

Webcrawling til analyse af hjemmesider

Afgangsprojekt udført hos Morning Train Technologies ApS

Macintosh HD:Users:rubatharisan:Downloads:morning-train-logo-left-black-big.png

**Udarbejdet af:**

Rubatharisan Thirumathyam

[ruthi12@student.sdu.dk](mailto:Ruthi12@student.sdu.dk)

[rt@morningtrain.dk](mailto:rt@morningtrain.dk)

Eksamens nr.: 324539

**20. maj 2017**

**Vejledt af:**

Ole Dolriis **(SDU)** [od@mmmi.sdu.dk](mailto:od@mmmi.sdu.dk)

Bjarne Bonde **(Morning Train)** [bb@morningtrain.dk](mailto:bb@morningtrain.dk)

**Synopsis**

**Forord**

Dette projekt er udarbejdet som et afslutningsprojekt, på uddannelsen som diplomingeniør indenfor Information- og kommunikationsteknologi. Projektet er udarbejdet i perioden 01-02-2017 til 22-05-2017.

Projektet er lavet i samarbejde med Morning Train Technologies ApS (ref. Morning Train), hvor jeg har i fællesskab med Bjarne Bonde (CTO v/Morning Train) defineret projektets vision. Dette projekt er blevet vejledt af Ole Dolriis (lektor v/SDU) og Bjarne Bonde. Under projektforløbet har jeg fået råd fra alle konsulenter fra webbureaet Morning Train. Den primære interessegruppe for dette projekt er webbureauer og webudviklere.

Som led i mit uddannelsesforløb har jeg haft et praktikophold på 6 måneder hos Morning Train. Morning Train ledelsen har herefter valgt at ansatte mig som serveransvarlig, hvor mit primære ansvarsområde har været serveropsætning, servervedligeholdelse og serversikkerhed. Mit sekundær ansvarsområde har været fullstack udvikling af systemer. Da jeg startede på dette projekt, havde jeg allerede været ansat i virksomheden i 1 år og 7 måneder.

Tak til Morning Train for at stille ressourcer til rådighed så dette projekt kunne blive et realitet. Specielt tak til Bjarne Bonde og Ole Dolriis for at have vejledt mig igennem de gode og de dårlige perioder i projektet.

- Rubatharisan Thirumathyam

ruthi12, eksamens nr. 324539

**Læsevejledning**

Denne rapport er struktureret og skrevet så det giver læseren en fornemmelse af hvordan processen har været for at gennemføre dette afgangsprojekt.

Afgangsprojektet indledes i kapitel 1 med en indledning hvor jeg præsentere projektets problemområde, problemformulering og problemafgrænsning.

I kapitel 2 beskriver jeg hvordan mine planer og planlægninger har været for at gennemføre projektet.

I kapitel 3 beskriver jeg hvordan kravspecifikationerne er blevet dannet på baggrund af problemformuleringen og problemafgræsningen.

Kapitel 4 består af research hvor jeg gøre rede for hvorfor jeg har valgt de teknologier og værktøjer til dannelse af produktet.

Kapitel 5 består af udviklingsmetode og udvikling, hvor jeg gør rede for hvordan jeg har udviklet produktet, hvilke processer jeg har anvendt for at kunne producere produktet.

Kapitel 6 består af konklusion og perspektivering, hvor jeg samler mine tanker, oplevelser og refleksion over projektet som helhed.

**Indhold**

1. Indledning 6

1.1 Problemområde 6

1.2 Problemformulering 7

1.3 Problemafgrænsning 9

2. Planlægning 9

2.1 Projektplan 10

2.2 Udviklingsmetode af produktet 12

2.3 Agil softwareudvikling 12

2.4 Scrum 13

3. Kravspecifikation 15

3.1 … 17

3.2 … 17

4. Design 18

4.1 … 18

4.2 … 18

5. Implementation og test 19

5.1 … 19

5.2 … 19

6. Evaluering 20

6.1 … 20

6.2 … 20

7. Konklusion 21

7.1 … 21

7.2 … 21

8. Referencer 22

8.1 … 22

8.2 … 22

9. Appendiks 23

9.1 … 23

9.2 … 23

# 1. Indledning

Målet med dette projekt er at anvende min viden, kompetencer og færdigheder tilegnet gennem min uddannelse. Dette projekt skulle være baseret på et problematik i en virksomhed, hvorved skulle jeg anvende min viden, kompetencer og færdigheder til at fremdrage en løsning. På grund af mine rolle i Morning Train, fik jeg lov til at medvirke til at forme et projekt, hvor jeg følte mine færdigheder kunne være brugbart.

