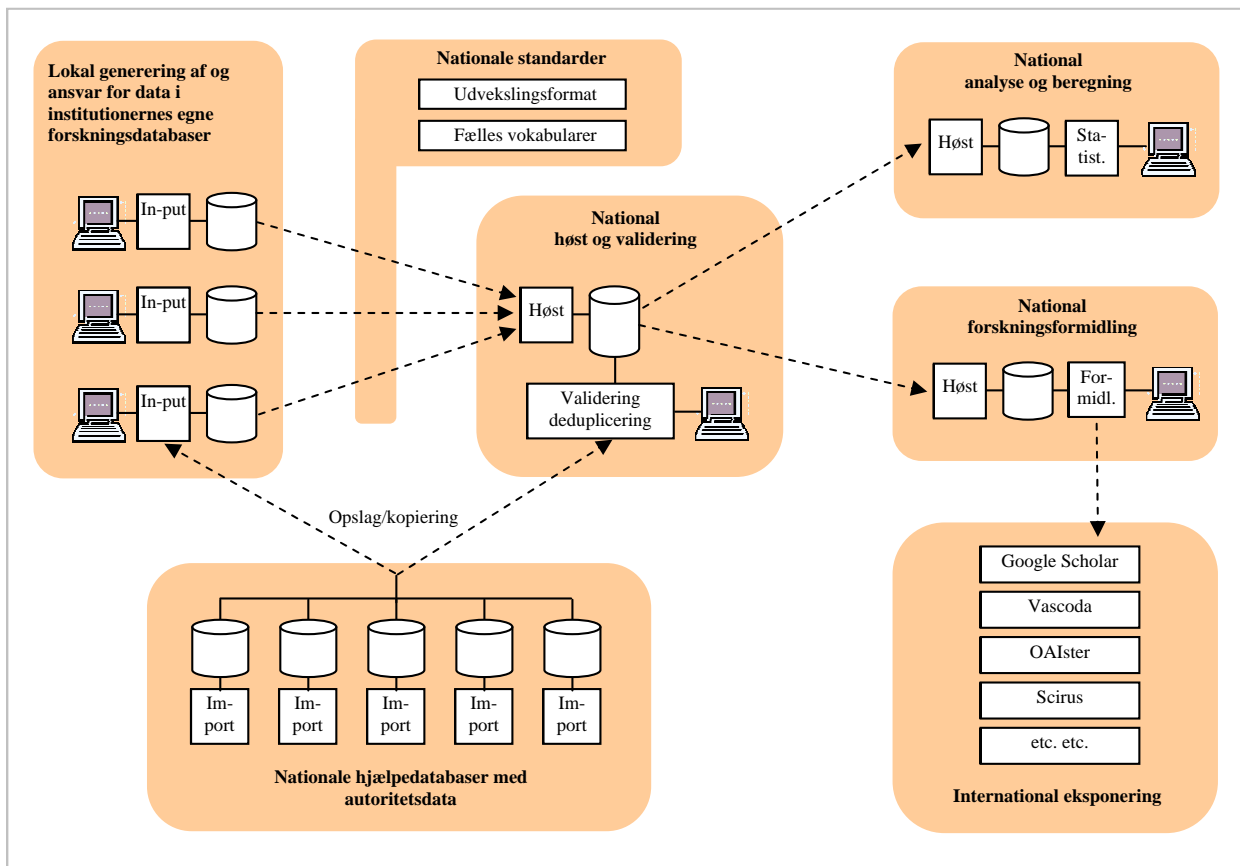
I Danmark er der en lang tradition for at registrere forskernes publikationer lokalt på universiteterne. Der har været forskellige formål og forskellige parter involveret i denne registrering. På den ene side har været universitetsbibliotekerne, som har registreret for at kunne udstille bibliografiske data om forskernes publikationer og for at kunne genfinde publikationerne og på den anden side administrationerne, som vil have information om forskernes output til ledelsesinformation. I dag har de fleste universiteters forskningsdatabaser en meget høj dækningsgrad, men det har krævet en stor indsats at forankre forskningsregistreringssystemerne i universiteterne og nå dette. Den tekniske og den proceduremæssige viden i dette arbejde har udviklet sig løbende og akkumuleret på hver sin måde i de to systemer Orbit og PURE. Men kendetegnet for begge er at de spiller sammen i form af enighed om en åben udvekslingsformat for metadata. Nedenstående systemlandskab er beskrevet ud fra at de eksisterende systemer og infrastruktur anvendes som en platform, der fra starten sikrer en national forskningsevaluering lokal forankring og en høj dækningsgrad.

1. Systemlandskabet

Diagrammet nedenfor illustrerer det samlede landskab opdelt i en række funktionsområder, hver med deres egen problematik og faglighed. I en driftssituation kan nogle af disse samles, men her behandles de separat for at sikre transparens og opretholde fleksibilitet for de senere faser i arbejdet.

- A. **Universiteternes lokale forskningsdatabaser.** Disse er basis for datagenereringen og dermed for det samlede overbygningssystem. I de senere år har alle universiteter installeret moderne systemer på området, nemlig enten PU:RE eller Orbit. Med en ny tjeneste med nationale hjælpedatabaser vil det lokale inddateringsarbejde kunne lattes og datakvaliteten øges.
- B. **Nationale udvekslingsformater og vokabularer for forskningsmetadata** sikrer en præcis og korrekt udnyttelse/repræsentation af universiteternes lokale data i det videre nationale system
- C. **National indsamling (høst) samt teknisk/bibliografisk validering af forskningsmetadata.** Denne centrale komponent indsamler løbende og automatisk de lokale data samt sikrer deres validitet i forhold til udvekslingsformat og vokabularer. Med en ny tjeneste med nationale hjælpedatabaser vil den automatiske validering kunne udvides væsentligt f.eks. hvad angår data om publiceringskanaler. I tillæg til validering vil der antagelig blive behov for deduplicering eller sammenkædning af metadata for den samme publikation - typisk når forfattere fra flere institutioner har arbejdet sammen. Dette kan understøttes af software, men vil givet også kræve menneskelig indsats for at opnå den tilstrækkelige præcision.
- D. **Nationale hjælpedatabaser med autoritetsdata.** Disse vil givet omfatte metadata om publikationskanaler som tidsskrifter, bogforlag og konferencer, hvortil yderligere kan tænkes bibliografiske data (fra ISI etc.) samt persondata.
- E. **National beregning af forskningskvalitetsindikator og -statistik.** Data udtrækkes af den validerede nationale indhøstning til forskningsadministrative formål i ministerieregi. Data behandles statistisk med henblik på at understøtte formålene for kvalitetsindikatoren og stilles til rådighed for universiteterne og de relevante administrative systemer i ministerieregi.
- F. **National forskningsformidling med internationale aspekter.** Data udtrækkes af den validerede nationale indhøstning til støtte for formålet national og international eksponering og formidling af danske forskningsresultater. Den nationale portal og søgemaskine indgår i samarbejde med Google Scholar og andre internationale platforme, der kan styrke eksponering og formidling.

Figur 1: Systemlandskabet



2. Høst til nationale formål

Beskrivelse af de nuværende erfaringer

Til høstning af bibliografiske metadata fra danske universiteters forskningsdatabaser anvendes i dag OAI-PMH¹ som standard for datakommunikation (protokol). Valget af OAI-PMH er i høj grad animeret af international erfaring og udvikling, samt at det er en protokol der er let at implementere for dataleverandørerne.

Med OAI-PMH kan én server hørste (service provider) opdaterede (nye, redigerede og slettede) poster fra én anden server (data provider). Protokollen tillader ligeledes at opdele data i særskilte sæt, som kan have forskellige egenskaber, et eksempel derpå er afsluttede års registreringer som vil kunne opdeles i sæt. Således kan høstningen være eksklusiv eller inklusiv forskellige sets.

OAI-PMH har været anvendt af Den Danske Forskningsdatabase (DDF) siden 2006 til at hørste data fra en række universiteter. Høstningen foregår i dag uden problemer, og de problemer der har været kan tilskrives fortolkninger af selve XML validerings schemaet DDF-MXD og ikke selve OAI-PMH standarden (valideringsproceduren beskrives i afsnit 4.).

