

Intervju med Johan Rung, SciLifeLab. Digitala innovationer i skuggan av Corona. 2020-05-20

Intervju med Johan Rung, chef för Data center vid SciLifeLab. Johan Rung var även samordnare för forskningsprogrammet "Data driven research: models and AI" i samband med Coronakrisen. Intervjun genomfördes via länk och filmades. Ljudet transkriberades av företaget Rappa tag.

I:	Intervjuare, Peter Du Rietz
JR:	Johan Rung
[Ohörbart]=	Hör inte flera ord eller mening/ar
[namn]=	Uppfattar inte vilket namn som sägs
?=	Hör inte enstaka ord
...=	Paus, avslutar inte meningen, blir avbruten
[utskriftskommentar]=	Rappa Tags kommentar
Gulmarkerat =	Kontrollera ord/mening

I: Nu sätts det igång och spelar in, vad trevligt. Nu har du berättat väldigt många kloka saker som vi inte har fått med på den här inspelningen och jag vet inte om vi ska ta en recap på det litegrann.

JR: Jo men visst, det kan vi göra. Jag kan ju börja kort med det jag sade om SciLifeLab så har du det. Alltså att SciLifeLab är en nationell infrastruktur med en rad olika teknikplattformar, genomik, proteomik, bilder och så vidare och att vad vi gör, det är att vi stödjer svenska forskare som kommer med prov till oss för att kunna få experiment gjorda och få data tillbaka så att dom ska kunna bedriva sin forskning. Och då sitter vi på... centralt då när den här coronavirussituationen kom, så tänkte vi att ja men vi är i en unik sits här att kunna göra någonting speciellt för landet i det här. Och eftersom vi har haft diskussioner då med Wallenbergstiftelsen om det här så gick dom in och sade att ja men,

vi stöder ett program som har olika komponenter. Och en del av det stödet har gått då till den här Lars Engstrands testverksamhet som byggs upp på Karolinska, det som heter CTMR, det är alltså Centrum för Translationell Mikrobiell... någonting. Och vad dom gör, det är att dom testar virussekvenser då. Så dom får prover från sjukvården som dom testar och rullar ut, som säger i princip om man är infekterad eller inte. Sedan en annan del av det här som har finansierats nu som också är knutet på ett sätt till SciLifeLab, det är det som är Hans-Gustaf Ljunggrens verksamhet på KI som testar... tar in mera långsiktig provinsamling för att studera sjukdomsförloppet och för att studera biomarkörer i blod till exempel. Sedan är det ju så att en del av det här som kommer från Wallenbergstiftelsen, det har gått då till att vi byggde upp det här forskningsprogrammet för covid-19 där vi också går in med egna medel från vår nationella finansiering som vi har. Och vi byggde upp det här programmet och vi gick ut med en utlysning och vi fick tvåhundraåttiofem ansökningar på fem dagar. Och det var en ganska stor insats där för att försöka gå igenom och prioritera och att göra det här urvalet av hur det här faktiskt skulle sättas upp.

I: Hade ni någon särskild strategi kring vad för sorts projekt ni ville få in och premiera så att säga.

JR: Ja. Det som var primärt då, det var ju... vi ville ju kunna se, vad har det här för impact på covid-19-situationen? Det var ju förstås det primära. Men sedan hade vi ett antal kriterier som var... och förstås ja, liksom välja dom bästa projekten... det var också så att projekten måste ha en möjlighet att komma igång och börja jobba snabbt. Så att till exempel att det inte skulle vara någon rekrytering i början till exempel och att det redan skulle finnas medfinansiering från olika håll, så att det var en existerande verksamhet så att säga. Och även det här samarbetsnätverket liksom, att det fanns ett nätverk vi kunde jobba med var viktigt. För det här är ju då, som jag sade, en byggsten i hela det här programmet det är just när vi säger team science. Att jobba kollaborativt, att jobba med någonting som ska bygga upp kapacitet som är långsiktig, som inte bara är att man gör sina projekt som man har fått finansiering till utan att man samtidigt bygger upp kapaciteter och möjligheter för landet att möta en sådan här pandemi om den kommer tillbaka.