## 1.1 Problemområde

Morning Train er et webbureau lokaliseret i Odense, drevet af kvalitetsbevidste webnørder. Morning Train specialisere sig primært indenfor tre kerne områder: digital design, intelligente webløsninger og online markedsføring.

Indenfor disse tre kerneområder er der mange problematiske udfordringer, som hver kunne være et projekt i sig selv. Jeg søgte et projekt som strækkede sig over alle tre kerne områder, et projekt der ville kunne være til gavn for alle tre kerne områder. Men for at kunne gøre dette, blev jeg nød til at finde ud af præcis hvordan de forskellige afdelinger fungere.

Kort beskrevet så når Morning Train udvikler en ny hjemmeside til en kunde, så er der tre afdelinger som tager hånd om opgaven. Morning Train har front-end udviklere som varetager det digitale design og layout opsætning, de har back-end udviklere der får programmering og funktionaliteten på plads. De har online markedsførings konsulenter som analysere og vurdere hjemmesiden for at efterfølgende at optimere på det de føler der vil gøre en forskel for hjemmesidens online tilværelse.

For at forstå hvad mit projekt skulle handle om, besluttede jeg for at snakke med medarbejdere fra de forskellige afdelinger, for at høre dem om hvordan et afgangsprojekt ville kunne bidrage til deres afdeling. Frontend afdelingen kunne ikke rigtigt se at jeg kunne lave noget brugbart for dem, deres problemer er primært grafisk orienteret. Online markedsførings afdelingen ville have mig til at lave et system som skulle systematisk søge forskellige søgeord på Google og finde placeringer af en bestemt hjemmeside, dette projekt kunne ikke lade sig gøre, da Google ikke tillader systemer at crawle deres søgemaskine – ret banalt. Backend afdelingen havde nogle spændende projekter som alle kunne have været det perfekte projekt at bruge som afgangsprojekt. F.eks. et projekt hvor man samlede data om brugeradfærden på en hjemmeside, for at derefter bruge dette data til at forbedre hjemmesiden – hvor brugeren bliver centrum.

Men det projekt jeg syntes der vil være til gavn for alle tre afdelinger blev ”webcrawling for analyse af hjemmesider”. I en mere forstående term: en bot som systematisk besøger hver eneste underside på en hjemmeside, for at vurdere om der nogle punkter som der kan forbedres. Vurderinger af siden, kan herefter udvides, til at kunne tage hensyn til flere ting. En vurdering kunne være f.eks. en undersøgelse om alle grafiske billeder ikke overgår en bestemt filstørrelse, da større en billede er, jo længere tid skal en besøgende vente på at billedet bliver fremvist.

## 1.2 Problemformulering

En hjemmeside har en værdi for mange virksomheder, nogle virksomheder lever af deres hjemmeside. Men vi ser fejl i alle hjemmesider, nogle af disse fejl bliver aldrig opdaget, da der mange som ikke kigger specifikt efter disse specifikke fejl.

I samarbejde med Morning Train vil der blive udviklet et framework til crawling af hjemmesider. En webcrawler er en robot som systematisk besøger alle links på en hjemmeside. Man skal scripte sig op imod dette framework for at viderebehandle de data der opsamles under besøget.

Der opstilles en række testcases for at teste frameworket. Dette framework skal udvikles arkitektonisk korrekt så det er fleksibelt, skalerbart og ligetil at scripte op imod, uden at skulle koncentrere sig om skalering.

Frameworkets primære formål er at være så fleksibel at den kan understøtte en bred vifte af testcases.

Frameworket skal til slut lave en teknisk rapport som påtaler alle forbedringsmuligheder den kan finde.

Følgende testcases kan være relevant til at validere frameworkets fleksibilitet ved at belyse forskellige brugstilfælde. Testcasene er defineret på baggrund af både menneskelige og systemmæssige fejl.