Pt. understøtter alle PURE, Orbit/InSight institutioner (se mere om systemerne i afsnit 8.) høstning via OAI-PMH. Det tæller følgende institutioner:

¹ Open Archives Initiative Protocol for Metadata Harvesting. Siden besøgt den 11. september 2007: [http://www.openarchives.org/OAI/openarchivesprotocol.html]

| | | |
|---|---|--|
| Aalborg Universitet (PURE) - Statens Byggeforskningsinstitut | Aarhus Universitet (PURE) - Det Jordbrugsvidenskabelig Fakultet (Danmarks JordbrugsForskning) - Aarhus School of Business - Danmarks Pædagogiske Universitetsskole - Danmarks Miljøundersøgelser | Copenhagen Business School (PURE) |
| Danmarks Tekniske Universitet (Orbit) - Risø - Danmarks Fødevareforskning - Danmarks Rumcenter - Danmarks Fiskeriundersøgelser - Danmarks Transportforskning- | Københavns Universitet (PURE) - KU Life (KVL) - Farma (Danmarks Farmaceutiske Universitet) | IT-universitetet (I forhandling om anskaffelse af PURE) |
| Roskilde Universitet (PURE) | Syddansk Universitet (PURE) - Statens Institut f. Folkesundhed | |

Erfaringer internationalt

Specielt Holland og Norge har erfaringer med nationale forskningsregistreringssystemer. Ingen af disse lande har anvendt OAI-PMH som protokol til høstning ind i deres centrale forskningsdatabaser.

I Holland anvendes Metis systemet som forskningsregistreringssystem, da alle hollandske universiteter anvender samme system anvendes en "native" Oracle protokol til udveksling af data systemerne i mellem. Men til data udveksling med andre systemer, især formidlingsdelen, af det hollandske forskningsregistreringssystem anvendes OAI-PMH.

Den norske model bygger op omkring en række hjælpe databaser, der leverer data fra forskellige autoritetskilder, til de "lokale" registreringssystemer (Frida/forskdok) (se afsnit 5.). Disse data suppleres med manuelle inddateringer og validering af decentrale aktører. Systemet er implementeret i en Oracle database, som forklarer hvorfor lokale står i anførelsestegn, og hvert universitet er tildelt et lukket område i database, der kaldes Virtual Private Database, der er altså tale om én central database med lokal netbaseret adgang. Data til DBH (der er udgør datagrundlaget for fordelingen af midler) foregår ved manuel upload af data direkte til den centrale database. Der er dog planer om at systemer i fremtiden skal kommunikere ved hjælp af webservices.

OAI-PMH omtales generelt positivt internationalt, men der har været kritik af OAI-PMH, overvejende pga. sparsomme metadata af varierende kvalitet, det kan ses som et resultat af OAI's "low-barrier" politik hvor Dublin Core er et minimumskrav. I et nationalt set-up hvor metadatakrav og validering er centralt bestemt, vil problemet formentlig ikke være til stede.

Udviklingsbehov

Det vurderes ikke der er behov for udviklingen af en selvstændig protokol til høstning af data fra lokale forskningsdatabaser. OAI-PMH kan uden problemer skaleres og håndtere det antal universiteter og institutioner som findes i Danmark

Der skal ske en vurdering af hvordan de nationale hjælpedatabaser og autoritetsdata kan stilles til rådighed for de lokale databaser.

3. Udvekslingsformat og vokabularer

Beskrivelse af nuværende situation og erfaringer

I Danmark har man siden 2006 anvendt udvekslingsformatet DDF-MXD (Danish Research Database - Metadata Exchange Format for Documents) til overførsel af metadata fra de lokale forskningsdatabaser til den nationale. Formatet, der også omfatter et antal kontrollerede vokabularer, baserer sig på XML og er implementeret med en menneskelæsbar dokumentation² samt et maskinlæsbart XML skema³, der især spiller en rolle i forbindelse med automatisk validering af overførte metadata.

Formatet er udviklet i DEFF-regi i årene 2004-05 i et samarbejde mellem universiteter mv. og den nationale forskningsdatabase. Dette samarbejde bærer ligeledes formatets og skemaets videre udvikling.

DDF-MXD omfatter følgende standardiserede vokabularer:

- **Dokumenttyper** - pt. koder for 25 forskellige typer fra "tidsskriftartikel" til "udstillingskatalog"
- **Sprog** - i praksis alle koder i henhold til ISO standarden 639-2
- **Lande** - i praksis alle koder i henhold til ISO standarden 3166
- **(Peer)Review typer** - pt. koder for "peer review", "andet review", "intet review"
- **Litterært niveau/målgruppe** - pt. koder for "videnskabelig", "undervisningsrettet", "populærvenskabelig" og "administrativ"
- **Nøgleords- og klassifikationssystemer** - en række koder, der præciserer den anvendte tesaurus eller klassifikationsnøgle
- **Bidragyderroller** - en række koder, der nærmere definerer personers eller organisationers rolle og ansvar i forhold til en publikation
- **Publiceringsstatus** - pt. "indsendt", "accepteret", "i trykken", "publiceret"

De tre vokabularer for hhv. Dokumenttype, Review-type og Litterært niveau udgør tilsammen grundlaget for Rektorkollegiets sammensatte publikationstyper, der konkorderer med specifikke kombinationer af disse tre simple typer. Herved kan DDF-MXD dels på fleksibel vis tilpasse sig en udvikling af de komplekse typer (uden at behøve at ændre alle data retrospektivt) og dels lette valideringsopgaven, da det ofte er enkelt at afgøre, hvorvidt en simpel type er anvendt korrekt eller ej.

Efter indkøringsperioden i 2006 er formatet fuldt tilpasset og understøttet af såvel PURE som Orbit systemer og fungerer ganske udmærket sammen med protokollen OAI-PMH

Erfaringer internationalt

DDF-MXD er et af de første eksempler på et moderne forskningsregistreringsformat, der sigter på at tilgodese såvel de forskningsadministrative som de formidlings- og biblioteksmæssige behov. De fleste formater sigter snævert på det ene af disse aspekter, med parallelle systemer og parallelle/dublerede registreringsindsatser til følge. Typiske eksempler på denne opsplnitning er på den ene side de forskningsadministrative systemer, ofte refereret til

² http://mx.forskningsdatabasen.dk/mxd/1.1.0/DDF_MXD_v1.1.0.pdf

³ http://mx.forskningsdatabasen.dk/mxd/1.1.0/DDF_MXD_Schema_v1.1.0.3.xsd

som CRIS⁴, og ofte baseret på eller inspireret af CERIF⁵ formatet. På den anden side de forskningsformidlende og -arkiverende systemer, ofte refereret til som OAR⁶ og typisk baseret på mere biblioteksorienterede formater som Dublin Core, MARC⁷ og MODS⁸.

Der har i de senere år i Nordeuropa være en del fokus på at forene disse parallelle system- og format-udviklinger og dermed undgå en unødigt dobbeltindsats. Således har emnet været behandlet af Knowledge Exchange, det nationale samarbejde mellem Storbritannien, Tyskland, Holland og Danmark i regi af JISC, DFG, SURF og DEFF. Der er publiceret en rapport fra en ekspert workshop om emnet i februar 2007, hvor der slås til lyd for konkrete tiltag i retning af forening af og/eller øget interoperabilitet mellem CRIS og OAR. Knowledge Exchange forbereder pt. et projekt, der med udgangspunkt i DDF-MXD skal udvikle et fælles format for CRIS og OAR.

Udviklingsbehov

Der er næppe nogen tvivl om, at det nuværende udvekslingsformat (inklusive dets vokabularer) kan videreudvikles til at håndtere de nye faciliteter som forskningskvalitetsindikatoren måtte kræve. Der er allerede gode erfaringer på begge sider af dataudvekslingen med at justere format, vokabular og skema "in mid-flight" dvs. med alle systemer sat i løbende produktion, og de nye krav synes umiddelbart ganske overskuelige:

- tilføjelse af en deklaration af fagligt hovedområde med tilhørende vokabular
- muligvis en justering af definitionen af indberetningsår, der indtil nu har fulgt kalenderåret
- muligvis en mindre revision af vokabularerne for hhv. Dokumenttype, Review-type og Litterært niveau

⁴ CRIS = Current Research Information Systems

⁵ Common European Research Information Format i regi af foreningen euroCRIS - <http://www.eurocris.org>

⁶ OAR = Open Access Repositories

⁷ MACHine-Readable Cataloguing - <http://www.loc.gov/marc/>

⁸ Metadata Object Description Schema - <http://www.loc.gov/standards/mods/>

4. Validering og deduplicering

Validering deles op i to dele. Først validering af de høstede data, Dernæst sikringen af at dubletter elimineres i den centrale database.