I: Och den här digitala innovationskraften, det finns projekt som arbetar med allt från tester och sedan till att bearbeta den data som testerna ger och forska på det. Som jag förstår det.

JR: Ja, precis. Så en del projekt är ju sådana då som primärt samlar in... en del projekt primärt samlar in prov. En del projekt primärt samlar in data och det är ju från olika tester då som kommer. Det kan ju vara till exempel att man försöker utveckla nya metoder att testa till exempel antikroppshalten i blod eller man tester till exempel virus-DNA så man ska kunna följa den utvecklingen. Eller man testar DNA:t på dom personer som har blivit infekterade så man kan se, vilka är det som blir sjukast, finns det en genetisk bakgrund i dom ?. Så det innebär att en del projekt är ju mer eller mindre datatunga, och en av dom på det sista här området som är med AI som vi kallar det för och det som är informatikprojekt... det finns ju också projekt inom mer drug discovery-hållet som är datatunga... och det vi ser där är ju att dom behöver ju extra mycket stöd i att kunna liksom arbeta med dom här stora datamängderna och utveckla nya metoder som är liksom tekniska metoder, analysmetoder som ofta baserar sig på den senaste forskningen inom AI för att man ska kunna dra ur så mycket information som möjligt ur det här.

I: kan du ge något exempel på hur det kan se ut, den typen av nya analysmetoder och så.

JR: Ja, någonting som är ganska stort just nu, det är ju att man använder AI i bildbehandling. Mikroskopi av olika sätt är ju någonting som har utvecklats väldigt mycket på senare tid. Och där har vi ju en bildplattform som har en lång rad olika typer av tekniker för att göra mikroskopi, ljusmikroskopi, cryo-EM, olika typer av metoder där. Och dom bilderna som genereras där, från dom kan man ju dra ur väldigt mycket information och det här görs mer och mer med väldigt starka maskin learning-metoder. Till exempel är det ju såhär, du kan tänka dig att du har en bild där man vill se ett snitt på en vävnad till exempel och så vill man kunna, från bilden, hitta vilka celler är det som uttrycker vissa proteiner. Och det här är ju ett typisk problem som är maskin learning-problem, att där kan vi ha en lång rad olika bilder som vi tränar upp en modell för att kunna hitta, identifiera var cellen är och vilka delar av dom här... och det är även liksom in på specifika delar av celler liksom, var proteinet uttrycks.

I: Är det mönsterigenkänning då, kan man säga?

JR: Det är väldigt mycket mönsterigenkänning och det är ju det som ligger till grund för den här... när man pratar om deep learning. Liksom att man lär datorerna att lära sig själva hur man ska hitta den där typen av features.

I: Och vad är det för projekt som jobbar med den typen av data, det typen av metodik?

JR: Ja, det finns ju... nu håller ju projekten precis på att komma igång så vi ser inte exakt vilka av dom här projekten såhär... men det finns ju väldigt mycket. Det finns någonting som heter Human cell atlas som är ett stort internationellt projekt som jobbar väldigt mycket med den här typen av tekniker och det är någonting som drivs från partners runtom i hela världen och i Sverige också. Och vi ser just att för både så att säga projekt som är translationella i det att man vill ha metoder inne på den kliniska sidan för att kunna ta... säg till exempel att man för diagnostik ta ett prov till exempel från en biopsi och kunna identifiera, vad är det i det här provet som kan ge oss information om vad det faktiskt är för någonting för att kunna göra en diagnos. Det här blir också mer och mer stött av bildbehandling och sådant som är liksom beräkningstekniker för att kunna stödja det här. Förut har det ju varit väldigt mycket att man har jobbat som patologer och tittar på bilder och att man ser... och fortfarande så att det är väldigt svårt att slå patologer, dom är jättebra. Men det är det här arbetet och även det som kommer in på andra delar av sjukvården, jobbar man ju mer och mer med beslutsstöd och att dom här teknikerna för beslutsstöd baserar sig ju mer och mer på, så att säga, stora mängder av data, inte liksom bara enskilda mätningar liksom. Att ja, men om du har ett blodtryck som är förökad liksom, då gör man någonting annat men här har vi kanske en karta av tjugotusen olika gener med olika mutationer i och att samtidigt kanske vi har en karta med femhundra proteiner som uttrycks samtidigt och det här... hela det här data-setet ska man använda sig av för att kunna få en bild av vad som händer i just den här personen. Det är här vi pratar om det... det här kallas ju precision medicine. Det är det att man använder sig av mer och mer data för att kunna göra en väldigt specifik behandling för enskilda personer.

