Styregruppen for National Data Management
Afdækning E: Opsummering af de faglige miljøers behov og præferencer

Introduktion

Denne afdækning informerer Styregruppens arbejde med planlægnings-dimensionerne "Infrastrukturer" og "Kompetenceudvikling og forskerstøtte". Således berører afdækningen eksisterende forskningsdata management relateret praksis såvel som behov for systemer/infrastrukturer og støttefunktioner (fx rådgivning af juridisk, teknisk, informations-organisatorisk art og kompetenceudvikling mv.) indenfor hovedområderne humaniora (HUM), samfundsvidenskab (SAM), sundhedsvidenskab (SUND) og naturvidenskab (NAT) samt teknisk videnskab (TEK).

Afdækningen baserer sig på kvalitative interviews (fokusgruppeinterviews og individuelle interviews) gennemført parallelt på de fem hovedområder, med 5-9 forskere per område, ud fra en spørgeguide opbygget over de 8 faser, den anvendte forskningsdata livscyklus model operer med. Respondenterne blev udvalgt med så bred spredning ift. akademisk niveau, fagområde og alder, hvilken institution de tilhører mv. som muligt. Nogle respondenter repræsenterede større forskningsgrupper, mens andre repræsenterede solo forskningindsatser.

Antallet af respondenter afspejler den korte tidsramme, der var givet til gennemførelse af afdækningen. Datagrundlaget er således begrænset, og der må tages forbehold for dækningen/ repræsentativiteten og dermed entydigheden af de konklusioner, der kan drages på det grundlag. Afrapporteringen fra de enkelte hovedområder er bragt som underbilag.

Nedenfor er opstillet to tabeller: Tabel 1 opsummerer infrastrukturønskerne og behovene for støtte, som respondenterne fra hvert hovedområde har givet udtryk for, samt supplerende kommentarer af strategisk betydning. Tabel 2 gennemgår for hver enkelt af de 8 faser hhv. den etablerede praksis og behovene for infrastruktur og støtte.

Tabel 1: Opsummering af hovedområdernes infrastrukturønsker og behov for støtte

	HUM	SAM	SUND	NAT	TEK
Infrastrukturønsker (opsumeret)	• System der er "tæt på forskerne" og deres faglige behov (midlertidig lagring af data), men er centraliseret hvad angår bevaring og deling (slutlager). • Centrale funktioner ses som national opgave, men data skal kunne tilgængeliggøres/deles globalt. • Bevaring af mange forskellige typer data i store mængder + metadata + annotationer. • Deling af data + metadata + annotationer både under og efter forskningsprocessen. • Håndtering af adgangsstyring ifm. følsomme data. • Mulighed for citering af datasæt med tildeling af DOI eller PID. • Formidling tværfagligt. • Evt. etablering af service, hvor man kan finde ældre software, eller en	• Udbygget infrastruktur på universitets- eller nationalt niveau, med ○ Adgang til (ekstern) storage. (behovet er for størstedelen under 2.000 GB, og for halvdelen under 500 GB). ○ Backup. ○ Versionering. ○ Mulighed for adgangsstyring. ○ Mulighed for deling mellem projektdeltagere både under og efter endt forskningsprojekt. ○ Evt. mulighed for adgang for studerende (til købte datasæt). ○ Adgang til regnekraft til tunge beregninger. ○ Nærhed mellem regnekraft og datalager ligegyldigt, så længe fjernadgang til server er velorganiseret med kontrol af adgang og brug. ○ Behov for dokumentation og	• Udvikling af sygehusenes systemer, så der på tværs af firewalls kan flyttes personfølsomme data mellem lager og HPC eller bare mellem sygehus og universitet. • Storage – helst lokalt. Evt. med Redcap. Ofte er der ikke plads nok på netværksdrevene. Forskergruppernes egne servere er dimensioneret til de enkelte projekter. • Langtidsbevaring – helst løst lokalt – man vil ikke betale for store eksterne systemer (der er ikke en bred forståelse for hvad langtidsbevaring egentligt indebærer). • Data bør gemmes for altid. Dels fordi data vedbliver at have værdi (fx undersøgelser af befolkningsgruppers helbredstilstand), dels fordi udvælgelse af relevante/interessante data er svært og tidskrævende.	• Fase 2: ○ Storageplads til de voksende datamængder er en stor udfordring, men der er ikke enighed om, hvorvidt det bedst løses på lokalt/ nationalt/internationalt niveau. ○ De modne DM-grupper ønsker skræddersyede løsninger til fagområderne og er villige til at betale med forskningsmidler. Dette billede er mindre entydigt ved de mindre modne DM-grupper. ○ Backupløsninger efterlyses af de mindre modne DM-grupper. • Fase 3: ○ Nogle bruger i dag 2.500 CPU kerner – andre har mindre behov men specielt adgang til maskiner med meget RAM (typisk 1TB). ○ Løsninger må kunne anvendes fra	• Fase 1: For at undgå dobbelt administration og fejl er det vigtigt med en kobling (men med mulighed for undtagelser) til eksisterende systemer, hvor fx projekter og medarbejdere registreres mhp. økonomistyring. • Fase 2: ○ Adgangskontrol/-begrænsning både indenfor institutionen selv og med samarbejdspartnere (virksomheder oa. institutioner). ○ At institutionens IT-service kan sikre indsamlet data til en given standard accepteret af en eventuel samarbejdspartner (mhp. senere databehandling med andet HW og SW end det samarbejdspartneren stiller til rådighed). ○ Flere indgange til data (fx SFTP, webbaseret interface mv.).

	HUM	SAM	SUND	NAT	TEK
	emuleringsservice, så man kan genåbne ældre filer. • Data er fx GIS data, data til beskrivelse af arkæologiske genstande, C-14 data, opmålinger, data knyttet til web-aktiviteter, kvantitative opgørelser af variable indenfor TV-produktioner, tekstfiler mm.	metadata er så specifikt, at det er svært at standisere. ○ Formidling af datas kontekst, specifikt link mellem data og projektbeskrivelser, projektansøgninger, publikationer eller andre formidlede præsentationer af data.	• Centralt tilbudte systemer og viden opfattes som en god idé. Dog også frygt for at et fælles system ikke tilfredsstiller den enkelte forskers behov og i værste fald tvinger ham til at anvende ringere systemer. Ingen tiltro til at en central løsning har tilstrækkelige faglige kompetencer eller tidsmæssige ressourcer.	alle udbredte OSer, inkl. Windows. • Fase 4: Mindre behov for computeressourcer end i databehandling. • Fase 7: De mindre modne DM-grupper, der finder bevaring vigtigt, mener at den skal ske lokalt, men deles internationalt gennem netværk og standardisering. Ingen villighed til at finansiere bevaringsløsning.	• Fase 3 og 4: Adgang til HPC ressourcer. • Fase 5: Evt. deponering ved kopiering til lokalt institutionelt repository. • Fase 5 og 7: Genanvendelse kan være svært uden compilere, SW, specifikke OSer og data ikke er gemt samlet. Løsning: Deponering/ bevaring af komplette kørende systemer (image af virtuel maskine).
Støttebehov	• Teknisk, fagligt, organisatorisk. • Faglig relateret supportfunktion. • Aflastning ift. rutineopgaver. • Metadatering.	• Fase 1: Skræddersyet support i form af rådgivning på universitetsniveau eller nationalt niveau med lokalt forankret repræsentant, der kan vejlede i forskerne i DM. • Fase 2: Projektspecifik juridisk rådgivning (persondata og køb af data/kilder).	• Fase 1: Kurser og rådgivning på institutniveau ifm. udarbejdelse af DMPer. Gerne betalt med forskningsmidler. • Fase 2: "Manpower" i form af data manager. Villighed (men ikke nødvendigvis evne) til at betale. • Fase 3: Data manager til dokumentation af datatransformationer. • Fase 4: ○ Efteruddannelse i håndtering af "big data", fx mønstergenkendelse (udbudt på nationalt niveau). ○ Kurser i statistik på højt niveau (udbudt på fakultetsniveau). ○ Fagstatistiker (institutservice). • Fase 5: Kurser i organisering af data. • Evt. hjælp til upload af store datasæt, både systemer og rådgivning nævnes (nationalt niveau), men vurders ikke som så vigtigt.	• Fase 1: Bred interesse for kurser i DMP og evt. rådgivning. Hvem der varetager opgaven bedst, er der ikke enighed om. Halvdelen er villige til at betale med forskningsmidler. • Fase 2: Enkelte ønsker støtte til databehandling, men koordineret internationalt og fagspecifikt. • Fase 3: Enkelte ønsker støtte til databehandling, men koordineret internationalt og fagspecifikt. • Fase 4: Enkelte så gerne kurser og vejledning, men vil ikke finansiere med forskningsmidler. • Fase 5: Interesse for hjælp til deponering på nationalt niveau – nogle finder det det nødvendigt med skræddersyet hjælp, andre mener det kan være generelt. Ingen finansiering med forskningsmidler.	• Fase 1: ○ Skabeloner (byggeklodser) til projektansøgninger m.v. - baseret på standarder (fx ISO). ○ Evt. mulighed for at andre overtager koordinering af DMP. ○ Juridisk vejledning ifm. fx personfølsomme data og kommercialiseringsinteresser. • Fase 6: ○ Mulighed for at citere data, fx i form af DOI. ○ Stabile links til filer, der hvor de allerede ligger, fx på et fileshare.
Andre behov	• Kassationspolitik for data (fx data, som ikke må bevares ud over 2 år). • Central financiering af grundydelser med mulighed for tilkøb af ekstra ydelser til lokale behov.	• Behov for politikker for (eller måske endda krav om) offentliggørelse af forskningsdata.	• N/A	• N/A	• N/A
Andet	• Hovedårsagen til at DM ikke foretages er, at der ikke findes en infrastruktur med mulighed for især	• Undersøgelsen anbefaler, at der udvikles et metadata format, der kan beskrive og formidle status og karak-	• Frem for krav fra bevillingsgivere ser man hellere overordnede reglsæt, som skitserer et antal	• Mange grupper er bekymrede for udsigten til at skulle langtidslagre stadigt større datasæt, men de ser	• 3. partsløsninger eller egne løsninger vælges pga. at det giver et umiddelbart overblik over udgifter, nem til-

HUM	SAM	SUND	NAT	TEK
bevaring af forskningsdata. Dette kæder repondenterne sammen med fraværet af en overordnet politik og en dertil hørende institutionel støtte, hvilket betyder, at ansvaret ligger hos forskerne. • Man ser gerne nationale løsninger af frygt for besparelser på universitetsniveau. • De lokale tjenester, der findes til backup, fx fx særlige drev for forskere, anvendes ikke eller kun i begrænset omfang. Der savnes en infrastruktur til at lette backup opgaven.	ter af data og datasæt. Udover generel metadatafunktionalitet, skal formatet specifikt: ○ Skelne mellem data og datasæt. ○ Præcisere kontekstuelle aspekter ved data/datasæt, så der opnås en faglig og mere omfattende fremstilling af, hvad et givent datasæt er baseret på fx projektbeskrivelse og -ansøgning eller publikationer. ○ Eksplicitere adgangsmuligheder, så data/datasæt ikke tilgås af uvedkommende (forudsætning for infrastrukturens troværdighed). ○ Tydeliggøre indholdsmæssige og administrative aspekter, fx versioner, kodningsomfang, ejerskab mv., så identifikation og vurdering af data effektiviseres.	minimumskrav, som lokalt forankrede systemer skal overholde, og som det er op til de enkelte institutioner eller grupper at implementere det, der bedst løser opgaven. • Der er et generelt behov for en øget bevidsthed om korrekt behandling af data samt et rådgivende organ enten lokalt eller nationalt. • Korrekt metadatering mhp. genfinding og hurtig forståelse af da er ønskværdig.	det som en teknisk udfordring der bedst løses lokalt. De modne DM-grupper giver udtryk for, at deres løsninger er langt billigere end hvad de kan købe services til på det frie marked, og hvad deres internationale kolleger rapporterer, de betaler for centrale løsninger.	gængelighed, overblik over løsning der er let at ændre konfiguration på. • Ansvaret for langtidsbevaring bør ligge hos den instans, der stiller krav eller den institution, det påhviler at opfylde kravet. • Finansiering: Kravstiller bør også have det finansielle ansvar. Basalt set bør udgiften til opbevaring, sikring og langtidsbevaring afholdes af universitet. Hvis ikke det financieres af basismidler, så er der stor risiko for at data forsvinder ved projektets ophør eller hurtigt derefter. Skal egenbetaling ske, så skal det som minimum være på institutniveau eller højere. • Pris er en faktor – det må ikke være dyrere end at gøre det selv (købe en usb-harddisk).

Tabel 2: Hovedområdernes etablerede praksis og behov for infrastruktur ift. hver fase

Område / Fase	HUM	SAM	SUND	NAT	TEK
Fase 1: Planlægning og ansøgning	**Praksis**: • Planlægning sker primært ifm. indsamling af data. **Behov**: • Fagligt relateret supportfunktion efterspørges. • Aflastning i forhold til rutineopgaver ønskelig.	**Praksis**: • Der gøres overvejelser som backup, sikring og tilgængeliggørelse – det er nødvendigt pga. datamængden. • Stigende krav om at stille data til rådighed (eksternt finansierede projekter). **Behov**: • Skræddersyet support; rådgivning på universitetsniveau eller nationalt niveau med lokalt forankret repræsentant til at vejlede i forskere i DM. • Ingen behov for kurser/støtte til udarbejdelse af DMPer.	**Praksis**: • De fleste udarbejder kun lejlighedsvis DMPer og i så fald efter forskellige standarder. **Behov**: • Kurser og rådgivning på institutniveau ifm. udarbejdelse af DMPer. Gerne betalt med forskningsmidler.	**Praksis**: • Halvdelen bruger i udpræget grad en struktureret DMP. **Behov**: • Bred interesse for kurser i DMP og evt. rådgivning. Hvem der varetager opgaven bedst, er der ikke enighed om. Halvdelen er villige til at betale med forskningsmidler.	**Praksis**: • Både eksempler på ad hoc styring af forskningsdata og detaljeret DMP. Projekter, hvor deling er et eksplicit mål, planlægger i større grad. **Behov**: • Skabeloner (byggeklodser) til projektansøgninger mv. - baseret på standarder (fx ISO). • For at undgå dobbelt administration og fejl er det vigtigt med en kobling (men med mulighed for undtagelser) til eksisterende systemer, hvor fx projekter og medarbejdere registreres mhp. økonomistyring. • Evt. mulighed for at andre overtager koordinering af DMP. • Juridisk vejledning ifm. fx personfølsomme data og kommercialiseringsinteresser.
Fase 2: Indsamling/ generering af data	**Praksis**: • Decentral lagring (PC, eksterne lagringsmedier mm) af mange forskellige typer data, fx GIS data, data til beskrivelse af arkæologiske genstande, C-14 data, opmålinger, data knyttet til web-aktiviteter, kvantitative opgørelser af variable indenfor TV-produktioner, rene tekstfiler mm. • Data kan være bundet til forskellige typer SW. • Begrænset metadatering. **Behov**: • Lagring med back up "tæt på de	**Praksis**: • Indsamler og anvender mange forskellige typer kilder og forskningsdata, fx statistikker, spøgeskemaer, interviews, feltnotater, lovgivningstekster, simuleringer, lyd, video, indhold fra sociale medier og websider. • Decentral lagring (PC, eksterne lagringsmedier, fælles netværksdrev, Dropbox, Google Docs, iCloud), fordi institutionens fællesdrev er ikke gearet til store datamængder. • Grundet begrænset lagerplads anvendes Dropbox mest til udveksling af data.	**Praksis**: • Blanding af analoge og digitale data. Sidstnævnte er primært data fra spørgeskemaer, blodprøve- og biopsi-resultater, data fra instrumenter, fx accelerometermålinger, transskriberede interviews, excel-ark og scannings- og røntgenbilleder. • 100 GB – 2-4 TB/år (enkelte studier flere TB/studie). • Data lagres i høj grad på netværksdrev – enten universitetets eller hospitalets. • Enkelte lægger deres data i OPEN (Region Syddanmarks forsknings-	**Praksis**: • Produktion af data: 3-5 TB/dag til 500 TB/dag, og inden længe PB. • Eksponentiel vækst i data. • Ikke alle data lagres; 3-100 TB/dag. • Digitale billeder (2D matricer), 1D sekvenser, 4D film. • Store dataindsamlere anvender velorganiserede storagesystermer og int. netværk og databaser. Personfølsomme data bliver dog altid lokalt, lagret i middelstore, professionelt drevne installationer. • Modne DM-grupper metadaterer alle data efter int. standarder;	**Praksis**: • Mange forkellige datatyper; måledata, lyd, video, billeder, surveydata. Blanding af proprietære og ikke-proprietære formater. • Lagres på egen PC eller måleinstrument eller ekstern enhed (fx kamera, mobiltelefon), netværksdrev (lokale fx i forskningsgruppen eller institutionsbaserede), offline lager (fx DVD, USB-HD), samarbejdspartneres infrastruktur, 3. partssystemer (fx DropBox, Amazon m.fl.) og non fil-baserede løsninger (fx

Område / Fase	HUM	SAM	SUND	NAT	TEK

	faglige miljøer" for at sikre faglige behov.
	• Sikker og fleksibel adgang til data.
	• Mulighed for deling af
	  ○ data
	  ○ metadata
	  ○ annotationer
	• Udtræk af data fra Netarkivet til analyse kræver særlig opsætning at få adgang til pga. personfølsomme data.
		• Metadatering er etableret praksis.
		