Beskrivelse af nuværende erfaringer med validering:

Høstede data til Den Danske Forskningsdatabase valideres automatisk via DDF-MXD XML-Schemaet⁹. Høstningen foregår løbende og der er indarbejdet feedback rutiner mellem høster/validator og lokale forskningsdatabaser. Databaseadministratoren får automatisk besked fra systemet når poster fejler valideringen og forsøger i fællesskab med dataleverandørerne at finde årsagen til at posterne afvises af høsteren og løse problemet. Løsning af problemet kan enten være fejl på dataleverandørens side eller også en for striks validering på høsterens side. I forbindelse med DDF-MXD blev der i indkøringsperioden justeret på schemaets validering mod en lidt løsere validering for at nå frem til et niveau der fungerer for alle parter. Generelt forekommer valideringen robust, når ellers der er konsensus om fortolkningen af valideringsreglerne i schemaet.

De største problemer i forhold til validering ligger i det som ikke kan opdages via en automatisk validering i forhold til et schema. Dvs. forkert anvendelse af semantikken i XML specifikationen, men hvor syntaksen er præcis nok. Et eksempel kunne være forkert anvendelse af publikationskategorier. Pt. rettes fejl i anvendelsen af vokabularer kun på baggrund af henvendelser fra personer der har opdaget fejl i søgedelen af DDF.

Erfaringer internationalt

I den norske model valideres på baggrund af XML-schemaer. Først eksporteres ITAR-data med autoritetslister i sit eget XML-format¹⁰ til Frida, derudover har Frida sit eget XML-format¹¹, ligesom der til DBH, hvor de aggregerede publikationsdata uploades til, findes et XML-format¹². Erfaringen fra Norge viser at man kun til et vist niveau kan opnå kvalitetssikring via validering i forhold til schemaer og ved hjælp af autoritetslister, der må også indgå procedurer hvor personer er ansvarlige for vurdering af enkelte publikationer. Modellen fungerer sådan at forskerne selv må godkende og validere poster hentet fra ITAR til den lokale Frida-base, publikationer som ikke findes i ITAR må forskeren selv registrere, og tæller ikke med i DBH før de er blevet oprettet. Efterfølgende kvalitetssikring af de enkelte poster er forankret i forskerens eget institut¹³.

I Metis findes der i dag ikke nogen valideringsmekanisme af de data der uploades til den centrale forskningsdatabase eller høstes af Narcis¹⁴. Data kvaliteten der høstes fra Metis databaserne er derfor heller ikke af tilfredsstillende kvalitet, i forhold til søgning og verifikation. Udover den manglende validering af data kan problemet også skyldes at der mangler services som autoritetsbaser til at sikre ensartet brug af vokabularer. Dog har man i Holland meget godt styr på forfatteridentifikationen (se også afsnit 5.).

⁹ http://mx.forskningsdatabasen.dk/mxd/1.1.0/DDF_MXD_Schema_v1.1.0.3.xsd

¹⁰ <http://frida.usit.uio.no/itar/dokumentasjon/eksport/schema/>

¹¹ <http://frida.usit.uio.no/eksport/forskningsresultater/>

¹² <http://dbh.nsd.uib.no/dbhvev/dokumentasjon/vitpub/format/>

¹³ <http://www.uib.no/frida/2006info.htm>

¹⁴ <http://www.narcis.info/?wicket:interface=wicket-0:1::>

Udviklingsbehov

Erfaringen både nationalt og internationalt viser at der bør være en automatisk valideringsmekanisme for de data der indtastes i systemet og som høstes fra lokale baser til en eller flere centrale databaser. En måde er validering via et XML Schema, der i princippet kunne være DDF-MXD udvidet og tilpasset de nye krav. Men erfaringen viser også at den automatiske validering ikke kan stå alene, specielt hvis databasens succes afhænger af en høj grad af tillid til at data er korrekte. Den automatiske validering kan ikke sikre at et kontrolleret vokabular er korrekt anvendt. Datakvaliteten på et semantisk niveau kan sikres ved at uddelegere ansvar på den ene side lokalt på universiteterne og institutniveau, som må udpege ansvarlige for den lokale registrering, og på den anden side må der centralt være en eller anden form for revision der sikrer valide data.

Teknisk må systemarkitekturen bygges sådan at den understøtter de nødvendige forretningsprocesser der vil være i forbindelse med valideringen af data. I de lokale forskningsdatabaser Orbit/PURE findes der i dag workflows der kan understøtte denne validering. Først og fremmest vil nationale hjælpedatabaser med autoritetsdata indgå som et vigtigt element i systemet, der sikrer at data er konsistente og genbruges systemerne igennem (se afsnit 5) og gør det let for brugerne at indtaste data i systemet og samtidigt giver færrest mulige risiko for fejl tastninger.

- Der skal tages stilling til om DDF-MXD kan anvendes som udvekslingsformat mellem de lokale forskningsdatabaser og den centrale database. Hvis dette er tilfældet skal det vurderes hvilke metadata, specielt administrative metadata, DDF-MXD skal udvides med
- Der skal der tages stilling til om lokal human validering skal suppleres med en central validering. Der er samtidigt behov for en generel vurdering af hvordan valideringen af forskningsregistreringen kan foregå økonomiskforsvarligt og tillidsvækkende
- Det skal undersøges hvordan notifikation systemer fra den centrale database til de lokale databaser kan udvikles. Manuelle procedure eller automatiske beskeder.

Beskrivelse af nuværende erfaringer med deduplicering

I dag findes der ingen dedupliceringsmekanisme i Den Danske Forskningsdatabase. Det er et problem fordi der ikke sjældent publiceres publikationer med forfattere fra flere forskellige danske universiteter, det betyder at der med sikkerhed vil være publikationer som bliver registreret to gange eller flere og dermed opstår redundans i DDF. Også i de lokale systemer forekommer dubletter fordi personer der indberetter forskningen kan overse publikationer der allerede er blevet indberettet.

På dansk grund har Danbib, der også er datagrundlaget for netpunkt.dk og bibliotek.dk, udviklet teknikker til at matche poster. Matchingen/clustering er ikke en egentlig deduplikering, men en række prioriterede regler der sikrer at poster der *matcher* hinanden i forhold til disse regler er søgbare men samles og vises som én post i søgeresultatet¹⁵.

I den norske model fødes (Frida/forskdok) med data fra den centrale ITAR autoritetsbase, det betyder at de lokale institutioner kan søge i de allerede eksisterende data, dvs. at der er en større chance for at undgå at lave dobbelt registreringer. På trods af at man i udgangspunktet har adgang til de samme grunddata oplever nordmændene alligevel problemer med dubletter i

¹⁵ <http://www.danbib.dk/index.php?doc=match>

deres database. I evalueringsrapporten af erfaringerne fra 2006 efterspørges således procedure der kan opfange dubletter og søgefunktioner der let kan identificere dubletterne.

Ifølge KNAW¹⁶ der er vært for den hollandske forskningsportal Narcis, så oplever man også her dubletter fra deres høstning af Metis data. Og man har endnu ikke fundet en offentliggjort løsning på problemet.

Udviklingsbehov

I DDF har problemet med dubletter været overskueligt da der alene er tale om en base til formidling og eksponering af dansk forskning. Men når der kommer et evalueringskriterium indover data er der pludselig et helt andet krav til at redundans elimineres.

Der er som et minimum behov for at datakilderne til den centrale base hjælpes til at levere metadata af høj og konsistent kvalitet. Det vil kunne ske ved at de lokale databaser i høj grad trækker på hjælpedatabaser som beskrevet i afsnit 5. En rigere beskrivelse af de enkelte poster vil samtidigt gøre det væsentlig lettere automatisk at finde dubletter i systemet.

