e-Science og forskningsbibliotekerne

I dagens forskningsverden anvendes computere i stort set alle hjørner af videnskaben. Forskningsdata er mange steder "født digitale", og analyser og publikationer er ligeledes næsten helt og holdent digitale. Forskningsbibliotekerne må forholde sig til de udfordringer, som denne udvikling stiller til egne kompetencer og services over for forskerne. Velkommen til e-Science.

Af Søren Bertil F. Dorch, Det Kongelige Bibliotek / KUBIS bfd@kb.dk

MINITEMA
ENDNU
"SO EIN DING?"

Hvad er e-Science i grunden – er det blot endnu et buzzword? Begrebet er ikke veldefineret. Er "eScience" for eksempel et registreret varemærke fra American Chemical Society? Betyder det højenergifysik på grid-netværk? Handler det om forskningsregistrering af data? Ordbøgerne giver ikke svaret, men det er faktisk "ja", til alle tre ting. De oftest anvendte beskrivelser på nettet stemmer ordret overens med Wikipedia, der lige som andre kilder, fører begrebet e-Science tilbage til John Taylor (1999), generaldirektør for det britiske "Office of Science and Technology". På det britiske e-Science Centers website kan man læse, at "e-Science handler om globalt samarbejde inden for videnskabelige nøgleområder, og den næste generation af infrastruktur, der muliggør dette samarbejde". Men det er ikke den eneste definition.
Det viser det sig nemlig, at e-Science kan defineres på mindst to måder, og man kan tale om enten ekskluderende eller inkluderende definitioner. En "ekskluderende definition" er, at e-Science udgøres af de forskningsområder, der "genererer digitale data", fx via numeriske simuleringer på supercomputere. En variant er, at det kun handler om "digitalt fødte data". Sådanne ekskluderende definitioner findes typisk i snævre fagkredse, hvor data netop er af denne type. "Inkluderende definitioner", betegner e-Science som de videnskaber, der helt eller delvis er afhængig af computere og informationsteknologi. Det vil sige, stort set al moderne forskning. Dette er eksempelvis præcis definitionen af e-Science, på Københavns Universitets webside for kandidatuddannelsen i e-Science, ved det Naturvidenskabelige Fakultet.

Forskningens vilkår
Hvordan man end definerer det, er e-Science ét af den moderne forsknings vilkår. Tony Hey, vicepræsident for Microsofts forskningsafdeling, taler om "the data deluge" - den digitale oversvømmelse af forskningen: Når det kommer til dataindsamling, -bearbejdelse og -arkivering, så udgør de digitale mængder en veritabel syndflod. Både for producenterne, for infrastrukturen, og for brugerne. ET udgangspunkt er Moores Lov, som er en empirisk sammenhæng, der illustrerer, at computernes beregningskraft fordobles cirka hvert andet år – så der kan foretages stadig større og mere komplicerede computerberegninger – med tilsvarende datamængder til følge. Derudover er priserne på IT-udstyr faldet, lagringspladsen er blevet større, internettet en smule hurtigere og en del mere udbredt, og alle grene af videnskaben har taget IT til sig.

Hey mener, at man kan tale om et nyt forskningsparadigme: Forskningen er blevet data-centrisk og eksperiment, teori og simulering forenes i e-Science. Samtidig mener han, at de værktøjer, der skal til for at hjælpe forskerne til at overkomme oversvømmelsen, skal komme fra anvendt datalogisk viden (og fra Microsoft, selvfølgelig). Disse værktøjer skal blandet andet hjælpe forskerne med automatisk dataindsamling, metadata, annotering, dataplads, datamining, databevaring, analyse m.m. Det vil sige, det er værktøjer, som kan understøtte og effektivisere dele af, eller hele forskningens *workflow*.