		Behov:
		• Udbygget infrastruktur, med
		  ○ tilstrækkeligt storage (behov for er for størstedelen under 2.000 GB, og for halvdelen under 500 GB).
		  ○ backup
		  ○ versionering
		  ○ mulighed for deling mellem projektdeltagere
		  ○ mulighed for adgangsstyring
		• Juridisk rådgivning i enkelte tilfælde (persondata og køb af data/kilder).
			infrastruktur).
			• Enkelte afdelinger har egne dedikerede systemer (PACS - picture archiving and communication system til medicinske billeddata).
			• Data lagres også på laptops (men kun anonymiseret).
			• Metadatering kun i begrænset omfang – evt. automatisk opsamlet. Oplysninger om indsamlingsmetoder mv. lægges ikke ved data, men oplyses ved publicering.
			
			Behov:
			• "Manpower" til håndtering af data, dvs. data manager. Villighed (men ikke nødvendigvis evne) til at betale.
			• Storage – helst lokalt. Evt. med Redcap. Ofte er der ikke plads nok på netværksdrevene. Forskergruppernes egne servere er dimensioneret til de enkelte projekter.
				mindre modne DM-grupper metadaterer også, men her anvendes primært logbøger.
				• Mindre modne DM-grupper lagrer næsten kun lokalt – lokal IT-afd. og eksterne harddiske.
				
				Behov:
				• Storageplads til de voksende datamængder er en stor udfordring, men der er ikke enighed om, hvorvidt det bedst løses på lokalt/nationalt/internationalt niveau.
				• De modne DM-grupper ønsker skræddersyede storageløsninger til fagområderne og er villige til at betale med forskningsmidler. Dette billede er mindre entydigt ved de mindre modne DM-grupper.
				• Backupløsninger efterspørges af de mindre modne DM-grupper.
					databaser).
					• Data forsynes ofte med metadata, men nogle domæner har flere muligheder for struktureret brug af metadata end i fx eksperimentelle ingeniørfag, som ikke har standarder for beskrivelse fysiske opstillinger (her er tekniske rapporter metadata).
					
					Behov:
					• Adgangskontrol/-begrænsning både indenfor institutionen selv og med samarbejdspartnere (virksomheder oa. institutioner).
					• At institutionens IT-service kan sikre indsamlet data til en given standard accepteret af en eventuel samarbejdspartner (mhp. senere databehandling med andet HW og SW end det samarbejdspartneren stiller til rådighed).
					• Flere indgange til data (fx SFTP, webbaseret interface mv.).

Fase 3: Behandling af data	**Praksis**: • Databehandling foretages med såvel standard- og special SW. • Begrænset metadatering. **Behov**: • Ingen særlig behov for regnekraft undtagen til 3D-modellering (løses af DeIC).	**Praksis**: • Data underkastes alt fra simpel validering over avanceret processering til beregningstung transformering. **Behov**: • Adgang til regnekraft til tunge beregninger. • Adgang til (ekstern) lagerplads. • Nærhed mellem regnekraft og datalager ligegyldigt, så længe fjernadgang til server er velorganiseret med kontrol af adgang og brug. • Behov for dokumentation og metadata er så specifikt, at det er svært at standisere.	**Praksis**: • Der opereres sjældent med behandling af data, som kræver HPC. De fleste databehandlinger foretages lokalt på alm. PC. **Behov**: • Udvikling af sygehusenes systemer, så der på tværs af firewalls kan flyttes personfølsomme data mellem lager og HPC eller bare mellem sygehus og universitet. • Data manager til dokumentation af datatransformationer.	**Praksis**: • Intensiv databehanling. • Alle anvender Unix-baserede systemer. **Behov**: • Nogle bruger i dag 2.500 CPU kerner – andre har mindre behov men specielt adgang til maskiner med meget RAM (typisk 1TB). • Løsninger må kunne anvendes fra alle udbredte OSer, inkl. Windows. • Enkelte ønsker støtte til databehandling (koordineret internationalt og fagspecifikt).	**Praksis**: • Data behandles med eget eller standard SW på egen PC (evt. egne clusterløsninger) medmindre der er tekniske eller datasikkerhedsmæssige hindringer. • Data metadates så vidt muligt. **Behov**: • Adgang til HPC ressourcer.

Område / Fase	HUM	SAM	SUND	NAT	TEK
Fase 4: Analyse af data	**Praksis:** • Databehandling foretages med såvel standard- og special SW. • Begrænset metadatering. **Behov:** • Standardprogramel til fx statistiske analyser vurders tilstrækkeligt. • Støtte til metadatering.	**Praksis:** • Data underkastes statistisk analyse, data mining, modellering og fortolkningsprægede analyser. **Behov:** • Samme som i fase 3.	**Praksis:** • Overvejende statistiske analyser på alm. PC. Derudover visualisering og tekstmining og -analyse. **Behov:** • Efteruddannelse i håndtering af "big data", fx mønster genkendelse (udbudt på nationalt niveau). • Kurser i statistik på højt niveau (udbudt på fakultetsniveau). • Fagstatistiker (institutservice).	**Praksis:** • Alle anvender Unix-baserede systemer og næsten kun open source SW. **Behov:** • Løsninger skal kunne anvendes fra alle udbredte OSer. • Mindre behov for computeressourcer end i databehandling. • Enkelte så gerne kurser og vejledning, men vil ikke finansiere med forskningsmidler.	**Praksis:** • Data analyseres med eget eller standard SW på egen PC (evt. egne clusterløsninger) medmindre der er tekniske eller datasikkerhedsmæssige hindringer. • Data metadates så vidt muligt. **Behov:** • Adgang til HPC ressourcer.
Fase 5: Deponering af data	**Praksis:** • Decentral bevaring, men der er ingen tradition for langtidsbevaring og deling af data. • Bevaring vanskeliggøres af, at data kan være bundet til bestemt SW. **Behov:** • Bevaring og deling også af metadata og annotationer.	**Praksis:** • Deponering (fase 5) og bevaring og deling af data (fase 7) fremstår som tæt forbunde for respondenterne. • Deponering sker i repositories sker i nogle, men ikke alle tilfælde – det er projektspecifikt. • Forskerne mødes generelt med krav om at gøre data tilgængelige efter forskningsprocessens afslutning, fx ifm. publicering. **Behov:** • Infrastruktur der kan understøtte deponering og deling af data – også gerne adgang for studerende (specielt til købte datasæt).	**Praksis:** • De fleste deponerer ikke data i et repository, men på netværksdrev. • DDA og OPEN anvendes. • Hvor data gemmes på lokale servere, foretages ikke langtidsbevaring selvom der er data tilbage fra 1994. • I nogle tilfælde foretages backup på privat indkøbte ekstern HD i hjemmet. • Ikke nok datadokumentation. • Data er typisk ikke citérbare. **Behov:** • Langtidsbevaring (for altid) – helst løst lokalt. • Evt. kurser i organisering af data.	**Praksis:** • Modne DM-grupper deponerer – har veletablerede formater og procedurer for det. Nogle deponerer alle resultaterne mens andre deponerer kun data, der ligger til grund for publicering. **Behov:** • Interesse for hjælp til deponering på nationalt niveau – nogle finder det det nødvendigt med skræddersyet hjælp, andre mener det kan være generelt. Ingen finansiering med forskningsmidler.	**Praksis:** • Deponering i de infrastrukturer, hvor data er i forvejen (fælles/ eget netværksdrev, egen PC, lab'et) mhp. egen eller gruppens genbrug/ validering. Andres adgang er sekundært. • Data fra egenbyggede fysiske modeller er tæt på umulige at anvende/ fortolke uden den fysiske model. **Behov:** • Genanvendelse kan være svært uden at compilere, SW, specifikke OSer og data er gemt samlet. Løsning: Deponering af komplette kørende systemer (image af virtuel maskine). • Evt. deponering ved kopiering til lokalt institutionelt repository.
Fase 6: Publicering og citering af data	**Praksis:** • Datasæt tilgænegliggøres pt. ikke ifm. publicering af artikler. **Behov:** • Mulighed for citering af datasæt med tildeling af persistent identifikator, fx PID eller DOI.	**Praksis:** • At gøre data citerbare er ikke hovedfokus hos respondenterne. • Krav om tilgængelighed af data mødes i nogle tilfælde ved at formidle data via e-mail. **Behov:**	**Praksis:** • Kun i begrænset omfang krav om publicering af data. • Et fåtal publicerer data sammen med artiklen, og ikke via datatidsskrifter. • Datasæt citeres oftest ikke – og kun gennem artiklen, de er anvendt i. **Behov:**	**Praksis:** • Modne DM-grupper er underlagt regler for publicering af data ifm. publikationer – disse anvender også citeringer af data. • Ingen brug af DOI. I stedet anvendes fagspecifikke dataciteringsformater. **Behov:**	**Praksis:** • Projekter, der har datadeling som mål, gør det ofte via websites med data og kode mv. alternativt open source fora. Ellers deles data mere uformelt via mail, Dropbox mv. på baggrund af personlig henvendelse. **Behov:**

Område / Fase	HUM	SAM	SUND	NAT	TEK
		• Ingen direkte behov.	• Evt. hjælp til upload af store datasæt, både systemer og rådgivning nævnes (nationalt niveau), men vurders ikke som så vigtigt.	• Ingen behov.	• Mulighed for at citere data, fx i form af DOI. • Stabile links til filer, der hvor de allerede ligger, fx på et fileshare.

Fase 7: Bevaring og deling af data	**Praksis**:
	• Decentral bevaring, men der er ingen tradition for langtidsbevaring og deling af data.
	• Bevaring vanskeliggøres af, at data kan være bundet til bestemt SW.
	
	Behov:
	• Bevaring og deling også af metadata og annotationer.
		Praksis:
		• Fsva. deling:
		  ○ Forskningsdata deles gennem både nationale og internationale dataportaler.
		  ○ Data journals er der ikke tradition for at bruge.
		  ○ Også mail og fildelingstjenester som Dropbox anvendes.
		
		Behov:
		• Infrastruktur der kan understøtte deponering og deling af data – også gerne med adgang for studerende (specielt til købte datasæt).
			Praksis:
			• Bevaring: Nogle anvender DDA.
			• Deling: Data udleveres case-by-case til dem, der henvender sig.
			
			Behov:
			• Data bør gemmes for altid. Dels fordi data vedbliver at have værdi (fx undersøgelser af store befolkningsgruppers helbredstilstand gennem et helt liv), dels fordi en udvælgelse af strengt relevante/interessante data er umådeligt svært og tidskrævende.
			• Lokalt system – vil ikke betale for store eksterne systemer (der er ikke en bred forståelse for hvad langtidsbevaring egentligt indebærer).
				Praksis:
				• Modne DM-grupper har stringente opbevaringsplaner og –regler samt systemer til deling.
				• Mindre modne DM-grupper har ingen databevaringstraditioner.
				• Modne DM-grupper deler i nogen udstrækning data – nogle deler helt generelt resultater, andre deler kun efter personlig henvendelse.
				
				Behov:
				• De mindre modne DM-grupper, der finder bevaring vigtigt, mener at den skal ske lokalt, men deles internationalt gennem netværk og standardisering. Ingen villighed til at finansiere bevaringsløsning.
					Praksis:
					• Fsva. deling af data: Se fase 6.
					• Ingen repondenter langtidsbevarer i særlige infrastrukturer til dette. Data lever i de infrastrukturer, hvor de anvendes under behandling og analyse.
					• I tilfælde af store data mængder flyttes data offline.
					
					Behov:
					• Genanvendelse kan være svært uden at compilere, SW, specifikke OSer og data er gemt samlet. Løsning: Bevaring af komplette kørende systemer (image af virtuel maskine).

Fase 8: Opdagelse af data	**Praksis**: • Forskningsdata opdages gennem personlige kontakter og netværk. **Behov**: • Tværfaglig formidling.	**Praksis**: • Forskningsdata opdages gennem personlige kontakter, netværk og tidsskrifter. **Behov**: • Bedre formidling af datas kontekst, specifikt link mellem data og projektbeskrivelser, projektansøgninger, publikationer eller andre formidlede præsentationer af data.	**Praksis**: • Opdagelse sker via litteraturen og via den enkelte forskers netværk. • Der søges kun i begrænset omfang efter data, og da i kendte databaser eller via Google. **Behov**: • Ingen særlige behov.	**Praksis**: • Datasæt findes hovedsageligt gennem litteraturen. • For de modne DM-grupper er dataportaler en del af miljøet. **Behov**: • Ingen behov.	**Praksis**: • Genopdagelse sker typisk via publikationer og networking – data som deles via hjemmesider forsvinder ofte over tid, eller adgangen kan være begrænset eller licensbelagt. **Behov**: • Ingen behov.

DeIC og DEFF
Styregruppen for National Data Management
Afdækning E: De faglige miljøers behov og præferencer
Humaniora
Undersøgelse gennemført april og maj 2014

Undersøgelsen er foretaget af Filip Kruse (Statsbiblioteket), Jesper Boserup Thestrup (Statsbiblioteket), Birgitte Munk (Det Kongelige Bibliotek) og Laura Riis Christensen (Det Kongelige Bibliotek)

1. Konklusioner og anbefalinger for Humaniora

1.1. Konklusioner - executive summary:

De gennemførte interviews gav mulighed for at drage en række foreløbige konklusioner på status og tendenser på området i forhold Data Management (DM). Interviewene gav overordnet indtryk af, at det humanistiske forskningsparadigme er i en udvikling, hvor digitaliseringen af kilder og bedre muligheder for indsamling, bevaring og deling af digitale forskningsdata skaber helt nye analytiske muligheder. Den traditionelt kvalitativt dominerede forskning indenfor humaniora er ved at blive suppleret med de nye muligheder for især kvantitative analyser, som anvendelsen af IT åbner op for.