Derudover vil det være en stor hjælp hvis det på systemsiden gøres let først lokalt at finde dubletter ved at matche publikationer ud fra en række parametre (dublet-clustering / matching) og dernæst centralt opsamle dubletter der kan sammenføjes i klynger, således at de kun tæller en gang. Sidst vil det være nødvendigt at der indarbejdes procedure som understøttes af systemerne når dubletter opstår og rettes af en central instans, som sikrer størst mulig gennemsigtighed i processen.

Som en opsamling udviklingsbehov opstilles en model, som beskriver en mulig løsning for sikringen af datakvalitet i et løst koblet it-arkitektur baseret på distribuerede dataleverandører høstet til en central base

De lokale forskningsdatabaser høstes til den centrale database, hvor der sker en central matching og sammenføjning af dubletter. Et use case kan beskrives som følger: En post oprettes som det sker i dag i den lokale forskningsdatabase. Forskeren eller en medarbejder indberetter i den lokale database hvorefter posten bliver valideret lokalt og derefter udstilles klar til høstning. Den centrale database høster løbende poster. Når posten er høstet over i den centrale database matches denne med evt. andre poster ud fra nogle forud definerede regler. En administrator og ansvarlig for den centrale base gøres opmærksom på at der er en post der matcher en anden post i basen. Her kan man tænke sig to scenarier; 1. den central administrative person tager beslutningen om der er tale om en dublet, eller 2. personen som oprindeligt har oprettet dubletten bedes tage stilling til om de to poster, som matcher kriterierne er dubletter. Derefter følger at hvis der sker efterfølgende ændringer til posten i et af de lokale systemer, så kan ændringerne enten tilføjes den sammenføjede post i den centrale database, det sker ved at systemet kan holde sammen på posterne vha. den lokale databases unikke post id kopieres ind i den sammenføjede posts administrative metadata i den centrale database.

Som et supplement til hjælpe data kan der også slås op i den centrale database via en webservice fra de lokale forskningsdatabaser, findes den post som man vil oprette allerede vil den kunne hentes over i det lokale system. I de lokale poster tilføjes et unik-id i de administrative data, som posten har med fra den base hvor posten blev født eller fra den

¹⁶ <http://www.niwi.knaw.nl/english/index.html>

centrale database, derved kan posterne sammenkædes igen, når dubletten bliver høstet af den centrale database. Det vil ikke synkronisere data men vil gøre det lettere at matche data centralt og dermed fjerne dubletter fra en statistik.

Eksemplet præsenteret her, samt erfaringen fra udlandet viser at uanset hvilken løsning der vælges vil der være behov for en menneskelig indsats for at sikrer færrest mulige dubletter i den centrale database.

- Der skal tages stilling til hvilke administrative metadata der er brug for, for at håndtere dubletter i systemerne
- Det skal undersøges nærmere hvordan dubletter, håndteres bedst muligt i en central database herunder:
 - Hvilket software der kan håndtere opgaven?
 - Hvilke regler matchingen/clusteringen skal følge?
 - Hvilke procedure og notificeringsmekanismer skal indgå?
 - Og hvordan skal den menneskelige indsats anvendes?

5. Hjælpedatabaser og autoritetsdata

Til støtte for den lokale forskningsregistrering på universiteterne, skal der opbygges en infrastruktur med en række hjælpedatabaser og autoritetsdata. Disse skal anvendes af universiteterne, når de registrerer forskernes publikationer i lokale forskningsregistreringssystemer.

De overordnede formål med denne infrastruktur er:

- Sikring af datakvalitet på tværs af universiteterne. Der skal sikres et ensartet og korrekt datagrundlag, som er forudsætning for en korrekt national beregning af forskningsindikator.
- Workflow og indberetningspraksis er forskellig fra universitet til universitet. Der er mange forskellige personer, som er involverede, og ikke alle universiteter har mulighed for at tilbyde stringent bibliografisk kontrol med indberetninger. Det er derfor vigtigt, at der etableres en teknisk infrastruktur, som vil understøtte en korrekt registrering af publikationer.
- Hjælpedatabaser/autoritetsdata, som skal kobles til det system som anvendes til registrering, vil sikre at opdaterede og vedligeholdte data altid er tilgængelige lokalt hvor publikationer registreres. I afsnit 8 beskrives hvordan de 2 systemer i brug (PURE og Orbit) kan anvende hjælpedatabaser og autoritetsdata.
- Hjælpedatabaser/autoritetsdata vil være en vigtig ressource for de centrale valideringssystemer, der med træk på disse vil kunne forbedre den automatiske validering meget væsentligt

I den lokale registrering af publikationer skal hjælpedatabaser og autoritetsdata også bruges til at afhjælpe registreringsprocessen, ved at f.eks. centrale hjælpedatabaser kan kobles til lokale registreringssystemer, og sørger for at bestemte felter kan udfyldes automatisk. På denne måde sikres en kvalitetsforbedring af data og en hurtigere registreringsproces.

Typer af hjælpedatabaser og autoritetsdata:

Hjælpedatabaser/autoritetsdata kan inddeles i 3 overordnede typer:

- A. Publiceringskanaler
- B. Data genbrug
- C. Personer og organisationer

A. Publiceringskanaler:

I en central database med ”publiceringskanaler” er registreret de titler, hvori forskernes forskningspublikationer er publiceret, såsom tidsskrifter, monografier og monografiserier, og som skal anvendes i den nationale forskningsindikator. Databasen skal bruges til korrekt og kontrolleret angivelse af f.eks. titler (tidsskriftets titel, serietitler osv.), status vedr. peer reviewed/ikke peer reviewed, data til brug i beregning af indikator (f.eks. hvis publiceringskanaler inddeles i grupper efter niveau eller lign.), emne, impact factor, sprog, land, Open Access betingelser og rettigheder.

En central database over publiceringskanaler skal kunne kobles til de lokale systemer, så ajourførte data altid er tilgængelige i den lokale registreringsproces.

Der skal udarbejdes et metadataformat for databasen for publikationskanaler, som dækker behovet for både den nationale forskningsindikator og lokale behov, f.eks. medtagelse af metadata elementer som kan anvendes i andre bibliometriske analyser, såsom impact factor, sprog mm.

Metadataformatet for et centralt register over publiceringskanaler skal fastlægges, registret skal oprettes og der skal etableres vedligholdsrutiner for basen og rutiner for synkronisering med decentrale systemer. Det er vigtigt at decentrale systemer inddrages i etablering af basen, både mht. metadataformatet og behovet for løbende synkronisering.

Tidsskriftsdata:

Alle lokale forskningsdatabaser skal registrere publikationer med standardiserede data vedr. publiceringskanaler. Det vil sikre, at publikationer registreres med den korrekte angivelse af tidsskriftstitel, ISSN (hvor der skal tages højde for forskellig anvendelse af ISSN versioner til e- og papir-versioner), at information vedr. peer review status for hvert tidsskrift anvendes ensartet og indgår i den nationale beregning på et ensartet grundlag for alle universiteter osv.

Der skal tages stilling til leverandører af tidsskriftsdata. Af muligheder kan nævnes det internationale ISSN register, Ulrichs, Ebsco og Swets, som er store internationale database med tidsskriftsoplysninger, Sherpa/Romeo databasen vedr. Open Access rettigheder og ISI-basen fra Thomsen med cirka 5.000 titler, som vil være værdifuld aht. angivelse af peer review status. Antagelig skal flere datakilder importeres og flettes.

Der skal tages stilling til behov for berigelse af tidsskriftsdata, f.eks. kobling til autoritetsdata vedr. peer review status, data som kan bruges i andre bibliometriske analyser, lokale behov osv.

Bogforlagsdata:

Monografier og monografiserier håndteres som tidsskriftsdata, for at sikre ensartede registreringer, der er grundlag for de nationale beregninger. Der skal tages stilling til en leverandør af monografidata, som dækker denne type publiceringskanal. Det vil være relevant at undersøge bogforlagsdata, ISBN og national bibliografisk data.

Konferencedata:

Konferencedata er dels data for artikler publiceret i Proceedings. Status vedr. peer review vil også gælde for konferencedata. Der vil være overlap i data vedr. tidsskrifter, monografier, monografiserier og konferencedata og der skal sikres en komplet og autoritativ dækning.