Nedenstående fejl kan kategoriseres som testcases:

1. en side som ikke eksistere længere, men stadig linkes til fra en anden side på hjemmesiden. (404 fejlen).   
a. besøgende kan føle sig misledt og få et dårligere forhold til hjemmesiden.

2. en side som udløser en systemfejl. (500 – Internal server error, såkaldte http 5XX fejl)   
a. besøgende kan blive frustreret over at ikke få lov til at udføre det tiltænkte.

Weboptimerende testcases:

3. en side som har billeder der overskrider en bestemt størrelse, hvilket gør at den besøgende skal vente uretfærdigt længe på at siden bliver fremvist.   
a. den besøgende kan blive frustreret over ventetiden, og forlader hjemmesiden.

4. en side som er for svær at finde pga. den ligger for langt væk fra forsiden. (Forside -> Side A -> Side B -> Side C -> Side D)   
a. besøgende skal igennem mange sider for at komme til sit mål.

5. en side som ikke er mobil-, tablet- eller desktopvenlig (responsiv).   
a. besøgende vælger at forlade siden, pga. elementer på hjemmesiden som ikke er optimeret til dens enhed.

## 1.3 Problemafgrænsning

Da det ikke er umiddelbart muligt at estimere hvor langtid dette projekt kommer til at tage, og med det faktum at jeg har en deadline til den 22. maj 2017, har vi som udgangspunkt besluttet at jeg kun skal lave frameworket til at kunne understøtte de første to test cases (punkt 1. og 2. fra problemformuleringen) – og lade frameworket fremtidig understøtte de resterende testcases. Hvis det viser sig at projektet holder sin tidsplan, vil dele af de resterende testcases blive implementeret og redegjort.

# 2. Planlægning

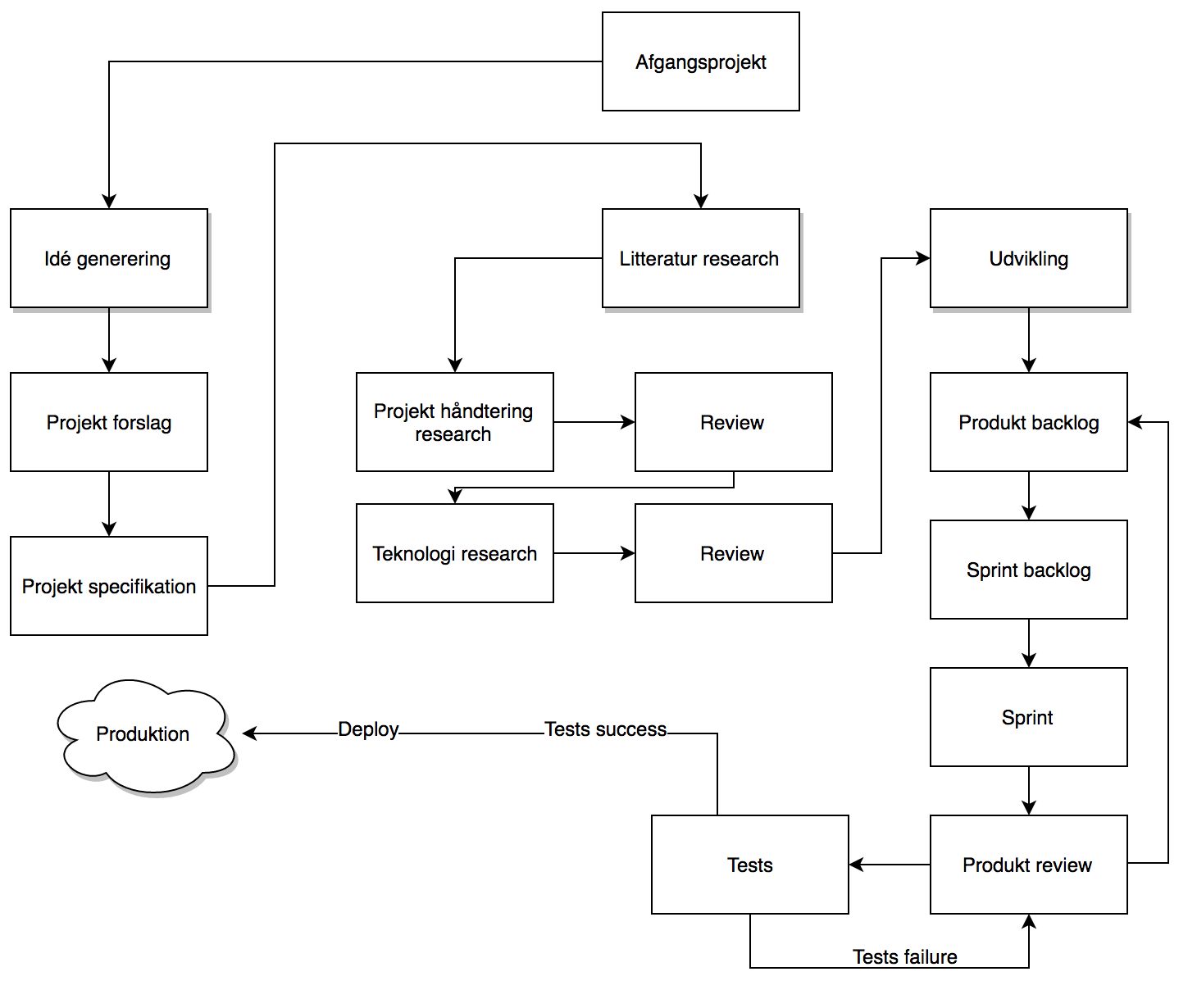
For at starte på dette projekt, blev jeg nød til at lave en projekt plan og følge den. Det hele virkede ret uoverskueligt i starten, for jeg er alene om at udvikle projektet. Da jeg startede med projektet, havde jeg intet dokumenter som beskrev hvordan fremgangsprocessen og ”starts-skuddet” skulle være for både rapporten og projektet. Jeg prøvede at lede mig frem til et dokument, via SDU.dk og vores læringsportal (e-Learn), men uden held.