- Datamængderne indenfor Humaniora er voksende, hvilket bl.a. skyldes nye muligheder for dataindsamling og behandling.
- De indsamlede data er af mange forskellige typer.
- Bevaring og lagring foretages aktuelt decentralt, af forskerne selv og på forskellige medier.
- Disse aktiviteter vanskeliggøres af, at data kan være bundet til bestemte typer software.
- DM i sin udfoldede form, som beskrevet i undersøgelsens fasemodel, er ikke udbredt.
- Planlægning foretages på enkeltprojektniveau, primært i forbindelse med indsamling af data.
- Behovet for støtte i forbindelse med DM (herunder også Data Management Planning – DMP) beskrives dels som en faglig relateret supportfunktion, gerne tæt på forskerne, dels som et behov for aflastning i forhold til mere rutineprægede opgaver.
- Der er ikke tradition for, at tidsskrifter tilgængeliggør datasæt i forbindelse med artikler.
- Data 'opdages' ikke på nettet. Det sker gennem personlige kontakter og netværk.
- Respondenternes krav til et fremtidigt system til støtte for DM er, at det skal kunne bevare data af mange forskellige typer og i store mængder, muliggøre deling af data både undervejs og efter forskningsprocessen, håndtere adgangsstyring, så følsomme data beskyttes og give muligheder for deling og bevaring af metadata samt annotationer til forskningsdata.
- Systemet skal på den ene side være tæt på forskerne og deres faglige behov og på den anden side være centraliseret, specielt hvad angår mulighederne for langtidsbevaring og deling af data. De centrale funktioner ses som en national opgave, men data skal efter behov kunne gøres tilgængelige globalt.

- Finansieringen af grundydelserne bør være central, DeIC, Uddannelses- og forskningsministeriet eller andre relevante ministerier ses som muligheder. Der skal være mulighed for at tilkøbe ekstra ydelser til lokale behov

- Der er behov for at etablere en kassationspolitik for data.

1.2. Anbefalinger

- DeIC og universiteterne skal skabe et system i fællesskab, som skal kunne rumme hovedområdernes forskellige behov, respektere de faglige traditioner og leve op til de tekniske krav, der stilles både på overordnet nationalt plan - og lokalt. Slutlageret skal være centralt eller nationalt, midlertidig lagring af data skal kunne foretages tæt på de faglige miljøer.

- Støttefunktioner skal være fag-fagligt funderede og formidlet af personer med kendskab til forskning på de enkelte områder.

- Systemet skal i sin helhed kunne bevare og formidle forskningsdata under forskningsprocessen, langtidsbevare og muliggøre deling af data. Tilsvarende gælder for metadata og annotationer til forskningsdata.

- Systemets basisydelser skal være centralt finansieret og dermed ikke udsat for potentiel risiko i form af lokale besparelser.

- Det følger af ovenstående, at opgave- og ansvarsfordelingen mellem centrale og lokale aktører skal overvejes nøje.

- Systemet skal være simpelt og brugervenligt.

- Etablering af et system forudsætter, at der findes en forpligtende politik på området.

- Showcases vil være relevante i en markedsføring overfor forskerne af nytten ved at bruge systemet.

2. Undersøgelsesmetode og udformning af afrapportering

Undersøgelsen er gennemført som en række interviews ud fra den spørgeramme som Styregruppen udformede. Interviewene blev planlagt i april og gennemført i maj 2014 og varede fra 1 time til 1,75 timer.

I alt blev der gennemført 5 interviews, der fordelte sig således på interviewpersonernes universitets- og instituttilhør:

Universitet	Institut	Titel	Køn
Aarhus Universitet	Forhistorisk Arkæologi	Lektor	Mand
Aarhus Universitet	Institut for Æstetik og Kommunikation	Lektor	Mand
Aalborg Universitet	Institut for Kultur og Globale Studier	Professor	Kvinde
Københavns Universitet	Saxo Instituttet	Ph. d. stipendiat	Kvinde
Københavns Universitet	Det Teologiske Fakultet	Postdoc	Mand

De fem respondenter dækker bredt i forhold til et ideelt karriereforløb for en forsker fra Ph. d. til professor. Interviewpersonerne er ansat på tre forskellige universiteter og fordeler sig bredt indenfor Humaniora. Dog er to forskere aktive indenfor medieområdet i bred forstand, mens to er indenfor arkæologi - om end indenfor vidt forskellige områder - hvilket mindsker spredningen. En mere repræsentativ undersøgelse ville have krævet respondenter fra flere humanistiske fagområder. Der er dog så god overensstemmelse imellem besvarelserne fra de fem personer, at vi mener at kunne påvise generelle tendenser.

Flere forhold gjorde, at interviewpersonerne ikke kunne svare uddybende på alle spørgsmål indenfor de 8 faser eller hovedområder, som spørgerammen dækker. Dette skyldes fravær af de angivne aktiviteter eller, at spørgsmålene manglede relevans for forskeren. Dette medførte dels fravær af svar, dels tematiske skift i svarene. For at sikre en sammenhæng og en rød tråd i fremstillingen er nogle af spørgsmålene fra spørgerammen i nedenstående bearbejdning af interviewene slået sammen. Det drejer sig om spørgsmålene til fase 1, 2, 3 og 4 (Planlægning og ansøgning, Indsamling og generering af data, Databehandling og Dataanalyse) samt spørgsmålene til fase 5, 6, 7 og 8 (Data deponering, Publicering og citering af data, Bevaring og deling af data og Opdagelse af data).

Som det fremgår ovenfor, er konklusionerne snævert bundne til vores interviewresultater og giver en bredere forståelse af den forståelsesmæssige kontekst. Anbefalingerne er handlingsrettede formuleringer ud fra interviewresultaterne, primært forskernes egne.

På grund af den korte tidsfrist har der ikke været tid til at få godkendt diverse transskriptioner af interviews samt brugen af udvalgte citater fra de enkelte interviews, som derfor optræder anonymt. Citaterne er let grammatisk redigerede for at sikre forståelsen.

3. Fase 1 - 4: Planlægning og ansøgning, Indsamling og generering af data, Databehandling og Dataanalyse

3.1. Hovedtendenser

- DM i sin udfoldede form er ikke udbredt, planlægning foretages på enkeltprojektniveau, primært i forbindelse med indsamling af data.

- Behovet for hjælp og støtte til DM (herunder også Data Management Planning – DMP) beskrives dels som en faglig relateret supportfunktion, gerne tæt på forskerne, dels som et behov for aflastning i forhold til mere rutineprægede opgaver.

- Mange forskellige datatyper indsamles og datamængderne er voksende, hvilket bl.a. skyldes nye muligheder for indsamling.

- Databehandling og -analyse foretages med såvel standard-, som mere specielle programmer, afhængig af det enkelte projekt.

- Bevaring og lagring af data foretages både digitalt og trykt. Opgaverne udføres aktuelt af forskerne selv.

- Digital bevaring vanskeliggøres af, at data kan være bundet til bestemte typer software.

- Der er behov for faciliteter til bevaring og lagring både tæt på forskerne - decentralt - og mere centralt.

- Metadatering af data under indsamling, behandling, analyse og bevaring foretages kun i begrænset omfang.

3.2. Uddybende kommentarer

- Hovedårsagen til, at DM som omfattende dataindsamling, -behandling, -analyse, -arkivering osv. ikke foretages er, at der ikke findes en infrastruktur med mulighed for især bevaring af forskningsdata. Dette kæder interviewpersonerne sammen med fraværet af en overordnet politik og en dertil hørende institutionel støtte, der igen betyder, at ansvaret ligger hos forskerne. Videre vurderes det, at traditionerne for DM indenfor humaniora generelt er svagere end inden for andre fag. Men som aktivitet betragtet, anses DM som nødvendig, ikke mindst i medfør af det voksende omfang af digitalt fødte data.

  ”Jeg har ikke gjort det, men jeg kan se, at der er problemer ved ikke at have gjort det.”

- Der er forskelle på, om hjælp til DMP anses for ønskeligt: i det omfang, forskerne allerede arbejder meget med digitale data, er der ikke et særligt behov herfor. Men hvor data ikke eller kun i begrænset omfang er digitale, er behovet mere udtalt. Her ses støtte gerne lokalt, f.eks. på fakultetsplan. Jo tættere på selve forskningsprocessen støttebehovet er, jo mere fagrelateret skal denne støtte være. Der er tilfredshed med IT-støttefunktionerne knyttet til de enkelte fag - der hvor de findes.

  ”Jeg har ikke brug for støtte til at lave DMP’er, det kan jeg godt finde ud af.”

- Data er stort set af alle typer: GIS-data, data til beskrivelse af arkæologiske genstande, C-14 data, opmålinger, data knyttet til web-aktiviteter, herunder sociale medier, kvantitative opgørelser af forskellige beskrivende variable for TV-produktioner, rene tekstfiler mm.

  Forskellen mellem forskningsdata og kilder ses primært som bestående i, at det er kilderne (bøger, konkrete objekter, arkiver, registre, interviews osv.), som gennem videnskabeligt arbejde kan resultere i forskningsdata.

- Bevaring af data under forskningsprocessen vanskeliggøres af de mange forskellige typer af back-up medier. Da ansvaret hviler på den enkelte forsker, er metoderne tilsvarende individuelle: data gemmes på egen pc, på eksterne harddiske, usb-sticks osv. Det siges direkte, at bevaring og arkivering ikke er meriterende for den enkelte forsker, det er kun publiceringen af forskningsresultater. De lokale tjenester, der findes, f.eks. særlige drev for forskerne, anvendes ikke eller kun i begrænset omfang. Det er et gennemgående træk, at der savnes en infrastruktur, der kunne lette dette arbejde, som trods nedprioriteringen betragtes som vigtigt.

  ”Vi har back-up medier som laserdiscs, MB-floppyer og sågar kassetter, hvad gør vi ved dem? Vi har brug for støtte, teknisk, fagligt og organisatorisk.”

  ”Jeg har en registrant over dansk TV-drama, der kun foreligger i trykt form, som sådan er den offentlig tilgængelig. Den kunne godt være digitalt tilgængelig, men hvor pokker skal jeg lægge den?”

  ”Arkæologer laver kataloger. Vi har tekstdelen og et katalog over hver enkelt genstand, fordi det er et genstandsbaseret studie.”

- Behandling og analyse af data kan kræve både regnekraft og lagerkapacitet, men der angives ikke at være behov for særlig regnekraft, undtagen til 3-D-modellering og her tilvejebringes denne i et samarbejde med DeIC. Standardprogrammer til f.eks. statistiske analyser vurderes generelt at være tilstrækkelige.

- Vanskelighederne ved de specifikt software-bundne data er mest udtalte ved ældre data og ved mere kompliceret software. Data i regneark og i tekstfiler, frembyder ikke store problemer.

  ”Datas overlevelse bestemmes jo af værtsprogrammerne, data gemt i en gammel Access-fil kan jeg ikke bare åbne igen. Det samme gælder data fra andre programmer.”

  ”Meget er håndholdte løsninger. Noget software er lavet til Facebook, andet til Twitter. Hvis der ændres i dem, skal softwaren til vores analyser også ændres.”

- En decentral lagring af data ses som nødvendig for at sikre forbindelsen til fagene. De funktioner, der med fordel kunne centraliseres, ville være vedligehold, oprettelse af et arkiv eller bibliotek for software og for data, der ikke er afhængige af specifik software.

  ”Forskerne skal være sikre på, at de får det rigtige. Der er grund til at frygte en centraliseret one-size-fits-all-løsning, uden vægt på faglige hensyn.”

- Metadatering anses klart som nødvendig af hensyn til både deling og bevaring, men foretages ikke i det omfang, som forskerne selv kunne tænke sig. Behovet for støtte til metadatering er udtalt, en enkelt respondent giver dette topprioritet.

  ”Metadata ville jo også gøre data søgbare og dermed sikre, at nogen får øje på dem.”

  ”Metadatering undervejs i analysen, det er et ømt punkt. Jeg ville nok ikke kunne sige nej til lidt støtte til det.”

  ”Jeg tror der ville være brug for støtte på et grundlæggende plan. Hvordan opbevarer man metadata.”

  ”Det vil være svært med støtte på fakultetsniveau, fordi man som dekan hytter sit eget skind. Så det vil måske nok være et niveau over. Til at starte med nationalt, men svært at tænke nationalt uden at tænke globalt.”

4. Fase 5 - 8: Deponering, Publicering og citering af data, Bevaring og deling af data og Opdagelse af forskningsdata

4.1. Hovedtendenser

- Der er ikke i dag tradition for digital langtidsbevaring, deling og citation af datasæt i følge respondenterne. Der er heller ikke tradition for, at tidsskifter tilgængeliggør datasæt i forbindelse med udgivelse af artikler. Der er behov for et system, der kan sikre håndtering af forskningsdata både under forskningsprocessen og derefter. Et sådant system findes ikke i dag.

  Respondenterne ønsker, at systemet kan:
  1. Bevare data, så de kan genbruges af andre.
  2. Bevare data, så værdifulde oplysninger for fremtidig forskning sikres.
  3. Rumme meget forskellige datatyper og datamængder.
  4. Muliggøre deling af data både under og efter forskningsprocessen.
  5. Håndtere adgangsstyring, så følsomme data beskyttes.
  6. Formidle tværfagligt og dermed sætte forskningsdata i spil.

- Arkivet eller arkiverne, der skal håndtere langtidsopbevaring af og muliggøre deling af data, skal som minimum være på nationalt plan. På visse fagområder og eller i forbindelse med visse projekter kan det være nødvendigt med globale adgange.

- Støtte i forhold til anvendelse og tilpasning af systemet skal ydes af personer med faglig indsigt og så tæt på forskerne som muligt.

- Systemet skal være finansieret af DeIC, Uddannelses- og forskningsministeriet eller andre relevante ministerier, da mange humanistiske forskere ikke har ressourcer til at betale for servicen. Systemet skal indeholde mulighed for at købe ekstra ydelser.

- Der er behov for at etablere en kassationspolitik for data.
- Der er behov for, at arkivet eller arkiverne giver mulighed for citering af datasæt, i form af tildeling af enten DOI eller PID.

- Metadatering og annoteringer til data bør også kunne bevares og tilgængeliggøres.

- Universiteterne skal udvikle services, der kan sikre data under forskningsprocessen, i form af et velfungerende personligt drev eller backup-system, evt. inspireret af Time Machine.

- Data 'opdages' ikke på nettet. Det sker gennem personlige kontakter og netværk.

4.2. Uddybende kommentarer

- Respondenterne oplever generelt mangel på et system, der kan arkivere og tilgængeliggøre forskningsdata efter forskningsprojekter er afsluttede og resultater publiceret. Der er generel enighed om, at sådanne arkiver ikke bør drives af det enkelte universitet, men samtidig ønskes nærhed til de faglige miljøer for at sikre de faglige behov. Begrundelserne er, at man vil sikre, at den fornødne specialistviden om databevaring er tilknyttet arkivet, man vil undgå, at samfundet skal investere i samme slags udstyr og tjenester flere gange og endelig frygter man for et universitetsbaseret systems større udsathed for besparelser. Generelt angiver respondenterne i forbindelse med alle spørgsmål, at de gerne ser en form for central finansiering af et system, hvor man så kan købe særlige services, hvis der er behov. En enkelt giver højeste prioritet til etablering og drift af et sådant system, en anden næsthøjeste.