Konferencedata er desuden præcise data om selve konferencen, som titel, sted, dato, peer review status, arrangør mv. Disse data kan antagelig importeres fra en eller flere databaseproducenter som British Library, InterDok, Allconferences.com, Confabb etc. og stilles til rådighed som en hjælpedatabase, der sikrer ensartede og korrekte registreringer.

Konferencebidrag volder ofte problemer i registreringssystemer, grundet upræcis identifikation af konferencen samt bidragenes meget forskelligartede form og publicering.

I Norge er der etableret en central database over publiceringskanaler¹⁷, som kan indgå i de undersøgelser vedr. oprettelsen af en dansk central database over publiceringskanaler, som dækker ovenstående typer.

B. Data genbrug:

Spørgsmålet om genbrug af publikationsdata fra andre databaser rejstes jævnligt. Den første kilde til datagenbrug bør være forskningsdatabasen selv. Denne udnyttelse vil være relevant i forbindelse med forfatterskaber, der går på tværs af institutioner og vil lette såvel det lokale registreringsarbejde som den centrale deduplicering.

De relevante eksterne databaser, hvorfra der kan købes/hentes data, er bibliografiske poster fra ISI's databaser (Science Citation Index, Social Sciences Index, og Arts and Humanities Index), PubMed og mange andre domænespecifikke databaser samt DEFF's kommende nationale "databrønd" for artikeldata, der vil dække på tværs af forlag og licenspakker.

I den norske model bliver bibliografisk data købt fra bl.a. ISI og overført til den centrale database ITAR (Importtjeneste og Autoritetsregistre). Efter central validering med autoritetsdata bliver bibliografisk data distribueret til de lokale systemer.

Et vigtigt element i genbrugsdata vil være angivelse af **publikationstyper** (peer reviewed artikler, bøger osv.) i de databaser hvor fra der hentes data. Der skal mappes fra de publikationstyper, som anvendes i disse baser til de publikationstyper, som anvendes i beregning af den nationale indikator.

Der skal tages stilling til, om genbrug af data skal anskaffes centralt eller om det skal være op til universiteterne selv at etablere et workflow, som sikrer dækning og validering op imod autoritetsdata, som stilles til rådighed i ministerieregi.

Mht. validering af genbrugsdata (uanset hvordan data fremkommer) skal en valideringsproces omfatte forfatternavne, publiceringskanaler, publikationstyper, organisationsnavne mm. og posterne skal forsynes med data i standardiseret form (f.eks. publiceringskanaler, person- og organisationsnavne tilføjes i den form, som anvendes i autoritetslisterne).

De faktorer som kan indgå i beslutningen vedr. central/decentral anskaffelse og validering af genbrugsdata, er f.eks. de eksisterende workflow på universiteterne, hvor systemerne efterhånden er meget standardiserede (og i mange tilfælde ens). Der arbejdes lokalt på en løsning, hvor forskerne selv kan downloade og arbejde videre på data fra andre databaser. Der skal også undersøges om en central løsning vil medføre forsinkelser i den lokale registrering, som ikke opvejes af fordele.

C. Personer og Organisationer:

Entydig identifikation af personer og sikring af, at en bestemt forfatter er den samme forfatter som har produceret et bestemt sæt publikationer ("disambiguation") er af afgørende betydning for levering af valide data. Personnavne udformes og staves på mange forskellige måder og de ændres over tid.

¹⁷ <http://dbh.nsd.uib.no/kanaler/>

Et vigtigt element i den korrekte registrering af forskernes publicering, er valide organisatoriske betegnelser og der skal sikres valide tilknytninger mellem personer og organisationer uafhængig af publikationerne. Organisationer kan være anført forkert i publikationerne og kan dermed forveksles med andre institutioner osv.

Der skal udarbejdes et metadataformat for en autoritativ liste over organisationsnavne, og der skal sikres samspil og synkronisering med decentrale systemer. Det er vigtigt, at organisationsnavne fastlægges på både dansk og engelsk, da publicering ofte er international.

Der skal tages stilling til om hjælpedatabaser og autoritetslister vedr. personer og organisationer skal placeres centralt, decentralt eller begge dele. I Norge anbefalede Universitets- og højskolerådet (UHR) i deres indstilling oprettelsen af et centralt register over ansatte (forfattere), men det norske tekniske udvalg valgte ikke at forfølge sagen pga. usikkerhed vedr. persondataloven. I Danmark i dag varetages den entydige kobling mellem person-publikation-organisation decentralt af universiteterne i egne systemer. I Holland har 12 universiteter samt en række organisationer gået sammen om at etablere en central tjeneste - DAI – Digital Author Identification - som entydigt identificere forskere og forfattere.

I Danmark, som nævnt i det forrige afsnit, er systemerne meget standardiserede, i mange tilfælde ens og er i forvejen koblet til lokale matrikelsystemer (eller kan blive koblet), hvor der hentes og opdateres data automatisk. Her er systemer mindre sårbare overfor manuel registrering af forskere og forfattere.

Vedr. alle hjælpedatabaser og autoritetsregistre:

I opbygning af den tekniske infrastruktur som skal støtte forskningsregistrering, skal der tages stilling til, hvad der skal håndteres centralt (i ministerieregi) og hvad der skal håndteres decentralt (på universiteterne, evt. også på "systemniveau"). Der skal træffes en beslutning om den endelige model samt fordeling af ansvar for etablering, vedligeholdelse og synkronisering af de forskellige elementer. Der skal tilstræbes en opbygning, som tager hensyn til eksisterende forhold og eksisterende workflows, som minimere afhængighed og maksimere validitet i data. Mht. det decentrale niveau, vil det være op til systemleverandører og deres kunder at specificere, hvordan de decentrale instanser skal integrere hjælpedatabaserne og autoritetsdata i universiteternes registreringsmiljøer.

Brug af unikke identifikatorer (UI)

Unikke identifikatorer kan bruges til entydig og vedvarende identifikation af entiteter på tværs af systemer og organisatoriske enheder. Der kan med fordel genereres identifikatorer for f.eks. personer (forfattere i Danmark), til at afhjælpe problemer forbundet med forskellige navneformer, stavemåde, navneskift osv. Der skal undersøges, om et personregister kan oprettes centralt eller decentralt og sammenhæng med CPR-nr.

UI kan også bruges til entydig identificering af organisatoriske enheder over tid og f.eks. til entydig kobling af personer (forfattere) til organisatoriske enheder. Et tæt beslægtet område som anvender UI er i forlagsbranche, hvor anvendelsen af Digital Object Identifiers (DOI) er udbredt til vedvarende identificering og lokalisering af f.eks. artikler i fuldtekst og til adgangskontrol til licensbelagte ressourcer.

En arbejdsgruppe nedsat af IT- og Telestyrelsen har udarbejdet et forslag til anvendelsen af Universal Unique Identifier (UUID) til identifikation af digitale objekter. I december 2006 fik arbejdsgruppen forslaget godkendt i Datastandardiseringskomiteen som en OIO standard på området.¹⁸ I det videre forløb skal behov for ID'er undersøges og denne standard skal tages i betragtning, selvom anvendelsesområde ikke er det samme.

Udviklingsbehov

Et centralt element i beregning af den national forskningsindikator er autoritetslisten over publiceringskanaler. Det vurderes, at denne eller disse hjælpedatabaser skal etableres centralt. I denne forbindelse er der behov for følgende udviklings- og undersøgelsesaktiviteter:

- Der skal udarbejdes et metadataformat for en database over publiceringskanaler som tidsskrifter, bogforlag og konferencer.
- Metadataformatet skal dække behovet for beregning af den nationale forskningsindikator, men da data vedr. publiceringskanaler er vitale i lokale systemer, vil det være nærliggende at afdække de behov, som lokale systemer måtte have, som med fordel kunne placeres i den centrale database og hentes derfra til anvendelse i de lokale registreringsprocesser.
- Det skal undersøges, hvilke leverandører der kan levere data, som dækker det dokumenterede behov.
- Det skal udarbejdes et workflow for den løbende vedligeholdelse af den centrale database og med placering af ansvar for vedligeholdelse. Lokale systemer opdateres løbende og krav til ajourført lokal registrering vil formentlig vokse, efterhånden som der kommer stigende fokus på området.
- Der skal udarbejdes en kravspecifikation til den centrale hjælpedatabase, som fastlægger de forskellige grænsesnit til systemet såsom import/normalisering af data, vedligeholdelse, opslag, brug af data i lokale systemer.