I: Och du nämnde, patologerna är väldigt duktiga på det här, men är dom så många så att man ändå inte får... så man inte får en flaskhals. Innebär dom här nya metoderna att man kan undvika eventuella flaskhalsar därför att kompetensen inte är tillräckligt utbredd?

JR: Ja förstås, jo så är det ju. Det är ju ditåt... man jobbar ju för att kunna både stödja dom i arbetet... det är ju på något sätt alltid i slutändan en människa som har ett medicinskt ansvar. Men det gör ju

också att det går ju fortare för dom att behandla proverna. Så förut kanske det var så, vi har haft sjukhusgenetiker som har tittat på DNA-sekvens och jobbat igenom det och det tar väldigt lång tid och med hjälp av system som man nu för in i det här för att göra dataanalyser och för att liksom på bioinformatiskt håll kunna förbereda och plocka fram det som är signaler som är det som är viktigast liksom i den här informationen. Det gör ju att deras arbete går snabbare.

I: Så tekniken kan leda patologerna rätt och uppmärksamma dom på, här ser vi någonting som ser misstänksamt ut, vad tycker du om det. Lite så eller.

JR: Ja, precis. Så det är ju så att säga att beslutsstöd i någon mening liksom.

I: Jo, det verkar vara mycket som är beslutsstöd av dom här projekten också som jag har förstått det.

JR: Ja, nej men så är det ju förstås att det är både metodutveckling och det som är så att säga beslutsstöd i form av dom dataanalyser som vi har.

I: Och som jag har förstått, det här skulle... dom här projekten kommer... ni kommer finansiera dom fram till 2022.

JR: Just det.

I: Och kommer det vara någon sammanställning av det här sedan på något sätt?

JR: Ja, precis. Och det som har varit viktigt också... och det här har också varit ett kriterium och ett krav på dom här projekten... det är att dom ska dela data och jobba med open science från dag ett. Vilket innebär att dom måste publicera saker och dela sina data i takt med att dom genereras.

I: Underbart.

JR: Vi vill inte att dom ska komma liksom efter att jobba ett helt år på sin kammare liksom och sedan publicera någonting utan det här är någonting som ska gagna hela forskningsfältet från dag ett. Det ska vara delbart, man ska följa internationella standarder för hur man beskriver sina data och man ska beskriva hur man har kvalitetssäkrat genereringen av data och hur man jobbar med det här liksom. Man ska ha datahanteringsplaner där man redan från början tänker på hur man ska jobba med IT-aspekten, datahanteringsaspekten, datapublicering, den typen av frågor.

I: Och datahanteringsplan och open access, det var ett krav som ni ställde redan i utlysningen?

JR: Ja, precis. Det är ju såhär att vi inser ju, det här är ju en stor kulturförändring. Forskningen har ju inte riktigt fungerat såhär. Fortfarande idag så ser vi att det är väldigt många som vill hålla i sina data och man tänker att ja men, jag som har jobbat så hårt för att samla in mina data, varför ska jag dela med mig av det? Och det argument som vi brukar prata om det är ju att ja, men den impact du får på din forskning, den blir faktiskt högre om du delar med dig än om du håller kvar den här.

I: Ja, och inte minst forskaren kan ju själv använda sig av ett jättestort set av data, som ett globalt set vad. Men hur upplever du att Sverige ligger i det här med öppen data. Du nämnde internationella standarder, ligger vi i framkant eller ligger vi så att säga i stemmet eller hur ser det internationellt ut och finns Sverige där?