Jeg valgte at læse bogen ”Projects in Computing and Information Systems: A Student's Guide” af Dr. Christian Dawson. Denne bog fik jeg udleveret i 2. semester, som en håndbog til IT-projekter. Igennem min tid på uddannelsen, har jeg anvendt forskellige dele af denne bog til mine semesterprojekter. Min vejleder fandt et dokument[[1]](#footnote-1) frem til mig, som han troede jeg allerede havde fået udleveret fra min uddannelseskoordinator. Jeg begyndte at følge dokumentets forudsætninger og krav, og herved tilrette dele af min rapport til at stemme overens med dokumentet.

## 2.1 Projektplan

For at overskue projektet, lavede jeg på anbefaling af Dr. Christian Dawson’s bog, en ”Work Breakdown Structure” diagram, som havde til formål at brainstorme og visualisere alle vigtige opgaver. Herved at gøre det lettere at give et tidsestimat på hver enkelt opgave.

I figur 1 ser vi min ”work breakdown structure”. I tabel 1 ser vi min tidsestimater givet for hvert enkelt opgave.

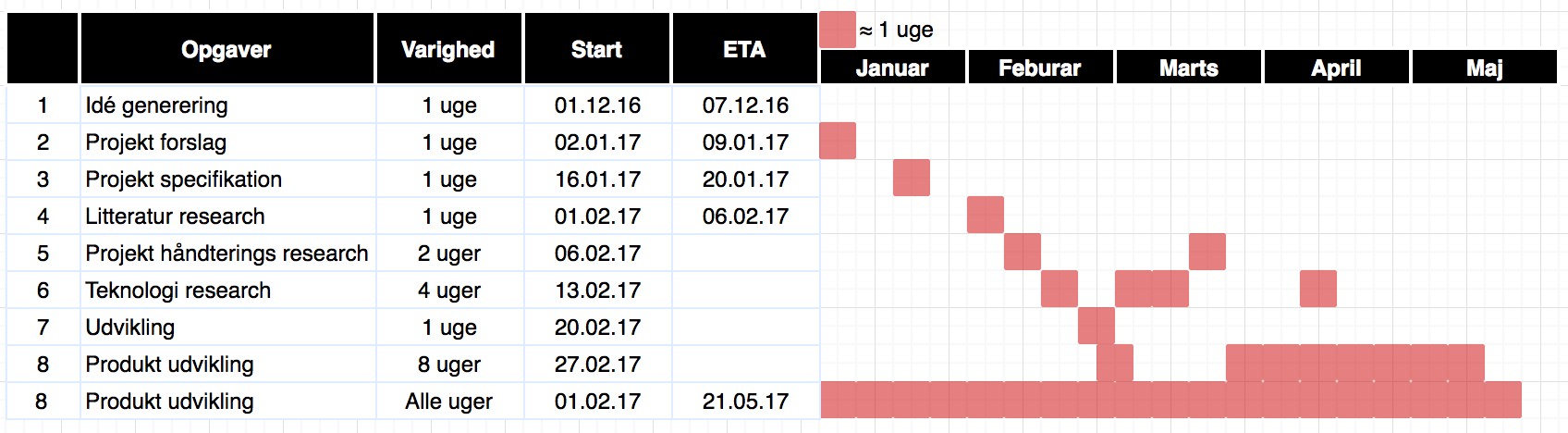


Figur – Opgave flow (Work breakdown structure)

|  |  |
| --- | --- |
| Opgave | Estimeret varighed |
| Idé generering | 1 uge |
| Projekt forslag | 1 uge |
| Projekt specifikation | 1 uge |
| Litteratur research | 1 uge |
| Projekt håndterings research  + Review | 2 uger |
| Teknologi research  + Review | 4 uger |
| Udvikling (opstart) | 1 uge |
| Produkt udvikling (iterativt forløb) + Sprint backlog + Sprint + Produkt review + Tests + Deployment | 8 uger |
| Rapport skrivning | Alle uger |

Tabel 1 – Tidsestimater for projektet

En tidslinje for alle vigtige opgaver kan ses i figur 2.



Figur - Tidslinje over projektet

## 2.2 Udviklingsmetode af produktet

Udviklingsmetoden som jeg har anvendt i dette projekt er Scrum, årsagen til at jeg har valgt Scrum er på baggrund af dets agile tilgang til software udvikling. For at kunne forklare hvordan jeg har implementeret Scrum, bliver jeg nød til at forklare hvordan min agile tilgang har været. Min agile tilgang har været baseret på manifesten for agil softwareudvikling.