I det mest ideelle af alle tilfælde, er forskningens workflow en lineær, fremadskridende proces med mange led, fx: En idé fødes, en hypotese opstår, et eksperiment foretages, data indsamles og analyseres, hypotesen verificeres (eller falsificeres), og den ny viden formidles gennem publicering. Denne simplificerede beskrivelse af forskningsprocessen, er typisk den vi får præsenteret i en videnskabelig artikel med afsnit, såsom introduktion og baggrund, teori og metode, dataindsamling, analyse, og konklusion. I praksis er forskningsprocessen dog noget mere ulineær, og kan bedre beskrives som en konstant tilbagevenden til ideer og antagelser, samtidig med gentagelse og forbedring af eksperiment, hypotese og analyse. Indimellem "drysser" der publikationer ud - opfang og frekvens afhænger af faget.

> "Hvordan man end definerer det, er e-Science ét af den moderne forsknings vilkår"

Hvor den overordnede forskningsproces er forholdsvis kaotisk, kan man i stedet vælge, at fokusere på isolerede workflows, fx på processen fra rådata til visualisering af data. Det vil sige, hvordan man kommer fra et datasæt til en graf, et billede eller et diagram. I sådanne workflows anvendes ofte specialiseret software til at analysere og *reducere* data. Når data reduceres bliver mængden af data mindre - i og med at information smides væk - men datatypen og dataformatet skifter også. For at bibeholde information om datareduktionen - fx om hvilke data, der er smidt væk - kan man tænke sig, at tilpasse dataformatet til workflow: At indtænke workflowet på denne måde, i beskrivelsen af data, giver i princippet mulighed for "bedre" forskning, i og med, at

det fx er lettere at reproducere forskningsresultater. Dette er et eksempel på såkaldt *workflow provenance*. Og det er værktøj til workflow provenance-*services*, som Tony Hey mener, skal redde forskningen fra den digitale oversvømmelse.

Biblioteksbriller

Hvis man med biblioteksbriller tænker på elektronisk videnskabelige litteratur, i stedet for elektroniske videnskabelige data, så er konklusionen nærliggende: Indsamling, organisering, arkivering m.m. er jo klassiske bibliotekskompetencer. Blandt andet derfor er biblioteker interesserede i e-Science. Og måske er det naturligt, at bibliotekernes definitioner af e-Science – undertiden også kaldet e-Research – er inkluderende. Et forslag kunne være:

"e-Science er forskning, der er muliggjort af samarbejde via internettet, som anvender digitale datasamlinger og -ressourcer, samt teknologier, der faciliterer denne forskning".

Denne definition knytter dermed e-Science tæt til andre begreber, fx Open Data, grid-teknologi (samarbejdende computere i globale netværk), "scientific computing", forskernetværk, forskerservices m.m.

I bibliotekerne fokuserer vi typisk på services, og hvis e-Science er selve forskningen, så er e-Science som genstandsfelt omgivet af en række understøttende services, værktøjer og teknologier, der til sammen udgør et forskningsmiljø for forskeren. Virtuelle forskningsmiljøer er nært beslægtede med begrebet cyberinfrastruktur m.fl. og søger ofte, at bidrage med services til at understøtte forskningens workflows.

Selvom tankerne om e-Science, digital oversvømmelse, forskningens workflow og virtuelle forskningsmiljøer til dels er udsprunget af problemstillinger inden for naturvidenskaberne, så er e-Science - specielt set fra bibliotekerne - ikke begrænset til naturvidenskab. De kompetencer, vi som biblioteker forbinder med e-Science, kan i princippet anvendes inden for alle former for videnskab, hvor der foreligger digitale data: Det vil sige, data der er digitalt genererede, digitalt indsamlede, eller digitalt bearbejdede efter indsamling.