JR: Det finns i Sverige väldigt goda ansatser. Vi kanske är... vi ligger inte först, det gör vi inte. Vi ligger i den så att säga... någon sorts mittenklunga som är ambitiösa och har en massa processer igång för att jobba med det här. Det finns en ganska stor medvetenhet om det här hos finansierarna, man arbetar ju till exempel... VR har ju ett uppdrag för att samordna öppen vetenskap i Sverige och det här har ju varit igång i flera år och det finns ett mål att det här ska liksom verkligen vara öppet nu om fem, sex år någonting. Men det är ju en bit att gå, men det finns väldigt mycket liksom ansatser och tryck uppifrån för det här. Men som sagt, det är en kulturförändring och i andra länder har man kanske varit lite bättre på att tidigt ta till sig det här. Det är också väldigt olika mellan olika fält. En del fält i fysik till exempel, partikelfysik, har ju jobbat väldigt mycket såhär med datadelning och det är ju så att olika vetenskapsfält arbetar på olika sätt och i life science och det som är biomedicinsk forskning, där har man varit väldigt... nästan över hela världen... hållit i data mer än vad man gjort på andra håll. Både för att man säger att ja men, det här är patientdata, vi måste ha en väldigt stark säkerhet kring hur man jobbar med delning av persondata och så vidare. Och det är helt sant, dom sakerna är väldigt viktiga att man får dit på liksom dom regulatoriska processer, lagstiftning och etik och sådant som har att göra med hur man kan dela den typen av data. Men ofta är det så att det där finns faktiskt på plats. Det finns redan idag ganska bra, jag menar... lagstiftningen är väldigt klar i hur man ska arbeta. Det som inte finns så mycket det är ju... dels så använder man sig ofta av det där litegrann som ursäkt att inte dela data, att man säger att nja men... jag håller i mina data därför att jag är mån om att deras information inte ska spridas, när man egentligen gör det för att man vill kunna hålla i och publicera mer själv liksom. Men det är ju så att vi ser att dom förutsättningar för att jobba med känsliga data, dom blir bättre och bättre. Vi vet hur man gör det här faktiskt och det [ohörbart] vi på. Men det jag ville säga, det var ju att processerna inom life science går framåt. Sverige är inte i framkant men vi är inte långt från framkanten om vi säger så. Vi är med i olika internationella samarbeten där vi gör det här jobbet och det här... just det, coronaarbetet driver på det här väldigt mycket.

I: Så det kan bli någonting av ett genombrott kring öppen data inom life science då eller?

JR: Ja, jag känner att det här är ju en av dom saker där man... och inte bara inom forskarfältet utan jag tror att nästan hela världen kan se... även från liksom allmänheten så att säga... att man ser dagligen. Man ser hur Folkhälsomyndighetens presskonferens-... liksom, dom visar upp data, bilder, diagram, data... man möts av siffror liksom, alla dagar är updates. Det är väldigt mycket det som är fokus på det här. Så man ser att den här öppenheten i data, att man beskriver vad som händer, man beskriver löpande en uppdaterad situation och det gör man baserat på data. Jag tror att det finns en starkare medveten om att... medveten om också det här att man vill basera beslut på data liksom. Man ska inte ta beslut bara på en hum utan man ska underbygga sina beslut med datainsamlingar och analyser och det här är ju förstås ett sätt att säga att ja men, hur ska vi kunna göra det om man inte delar... om man inte visar sina data öppet. Då vet vi inte vilka metoder man har använt, det finns ingen som kan reproducera det här. Liksom hela integriteten för processen blir lite skagig om man inte har faktiskt del av metoder och data men även liksom den kod, om man har haft analysmetoder som man har programmerat upp, då ska man kunna dela den koden.

I: Hur länge har det funnits den typen av datakraft som har gjort att det har varit möjligt att över huvud taget arbeta såhär storskaligt med data?