  "Der skal være en langtidsbevaring af data."

  "Det ville være ærgerligt, hvis de forsvinder ligesom når en computer går ned."

- Der i øjeblikket ikke tradition blandt de adspurgte for systematisk digital bevaring af data. I hvor høj grad data arkiveres og deles er specifikt for det enkelte forskningsprojekt og de specifikke arkiveringsmuligheder. Respondenterne gør dog generelt opmærksom på, at et arkiv skal give adgang til sine data. En enkelt overvejer, om man skal etablere en service, hvor man kan finde ældre software eller om emulering er løsningen for at kunne genåbne gamle filer.

- Respondenterne har forskellige holdninger og krav til hvilke data, der skal langtidsopbevares og kunne deles. Eksempelvis nævnte en respondent, at arkæologiske rådata fra udgravninger ikke må kasseres overhovedet. En anden gav udtryk for, at der er behov for en kassationspolitik, så man ikke gemmer data, der ikke kan bruges senere. Eller gemmer data, der ikke må arkiveres ud over en

bestemt periode. Det sidste kan eksempelvis være tilfældet med projektet Digital Footprints[1], hvor dele af rådata kun må bevares i 2 år.

- Ligesom forskningsdata normalt ikke uden videre stilles til rådighed online i arkiver er der heller ikke tradition for at tilgængeliggøre forskningsdata i forbindelse med publikationer, i form af åbne datasæt. Flere respondenter nævner, at data publiceres i form af tabeller, bilag etc. som del af publikationer. De angiver, at der kun er få tidsskrifter, der muliggør deling af forskningsdata via vedhæftede datafiler, men ingen af disse har som krav, at forskningsdata vedlægges indsendte artikler.

- Spørgsmålene om publicering og deling af forskningsdata rejser nogle principielle diskussioner for respondenterne. Der er behov for at kunne styre adgangen til data: data kan være personfølsomme, f.eks. interviewdata, data kan indeholde oplysninger om lokalisering af værdifulde fund, data kan være belagt med ophavsretslige begrænsninger ï forhold til adgang mm.

  ”Ved en udgravning fortsætter man tit i flere år for hvert år kunne man lave en foreløbig rapport, som man kunne lægge i et repository, som andre så kunne bruge. Det ville være enormt smart. Problemet er bare, at folk tit hæger om deres materiale. De vil ikke ud med det, før de kommer med en stor publikation.”

- Datamængderne kan også være så omfattende, at det vil kræve særligt udstyr for at kunne håndtere en sikker og tilstrækkelig fleksibel adgang. En af respondenterne beskriver, at vedkommende i sit arbejde kan få brug for at trække dele af Netarkivet ud til analyse. Disse data er både meget omfattende og personfølsomme og vil kræve en helt særlig opsætning for forskeren at få adgang til. Samtidigt vil udtrækning af data medføre, at andre forskere kan få problemer med at få adgang til dataene. Respondenterne fremhæver også behovet for - og problemerne i - citering af forskningsdata fra mere eller mindre offentligt tilgængelige arkiver. Der er behov for et system med persistente identifikatorer. Eksempelvis som PID eller DOI[2].

- Som tidligere nævnt, er respondenterne positive overfor at kunne dele forskningsdata. Men som nævnt først efter projektets afslutning og publicering af dets resultater og kun i det omfang lovgivning og andre regler om adgangsbegrænsning tillader. Tanken om at kunne genbruge data tiltaler alle respondenterne, formidling af arkiverede data nævnes her også som muligt indsatsområde. I forbindelse med udviklingen indenfor humaniora mener flere respondenter, at det vil styrke det tværfaglige arbejde.

  ”Der er i hvert fald mange, der er bange for hele den elektroniske transformation af data, som særligt vi humanister arbejder med. Man kan gå ind på Gutenberg og finde førsteudgaver af bøger, som man før i tiden skulle have tilsendt igennem det Kongelige Bibliotek og få dem herind. Nu ligger det der og det er formidabel kvalitet.”

[1] http://digitalfootprints.dk/
[2] http://infopid.dk/hvad-er-pid/

- Der er dog ikke kun behov for at kunne dele projektspecifikke forskningsdata. I flere interviews diskuteres muligheden for også at gemme og dele annotationer til forskningsdata. Disse annotationer kan berige det originale datasæt. Funktionen kendes f.eks. fra LARM.fm[3] og gør det lettere at søge data, idet annotationerne fungerer som metadata. Derfor ønskes der også adgang til at dele både forskningsdata og annotationer til disse.

- Der er forskellige forslag til hvem, der skal drive et centralt system. DeIC og DDA bliver foreslået. Endelig er der ikke et ønske om at flytte data fra eksisterede arkiver, som eksempelvis Netarkivet, til arkiver, der kunne tænkes oprettet i forbindelse med en forbedring af mulighederne for bevaring og deling af forskningsdata. Fagenes ønsker skal afgøre, om arkivering skal ske nationalt eller globalt.

  "Jeg tror det er svært på fakultetsniveau, fordi man som dekan hytter sit eget skind. Så det vil måske nok være et niveau over. Til at starte med nationalt, men det er svært at tænke nationalt uden at tænke globalt. Jeg tror det er for vigtigt et spørgsmål til at lægge ned på det enkelte fakultet, der har sine kernefags prioriteter".

- Generelt skal tjenester til bevaring og arkivering være koblet op til WAYF, være brugervenlige og tilbyde muligheder, der minder om eksempelvis kommercielle løsninger såsom Dropbox. Ellers vil forskerne finde andre services.

- Til spørgsmålet om opdagelse af data svarer respondenterne, at de ikke deler data på denne måde. I stedet nævnes konferencer og personlige faglige netværk som vigtige måder at få kendskab til andres data. I denne sammenhæng nævnes Academia.edu [4] , hvor forskerne deler artikler, data med videre. En form for søgetjeneste, hvor der kan søges information om afsluttede og igangværende forskningsprojekter efterlyses. Deling af data af foregår oftest igennem netværk, igennem personlige samtaler eller via e-mail. De årlige rapporteringer på fakultetsniveau sikrer et overblik over aktiviteter såsom konferencer og seminarer, der kan danne afsæt for forskeres eftersøgning af relevante data.

[3] http://larm.fm/indholdilarm.html
[4] http://www.academia.edu/

INSTITUT FOR KOMMUNIKATION
AALBORG UNIVERSITET

RESEARCH DATA MANAGEMENT: DE SAMFUNDSVIDENSKABELIGE FAGLIGE MILJØERS BEHOV OG PRÆFERENCER (SAMF AFDÆKNING E)

UDARBEJDET AF: TANJA SVARRE, METTE SKOV OG MARIANNE LYKKE, AAU, JUNI 2014

Indhold

UNDERSØGELSENS FORMÅL

Denne rapport har til formål at informere Danish e-Infrastructure Cooperation (DeiC) om infrastrukturer og kompetenceudvikling/forskerstøtte i forhold til at kunne supportere danske samfundsvidenskabelige forskere i deres arbejde med research data management. Der tages udgangspunkt i DeIC's 8-trinsmodel over data life cycle med henblik på at kunne identificere, hvorvidt der er bestemte faser i forskningsprocessen, hvor forskerne i særlig grad oplever et behov for support af enten systemmæssig eller rådgivnings-/vejledningsbaseret karakter.

OM AFDÆKNINGEN PÅ FAGOMRÅDET SAMF

Afdækningen er udført af forskellige aktører. Mette Skov og Tanja Svarre har stået for planlægningen af undersøgelsens struktur (spørgeskema og interview). Lars Nondal (CBS) og Anne Sofie Fink Kjeldgaard (Statens Arkiver) har rekrutteret undersøgelsens deltagere samt deltaget i planlægningen af undersøgelsen. Indsamling og analyse af data samt udfærdigelse af rapport er Mette Skov og Tanja Svarre ansvarlige for. Som det vil fremgå nedenfor, er ikke alle fagområder indenfor samfundsvidenskab dækket af undersøgelsen, ligesom der ikke er en ligelig fordeling imellem CBS- og KU-deltagere. Dette skyldes blandt andet undersøgelsens korte tidsvindue. Der har i udvælgelsen af deltagere været lagt vægt på, at de repræsenterer både forskellige underområder til samfundsvidenskaberne, men også forskellige måder at indsamle, anvende og formidle forskningsdata.

DATAINDSAMLING

Rapporten baserer sig på et introducerende spørgeskema og hhv. et fokusgruppeinterview med 3 deltagere og to individuelle interviews. Alle deltagere er forskere i samfundsvidenskabelige discipliner. Spørgeskemaet blev fremsendt til og besvaret af deltagerne forud for fokusgruppe/interviews. Formålet var at kunne henlægge faktuelle spørgsmål til spørgeskemaet og dermed frigive mere plads til mere diskuterende spørgsmål til interviews. Som eksempler på spørgeskemaspørgsmålene kan nævnes: "Hvilken type data/kilder baserer du din forskning på?", "Hvor lagrer du typisk dine forskningsdata?" og "Hvor/hvordan opdager du relevante forskningsdata?" Både spørgeskema og interviews bygger på en tilrettet version af spørgeguiden udarbejdet af DeIC. I tilretningen er der blandt andet fokuseret på at gøre spørgsmålene mere forklarende. Fx "Hvordan arbejder I i praksis med data management i planlægningen af et forskningsprojekt?" og "Hvordan arbejder I med databehandling og dataanalyse? Stiller dette specielle krav?" eksemplificerer spørgsmål fra interviewguiden.

Dataindsamlingen præsenteres således, at hver fase indledningsvis introduceres med de spørgeskemaresultater, der relaterer til den pågældende fase. Efterfølgende uddybes med resultater fra fokusgruppe/interviews.

DELTAGERE

Undersøgelsen havde 6 deltagere, som besvarede hhv. spørgeskema og deltog i enten fokusgruppe eller interview. Alle 6 deltagere besvarede spørgeskemaet, mens 3 CBS-forskere deltog i fokusgruppe interview og en CBS- og en KU-forsker deltog af logistiske årsager i individuelle interviews over Skype. Den sidste deltager måtte med kort varsel melde afbud til interviewet, og det var ikke muligt indenfor undersøgelsens tidsramme at gennemføre interviewet på et senere tidspunkt. Deltagerne er alle mænd og de repræsenterer et aldersspænd mellem 37-60 år. Heraf er 4 lektorer og 2 professorer. 2 deltagere er fra Københavns Universitet, mens de resterende 4 er fra CBS. Deltagerne repræsenterer følgende institutter:

- Deltager 1: Institut for Antropologi (KU)
- Deltager 2: Institut for Medier, Erkendelse og Formidling (KU)
- Deltager 3: Økonomisk Institut (CBS)
- Deltager 4: Department of Business and Politics (CBS)
- Deltager 5: Department of Finance (CBS)
- Deltager 6: Department of International Business Communication(CBS)

DATA LIFE CYCLE

I forbindelse med fokusgruppeinterviewet og de to interviews blev DeIC's "data life cycle" model udleveret til og præsenteret for deltagerne. Modellen udgjorde strukturen for interviewet og gav generelt god mening for deltagerne. Dog følges processen ikke nødvendigvis så lineært, som modellen indikerer. Både deltager 3 og 6 forklarer, at dataanalyse i nogle tilfælde kan gå forud for at kilder og forskningsdata indsamles, eksempelvis hvis en tidligere anvendt kodning af en dataanalyse kan genanvendes til analyse af et ny datasæt.

Indledningsvist i interviewene blev deltagerne spurgt, om de skelner mellem kilder og forskningsdata og hvorvidt følgende definition gav mening for dem: "Forskningsdata afgrænses til digitale data, der indsamles eller skabes mhp. forskning, i modsætning til offentlige registre og andre kilder, der også kan gøres til genstand for forskning og videnskabelig analyse"? Svarene viser, at en skelnen mellem kilder og forskningsdata kun i mindre grad giver mening for deltagerne. Flere deltagere ser forskningsdata som et overordnet begreb, som dækker over både det, der refereres til som henholdsvis kilder og data i den foreslåede definition. Deltager 5 kan godt følge definitionens skelnen mellem forskningsdata og kilder og understreger, at det er en del af forskningsprocessen at gøre data håndterbare. Deltager 2 er ikke enig i den foreslåede definition og foretrækker en skelnen mellem primære, sekundære og tertiære data.

I denne rapport anvendes, i tråd med definitionen foreslået af DeiC, følgende terminologi:

- Kilder: data indsamlet fra allerede eksisterende kilder (fx statistiske databaser eller offentlige registre)
- Forskningsdata: data der indsamles/skabes af forskerne selv med henblik på forskning. Som en del af forskningsdata medtages også aggregerede kilder (fx sammenlægning af statistikker fra forskellige databaser), idet denne proces er et udtryk for en del af forskningsprocessen, hvor der skabes nye data.

I forlængelse heraf skal vi dog gøre opmærksom på, at interviewene indikerer, at allerede i løbet af fase 2 og 3 i datalivscyklusmodellen transformeres kilder til forskningsdata som led i forskningsprocessen.

FASE 1: PLANLÆGNING OG ANSØGNING

Forskernes overvejelser vedr. data management i den indledende planlægnings- og ansøgningsfase relaterer sig først og fremmest til at sikre sig adgang til de ønskede kilder og forskningsdata. Fx hvorvidt man kan få adgang til de ønskede kilder i økonomiske databaser, og ligeledes har deltager 2 samarbejdet med erhvervslivet for at kunne få adgang til nødvendige kilder. Dernæst nævnes overvejelser om backup, sikring og tilgængeliggørelse af data på dette tidlige stadie i forskningsprocessen. Eksempelvis beskriver deltager 6, hvordan de i et aktuelt forskningsprojekt har lovet bevillingsgiver, at projektets rådata stilles til rådighed for andre forskere. Forskningsprojektet ved p.t. ikke hvordan løftet skal indfries, idet den indsamlede empiri fylder langt mere end man kan uploade ved hjælp af Google Drive, Dropbox osv. På tilsvarende vis oplever andre af deltagerne, at det i stigende omfang er et krav i forbindelse med eksternt finansierede forskningsprojekter, at forskningsdata stilles til rådighed ved projektafslutning.

Til spørgsmålet om, hvorvidt der er behov for støtte til fase 1, så nævner ingen behov for støtte til at udarbejde data management planer. Derimod påpeger flere, at det vil en stor hjælp at kunne få rådgivning i denne fase. Det skal ikke være i form af udbudte kurser men derimod en person man kan kontakte for at få skræddersyet support. Støtten kan enten ydes på universitetsniveau eller på nationalt niveau med en lokal kontaktperson tilknyttet. Deltager 3 har ikke brug for støtte i denne fase, da behovet i givet fald opstår i forbindelse med brug af registerdata, og i de situationer har de institutioner, som data rekvireres fra de den fornødne indsigt i juridiske forhold omkring dataanvendelse og -sikkerhed.