Virtuelle forskningsmiljøer

Rumforskning er et eksempel på et tværfagligt område, hvor forskningsbibliotekerne gerne vil på banen: Inden for rumforskningen kan data have mange forskellige former og være fremkommet på forskellige måder. Rumforskning spænder fx over både biologi, kemi, geologi, geodæsi, kommunikationsteknologi og rummedicin, til astronomi og astrofysik. Selv inden for astrofysik kan digitalt fødte data variere vidt og bredt i omfang, format og indsamlingsmetode: Data kan være indsamlet af et eller flere af højteknologiske instrumenter på en satellit, de kan stamme fra et apparat på en ballon, fra et avanceret teleskop i Chile, fra en gravko på Mars, eller data kan komme fra en computergenereret simulering af alt fra solen til The Big Bang. Det vil sige, at der er en ufattelig variation. Ikke desto mindre, er rumforskning og astronomi ét af de områder, hvor talen ofte falder på e-Science, virtuelle forskningsmiljøer og workflows.
Et eksempel på virtuelle forskningsmiljøer inden for astrofysik er ideen om *virtuelle observatorier*, der har været på astronomernes dagsorden siden 1990erne. Dette skyldes blandt andet den digitale oversvømmelse, samt den øgede konkurrence om adgangen til de stadig færre, men dyrere og større teleskopfaciliteter. Og for at kunne bygge virtuelle observatorier, kræves stan-

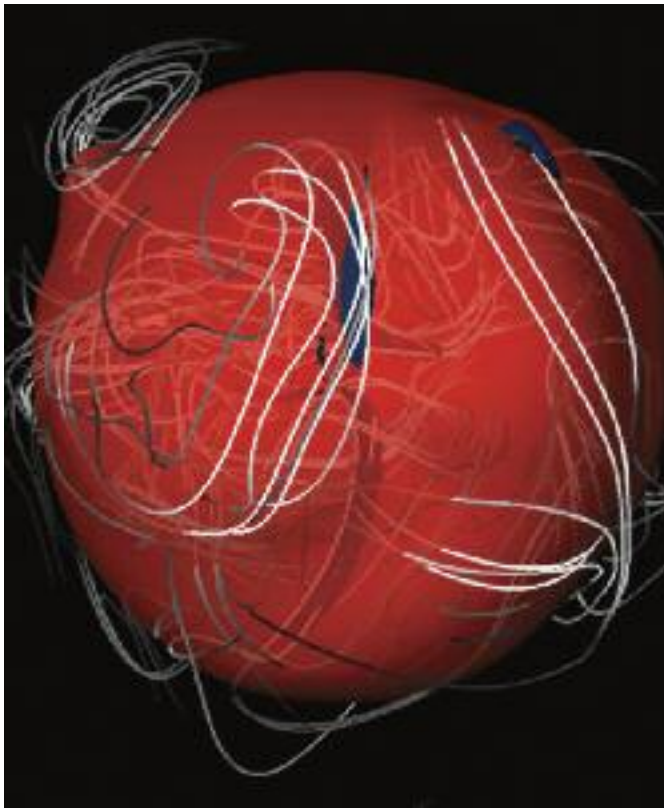
Kilde: NASA & T. Brown

Ved hjælp af IVOA lykkedes det et hold franske og tyske astronomer, at finde en større mængde galakser, af en ny type. Resultaterne er publiceret i tidsskriftet Science (2009)

dardiserede datatyper og formater, og inden for astronomi er der en lang og relativt veletableret tradition for dette. Et virtuelt observatorium er realiteten en platform, fx et website, hvor forskere kan få adgang til og analysere arkiverede data, fra flere forskellige astronomiske instrumenter verdenen over. Via et virtuelt observatorium kan man også forberede, bestille og foretage en observation eller beregning, hvorefter de indsamlede data bliver tilgængelige og kan analyseres. Det vil sige, et virtuelt observatorium er designet til at tage hånd om forskningens workflows.

Der findes i verdenen en lang række projekter om at etablere virtuelle observatorier og tilhørende standarder. De virtuelle observatorier var i første omgang nationale satsninger, fx National Virtual Observatory i USA, men i 2002 etableredes både et fælles europæisk virtuelt observatorium og det Internationale Virtuelle Observatoriums Alliance (IVOA). I de senere år, er IVOA og andre lignende projekters arbejde begyndt at bære synlig frugt, i form af flere publicerede forskningsresultater, der er baseret på data og analyser muliggjort af virtuelle observatorier.