JR: Ganska länge skulle jag säga. Det är ju så att det finns... i Europa så finns det ett centrum i Cambridge som heter European Bioinformatics Institute. Och det är dom som lagrar... där kan man som forskare gå och lagra sina data som kommer från biomedicinsk forskning och det här har funnits sedan 1994 tror jag. Och dom har beräkningskraft där och dom har lagring, jag tror dom idag lagrar dom uppåt trehundra petabyte.

I: Så all den här datan som era projekt resulterar... dom hamnar där sedan också?

JR: Dom hamnar där så långt det någonsin är möjligt därför att det jag ser det är ju att, ja men... det är ett av dom centra i världen där man går som forskare för att hitta sina data. Så det är en form av standard så vill man att andra ska kunna hitta vad man har gjort, då är det bra att kunna deponera det i dom databaserna där. Och du vet, ett krav som vi har haft också i det här, det är att vi ska bygga upp en svensk dataportal själva. Så det här håller vi på med nu. Och den här dataportalen, det är ju i princip en samling med information och tjänster för att svenska forskare ska kunna dela sina data och den här är knuten till den här europeiska infrastrukturen.

I: Och det är SciLifeLab som ligger bakom den som ni ska bygga upp?

JR: Ja, det är vi som bygger upp den och det gör vi på uppdrag av Vetenskapsrådet då.

I: Okej. När står den klar?

JR: På onsdag.

I: Onsdag, okej.

JR: Ja, står klar är väl kanske att ta i men det är ju förstås väldigt...

I: Invigs.

JR: Den invigs och den kommer innehålla information och länkar till det europeiska samarbetet och innehålla liksom hur man ska jobba med olika datatyper och dom tjänster som finns som vi kan bidra med.

I: Och då är det som en egen webbsida med inloggning där eller?

JR: Ja, den är öppen för alla. Den är öppen för allmänheten med liksom, så att den kommer att lanseras brett då.

I: Du får jättegärna skicka en länk.

JR: Ja, det ska jag göra. Det kommer ut ett pressmeddelande på onsdag också, men jag gör så, jag skickar det där till dig när den här kommer ut så kan du se. Det är ju så att det är inte en databas i sig. Det är liksom inte så att i Sverige samlar vi i en egen databas våra data här, att vi har en burk liksom utan det här är väldigt mycket gjort för att vi ska kunna få precis den här öppenheten och delbarheten att vi lägger våra data så att svenska forskare ska kunna delta i internationella samarbeten och kunna dra nytta av alla dom data som kommer in från andra länder också.

I: Okej. Har du någonting mer som du skulle vilja tillägga här som jag inte riktigt tagit upp eller som vi inte har tagit upp, innan vi börjar avrunda.

JR: Ja. Någonting som man ser ofta, man pratar ju om att man i takt med att datamängderna ökar så behöver man mer och mer datorkraft och man behöver större maskiner liksom och att i det så känner man kanske att det här gör att... initialt kanske man tänker att det här borde göra det svårare för enskilda forskargrupper att få saker gjort för att man behöver liksom mer kraft. Och det är väl kanske ett par saker jag skulle vilja säga där om det och det är ju dels att nu har vi ju infrastrukturer där vi stödjer forskare här. Som till exempel, på SciLifeLab så... eftersom vi har plattformar som har väldigt dyrbar utrustning och instrument och då personal som är upptränade på att liksom köra dom här grejerna dag efter dag, så gör ju det att forskare som kommer till en infrastruktur för att göra sin forskning, dom behöver ju inte själva köpa dom här instrumenten. Dom behöver inte själva anställa personal som ska kunna dom här tekniska detaljerna utan vad man gör, det är att man utnyttjar, man centraliserar... sådant som är dyrt och svårt centraliserar man i infrastrukturerna och genom att vi bygger nationella infrastrukturer så gör ju det att forskarna får en mycket större möjlighet för då har dels så har man ju en... dels så är det ju demokratifråga, vi gör det här tillgängligt för alla. Det spelar ingen roll var man är eller var det kommer från utan det här är tillgängligt för alla projekt. Och det gör ju att det går att göra den typen... även om man är en liten forskargrupp, en som precis har liksom avslutat sin post-doc på något universitet och har en doktorand, liksom en liten grupp... dom kan göra experiment på extremt dyrbar utrustning. Och det här är någonting som i princip har varit omöjligt innan man hade dom här infrastrukturerna. Och det här gäller även då datorkraften, i det att vi samlar datorkraft och delar ut genom nationella infrastrukturer. Så att liksom arbeta med och satsa på den här typen av infrastrukturer, det är ju väldigt viktigt när man vill stödja hela landet och liksom alla delar av forskningen.