Behovet for understøttelse i fase 1 kan opsummeres til at omfatte:

- Rådgivning på enten universitetsniveau eller nationalt niveau med en lokalt forankret repræsentant, der kan vejlede forskerne i data management

FASE 2: INDSAMLING/GENERERING AF DATA

Besvarelserne fra spørgeskemaundersøgelsen viser, at deltagerne indsamler og anvender mange forskellige typer af kilder og forskningsdata. Den mest udbredte *dataform* er statistikker, som alle seks deltagere anvender. Statistikkerne hentes både fra offentlige registre, så som Danmarks Statistik, men også internationale statistiske databaser og erhvervsdatabaser benyttes som kilder. Næst hyppigste datakilder er spørgeskemaer, interviews, feltnotater, samt tekster og dokumenter (inkl. lovgivning). Endelig baserer deltagerne deres forskning på meningsmålinger, simuleringer, målinger, lyd, video, indhold fra sociale medier/Internet, massemedier/-kommunikation og hjemmesider (1-2 deltagere per svarkategori). Svarene vidner om en stor bredde i, hvad den samfundsvidenskabelige forskning baseres på, og at kilder og forskningsdata forekommer i

forskellige medier og formater. Af interviewene fremgår det desuden at de fleste af interviewdeltagerne (undtagen deltager 6) arbejder med allerede eksisterende datakilder typisk i form af statistiske og økonomiske databaser samt registerdata. Disse kilder anvendes med forskelligt formål herunder fx til at generere en stikprøve på baggrund af en population eller med henblik på deciderede analyser. Samtidigt arbejder hovedparten af deltagerne (undtagen deltager 5) ligeledes med forskningsdata, som de selv indsamler.

Forskningsdata lagres som oftest på computerens harddisk, på en USB-stik eller andre eksterne lagringsmedier, i Dropbox eller på et fælles netværksdrev. Google Docs og iCloud anvendes i mindre omfang til lagring af data. Behovet for lagringskapacitet ligger for størstedelen af deltagerne under 2000 Gb, og for halvdelen under 500 Gb. Det fremgår tydeligt af interviewene, at omfanget af deres data udgør et gennemgåede problem med hensyn til lagring. De kan kun have mindre datamængder lagret på institutionens fællesdrev, og deltager 4 karakteriserer det generelle IT-system til ikke at være gearet til store datamængder. Således er det gennemgående for deltagerne, at de i stedet lagrer data forskellige steder herunder på eksterne harddiske. Deltagene beskriver, at det er besværligt at skulle holde styr på forskellige lagringsmedier og versionering. Samtidigt er deltagerne bevidste om, at eksterne lagringsmedier kan blive ødelagt eller stjålet. Af samme grund har én af deltagerne sin eksterne harddisk stående i sekretærernes pengeskab. Ved forskningssamarbejde med forskere fra andre institutioner anvendes typisk Dropbox eller lignende til deling af dokumenter og samarbejde omkring indsamling/generering af data. Men også her kan det være svært at arbejde med større datamængder og grundet den begrænsede lagerplads hos eksempelvis Dropbox, egner denne platform sig bedre til udveksling af data og mindre godt til datalager.

I fase 2 udtrykker hovedparten af deltagerne behov for en udbygget infrastruktur med hensyn til lagringskapacitet. De understreger, at adgangsstyring er vigtigt, så kun de rette og ikke uvedkommende får adgang til forskningsdata. De arbejder alle med at metadatere eller dokumentere deres data og oplever ikke behov for støtte eller standardisering på dette punkt. Blandt deltagerne er der forskellig holdning til spørgsmålet om, hvorvidt der er behov for juridisk rådgivning i denne fase. Deltager 2, 3 og 4 har ikke noget behov for støtte i denne fase. For en enkelt deltager (deltager 4) er dokumentationen en så integreret del af forskningsprocessen, at en standardisering ikke lader sig gøre, da dokumentationen kræver faglig indsigt og forståelse. Deltager 5 og 6 siger, at behovet ikke er stort, men i enkelte tilfælde kan der opstå behov for juridisk rådgivning fx omkring køb af data. Igen efterspørges der ikke kurser, men projektspecifik juridisk rådgivning.

Behovet for understøttelse i fase 2 kan opsummeres til at omfatte:

- En infrastruktur, der:
    - Giver tilstrækkelig lagerplads
    - Giver mulighed for deling mellem projektdeltagere
    - Muliggør adgangsstyring

- Juridisk rådgivning i enkelte tilfælde (fx vedr. persondatalovgivning og særlige regler vedr. købte data/kilder)

FASE 3 OG 4: BEHANDLING OG ANALYSE AF DATA

Fase 3 og 4 bliver her præsenteret samlet, da interviewbesvarelserne vedr. disse to faser hænger tæt sammen. I spørgeskemaet har vi spurgt om, hvilke former for databehandling, der typisk bliver udført. På et spektrum rangerende fra simpel validering til avanceret processering er begge ender af spektret repræsenteret, dog med en vis overvægt til avanceret processering og beregningstunge transformeringer. Dernæst blev der i spørgeskemaet spurgt til, hvilke specifikke former for dataanalyse der oftest bliver udført. Svarene viser, at statistisk analyse er den mest dominerende analyseform, men der foretages også data mining, modellering og fortolkningsprægede analyser.

I interviewet beskriver deltager 5, hvordan behandling og klargøring af data i fase 3 (forud for dataanalyse) er ressource- og tidskrævende. Det samme oplever deltager 2 i projekter med meget store datamængder. Deltager 2 beskriver et konkret projektsamarbejde med en privat virksomhed, hvor dataindsamling og databehandling måtte foregå hos virksomheden pga. meget store datamængder og krav til avanceret software, som han ellers ikke ville have adgang til. Ellers beskriver deltagerne generelt, at de har adgang til det nødvendige software til behandling og analyse af data. Deltager 3 tilføjer dog, at adgang til hurtige maskiner/regnekraft vil være en stor hjælp, når han laver avanceret processering. Pt. kan behandling af data tage dagevis. Derimod er nærhed mellem regnekraft og datalager stort set ligegyldigt så længe, der er fjernadgang. Det kræver dog, at adgang til serveren er velorganiseret med kontrol af adgang og brug.

Deltager 6 gør i interviewet opmærksom på, at han ser rækkefølgen af databehandling (fase 3) og dataanalyse (fase 4) omvendt. I hans forskningsarbejde er det nødvendigt først at foretage grundige analyser af den indsamlede empiri, før de kan gå videre til statistisk databehandling på tværs af materialet.

Med hensyn til anvendelse af metadata i fase 3 og 4, så er der generel enighed om, at det er meget projekt- og softwarespecifikt. Deltager 5 forklarer, hvordan mange oplysninger ligger i dokumentationen i de enkelte databaser, som de arbejder med. Derudover forsøger han, at automatisere processen med datarensning, så dokumentationen ligger i lagret i koden. Generelt er deltagerne enige om, at behovet for dokumentation og metadata er så specifikt, at det er svært at standardisere.

Behovet for understøttelse i fase 3-4 kan opsummeres til at omfatte:

- En infrastruktur, der har tilstrækkelig maskinkraft til at udføre tunge beregninger
- Adgang til (ekstern) serverplads

FASE 5 OG 7: DEPONERING, BEVARING OG DELING AF DATA

Fase 5 og 7 har både i spørgeskema og interviews fremstået tæt forbundne og de præsenteres derfor samlet her. Fra spørgeskemaet kan vi se, at opbevaring af forskningsdata i repositories eller datalagre sker i nogle, men ikke alle tilfælde. Deltagerne uddyber, at de generelt mødes af krav fra forskellige sider til at gøre data tilgængelige efter forskningsprocessens afslutning. Fra forlagene kan det være et krav, at data følger en afsendt publikation, så de kan kontrolleres i forbindelse med peer review-processen. Deltager 3 beretter desuden om, at det ofte i forbindelse med ansøgning om forskningsmidler kan være et krav, at forskningsdata gøres tilgængelige i forbindelse med afslutning af projektet. Disse krav kan ifølge spørgeskemaet ses afspejlet i de steder, forskerne vælger at dele deres data. Således deles forskningsdata gennem nationale dataportaler (2 deltagere) eller internationale dataportaler (1 deltager). Data journals er der blandt deltagerne ikke pt. tradition for at benytte til formålet.

Såfremt der ikke er stillet krav om deling af forskningsdata i et specifikt projekt, gøres det i nogle tilfælde, men ikke altid. For nogles vedkommende er der tale om deling af rådata alene, for andre deles også trinnene i data analysen (f.eks. den fulde syntaxfil i SPSS). Deles kun rådata, er det tanken, at publikationen formidler den dataanalyse, der har fundet sted.

At gøre data citerbare (dvs. at deponere og beskrive et datasæt i et repository og gøre datasættet citerbar) har ikke hovedfokus hos deltagerne. 4 ud af de 6 gør det sjældent eller aldrig, de resterende 2 gør det med en middel frekvens. Der er altså ikke en etableret måde at gribe dette an på hos forskerne.

I forbindelse med deponering, bevaring og deling af data udtrykker deltagerne behov for en bedre infrastruktur, der kan understøtte deres arbejde med deponering og deling af forskningsdata. Behovet udtrykkes på forskellig vis. Deltager 5 argumenterer for, at behovet primært opstår i forbindelse med egne indsamlede data og i mindre grad med data indsamlet fra eksisterende kilder som registerdata. På den anden side fremføres det, at det kan være en fordel for studerende at kunne tilgå datasæt, hvor ”købte” data har været genstand for en omfattende klargøring og databehandling. Flere ytrer (deltager 3 og 4), at et repository ikke må indeholde fortrolige data. Det mener deltager 3 dog i nogle tilfælde vil kunne løses ved en regenerering af analysedata. Et andet ønske fra deltager 3 er, at der laves en bedre formidling af de lagrede data. Det opleves som et problem, når data hos eksempelvis Dansk Data Arkiv i dag ikke er koblet til projektbeskrivelser, projektansøgninger, publikationer eller andre formidlede præsentationer af data. Det gør det svært at gennemskue data. Flere af deltagerne peger endvidere på, at det ville være en fordel, hvis den generelle procedure omkring upload af data blev mere standardiseret. Ønsket går fra, at der kunne være en politik på området (deltager 5) til at der stilles deciderede krav om det (deltager 3). Deltager 2 advarer dog om, at der udarbejdes standardiserede krav om deponering og deling af data. Han forudser, at det kan besværliggøre samarbejde med erhvervslivet om adgang til data. Hvilket organisatorisk niveau et sådant repository skal være på, er der dog ikke fuld enighed om. Nogle argumenterer for, at det skal være på universitetsniveau (deltager 4) m og her fremhæves et data repository på Harvard University (http://thedata.harvard.edu/dvn/) som et internationalt

eksempel på et institutionelt data repository. Andre mener, at det skal være på nationalt niveau med international adgang (deltager 3) og gerne i regi af Dansk Data Arkiv.

Behovet for understøttelse i fase 5 og 7 kan opsummeres til at omfatte:

- En infrastruktur, der:
  - kan understøtte deponering og deling af data
  - formidler datas kontekst
  - opererer på enten universitetsniveau eller nationalt niveau
- Politikker for eller måske endda krav om offentliggørelse af forskningsdata

FASE 6: PUBLICERING OG CITERING AF DATA

Som det fremgår af de ovenforstående afsnit, så er der et stigende krav om publicering af forskningsdata i forbindelse med generel publicering, enten til fagfællebedømmere eller til data repositories. Der er dog også deltagere repræsenteret, som ikke møder den slags krav fra tidsskrifterne (deltager 2 og 6). Dog er det stadig yderst vigtigt, at det formidles nøjagtigt gennem publikationer, hvordan dataindsamlingen er foregået i de præsenterede studier. Der kan også være krav til at data skal være tilgængelige. Dette krav møder forskerne i nogle tilfælde ved at formidle data via e-mail på forespørgsel (deltager 5). Derved spares den arbejdsindsats, der er forbundet med at klargøre forskningsdata til upload i et repository. Og netop arbejdsindsatsen forbundet med at klargøre forskningsdata til et repository betragtes som en belastning, der ofte udelades, hvis der ikke er deciderede krav til at dele data qua forlag eller bevillingsgiver. Adspurgt om, om det kunne være et incitament til at uploade forskningsdata, hvis det blev belønnet med BFI-point [1], svares der i fokusgruppen ja under forudsætning af meget klare retningslinjer for, hvad der udløser BFI-point.

Med hensyn til citering af forskningsdata, så er det ikke en etableret praksis i dag. Det forekommer, men er endnu ikke en fast bestanddel i forskningsprocessen. Det er en absolut forudsætning for citering af forskningsdata, at det er fuldstændigt klart, hvordan data er tilvejebragt (deltager 5). Almindeligvis er det formidlede data i artikler, der citeres i forbindelse med egen forskning, og i mindre grad fagfællers egentlige forskningsdata. Flere af deltagerne har dog oplevet at orientere sig i andres delte forskningsdata i forbindelse med egen forskning (deltager 4 og 5). Ligeledes er der flere, der går tilgængelige data igennem, for at tjekke andres arbejde indenfor forskningsfeltet (deltager 3 og 6). På samme måde er deling af egne forskningsdata med kolleger i og udenfor egen institution i nogen grad udbredt, hvad enten det sker via e-mail, fildelingstjenester som eksempelvis Dropbox eller gennem etablerede repositories.

I fase 6 udtrykkes der ikke direkte nogle behov for understøttelse i relation til publicering og citering af forskningsdata.

[1] Point, der bruges som baggrund for at tildele forskningsmidler til danske universiteter. Points tildeles på baggrund af den bibliometriske forskningsindikator, som måler publikationsaktivitet. Se evt.: http://ufm.dk/forskning-og-innovation/statistik-og-analyser/den-bibliometriske-forskningsindikator

FASE 8: OPDAGELSE AF DATA

Ifølge deltagernes svar i spørgeskemaet opdages forskningsdata primært igennem faglige netværk og kolleger samt ved omtale i tidsskrifter. På den måde betragter forskerne opdagelse af forskningsdata som en del af forskningsarbejdet (deltager 5).
Søgemaskiner og databaser i sig selv er i mindre grad anvendt til at identificere relevante forskningsdata. En forklaring på den begrænsede brug af databaser og søgemaskiner kan muligvis findes i den utilfredshed, deltager 3 udtrykte omkring manglen på metainformation, der følger data. Det er altså ikke tilstrækkeligt, at data stilles til rådighed i repositories. Skal de bruges til at opdage data, skal der være tilstrækkelige kontekstuelle metadata (også af faglig karakter) tilknyttet, for at det har en værdi for forskerne. En anden årsag kan være, at deltager 4 til tider oplever, at der er databaser, som han ikke har adgang til.

Behovet for understøttelse i fase 8 kan opsummeres til at omfatte:

- En infrastruktur, der formidler datas kontekst i form af metadata

KONKLUSION OG ANBEFALINGER

De samfundsvidenskabelige deltagere i denne undersøgelse er meget forskellige, hvad angår behov og præferencer i relation til data management. Denne konklusion vil opsummere en karakteristik af deltagernes praksis vedr. data management og herunder pege på de faser af datalivscyklussen, hvor der især er identificeret behov for understøttelse hos forskerne samt pege på hvilken form denne understøttelse skal have. Samtidigt er det vigtigt at understrege, at rapporten og herunder konklusionen skal læses med det forbehold, at undersøgelsen havde relativt få deltagere givet den korte tidsfrist for gennemførelse.

Undersøgelsen giver indblik i, at samfundsvidenskabelige forskere repræsenterer meget forskellige tilgange til at arbejde med empiriske data og herunder data management. Samtidigt arbejder deltagerne med en meget bred variation af data- og kildetyper, hvor de vigtigste som tidligere nævnt er statistikker efterfulgt af spørgeskemaer, interviews, feltnotater, samt tekster og dokumenter (inkl. lovgivning). Ligeledes repræsenterer deltagerne forskellige tilgange til dataanalyse. På tværs af disse variationer afspejler interviewene dog tydeligt, at empirisk arbejde med store datamængder og/eller tunge beregningsmodeller er en integreret del af deltagernes forskningspraksis. Deltagerne arbejder alle med så betydelige datamængder, at overvejelser vedr. data management er nødvendige allerede fra opstart af forskningsprojekter. Det vurderes, at to af deltagernes datamateriale er så omfattende, at det kan betegnes som egentlige big data.