I Danmark arbejder Det Kongelige Bibliotek blandt andet med et projekt om data fra NASAs Kepler-satellit, sammen med via partnere på Aarhus Universitet. Målet er både at langtidsbevare data, men også at tilbyde forskerne et virtuelt samarbejdsmiljø omkring data. Og Det Kongelige Bibliotek/KUBIS har faktisk en – i moderne sammenhæng – længere historie mht. rumforskningsdata: I 2002 oprettedes et virtuelt observatorium på Københavns Universitet, der primært var rettet mod studerende. I 2007 fik Det Kongelige Bibliotek/KUBIS midler fra Kulturministeriets Forskningspulje, til at opgradere og videreføre observatoriet under navnet Urania – Astrofysisk Virtuel Observatorium". Og i år fik et nyt KUBIS-projekt, til udvikling af et "Virtuel Space



Kilde: Bertil F. Dorch, i tidsskriftet Astronomy & Astrophysics, september 2004

Visualisering af data genereret via en 3d-computerberegning af magnetfelter på stjernen Betelgeuse. Et eksempel på e-Science med data, der er født digitale.

Science Environment", støtte fra Det Strategiske Forskningsråd. Et andet forskningsområde, der ligesom astronomi har en lang tradition for indsamling og klassifikation af store og forskelligartede mængder data, er arkæologien. Også her er man forholdsvis aktive mht. at understøtte arbejdet med digitale data. Et eksempel er Archaeology Data Service (ADS) støttet af det britiske forskningsråd for humaniora, der skal supportere forskning og undervisning med digitale ressourcer, blandt andet ved hjælp af langtidsbevaring af digitale data, og ved at bidrage til formidlingen af arkæologiske data. Derudover, er det ADS' opgave, at udvikle "gode" workflows mht. brugen af digitale data inden for arkæologi, samt at understøtte forskningen med vejledning og implementering af IT i forskningen.

Hvor ADS fokuserer på services over for arkæologer, der arbejder med digitale data, er der flere eksempler på virtuelle forskningsmiløjer, tilknyttet diverse forskningsprojekter. Eksempelvis Virtual Environments for Research in Archaeology (VERA) et projekt, der ønsker at skabe en virtuel forskningsverden for arkæologer, blandet andet ved at skabe en webportal med værktøjer til kommunikation om forskning, forskerne imellem.

Også her er Det Kongelige Bibliotek/KUBIS på banen: Et nystartet projekt sigter på at beskrive kravene til et system, der kan bevare og formidle arkæologiske data, som forskere ved Københavns Universitet indsamler, fx under projekter i udlandet.

e-Science i Danmark

På trods af at hverken ordbøger eller aviser skriver ret meget om e-Science, så sker der alligevel en del i Danmark. Gennem de sidste fem år er begrebet gradvist blevet anvendt mere på nettet og i medierne. Som nævnt findes der en tværfaglig kandidatuddannelse i e-Science på Københavns Universitet, der er beskrevet af Undervisningsministeriet på "UddannelsesGuiden.dk", hvor der står, at "eScience handler om at bruge computere og it i forskning og videnskab fx til at udvikle nøjagtige vejrudsigter eller kortlægge menneskets arvemateriale".