Mht. data genbrug er der flg. udviklings- og undersøgelsesbehov:

- Det skal undersøges, om data med fordel kan stilles til rådighed for lokale systemer fra en central hjælpedatabase, eller om anvendelse af genbrugsdata skal håndteres lokalt.
- Det skal undersøges hvilke krav lokale workflow stiller til anvendelse af genbrugsdata, f.eks. om optagelse i de relevante internationale databaser er hurtig nok ift. lokale behov og om genbrug vil medføre en reel tidsbesparelse i den lokale registrering.
- Der skal udarbejdes en kravspecifikation til den centrale hjælpedatabase, som fastlægger de forskellige grænsesnit til systemet såsom import/normalisering af data, vedligeholdelse, opslag, brug af data i lokale systemer, synkronisering med database over publikationskanaler osv.

Mht. hjælpedatabaser og autoritetsdata vedr. personer og organisationer, er der flg. udviklings- og undersøgelsesbehov:

- Det skal undersøges, om der med fordel kan etableres en central database over personer (forsker/forfattere i Danmark) eller om problematikken kan håndteres lokalt. Her kan inddrages erfaringer med det hollandske Digital Author Identification (DAI) system.
- Mht. alle hjælpedatabaser og autoritetsdata skal der undersøges, om der med fordel kan anvendes unikke identifikatorer til at fastlægge entydighed.

¹⁸ <http://www.oio.dk/index.php?o=2440a0184c2969dbbdcd523b2ac57e01>

6. National analyse og beregning (bibliometrisk kvalitetsindikator)

Den bibliometriske kvalitetsindikator skal bidrage til VTUs kvalitetsfinansieringsmodel til fordeling af basismidler til forskning i forbindelse med udmøntning af FL 2010.

Det foreløbige udgangspunkt for kvalitetsfinansieringsmodellen, er at der etableret et system, der årligt omfordeler en begrænset del af basismidlerne efter let gennemskuelige kriterier. Målet er, at der med udgangspunkt i en årlig rangordning af universiteternes kvalitet målt på en række indikatorer overføres en andel af basismidler fra de universiteter, der klarer sig dårligst i sammenligningen, til de universiteter, der klarer sig bedst.

Den bibliometriske kvalitetsindikator udgør således kun én ud af flere indikatorer i den endelige kvalitetsfinansieringsmodel. På nuværende tidspunkt er der ikke taget endelig beslutning om hvilke indikatorer der skal indgå i modellen. Der opereres imidlertid med ti forskellige indikatorer inden for områderne forskning, uddannelse og videnspredning (se nedenstående tabel).

Tabel 1. Eksempel på indikatorer, som kunne indgå i kvalitetsfinansieringsmodellen

| Forskning | Uddannelse | Videnspredning |
|---|---|---|
| <i>Bibliometrisk kvalitetsindikator</i> | <i>Gennemførelstid på kandidatuddannelsen</i> | <i>Ikke-finansielt samarbejde med erhvervslivet</i> |
| <i>Eksterne forskningsindtægter</i> | <i>Studieprogression på 1 år af bacheloruddannelsen</i> | <i>(Erhvervsbarometer)</i> |
| <i>Ph.d.-beståelsesprocent</i> | <i>Kandidaternes beskæftigelsesgrad</i> | <i>Økonomisk omfang af relation til omverdenen</i> |
| <i>Forsker-internationalisering</i> | <i>Studerendes studieophold i udlandet</i> | |

De enkelte indikatorer vil i den endelige kvalitetsfinansieringsmodel blive vægтет forskelligt. Det betyder, at nogle indikatorer får større betydning på omfordelingen af basismidler end andre. Der er endnu ikke taget stilling til den indbyrdes vægtning af indikatorerne.

Udvælgelsen af indikatorer er sket med udgangspunkt i arbejdet i de tre indikatorgrupper for henholdsvis forskning, uddannelse og videnspredning, som Universitets- og bygningsstyrelsen og Forsknings- og Innovationsstyrelsen har etableret i samarbejde med Rektorkollegiet.

Hovedområdekorrektion

For at lette sammenligningen af universiteterne er det valgt at sammenligne hovedområderne med hinanden på tværs af universiteterne. Hovedområdekorrektionen foregår i praksis ved at sammenligne universiteternes præstationer på fire hovedområdeniveauer: Samfundsvidenskab, sundhedsvidenskab, humaniora samt natur og teknik. Denne opdeling på fagområder er identisk med opdelingen i Rektorkollegiets nøgletal.

Enhedspræstationer

Indikatorerne opgøres som universiteternes præstationer pr. enhed, altså en brøk. I relation til "uddannelsesindikatorerne" sættes antal beskæftigede i forhold til antallet af kandidater,

antallet af studerende på udlandsophold i forhold til antallet af ressourceudløsende studerende osv.

For de indtægtsbaserede indikatorer og for den bibliometriske indikator sættes præstationen i den enkelte indikator i forhold til universiteternes anvendelse af videnskabeligt personale (VIP-lønsum). Det betyder for den bibliometriske indikator at mængden af publicerede artikler måles i forhold til størrelsen af VIP-lønsummen på det enkelte universitet.

Rangordning af universiteterne

I kvalitetsfinansieringsmodellen bygger omfordelingen af basismidler på en rangordning af universiteterne inden for de enkelte indikatorer. For den bibliometriske indikator betyder det, at der ikke er en direkte sammenhæng mellem antallet af registrerede forskningspublikationer og størrelsen af den bevilling universitetet modtager. Derimod fordeles basismidlerne på baggrund af universitetets placering i forhold til de øvrige universiteter.

Når universiteternes præstationer og den heraf følgende rangordning er beregnet, omsættes resultatet til en omfordeling af basismidler. Et centralt krav til modellen har i den sammenhæng været, at kvalitetsvurderingen ikke må føre til meget store ændringer i basismiddelbevillingen fra det ene år til det næste. Det ville påføre de dårligst placerede universiteter en urimelig omstillingsbyrde. Som et første skridt i beregningen af det enkelte universitets basismidler i den nye periode, fastlægges en procentuel overgrænse for det tab af basismidler, som det dårligst placerede universitet i kvalitetskonkurrencen må opleve. Den kan f.eks. sættes til 5 pct.

Med dette udgangspunkt beregnes på baggrund af universiteternes placering i kvalitetsopgørelsen årets omfordeling af basismidler.

Udviklingsbehov

Det er vurderingen, at kvalitetsfinansieringsmodellen kun i begrænset omfang får direkte konsekvenser for den måde den bibliometriske kvalitetsindikator indrettes og organiseres på. Der er dog to områder hvor kvalitetsfinansieringsmodellen kan få betydning:

Hovedområdekorrektionen i kvalitetsfinansieringsmodellen har for det første betydning for den måde data i den bibliometriske kvalitetsindikator registreres og indberettes. Det er i den forbindelse vigtigt, at de enkelte indberetninger entydigt kan kategoriseres inden for de ovenstående fire hovedområder. Det betyder sandsynligvis, at der vil være behov for, at forskerne i forbindelse med deres registrering af publikationer i Pure og Orbith også skal angive hvilket hovedområde publiceringen skal kategoriseres under. En alternativ løsning kunne være at kategoriserer alle danske forskere eller de ”godkendte” publiceringskanaler inden for de fire kategorier.

Indberetningsperiode. Fagligt Udvalg eller Styregruppen skal tage stilling til hvilken indberetningsperiode, der skal indgå i de årlige opgørelser. På nuværende tidspunkt vil fordelingen af basismidler i 2010 ske på baggrund af indberetninger for kalenderåret 2008. Hvis Styregruppe/universiteter ønsker at fordelingen af basismidler skal ske på baggrund af nyere indberetninger, kan der eventuelt være behov for at ændre indberetningsperioden, så den ikke følger et kalenderår, men går fra sommer til sommer (eksempelvis fra august 2008 til juli 2009). Det kræver imidlertid, at flere universiteter ændrer på den eksisterende indberetningspraksis.