I: Hur länge har dom här infrastrukturerna funnits?

JR: Ja, SciLifeLab har ju funnits i tio år nu och det har varit såhär nationell infrastruktur i fem år.

I: Och datacentret på SciLifeLab har funnits lika länge eller?

JR: Vi har funnits där i tre år och vi har precis liksom blivit en... vi har varit lite experimentella men från nu i år så befäster vi oss och blir liksom en del i hela verksamheten. Och tanken från vår sida, det är ju att vi ska kunna stödja... vi vill kunna flytta dom saker som är svåra från enskilda forskare till infrastrukturen. Det här är ju egentligen någonting genomgående i allting vi gör, både med instrumentering och med hantering av experiment men även på datorsidan då. Så att det här som har att göra med fair, som vi sade, att producera metadata... alltså, data som beskriver dom data som vi arbetar med... för att du ska kunna analysera ett dataset... säg att du publicerar ett dataset, ingen annan kan ju analysera det där om dom inte vet exakt hur har det här genererats, vad är det för instrument, vad är det för prov som har körts, hur har dom proven behandlats. Hur har man liksom gjort. Och allt det där metadata som man samlar in för att beskriva dataseten, det här är extremt viktigt och det är också extremt viktigt att man använder sig av internationella standarder för att beskriva det här. Så om man säger till exempel att ja men, det här är ett prov från en person som har diabetes, vad är diabetes liksom såhär? Finns det en definition? Ja men det gör det. Det finns en internationell standard för det vad det betyder. Då applicerar man det, så det gör att när man delar data så betyder det som jag menar... förstås av en forskare från annat håll i världen för att det följer en given standard. Så vad vi försöker göra nu, det är att lyfta dom här frågorna som har med... liksom dom tekniska frågor kring databehandling så att enskilda forskare ska inte behöva ha lika mycket liksom... arbeta lika hårt och sätta sig in i dom här tekniska detaljerna kring data så mycket

utan dom ska kunna fokusera på sina forskningsfrågor. Och istället försöker vi lyfta dom här tekniska delarna upp till dom infrastrukturer som producerar data. Så att redan från start ska vi kunna generera allting som behövs för att man ska kunna dela data med forskare i andra världen. Och det gör ju liksom att ju högre upp i kedjan som vi kommer så underlättar det ju för alla nedströms så att säga.

I: Just det, precis.

JR: Så det här är lite filosofin.

I: Var sitter ni någonstans.

JR: Jag sitter i Uppsala främst men vi har kontor i Stockholm också, i Solna då. Så SciLifeLab har ett campus som ligger på KI, det är KTH som är huvudman för det här. Vi har även i Uppsala och sedan så har vi... vi styrs då av en... vi har en styrelse som består av folk från tio olika universitet tror jag det är och vi har verksamhet vid nio eller tio olika universitet i landet. Så det ska vara nationellt liksom, verkligen.

I: Johan, innan vi rundar av, kan du bara säga ditt namn och din funktion.

JR: Så jag heter alltså Johan Rung och jag är chef för data center på SciLifeLab.

I: Jättebra. Du ska ha stort tack för samtalet. Och jag hör av mig så småningom när vi har kommit lite längre i projektet.

JR: Ja men gör det, när som helst. När det är någonting du behöver från mig, så det är bara att höra av sig.

I: Toppen, jättekul.

SLUT INTERVJU/LW

Transkriptet redigerat av PDR 2020-11-16