En mængde danske initiativer og projekter falder inden for den brede definition af e-Science: Stort set al videnskab foretaget med computere ved Danish Center for Scientific Computing (DCSC) er født e-Science, og inkluderer astrofysik, bioinformatik, kemi, klimaforskning og økonomi m.m. Derudover er DCSC styrende i forhold til grid-infrastruktur i Danmark: Grid-begrebet, der omhandler beregninger på vidtspredte computere over nettet, er beslægtet med, men ældre end e-Science, og grid-aktiviteterne stammer fra behovet for computerkraft, hos de højenergifysikere, der deltager i eksperimenter på CERN i Schweiz. Udvikling begyndte i 2000 da Niels Bohr Institutet, blev aktivt involveret i det CERN-ledede EU-projekt DataGrid. I 2002 blev "Nordic Data Grid Facility" (NDGF) etableret med finansiering fra Nordisk Ministerråd. Efterfølgende blev "Danish Center for Grid Computing" (DCGC) etableret, for at fokusere de danske grid-aktiviteter. Danmark har således fremragende kompetencer og infrastruktur, i forhold til e-Science, der benytter sig af gridteknologier.

Udover disse aktiviteter, der genererer digitale data i stort omfang, findes der også en mængde mindre projekter og aktiviteter, men e-Science i bredere forstand er stadig et relativt uudforsket område.

Et andet ord, der i Danmark ofte høres, når talen falder på e-Science - fx hos NordForsk eller Forsknings- og Innovationsstyrelsen (FI), er "forskningsinfrastruktur". Dette skyldes, at EU-Kommissionen mener, at forskningsinfrastruktur er "faciliteter og ressourcer, der leverer væsentlige ydelser til det offentlige og private forskersamfund".

Derfor er forskningsinfrastruktur også en del af regeringens globaliseringsstrategi, og man ønsker at øge investeringerne på området: I følge FIs hjemmeside anvendes forskningsinfrastruktur på samtlige videnskabelige hovedområder og omfatter en bred vifte af avanceret udstyr, databaser, laboratoriefaciliteter, forsøgsanlæg samt andre værktøjer og faciliteter, der er nødvendige for forskningsprocessen.

Også bibliotekernes eget ministerium og organisationer er begyndt at indtænke e-Science, især mht. problemerne omkring adgang til og organisering af forskningsdata: I sit høringssvar til Nordforsks "Nordic eScience Action Plan" skriver DEFF i 2009, at "DEFF finder det essentielt, at forskningsbibliotekernes vitale rolle i forskningens og de højere uddannelsers infrastruktur udnyttes optimalt i forbindelse med planlægning af eScience. Hvordan begrebet eScience end anskues, ender vi i sidste instans med store sæt af primære forskningsdata som skal behandles som bibliotekariske objekter [...] Det er præcis dét, forskningsbibliotekerne gør i dag med dokumenter."

DEFFs Programgruppen for Informationsforsyning gjorde det i sin handlingsplan for 2009-2010, til en målsætning, at gennemføre pilotprojeker, der skal bidrage til en afklaring af bibliotekernes rolle mht. repositories for forskningsdata. Derudover lægger

DEFF vægt på open-access til data, på samarbejde med eksisterende datacentre og med aktiviteterne i Knowledge Exchange og NordForsks eScience initiativer.

Har universitetsbibliotekerne en rolle?
Via DEFFs internationale samarbejde i Knowledge Exchange, indgår de danske forskningsbiblioteker i arbejdsgrupperne for "Primary Research Data" og "Virtuel Research Environments". De andre partnere i samarbejdet er på samme måde som DEFF, ved at finde sine fødder mht. e-Science; allerede i 2003 udgav JISC rapporten "e-Science Curation Report", der angav bibliotekerne som en vigtig aktør mht. e-Science. JISC indgår også i en mængde samarbejder omkring e-Science, og finansierer projekter, som fx "Arts and Humanities e-Science Support Center", sammen med det britiske forskningsråd for humaniora. Det virker givet, at e-Science vil påvirke relationen mellem forskningsbiblioteker og deres brugere. Men i øjeblikket faciliteres denne type af services og funktioner delvist af andre organisationer, fx datacentre, grid-infrastrukturer, samt mere eller mindre permanente forskningsorganisationer.