7. National forskningsformidling (og international eksponering)

Siden 1988 har Danmark haft en national forskningsdatabase, Den Danske Forskningsdatabase. Databasen har indtil fornyeligt eksisteret uden de store ændringer. På baggrund af Rambøll Managements evaluering af DDF i 2004¹⁹ blev der fra 2005 i gang sat en række ændringer af basen. Først og fremmest blev der i fællesskab med flere af databasens interessenter udviklet et nyt xml baseret udvekslingsformat, DDF-MXD²⁰. Samtidigt blev strategien for at få data i basen ændret fra tidligere at medtage data både som batch uploads og som direkte indtastning i et online katalogiseringsmodul til at være udelukkende baseret på høstning via OAI-PMH. Med den nye strategi sikredes det at datakvaliteten i basen er blevet væsentlig rigere og højere end tilfældet var i den gamle udgave, samtidigt sikres der løbende opdatering af posterne i basen, således at der er overensstemmelse mellem data i hhv. DDF og dataleverandørernes databaser. Den ændrede strategi har dog samtidigt betydet at der er færre, som leverer data til databasen end tidligere. Årsagen til dette skyldes to ting for det første er det ikke er alle, specielt ikke de mindre, institutioner som kan levere data via OAI-PMH høstning og for det andet er der andre universiteter, som har problemer med at kunne få gjort data kvaliteten lokalt klar til høstning på trods af at de har systemerne der kan høstes.

Det forbedrede datagrundlag betyder at der er grundlag for en række nye søgemuligheder som ikke udnyttes i det nuværende brugergrænseflade til DDF, der samtidigt trænger til en fornyelse. En forbedring af faciliteterne i DDF samt en relanceringen af brugergrænsefladen forventes lanceret i november 2007.

Allerede i dag er DDF eksponent for at de data som DDF høster spredes til en række internationale søgemaskiner, således er det muligt at finde poster fra DDF i Google Scholar ligesom man har en aftale med MSN Live Search Academic om en lignende eksponering. Derudover stilles et afgrænset sæt af DDF-MXD data til rådighed for OAI-PMH høstning i Dublin Core for andre internationale søgemaskiner. Projektstyringen af basen er generelt på udkig efter nationale og internationale samarbejdspartnere og senest har WorldWideScience.org inkluderet DDF i sin internationale forskningsportal²¹.

I Danmark er der en god tradition for at integrere forskningsregistreringssystemer med institutional repositories, det betyder at man både finder bibliografiske data og fuldtekst dokumenter i de danske forskningsdatabaser. Der betyder samtidigt en betydelig værdiberigelse for brugerne af både de enkelte institutioners databaser og for dem der søger i forskningsportaler som eksempelvis DDF.

I flere lande har findes der lignende forsknings portaler. En af de mere udbyggede er Narcis²² fra Holland der kombinere data fra forskningsdatabaser, institutional repositories (fuldtekst arkiver) og web crawling i en portal. Ifølge Knaw der administrerer portalen er kvaliteten af de da der kommer fra de hollandske forskningsdatabaser af meget svingende kvalitet. I Flandern i Belgien er man lige nu i gang med at opbygge en forskningsportal baseret på Cerif datamodellen²³, interessant ved integrationen af Cerif i dette tilfælde er den stærke integration af klassifikationsskema i systemet der gør det muligt i langt højere grad at inddele forskningen i fagområder end det er tilfældet i noget tværfagligt system i Danmark i dag.

¹⁹ <http://www.deflink.dk/nyheder/nyheder2.asp?id=1520>

²⁰ http://mx.forskningsdatabasen.dk/mxd/1.1.0/DDF_MXD_v1.1.0.pdf

²¹ <http://worldwidescience.org/>

²² <http://www.narcis.info/>

²³ <http://www.iweto.be/Home.html>

Disse baser viser også en tendens til at forskningsdatabaser eller portaler ikke bare er publikationsdata, men også data om projekter, eksperter m.m.

Udviklingsbehov

I praksis vil DDF kunne fortsætte med at høste fra de lokale forskningsdatabaser for at stille udstille dansk forskning i en samlet dansk forskningsdatabase som det har været tilfældet siden 1988. Selvom der kommer et nyt fokus på forskningsregistrering i Danmark, så vil det i første omgang betyde at dækningen af den danske forskningsregistrering lokalt på universiteterne øges, ligeledes vil anvendelsen af nationale hjælpe databaser sikre en endnu højere metadata kvalitet end det allerede høje niveau Danmark har i dag. Arbejdet vil således være med til at skabe et godt udgangspunkt for at Danmark kan få en forskningsdatabase meget høj international klasse.

En ny konstruktion hvor høstningen til en central database, med det formål at måle kvaliteten af dansk forskning, vil samtidigt stille spørgsmål til DDFs rolle og funktion. Skal DDF fortsætte med at høste data direkte fra de lokale databaser? Eller skal DDF hente sine data direkte fra databrønden (de samlede høstede data til den centrale forskningsdatabase)? Uanset hvilken model der vælges er der stadigvæk en interesse i at have en dansk forskningsdatabase, specielt set i lyset af det potentiale der ligger i at datamængden og kvaliteten øges ved den øgede fokus på forskningsregistreringen i Danmark. Hvor data kommer fra til en dansk forskningsdatabase er kun en af mange opgaver omkring vedligeholdelse og udviklingen af databasen. Uanset hvor arbejdet og ansvaret placeres er der brug for at gøre en indsats omkring videreudviklingen af grænseflader, funktioner og synliggørelsen af den nationale forskningsdatabase internationalt, samt indgåelse i internationale samarbejde omkring forskernes informationsforsyning.

I betragtningen af at øvelsen med at få en national bibliometrisk kvalitetsindikator skal være på plads på relativ kort tid vil det være fornuftigt at holde fokus på løsningen af den del af opgaven og efterfølgende vurdere fremtiden for en dansk forskningsdatabase og placering af ansvaret for denne. Og således fortsætte med DDF i sin nuværende form, baseret på OAI-PMH høstning direkte fra de lokale forskningsdatabaser, og inden for 1-2 år fra etableringen af den nationale forskningskvalitetsindikator afgøre DDFs rolle og placering i det nationale forskningsformidlings arbejde.

For at skabe klarhed for interessenter omkring DDF og formidlingen af dansk forskning er der behov for at ansvarlige beslutningstagere tager eksplicit stilling til DDFs rolle i forhold til den centrale forskningsdatabase.

8. Universiteternes lokale forskningsdatabaser

Der findes i dag to systemer i Danmark til registrering af universiteternes forskningsproduktion, PURE og Orbit. I dette afsnit beskrives de to systemer kort ift. deres nuværende workflow i registreringsprocessen og eventuelle ændringer som følger af indførelsen af en national forskningsindikator.

8.1 PURE

PURE er et kommercielt system, som anvendes af alle universiteter (undtagen DTU) samt en række hospitaler, til at registrere institutionernes forskningspublikationer. I PURE findes moduler til at registrere publikationer, forskningsprojekter, forskningsaktiviteter, studenterprojekter og bibliometriske data (citationer) for publikationer. Ikke alle moduler anvendes af alle universiteter.

I PURE findes en fleksibel rapportgenerator, som kan bruges til f.eks. løbende afrapportering fra institutter/faggrupper, at producere grundlaget for en årsberetning og til levering af Rektorkollegiets nøgletal vedr. publikationer og aktiviteter.

De universiteter som anvender PURE har etableret et samarbejde med en styregruppe, og en arbejdsgruppe som koordinerer videreudvikling i tæt samarbejde med leverandøren. Det overordnede formål med samarbejdet har været at fastholde et fælles metadataformat, som vil sikre datagrundlaget på tværs af universiteterne og giver mulighed for sammenligning på tværs. Af afgørende betydning her er anvendelse af det samme sæt publikationstyper (artikler, bøger osv.). Fastlæggelse af disse publikationstyper er også udarbejdet i tæt samarbejde med DDF.