Med hensyn til deponering af data, publicering og citering af data samt bevaring og deling af forskningsdata (fase 5, 6 og 7 i datamodellen) så repræsenterer deltagerne igen forskelle i deres forskningspraksis. Undersøgelsens resultater indikerer, at der indenfor det samfundsvidenskabelige forskningsområde ikke er et standardiseret krav om deponering af data og deltagerne udtrykker, at det er projektspecifikke forhold der afgør, hvorvidt de deponerer data i en form for arkiv/repository. I forlængelse heraf er det generelt ikke er så udbredt at publicere data,

og dermed gøre dem citerbare. Samtidigt er det dog tydeligt, at deltagerne er meget opmærksomme på den faglige udvikling indenfor deres forskningsmiljø, og de oplever det som et stigende krav og trend at gøre data tilgængelige i bred forstand.

På tværs af faserne blev der udtrykt forskellige behov for support i relation til data management, hvilket kan opsummeres som følger:

- **Behov for personlig rådgivning.** Behovet for personlig, skræddersyet support blev primært formuleret i relation til de første to faser. Her var der enighed om, at deres projekter og problematikker var så specifikke, at det ikke vil give mening at etablere kurser og lignende. I stedet skal der være adgang til projektspecifik rådgivning af fx juridisk eller teknisk karakter. Det kan enten være på institutionsniveau eller på nationalt niveau med en lokal kontaktperson.
- **Behov for support i form af infrastruktur.** I de fleste faser i datamodellen (fase 6 og delvist fase 8 undtaget) var der blandt deltagerne et kraftigt ønske om øget support i form af infrastruktur. Her var det specielt behov for øget lagerplads og forbedret mulighed for deling af data samt adgangsstyring. Derudover blev der identificeret et behov for øget maskinkraft til at udføre tunge beregninger.
- **Behov for support i form af overordnede politikker.** I forbindelse med publicering og tilgængeliggørelse af forskningsdata, blev der fra et par af deltagerne udtrykt ønske om en fælles eller overordnet politik eller endda krav om offentliggørelse af forskningsdata. Det er dog vigtigt at understrege, at deltagerne på dette punkt havde forskellige holdninger og oplevede forskellig praksis på området.

På baggrund af forskernes udtrykte behov anbefaler vi desuden, at der udvikles et sæt metadata, der kan beskrive og formidle status og karakter af data og datasæt i den ovenfor foreslåede infrastruktur. Udover en generel metadatafunktionalitet, skal formatet, for at understøtte samfundsvidenskabelige forskere i deres praksis omkring data management, specifikt kunne:

- give mulighed for at skelne mellem data og datasæt. Dette vil understøtte forskernes egen data management, men også deling af data med fagfællesskabet.
- præcisere kontekstuelle aspekter ved data/datasæt, så der kan opnås adgang til en faglig og mere omfattende fremstilling af, hvad et givent datasæt er baseret på i form af eksempelvis projektbeskrivelser, projektansøgninger eller publikationer. Dette vil assistere forskerne i deres vurdering af et datasæt, hvad enten det vurderes i forhold til bedømmelse eller anvendelse.
- eksplicitere adgangsmuligheder, så det sikres, at data/datasæt ikke tilgås af uvedkommende. Dette er en forudsætning for at sikre infrastrukturens troværdighed.
- tydeliggøre indholdsmæssige og administrative aspekter af indhold, så som versioner, kodningsomfang, ejerskab mv. Dette vil i kombination med de øvrige metadata bidrage til en mere koncis præsentation af data/datasæt, så identifikation og vurdering af data kan effektiviseres.

Rapport vedr. Afdækning E: De faglige miljøers behov og præferencer

Fagområde: Sundhedsvidenskab.

- Empiriske studier inden for bevægelse og biomekanik
- Empiriske studier inden for kardiologi
- Forskning i hormonsygdomme
- Translationel forskning, der danner bro mellem klinisk og mere traditionel forskning
- Epidemiologisk cancerforskning

Fase 1: Planlægning og ansøgning

Billedet omkring datamanagement planer (DMP'er) er meget broget. Nogle laver DMP'er afhængigt af det aktuelle forskningsprojekt. Disse DMP´er udarbejdes, da datatilsynet kræver det, og planerne følger deres standard. Andre anvender et system kaldet Odense Patient data Exploratory Network (OPEN) til datadeponering, og indbygget heri findes der planer til udfyldelse. De fleste udarbejder kun lejlighedsvis DMP'er, og de følger en egen standard. Disse indeholder typisk kun information om, hvor data kommer fra.

Der er generelt stor opmærksomhed på datatilsynets krav om beskyttelse af personfølsomme data samt guidelines fra UVVU hos mange af respondenterne.

I et tilfælde nævnes det, at der ved internationale samarbejder, udvikles planer for hvordan projektets database skal bygges op osv. men ellers ikke.

En enkelt respondent udtrykte stort behov for at kunne følge en standardiseret formular til udarbejdelse af DMP'er og for støtte i form af kurser og rådgivning. Dette skulle være på institutniveau, og man var villig til at betale for det (4 ud af 5). En enkelt anden vurderede vigtigheden af dette til 2 ud af af 5.

Fase 2: Indsamling/generering af data

Data består typisk af en blanding af analoge og digitale data. For de digitale data, som denne afdækning er mest fokuseret på, består data primært af data fra spørgeskemaer, blodprøve- og biopsi-resultater, data fra instrumenter, fx accelerometermålinger , transskriberede interviews, excel-ark og billeder (scanninger, røntgen). De analoge data som består primært af journaler, målinger og spørgeskema-data og nogle af disse scannes ind med henblik på analyse. Andre analoge data eller metadata findes i logbøger og laboratorie-bøger.

Der nævntes datamængder i størrelsesordenen 100GB til 2-4 TB pr. år (i et tilfælde nævntes data i størrelsesordenen 1 til flere TB pr. studie), men de fleste havde svært ved at sætte præcise tal på.

Data gemmes i høj grad på institutionens netværksdrev – enten universitetets eller hospitalets. Enkelte lægger deres data i OPEN, der er Region Syddanmarks forskningsinfrastruktur. Enkelte afdelinger har egne

servere, fx PACS. Data ligger ofte også på laptops, men kun i anonymiseret form. En enkelt respondent nævnte herudover USB-sticks som lager, da man løb tør for plads på netværksdrevene. Samme respondent nævnte, at arbejdet med at skulle fjerne såkaldt *følsomme data* fra netværksdrevene igen, er uoverskueligt og meget tidskrævende, så dette udføres generelt ikke, men da data ikke indeholder kontroversielle oplysninger, vurderes det som uproblematisk.

> ***"Jeg stoler blindt på universitetets IT-service/server. Der har aldrig problemer, så der er ikke tænkt yderligere på backup"***

Særlige behov knytter sig til plads. Der er som oftest ikke plads nok på netværksdrevene. Egne servere i forskergrupperne er dimensioneret til det enkelte projekt, så der er plads nok. Der er i nogle få tilfælde tale om samarbejde på tværs af landegrænser, hvor lagring er nødvendig. I et enkelt tilfældes nævnes Redcap, der er en webapplikation til opsamling af data. Der nævntes i et enkelt tilfælde, at der er behov for at virksomheder udvikler systemer, der er bedre end det, der benyttes nu. En anden respondent svarede, at data kun udveksles med udenlandske partnere i de tilfælde, hvor projektet er udført sammen med dem, og hvor data er skabt i forbindelse med det konkrete projekt. I disse tilfælde udarbejdes der en aftale, der klarlægger procedurer, ansvar og betingelser.

En respondent nævnte "manpower" til håndtering af data som et særligt behov, men at man mødes uden forståelse fra forskningsfondene, når der ansøges om midler til dette. Hvis der blev ansat en datamanager på instituttet, ville der blive rift om vedkommende, og man ville være villig til at betale for det, men man ville have svært ved at finde midlerne (5 ud af 5 på skalaen).

Det hænder med nogen hyppighed, at data forsvinder på grund af manglende backup.

Data metadateres kun i begrænset omfang i denne fase. I enkelte tilfælde laves der txt-filer til hver enkelt fil, men dette udføres automatisk af laboratorie-udstyret. Ved indscanning af analogt materiale opsamles scanningsmetadata. Metadatering udføres i et tilfælde ved hjælp af link til artikler, hvor data er beskrevne, hvis data er indsamlet fra andre. En anden respondent oplyste, at data mærkes, så der kan findes tilbage til den metode, der er anvendt ved dataindsamlingen. Det generelle billede i denne fase er, at oplysningerne om indsamlingsmetode m.m., ikke lægges ved data selv, men oplyses ved publicering af artiklen.

En respondent udtrykte ønske om at kunne ansætte en datamanager på instituttet, der kunne administrere data håndteringen, samt behov for rådgivning og hjælp, herunder både juridisk, teknisk og kompetencemæssigt på institutniveau.

> ***"Hele det juridiske og tekniske omkring databasefunktioner dvs. selve det tekniske, opbygning af databaser osv. er meget tung, og her kunne godt bruges hjælp. Der er heller ingen standarder for dette i de forskellige grupper trods lignende projekter."***

En ville gerne have et system til lagring, som fx Redcap.

Behov for ensretning fx i forhold til navnekonventioner (filnavngivning) og strukturering af data udtryktes også. Måske i form af etablering af enighed i gruppen – altså forstået som en kulturændring. Indtrykket ved denne respondent er, at den nødvendige viden og den tekniske infrastruktur er til stede, men der er behov

for, at man ændrer adfærd. Hjælp til de juridiske aspekter søges hos datatilsynet og det juridiske kontor på universitetet eller regionen.

I det omfang der nævnes behov for hjælp, foretrækkes lokale løsninger (institut eller fakultet), da de vil være hurtigere at implementere og udvikle på. Der nævnes en problematik omkring centrale løsninger, at de nogle gange drives af en organisation, der selv har interesser i de data, der lagres, og at der derfor mangler tillid til, at der er upartiskhed.

Fase 3: Behandling af data

Der opereres sjældent med behandling af data, der kræver supercomputere. Massespektrometri udføres f.eks. lokalt på egne servere. Det hænder, men der opleves problemer med at overføre personfølsomme data på tværs af firewalls (regionens/hospitalets og universitetets). Der nævnes et mis-match mellem regionens og universitetets sikkerhedsprocedurer, og at der pga. personalemæssige resurser ikke gøres noget ved dette. Det blev foreslået, at Danske Regioner bliver involveret i dette, da de enkelte sygehuse således ikke kan lave den udvikling af deres systemer, der skal til for at kunne udveksle data mellem lager og HPC.

De fleste databehandlinger foretages lokalt på almindelige computere. I de tilfælde hvor data udefra skal behandles, er det nødvendigt, at data skal kunne åbnes af den software, der anvendes.

En respondent oplyser, at det eneste krav/behov der er i forbindelse med databehandling, er tid til at transformere data. Datatransformationer dokumenteres generelt, og til disse procedure er der ligeledes behov for en datamanager hos flere af respondenterne. Opnås der støtte til en datamanager, der kan varetage dette, vil vedkommende kunne blive ansat på fuld tid i gruppen.

En enkelt respondent er ikke interesseret i en datamanager, da det er nemmere og hurtigere at gøre det hele selv.

Fase 4: Analyse af data

Langt overvejende udføres der statistiske analyser på de data, der skabes eller indsamles – typisk på almindelige computere. Visualisering af data foretages også. En enkelt nævnte, at der til projekterne er tilknyttet et antal humanister, der udfører tekstmining -og analyse, og at der til dette anvendes særligt værktøj. Det ville være værdifuldt at få kvalificeret hjælp til dette – det er svært at forstå de processer der ligger bagved f.eks. mønstergenkendelse. Denne støtte burde være på nationalt niveau – der er et stort efteruddannelsesbehov for håndtering af ”big data”. Al statistisk metode dokumenteres af de adspurgte.

Af behov for hjælp nævnes kurser i statistik på højt niveau men også personlig vejledning, som er konkret i forhold til det pågældende projekt.

Fakultetet nævnes som det niveau, der skal ydes hjælp fra, da den faglige kapacitet ikke findes på institutniveau. Andre nævner institutniveau som passende. Der er ikke i gruppen behov for en fagstatistiker på fuld tid, men vedkommende kunne så hjælpe andre grupper.

En enkelt respondent vurderede denne hjælp til 4 ud af 5 - en anden til 5 ud af 5.

Fase 5: Deponering af data

Af repositorier der anvendes, nævnes Dansk Data Arkiv (DDA) og OPEN, men de fleste deponerer ikke data i egentlige repositorier. De fleste lader data ligge på netværksdrevene, en enkelt oplyser, at data deponeres, hvis det er en betingelse, der knytter sig publicering i forbindelse med det enkelte projekt, men dette er sjælendt. I disse tilfælde er det forfatteren, der beslutter hvilke data, der skal deponeres.

En respondent nævnte GEO-baser (Gene Expression Omnibus) som deponeringssted. I dette tilfælde metadateres data som standard, men det er en udfordring, da filerne ofte er store, og det er svært at udføre. Der er behov for kompetente folk til at hjælpe med det. Nogle tidsskrifter kræver, at data lægges i en GEO-base. Det kan være forbundet med besvær at anvende andres data fra en GEO-base, da det nogle gange kræver et særligt program at læse filerne.

En respondent mener ikke, at der er faglige traditioner for deponering, men at det måske langsomt er ved at komme. En anden mente, at det er almindeligt i det pågældende fagområde, da det kunne give problemer med IP og kvalitetskontrol, hvis det ikke lægges åbent et sted, samt at det i stigende grad er et krav fra tidsskrifterne, men at det kræver mange resurser at foretage deponering i forbindelse med hver enkelt publikation.

Alle data bør gemmes for altid. Dette er den gennemgående holdning. Det nævnes endda som et etisk problem, at etisk råd kræver sletning af visse data, da det i de tilfælde så ikke er muligt for en forsker, at verificere sine resultater. Der er en hvis modstand mod at smide data ud, dels da de ofte fører til ny forskning, dels da det vil kræve for store resurser at skille de bevaringsværdige data fra dem der godt kan slettes. En enkelt nævner 10 år som standard, og 30 år for data, der har længerevarende interesse. En anden respondent nævnte 50 år som minimumsgrænse. De steder hvor data gemmes på lokale servere, foretages der ikke egentlig langtidsbevaring, men der ligger data fra 1994. I nogle tilfælde foretages der backup til privat indkøbte eksterne harddiske i hjemmet.

Datadokumentation af statiske data foretages ikke nok ifølge nogen af respondenterne. En enkelt respondent mente, at det bør indskrives i en Code of Conduct. En ønskede et system til dette. Nu bruges håndskrevne logbøger i laboratoriet. Dette er et problem i forhold til genanvendelse.

En anden oplyste dog, at de statiske data altid dokumenteres, og at det bl.a. foregår i forbindelse med publicering af en artikel. En respondent oplyste, at det i alle tilfælde beskrives hvordan data er deponeret.

Adgang til data er ofte kun for de personer, der har været involveret i projektet, hvor data blev skabt eller indsamlet. Derefter kan der gives tilladelse til andre, som skal bruge data i forbindelse med et samarbejde.