I sin artikel i REVY nr. 2 (2010), skrev Jakob Nedergaard Mortensen om forskningen i forandring, og citerede den nylig rapport af Hans Siggaard Jensen for at rejse spørgsmålet om forskningsbibliotekerne skal "være infrastruktur eller medskaber af viden?". På tilsvarende vis, skrev American Research Libraries i 2007, i en rapport om e-Science, at "forskningsbibliotekerne oftest ikke opfattes som del af den forskningsinfrastruktur, der er under udvikling"; rapporten foreslår, at bibliotekerne bliver mere involverede med forskerne, og stiller spørgsmålet, om "vi kan genopfinde forskningsbiblioteket for e-Science?" Et svar på spørgsmålet i Siggard-rapporten kan være "begge dele": Forskningsbiblioteket er nødt til, gennem nye forskningsinfrastrukturer og services, at understøtte forskernes vilkår i en e-Science-verden.

Alt i alt er der meget, der taler for, at hvis bibliotekerne ønsker en rolle mht. e-Science, så skal vi arbejde ihærdigt med at redefinere os selv som del af den ny forskningsinfrastruktur. Det handler dog ikke kun om vilje til at udvikle og tilbyde nye services til forskerne, men også om økonomi, og om den rette teknologi, samt om at have de rette kompetencer til at løfte opgaven. Heldigvis har bibliotekerne et udmærket fundament: Dels besidder vi de grundlæggende bibliotekariske kernekompetencer – og dels, som Tony Hey siger, spiller fag- og institutionelle repositories en nøglerolle i cyberinfrastruktur – de fleste universitetsbiblioteker bestyrer jo allerede arkiver for universiteterne.

Og universitetsbibliotekerne forsøger faktisk at komme på banen: I 2009 har KUBIS gennemført et eksplorativt projekt for at undersøge forskernes brug af videnskabelige data på det samfundsvidenskabelige område. Undersøgelsen viste, at der er et stort behov blandt forskerne, for at kunne arkivere, bevare og have adgang til forskningsdata på en systematisk og brugervenlig måde. På tilsvarende måde, gennemføres et pilotprojekt i 2010, via DEFFs Programgruppen for Informationsforsyning, om at skabe Open Access til forskningsdata i forskellige cases på DTIC, AUB og KB/KUBIS, blandt andet i samarbejde med fx Dansk Data Arkiv. Pilotprojektet skal bidrage til DEFFs strategi på området.

"Det er svært at spå, især om fremtiden"
Institutleder for Niels Bohr Institutet, John Renner Hansen, udtaler på websiden for eScience Center ved Det Naturvidenskabelige Fakultet (Københavns Universitet), at i år 2020 er alle naturvidenskabelige forskere "eScience forskere". Renner Hansens institut har også netop oprettet en særlig forskningsgruppe for e-Science, ligesom mange andre institutioner gør det internationalt.

Kilde: Sarah Court, Herculaneum Conservation Project

Billede af vægmaleri fra Herculaneum.
Fundet i Image Bank i Archaeology Data Service.

I den e-Science-verden, som vi øjensynlig bevæger os i retning af, er udfordringerne for både forskere og forskningsbiblioteker mange. Ingredienserne i det fremtidige forskningslandskab vil formentlig være udpræget brug af "cloud computing"-faciliteter, det semantiske web (jf. Esben Fjords artikel, REVY nr. 2, 2010), "semantisk computing", og som ophavsmanden til WWW - Tim Berners-Lee - forudser, vil vi kalde det hele for "Web 3.0" . Forskningens del af Web 3.0 vil være e-Science – og vi vil måske kalde det e-Research – under en bred kam.

Hvordan fremtiden end ser ud, så er det sikkert, at e-Science er kommet for at blive - men om føje år vil man måske slet ikke tale om e-Science, e-Research og virtuelle forskningsmiljø, fordi det vil være "sådan man gør". Ligeledes vil der - af tvingende nødvendighed - eksistere services og faciliteter, der understøtter forskningens brug af digitale forskningsdata - men hvilke institutioner der varetager disse opgaver afgøres nu!