Registreringsworkflow på universiteterne varierer fra universitet til universitet. Nogle universiteter har et bibliotek, som varetager den sidste validering af registreringer og på nogle universiteter (som regel hvor der ikke findes et centralt bibliotek), foretages validering på institut/fakultetsniveau. I PURE håndteres workflowet ved at tildele de relevante personer forskellige roller, hvorefter publikationer kan registreres (f.eks. af en institutsekretær eller forskere selv), sendes til godkendelse, og til sidst kan valideres. Efter validering bliver registreringer af de enkelte publikationer låst og kan kun ændres af en administrator.

I PURE er der en stram kobling mellem interne personer og deres organisationer. Denne kobling styres af PURE, som kan hente data fra f.eks. universitetets matrikel og kan anvende en unik 'matrikel ID' til kobling mellem en person og personens organisation/organisationer. Der bruges en række ID'er i PURE-systemet, f.eks. for hele poster, fuldtekst publikationer, personer, organisationer osv. men de er tildelt af systemet.

Der er en række kontrollerede autoritetslister, som kan bruges i registreringsprocessen, f.eks. valg af sprog, rolle (forfatter, redaktør osv.). Men der er andre metadatafelter, f.eks. eksterne organisationer, tidsskriftstitler, ISSN som lige nu er fritextfelter. Kobling af en central database over publiceringskanaler mm. vil medføre en betydelig forbedring i kvalitet af publikationsregistreringer.

Til KU er der udviklet nogle konvertere til import af data fra andre databaser (f.eks. ISI og PubMed). Disse konvertere er lige nu beregnet til brug for systemadministrator på vegne af forskerne, og vil derfor formentlig ikke blive taget i brug i større omfang. Der er forslag om at

udvikle nogle konvertere, som er tilgængelige for forskerne selv, som efter import kan tilrette det importerede data.

PURE-systemet skal kunne benytte den centrale hjælpedatabase over publiceringskanaler. PURE universiteterne, via PURE styregruppen/arbejdsgruppen, vil gerne bidrage til fastlæggelse af metadataformatet, da vi har ønsker om enkelte dataelementer, som med fordel kunne indkøbes til og hentes fra den centrale base til brug i den lokale registreringsproces.

Webservices anvendes i dag i vid udstrækning til at udstille data i det enkelte universitets PURE repository på f.eks. universiteternes hjemmesider.

Udviklingsbehov

Der er ikke noget til hinder for, at PURE-systemet vil kunne tilpasses til at hente data fra eksterne hjælpedatabaser og anvende disse data videre i registreringsprocessen, eller aflevere data til en central database, f.eks. med OAI-PMH.

8.2 Orbit systemet

Orbit og Insight (der bygger på Orbits platform) anvendes i dag på Danmarks Tekniske Universitet og Forsvars Akademiet.

Orbit bygger på DTVs egen udviklede open source platform til registrering af metadata MetaToo. Det er et værktøj der specielt er udviklet mhp. det digitale bibliotek, hvor der er behov for metadata registrering og fleksibilitet omkring implementeringen af forskellige metadataformater og skabeloner til registrering. Samt efterfølgende at kunne udstille data via webservices. Til Orbit er MetaToo specielt blevet udvidet for at kunne håndtere forskellige roller og arbejdsprocesser. Orbit udvides løbende for kunne i møde gå krav fra omverdenen og brugerne.

Systemet fødes ved en distribueret arbejdsgang, hvor institutterne har ansvaret for at forskerne på de enkelte institutter får indberettet deres forskningsoutput i Orbit. Det er op til de ansvarlige på institutterne selv at planlægge hvordan indberetningen til databasen foretages, og det er således forskelligt fra institut til institut, om det er forskerne selv der står for indrapporteringen, eller det er andre på instituttet der står for dette arbejde.

Efterfølgende valideres posterne centralt af personalet i ePub på DTV, her sikres det at posterne er registreret korrekt i systemet, der er dog ikke noget krav til at forskerne skal indsende publikationerne for på den måde at kunne verificere direkte med kilden. Enkelte institutter har selv kvalificeret personale og står for egen validering af posterne. I sidste ende ligger ansvaret for at den indberettede post er korrekt hos den først nævnte DTU forfatter i posten, også selvom denne ikke selv har indtastet posten.

Orbit er udstyret med en række skabeloner til forskellige publikationstyper. Inden for de forskellige publikationstyper er det muligt at registrere en række kriterier fra kontrollerede vokabularer, så som sprog, forskningsindikatorer, personroller, m.fl. Skabelonerne kan administreres af Orbit's systemadministratorer, som kan redigere og oprette nye publikationsskabeloner, samt oprette og redigere kontrollerede vokabularer.

Orbit kan håndtere autoritetsdata via web services og håndterer personoplysninger og organisationsdata fra DTU-basen i Orbit. Person kan slås op via en søgning på navn eller personalenummer, som også fungerer som unik-id, og derefter kan oplysninger hentes over i posten. Orbit har pt. ikke nogen autoritets database til tidsskrifter, men i forbindelse med Penlist projektet har DTV arbejdet med integration af data fra andre autoritetsbaser og vist at disse data kan implementeres på samme måde som person data i MetaToo og dermed Orbit.

Udviklingsbehov

Orbit anvender i vid udstrækning åbne grænseflader der let vil kunne indpasses i en service orienteret system arkitektur hvor lokale databaser kan hente autoritetsdata fra centrale services og hvor Orbit vil kunne levere data til en central database i OAI-PMH.

Der vil dog være behov for tilpasninger, da hver services som skal integreres med Orbit vil der på Orbits side skulle programmeres sådan at Orbit kan bruge disse data. Men så længe at de centrale databaser bygger på åbne standarder vil denne integration være mulig.

9. Opsummering

Grundpapiret beskriver en række centrale elementer og karakteristika ved det danske forskningsregistreringssystem. Papiret illustrerer, hvordan der inden for den seneste årrække er opbygget en betydelig og udbygget infrastruktur omkring forskningsregistrering, som på mange måder lever op til de behov som VTU måtte have i forhold til den tekniske organisering af en bibliometrisk kvalitetsindikator. Der findes således allerede et udbygget indberetningssystem på universiteterne, samt en omfattende erfaring med høstning af lokale publikationsdata.

Grundpapiret beskriver i den sammenhæng de to registreringssystemer, som allerede findes til registrering af universiteternes forskningsproduktion, PURE og Orbit. Det vurderes i papiret, at begge systemer som udgangspunkt kan benyttes som ”dataleverandører” til den bibliometriske kvalitetsindikator. Det er endvidere vurderingen, at der ikke i nævneværdigt omfang er problemer med høstning eller samkøring af data på tværs af de to systemer. Der peges i papiret dog på, at der vil være behov for mindre justeringer af indberetningssystemerne for at kunne håndtere de fremtidige indberetningsbehov til den centrale forskningsdatabase – eksempelvis i forbindelse med anvendelse af hjælpedatabaser og opdeling af forskningsregistreringer på hovedområder.

Grundpapiret viser samtidig, at der i dag sker en betydelig validering af data – såvel decentralt som centralt. Det er vurderingen, at der vil være behov for at kvalificere den eksisterende valideringspraksis yderligere - bl.a. med en manuel validering. Der er i den sammenhæng behov for at undersøge, hvordan en manuel validering kan organiseres samt at få konkretiseret indholdet i den manuelle validering. Ydermere er der behov for at få undersøgt, hvordan der kan etableres en validerings- og/eller kontrolpraksis, der kan mindske antallet af deduplikeringer.

Grundpapiret beskriver endvidere behovet for at udarbejde forskellige former for hjælpedatabaser. Det er vurderingen i papiret, at der er behov for at få etableret et antal forskellige hjælpedatabaser – bl.a. i forhold til autoritetslister for publiceringskanaler. Det er samtidig vurderingen, at der er behov for yderligere undersøgelser af hvilke hjælpedatabaser, der kan støtte institutionernes/forskernes indberetninger af data.

[mangler noget om Udvekslingsformat og vokabularer]

I grundpapiret belyses endeligt også sammenhængen mellem den bibliometriske kvalitetsindikator og VTUs kvalitetsfinansieringsmodel. Det er vurderingen, at VTUs kvalitetsfinansieringsmodel kun i begrænset omfang får direkte konsekvenser for den måde den bibliometriske kvalitetsindikator indrettes og organiseres på. Teknisk Udvalg skal dog være opmærksom på, at der fremover vil være behov for at kunne opdele indberetningerne på fire hovedområder.