> ***"Datatilsynet kræver ofte ved større projekter, at det er et offentligt register som er paraplyanmeldt via sygehuset, og så er det institutionen, der ejer data. Registrene bør ikke være privat-godkendte og dermed privatejede. I de tilfælde kan fx en Ph.d.-studerende nægte andre adgang til data – selv vejlederen."***

Data gøres som standard ikke citérbare. En respondent oplyser, at alle deres data i DDA er publiceret, i betydningen som artikler, der så kan citeres.

En respondent sagde, at rådata helst ikke deles, men det bliver gjort, hvis der er en god grund til det og kun på lovlig vis.

Af hjælp nævntes rådgivning om, hvordan man organiserer data, så de nemt kan bruges af andre, og hvordan kan man anvende andres data og et fælles sted til store datasæt, som kan hentes udefra. En enkelt respondent mente, at de selv har styr på alle aspekter af data management, så alle større initiativer til indførelse af fælles systemer eller tilbud om udefra kommende personlig assistance ikke er ønskværdig, da de vil blive hæmmet i deres forskning. Det skyldes, at et fælles system sandsynligvis ikke vil kunne bruges eller er dårligere end det, de selv har. Ekstern assistance vil være så input-krævende, at det er nemmere at gøre tingene selv.

Fælles lagring opfattes af nogen som et problem på grund af IP-issues, og et fælles system er derfor ikke ønskværdigt.

En nævnte DDA som en god støtte, så der ikke er behov for yderligere.

En anden sagde, at man ønskede lokale løsninger, bare de opfylder bestemte retningslinjer eller standarder. Også af sikkerhedshensyn. Det skal ikke være muligt for en enkelt medarbejder, at have adgang til alle mulige data om identificerbare personer. Landsdækkende systemer vil ikke være hensigtsmæssige, da de ikke udvikles hurtigt nok i forhold til behovene (en pendant til Pure, hvor udviklingsønsker ikke imødekommes hurtigt nok), og der er fare for, at de ikke understøtter alle forskeres behov.

Der blev både nævnt universitets- og fakultetsniveau som yder af støtte. Den eneste, der kunne sætte tal på, sagde 2 på skalaen.

Fase 6: Publicering og citering af data

Respondenterne oplever kun i begrænset omfang krav om publicering af data – kun fra et antal tidsskrifter. En enkelt ser det som noget, der er på vej. Data udleveres case-by-case til andre, der henvender sig. Der udarbejdes typisk en kontrakt mellem samarbejdspartnerne, der udstikker regler for, hvem data må deles med. Et fåtal publicerer data sammen med artiklen, men ikke gennem data-tidsskrifter. En enkelt mente ikke, at de resurser, der er forbundet med publicering er umagen værd. DDA benyttes kun i begrænset omfang i forbindelse med publicering.

Datasæt citeres oftest ikke, og i de tilfælde de gør, er det via den artikel, de er anvendt i. En mente ikke at direkte citeringer af datasæt kan gøres, da de ikke er peer review'ede.

Citering af datasæt regnes ikke for meriterende endnu, men det burde det måske ifølge en respondent.

Der kunne være behov for hjælp til upload af store datasæt, både systemer og rådgivning nævnes, og det kunne godt være på nationalt niveau, men det er ikke så vigtigt, så støtten vurderes som 2 på skalaen.

En respondent nævner, at *embedded librarians* kunne være godt, for de kunne indgå mere i de konkrete projekter. De kunne finansieres via forskningsmidlerne som et krav via fondene eller hospitalerne.

Fase 7: Bevaring og deling af data

Det er meget forskelligartede svar, der gives på dette område. Nogle anvender DDA, andre ikke. De bruger egne systemer med backup eller lader blot data ligge på netværksdrevene på universitetet eller hospitalet.

De fleste mener, at data bør gemmes for altid, en enkelt mener ikke, at de data der skabes, er interessante efter 10 år.

Nogle mener, at hvis der var et system, skulle det være lokalt, og vil ikke betale for store eksterne systemer. Nogle forskere synes at opfatte langtidsbevaring og lagerplads som ens. Der er ikke en bred forståelse for hvad langtidsbevaring egentlig er.

Der er enkelte, der mener, at der er behov for støtte i form af systemer og rådgivning, men at det skal være på lokalt niveau – dog ikke til udvælgelse. Igen foretrækkes lokale løsninger for at sikre en armslængde mellem udbyder og "dataleverandør". Der udtryktes bekymring for om en større løsning ikke ville være i stand til at levere samme arbejdsintensitet som der er i forskningsmiljøet.

> ***"Jeg har gamle data-filer tilbage fra først i 80´erne – jeg har gemt dem i ASCII format, således jeg kan læse dem ind i statistikprogrammer m.m. – det tager lang tid. Når data først er læst ind er det intet problem. I sin tid var data gemt på floppy-disks og nu på CD-rom… men disse forsvinder jo også. Man skal kende filerne for at vide, hvordan man åbner dem. Jeg kan heller ikke sende data på disse filer, men tager dem selv ind, og konverterer dem, inden jeg sender til evt. samarbejdspartnere.***
>
> ***Man vil støde på dette problem igen og igen, og jeg ved ikke, hvordan man skulle kunne løse dette. Jeg har aldrig oplevet større problemer med ældre data – kun mindre problemer, som figurer der ikke længere kan laves, da det pågældende program ikke findes mere."***

Fase 8: Opdagelse af data

Opdagelse af data sker via litteraturen – især artikler og via den enkelte forskers netværk. Der søges kun i begrænset omfang efter data. Dette sker nogle gange for at lave metaanalyser eller for at få et indtryk af et område. Men i hovedsagen skabes egne data, eller de indsamles fra kendte baser. Søges der efter data, benyttes af en enkelt Google, ellers anvendes kendte databaser og PubMed.

Der er ikke udtrykt behov for støtte til at finde data. De ved generelt hvor og hvordan, man finder data i det omfang, det er nødvendigt. De gør hinanden opmærksomme på datasæt, hvilket opfattes som mere effektivt, da man får mere relevant information den vej.

Der var en enkelt, der udtrykte behov for en portal, udviklet på nationalt plan til opdagelsen af datasæt.

Overordnede, afsluttende spørgsmål

Nogle behov udtryktes ikke som behov for systemer, rådgivning eller kurser, men som at der er et behov for kulturændring i de enkelte forskningsgrupper, og at de unge forskere er nemmere at få til at gøre tingene på en anden måde, mens de ældre forskere er sværere. Denne kunne ændres ved at fonde og andre begyndte at stille krav. Men dette synspunkt modsiges af en anden, der frygter at oppefra kommende krav skal forringe allerede eksisterende workflows og systemer. Her er holdningen den, at man

hellere via overordnede regelsæt bør skitsere et antal minimumskrav, som lokalt forankrede systemer skal overholde, og så lade det være op til de enkelte institutter eller grupper at implementere det der bedst løser opgaven. Det synes dog som om, at der er et generelt behov for en øget bevidsthed om korrekt behandling af data samt et rådgivende organ enten lokalt eller nationalt.

De adspurgte respondenter har, til trods for at de alle tilhører det sundhedsvidenskabelige område, meget forskelligartede behov, kompetencer og ønsker. Hvis man skal nævne et punkt, hvor de ligner hinanden, er det bl.a. på behovet for, at data (for nogle betyder det alle data) bevares for altid. Dels fordi data vedbliver at have værdi (fx undersøgelser af store befolkningspopulationers helbredstilstand gennem et helt liv), dels fordi en udvælgelse af strengt relevante/interessante data er umådelig svær og tidskrævende.

Generelt var der en holdning til, at korrekt beskrivelse af data med henblik på genfinding og hurtig forståelse af data, er ønskværdig, og endda at der måske burde være krav om det. Centralt tilbudte systemer og viden opfattes af mange som en god ide, da man så ikke behøver at opfinde den dybe tallerken alle steder. Omvendt er der også en frygt for, at et fælles system ikke tilfredsstiller den enkeltes behov og i værste fald tvinger forskerne til at anvende ringere systemer med mindre og ringere forskning som konsekvens. Ej heller er der en udbredt tiltro til, at faglige kompetencer eller tidsmæssige resurser vil være til stede i tilstrækkeligt omfang i en central løsning.

Nogle respondenter nævner behov for ansættelse af personer med faglig indsigt til at håndtere data, fx i forbindelse med klargøring til DDA, transformeringer og lign.

Håndtering ad forskningsdata: Naturvidenskab

Interview runden for naturvidenskab er gennemført med deltagere fra Århus Universitet, DTU og Københavns Universitet. Udvælgelsen af grupper er sket sådan at der er den størst mulige bredde mellem grupperne, og samtidigt er der valgt de større forskningsgrupper da det her er mest sandsynligt at de udvalgte spørgsmål har været debatteret. Forskningsgrupperne har selv haft indflydelse på hvilken person der deltog i interviewet, det har i alle tilfælde været en person der formelt eller uformelt har rolle som den centrale IT og data person. Alle de gennemførte interviews viste at der, som ventet, var reflekteret over de fleste af spørgsmålene der er opstillet i interview guiden. Flere af de interviewede ytrede dog at det for dem var vanskeligt at differentiere mellem databehandling og dataanalyse, i disse tilfælde blev de 2 områder slået sammen som et i diskussionen.

Interviews har været individuelle, og rapporten baseret på notater fra disse interviews.

Fase 1 Planlægning og ansøgning

De interviewede personer delte sig i to klare grupper, den ene – ca. halvdelen – brugte i udpræget grad en struktureret datamanagement plan der allerede er på plads ved ansøgningstidspunktet. De samme grupper havde internationale standarder for data-management og annotering af data. Selvom disse grupper allerede har styr på datamanagement og annotering ytrede de alle at formaliserede kurser på området ville være værdifulde. Hvem der bedst ivaretager opgaven med kurser var der ingen enighed om, alt fra fakultetet til internationale organisationer var i spil. Disse grupper var også grundlæggende villige til at betale for kurser med forskningsmidler.

Den anden gruppe brugte mere ad hoc data-management, de kendte kun i begrænset grad til eksistensen af standarder i deres felt og advokerede typisk en tilgang hvor en uformel beskrivelse af data var tilstrækkeligt. Disse grupper var også interesserede i kurser og evt. rådgivning – heller ikke her var der enighed om hvor ansvaret for at udbyde kurser burde ligge, men interessen i at betale for dem med forskningsmidler var langt lavere.

Fase 2 Indsamling/generering af data

Alle de adspurgte grupper indsamler ret store mængder data, der er generelt tale om dataopsamling der baseres på digitale billeder, om end ikke konventionelle kameraer men i alle fald 2D matricer af data. Alle områder beskriver at den øgede opløsning på chip baseret dataindsamling oversætter direkte til øgede datamængder for dem, i næsten alle tilfælde taler man op eksponentiel vækst i data, hvor de grupper der i dag producerer mindst taler om 3-5TB med data per dag mens de store data indsamlere taler om 500TB pr da og venter at tale PB inden længe. Ikke alle data lagres men de lagrede data rapporteres i området 3-100TB pr dag. Datas formater varierer fra 1D sekvenser til 4D film.

Differentieringen mellem grupperne fra fase 1 er også her tydelig; de grupper der allerede har et modent datamanagement setup er også dem der indsamler mest data, en gruppe ventede at vokse med 3PB i 2015 (biologi). Disse grupper har også velorganiserede lagrings systemer og rapporterer generelt om veletablerede internationale netværk. Data lagres både lokalt og på internationale databaser, ikke alle data sendes til de internationale databaser, specielt personfølsomme data bliver altid lokalt. Disse modne datamanagement grupper har

middelstore installationer lokalt som drives professionelt. Nogle har ingen begrænsninger eller forskrifter der dikterer omgangen med data, mens andre har juridiske betingelser der specielt retter sig mod personfølsomme data. Alle metadaterer alle datasæt konsekvent og har internationale standarder for at beskrive datasæt. De ser håndtering af de voksende datamængder som en stor udfordring men der er ingen konsensus om på hvilket niveau det bedst håndteres lokalt/nationalt/internationalt. Der er dog enighed om at løsningen skal være skræddersyet til deres fagområde og de er villige til at finansiere det med forskningsmidler.

De mindre modne datamanagement grupper har også store og voksende datamængder – typisk nogle få hundrede TB pr år. Deres data lagres næsten eksklusivt lokalt, nogle bruger den lokale IT afdeling mens andre primært bruger eksterne harddiske. Disse grupper er meget bekymret over væksten i data og er alle i overvejelser om hvordan de skal organisere datamanagement i fremtiden. Alle grupperne ytrer også at backup er en stor bekymring mens deres data generelt ikke er underlagt juridiske begrænsninger og kan distribueres frit. Disse grupper ser også gerne koordinerede løsninger på deres fremtidige problemer, men der er ikke konsensus om hvorvidt det skal være skræddersyet til deres fag eller mere generelt, og på hvilket niveau det bedst placeres, dog med en vis veneration for lokale løsninger. Villigheden til at finansiere en løsning med forskningsmidler varierer fra helt sikkert til absolut nej, dem med fungerende lokal lagring var ikke villige til at betale men dem der måtte løfte alt i gruppen var meget villige. Også disse grupper beskriver at data metadateres og de fleste referer til klassiske laboratorie logbøger som en værdifuld metode til metadatering.

Fase 3 Databehandling

Differentieringen mellem databehandling og analyse er ikke åbenbar for alle – et par grupper har alle deres svar i denne gruppe og udelades fra Analyse afsnittet.

Alle de interviewede personer rapporter store behov til databehandling, anvendelser varierer fra rensning af sekvensdata over tomografisk rekonstruktion til foldning af multispectrale datakilder. Beregningsbehov er meget store for nogle, en gruppe rapporterer at de i dag bruger 2500 CPU kerner, mens andre beskriver mindre behov men specielt adgang til maskiner med meget RAM, typisk 1 TB, som deres behov.

Alle bruger UNIX baserede systemer som deres basis, men flere udtrykker at deres løsninger nødvendigvis må kunne anvendes fra alle udbredte operativsystemer, inkl Windows. Nogle metadaterer de processeserede data, igen de grupper der kunne kaldes modne datamanagement grupper, mens andre ikke gør det. Sidstnævnte begrunder manglen på metadatering af processerede data med manglende tradition og krav. Kun en enkelt gruppe fandt behov for støtte til databehandling, de mente eksplicit at det skulle koordineres internationalt og fagspecifikt.

Fase 4 Datanalyse

Data analysen ses af alle som dybt fagspecifikt og mens behovet er stigende var alle af den mening at de i dag har styr på de behov. Også her gælder det at man baserer sig på UNIX baserede systemer og næsten eksklusivt open-source software, men at man oftest ønsker at kunne anvende alle operativsystemer. Computer resurserne til analysen er i alle tilfælde mindre end dem der behøves til databehandlingen.

En enkelt gruppe ytrede at de gerne så kurser og vejledning i metadatering af analyse resultater og mente at dets fravær var begrundet i manglende tradition i feltet, de var dog ikke villige til at finansiere en sådan løsning over forskningsmidler.

Fase 5 Deponering

Ved deponerings spørgsmålene så vi igen to typiske grupperinger, de samme som tidlige. De grupper der har et modent datamanagement setup anvender også deponering og i nogen udstrækning deling. De der anvender deponering har typiske veletablerede formater og procedurer for det. Nogle deponerer alle resultater mens andre deponerer data der er anvendt i artikler og afhandlinger. Nogle har helt generel deling af resultater, andre deler kun efter personlig henvendelse og andre igen deler ikke pga. juridiske bindinger.

To grupper fandt det interessant med hjælp til deponering, begge på nationalt niveau. Den ene gruppe fandt det skulle være skræddersyet til deres behov mens den anden mente det kunne være generelt. Ingen af de to grupper var villige til at finansiere det med forskningsmidler.

Fase 6 Publicering og citering af data

Igen var det mønsteret at de grupper med modne datamanagement planer også var underlagt regler for publicering af data ifm. publicationer, disse grupper anvender også citering af data, om end ingen var afklarede med hvorvidt citeringer af datasæt var meriterende.

Der var ingen grupper der brugte Data DOI, de der havde data journaler, kun 2 grupper en moden og en umoden, havde begge fagspecifikke formater for data citering.

Der var ingen ønske om støtte til publicering og citering af data fra nogen af de interviewede grupper.

Fase 7 Databevaring og –deling

Det tidligere mønster var igen meget tydeligt; de grupper med modne datamanagement planer har også stringente opbevarings planer og regler samt systemer for eventuel deling af data.

De mindre modne grupper havde ingen databevarings traditioner, nogle fandt det uvigtigt mens andre fandt det vigtigt. De der fandt det vigtigt mente at bevaringen skulle ske lokalt men deles internationalt gennem netværk og standardisering. Der var dog ingen er var opsatte på at betale for en databevarings løsning.

Fase 8 Opdagelse af data

Alle de adspurgte grupper var ret uforstående overfor dette spørgsmål og vi må antage at det ikke rammer naturvidenskab særligt godt. De rapporterede alle at interessante datasæt findes gennem videnskabelige artikler, de grupper der kom fra miljøer med modne data-managementplaner havde i miljøet også dataportaler. Nogle enkelte rapporterede om data-journaler men at disse ikke var centrale i deres forskning.

Ingen havde interesse i nationale løsninger.

Afsluttende kommentarer

Mange grupper var bekymrede for udsigten til at skulle langtidslagre stadigt større datasæt, men de så det som en teknisk udfordring der bedst blev løst lokalt, de grupper der kunne

kategoriseres som modne beskrev at deres løsninger var langt billigere end hvad de kunne købe services til på det frie marked, og hvad deres internationale kolleger rapporterer de betaler for centrale løsninger.

To grupper som indsamler data fra internationale laboratorier beskrev at de meget gerne så bedre netværksforbindelser til disse steder så de kan flytte data over nettet fremfor at bære dem på eksterne harddiske som de gør i dag.

En gruppe udtrykte at de gerne så at der blev afsat midler til at understøtte deres metadatering om end de så et problem i mandskab da de mente at kun fagfolk ville kunne metadatere og disse nok ikke ville være interesserede i opgaven.

Håndtering af forskningsdata på den tekniske videnskab

Denne del-rapport er afrapportering for undersøgelsen af "Afdækning E: De faglige miljøers behov og præferencer" for den tekniske videnskab. Resultatet af fremkommet ved en række individuelle strukturerede interviews baseret på den udsendte interview-guide. Interviewene er optaget, transskriberet og efterfølgende analyseret til denne rapport.

Udvælgelsen har været baseret på ønsket om at interviewe forskere fra de tekniske videnskaber. Vi har primært koncentreret os om de ingeniør-faglige videnskaber, for at hindre for meget overlap til de naturvidenskabelige videnskaber. Samtidig ønsker vi at interviewe personer der kender til deres egen forskningspraksis, men som samtidig har kontakt til eksempelvis Ph.d.-studerende som led i deres ansættelsesforhold og eventuelt projektsamarbejde.

Undersøgelsen tager udgangspunkt i dagens praksis. I nogle tilfælde har der været fremsat ønske om at ændre praksis, men ellers er det op til læserne af denne rapport at danne sig et indtryk af, om der er behov for at ændre praksis.

Resultaterne er inddelt i faser, og omfatter bruttoliste over behov og eventuelle konkrete løsningsforslag. Alle holdninger til finansiering er samlet i et separat afsnit til sidst.

Fase 1: Planlægning og ansøgning

Der er både eksempler på ad hoc styring af forskningsdata, hvor der ikke er nogen større planlægning, til detaljeret planlægning. Sidstnævnte er afstedkommet af krav fra:

- Bevillingsgivere, f.eks. EU i forbindelse med adgang til avanceret udstyr mod at indsamlede data publiceres og kravene relateret til Horizon 2020
- Samarbejdspartnere, typisk andre universiteter og firmaer, hvor der stilles krav om eksplicitte ønsker til indsamling og organisering af data.

De forskere der har oplevet skift internt i organisationen, f.eks. skift af institut eller faggruppe, har oplevet udfordringen med at de selv eller andre har flyttet data fra et system til et andet, hvilket øger deres fokus på behovet for planlægning af lagring.

Endelig er det en faktor om projektet har deling af data som et direkte mål, eller det blot sker ad hoc. I de projekter, hvor deling er et eksplicit mål, er der allerede en større grad af planlægning.

Ud af de interviewede er der én, der har haft hjælp udefra til DMP. Der bliver dog generelt efterspurgt muligheder for at trække på skabeloner som byggeklodser til projektansøgninger m.v., som gerne må være baseret på standarder (eksempelvis ISO), ligesom der nævnes muligheden for, at andre overtager koordinering af DMP i stil med den assistance, der tilbydes i.f.m. projektstyring på AAU. Vejledningen kan også omfatte assistance omkring juridiske forhold i de projekter, hvor der eksempelvis er persondata eller kommercialiserings-interesser, der stiller øgede krav til omgangen med data.

Det er vigtigt, at de, der stiller kravene til DMP, også er med til at stille nogle af løsningerne til rådighed, således at et krav også efterfølges af minimum en instruks om hvordan det efterleves. Endvidere er det vigtigt at der er koblinger til de eksisterende systemer, hvor eksempelvis projekter og medarbejdere registreres med henblik på at styre økonomi, og redundant administration skal undgås, både af hensyn til tid, men også for at minimere risikoen for fejl. Dog skal alle systemer kunne håndtere undtagelser, så der

eksempelvis kan være tilknyttet en medarbejder, fordi vedkommende er aflønnet af projektet, men det kan være at vedkommende ikke må tilgå data pga. eksempelvis involvering i andre projekter.

Fase 2: Indsamling/generering af data

Data indsamles i et væld af formater. De datatyper, der er nævnt, er måledata fra laboratorier, lyd, video, billeder og survey-data. Data foreligger i både proprietære og ikke-proprietære formater. Forskningsdata kan også lagres i andre strukturer end filsystemer eksempelvis databaser. Endvidere er der data fra øvrige kilder ud fra den skelnen, der er opsat i interviewguiden mellem forskningsdata og kilder.

Laboratorie-data lagres ofte lokalt i laboratoriet, inden de overføres til en anden it-infrastruktur, hvor de kan bearbejdes. Generelt kan lagringen af data opdeles i følgende:

- Egen computer, måleinstrument eller eksterne enheder (f.eks. kamera, mobiltelefon)
- Netværksdrev
    - Lokale installationer i eksempelvis forskningsgruppen
    - Institutionsbaserede løsninger
- Offline lagring, eksempelvis DVD'er, USB-harddiske
- Samarbejdspartneres infrastruktur
- 3. parts-systemer, eksempelvis DropBox, Amazon m.fl.
- (Ikke fil-baserede løsninger, f.eks. databaser)

Data flyttes løbende imellem disse former for lagring afhængig af f.eks. behovet for at kunne tilgå filer hurtigt, hvis f.eks. netværket ikke er hurtigt nok eller der ikke er den nødvendige kapacitet. Der er tillid til, at de institutionsbaserede systemer fungerer, som de skal. Der er generelt ikke meget viden om begrænsninger i eksempelvis kapacitet, back-up-rutiner o.lign. Ved behov for at gemme store datamængder, som ikke nødvendigvis skal indgå i den videre proces, men gemmes til eksempelvis senere brug, kan der allerede i denne fase deponering til billigere lagringsløsninger, f.eks. usb-harddiske.

Der er behov for deling med adgangskontrol/-begrænsning både inden for institutionen selv, og med samarbejdspartnere (firmaer, organisationer og andre uddannelsesinstitutioner), ligesom der bliver peget på muligheden for at have flere indgange til data, f.eks. SFTP, webbaseret interface mv. afhængige af de enkeltes behov og datatype. Andrew File System (AFS) blev nævnt som en mulig løsning, da man har erfaringer med dette tidligere. Årsagen til valg af 3. parts-systemer eller egne løsninger er et udemiddelbart overblik over udgifter, nem tilgængelighed, overblik over løsning og let at ændre konfiguration på, hvis kravene skifter.

I projekter, hvor data er af særlig fortrolig karakter og der samarbejdes med virksomheder, er data nogle gange pålagt store restriktioner pga. sikkerhed, at data ikke kan forlade virksomhedens systemer. Dette giver nogle særlige udfordringer både i denne fase, samt fase 3, 4 og 6. Dette opleves også ved samarbejde med andre universiteter, hvor lokale sikkerhedspolitikker forbyder brug af 3. parts-systemer.
Der nævnes et ønske om at institutionens it-service kan sikre data i henhold til en given standard accepteret af virksomheden, så data kan behandles med andet hardware og software end det virksomheden der har data, kan stille til rådighed.

Data forsynes ofte med metadata, men tit i et ad-hoc format, der gør, at data kan forstås inden for tidsrammen af projektet undtagen i projekter, hvor deling af data er en leverance. I nogle domæner – og deraf brugte filformater – er der flere muligheder for struktureret brug af metadata, end i eksempelvis de eksperimentelle ingeniørfag, hvor fysiske opstillinger ikke har standarder for beskrivelse. Traditionelt har tekniske rapporter fungeret som metadatabeskrivelse.

Fase 3: Behandling af data & Fase 4: Analyse af data

Da det umiddelbart inden for domænet kan være svært – på basis af interviewene – at skelne mellem fase 3 og 4 har vi valgt at behandle disse under et.

Som beskrevet i fase to, så lagres data mange forskellige steder. Med mindre der er tekniske hindringer, så bliver mindre datasæt ofte kopieret til egen computer for at arbejde på dem. Det opleves derfor ofte som et tilbageskridt, når data af den ene eller anden årsag skal opbevares i lukkede miljøer pga. datasikkerhed, da dette sætter begrænsninger på den mulige behandling og videre analyse af data.

Data analyseres og behandles med eget eller standard-software, hvor der ofte indgår modeller og algoritmer. Der er behov for at kunne versionere software, og i nogle tilfælde også data.

Muligheden for brug af HPC nævnes af flere, hvor nogle også har opsat egne cluster-løsninger for at øge beregningskapaciteten. Interviewet er ikke kommet nærmere ind på de specifikke behov for HPC, herunder infrastruktur, kapacitet, software mv. Det bør afklares i et særskilt projekt med fokus på HPC-behov.

Data påføres metadata om behandling og analyse i det omfang det er muligt, men ikke nødvendigvis med henblik af brug på data på længere sigt, hvor ophavsmanden ikke længere kan bidrage til fortolkningen.

Fase 5: Deponering

I dag foregår deponering typisk i de infrastrukturer, hvor data i forvejen findes eksempelvis fælles netværksdrev, eller i nogle tilfælde også på forskerens eget netværksdrev. I nogle tilfælde findes data kun på forskerens egen computer eller ekstern lagring i laboratoriet. Deponering sker primært for, at forskeren og dennes forskergruppen kan genfinde eller reproducere forskningsresultater. Det er sekundært, at andre interessenter kan få adgang/kopier af data til validering eller genbrug. Egentlig deponering initieres ofte af ansættelsesophør - ph.d.er primært.

Genanvendelse kan være vanskeliggjort af at compilere, software, specifikke operativsystemer data ikke er bevaret/gemt som samlet hele – og dermed kan data ikke ses/afvikles. En løsning på dette kan være deponering af komplette kørende systemer (image af virtuel maskine).

Der er fagmiljøer, hvor måledata, fra egen-byggede modeller der har en fysisk form, er tæt på umulige at fortolke/anvende uden nærvær af den fysiske model f.eks. en robotarm, bølgeenergianlæg o.lign.

Der er en større tradition for at deponere software via strukturerede systemer som eksempelvis SVN og GitHub.

Et konkret forslag går på deponering ved direkte kopiering over i lokalt *institutional repository* (VBN)

Fase 6: Publicering og citering af data

Der er ikke udtrykt en modstand imod at dele data med andre inden for det akademiske forskningsmiljø. Dog er der en række forhold, der gør, at forskningsdata ikke nødvendigvis kan eller skal deles. Her nævnes:

- Ønske om embargo-periode for selv at kunne udnytte forsknings- og udgivelsespotentialet i data.
- Ophavsret.
- Personhenførbare data
- Patent og kommercialiseringsinteresser
- Ønske om kontrol af hvem der tildeles adgang

For de projekter som har datadeling som mål skabes der ofte hjemmesider med data og kode mv. Flere forskere har dog oplevet at hjemmesiderne har en begrænset levetid pga. ikke kontinuerlig understøttelse af den infrastruktur, der er anvendt under publicering. Det kan skyldes, at det er projektet egen webserver.

Deling af data sker ofte af mere uformelle veje i form af henvendelser på baggrund af artikler, anden publikation, præsentation o. lign. Deling sker i så fald i form af mail, DropBox, egen hjemmeside og deslige. Andre deler i open source fællesskaber relateret til deres fagdomæne.

Der er et ønske om muligheden for at kunne citere data f.eks. i form af DOI. Dette er desværre ikke oplevet som en nem proces på datasæt i modsætning til en artikel. Ligeledes vil man gerne kunne lave stabile links til filer der hvor de allerede ligger f.eks. på et fileshare.

Det skal være mulig at fraskrive sig ansvar for andres brug af forskningsdata og softwarekode.

Noget data beriges også i denne proces med metadata, så de kan forstås og bruges af andre.

Fase 7: Bevaring og deling af data

Vi har ikke fundet nogle forskere, der langtidsarkiverer data i særlige strukturer beregnet til dette. Data lever i de strukturer hvor de anvendes under behandling og analyse. I nogle tilfælde, hvor der er store datamængder flyttes data til andre lagringsmedier, der er billigere men offline.

Hvis data skal flyttes, kan der være særlige forhold som eksempelvis ændring af sti-strukturer, som gør at flytning til et andet system gør data uanvendelige. Så hvis dette skal foregå, så skal det indgå i planlægningsfasen, så der vælges løsninger hvor flytning er muligt.

Ansvaret for langtidsbevaring bør ligge hos den instans der stiller kravet eller den institution det påhviler at opfylde kravet.

Det er mange af de samme udfordringer og pointer i denne fase som allerede er nævnt under fase 5.

Fase 8: Opdagelse af data

Genopdagelse/genanvendelse af forskningsdata sker typisk på baggrund af publikationer og networking. Data som deles via hjemmesider forsvinder ofte over tid. Adgangen til andres data kan være både offentlige, beskyttet eller licensbelagt.

Der menes ikke at være et umiddelbart behov for hjælp til at finde data. Erfaring med hjælp til patentsøgning har ikke været positive, og det spiller ind i holdningsdannelsen.

Finansiering

Der er en samstemmig holdning til, at dem, der stiller kravene, bør også være den finansielle ansvarshavende. Basalt set bør udgiften til opbevaring, sikring og langtidsbevaring afholdes af universitet. Hvis ikke det bliver finansieret af basismidler (universitet), så er der stor risiko for at data forsvinder ved projektets ophør eller for hurtigt derefter. Generelt er der en holdning til at hvis der skal forekomme egenbetaling, så skal det som minimum være på institutniveau eller højere.

Pris er en faktor – det må ikke være dyrere end at gøre det selv (købe en usb-harddisk).