## 2.3 Agil softwareudvikling

Manifesten for agil softwareudvikling afdækker bedre måder at kunne udvikle software på. Følgende værdier har en væsentlig betydning for hvordan agilt software arbejde bør udføres.

* Individer og samarbejde frem for processer og værktøjer

Ved at jeg kommunikere åbent med alle relevante personer for at få feedback, og ikke anvender unødvendige værktøjer til at planlægge, udvikle og iværksætte mit produkt.

* Velfungerende software frem for omfattende dokumentation

Ved at jeg laver et produkt som skal kunne gøre de nødvendige ting som Morning Train efterspørger, i stedet for at arbejde med omfattende kravspecifikationer, diagrammer og dokumenter.

* Samarbejde med kunden frem for kontraktforhandling

Ved at jeg hyppigt er i kommunikation med Morning Train, og ikke opsætter kontrakter om hvad der skal med, ikke med i produktet, og hvilke vilkår og forudsætninger der skal tages forbehold til.

* Håndtering af forandring frem for fastholdelse af en plan

Ved at jeg er tilbøjelig til at ændre på mine tidsplaner og planer. Tilbøjelig til at ændre, tilføje og fjerne krav efter Morning Trains ønsker og prioriteringer.

Disse fire værdier har markant indflydelse på udviklingen af mit produkt, men at følge nogle bestemte principper vil være til gavn for mit produkt og Morning Train. Herved har jeg anvendt disse principper inspireret fra manifesten for agil softwareudvikling.

* Min højeste prioritet er at tilfredsstille Morning Train igennem hurtig og vedvarende udvikling af produktet.
* Tilbøjelighed for ændrende krav, selv i sene stadier af udviklingen. Til gavn for Morning Train og dets konkurrerende marked.
* Løbende levering af et virkende produkt, jo hyppigere desto bedre.
* Dagligt samarbejde med forretningsfolkene bag Morning Train.
* Møder og samtaler ”face-to-face” hos Morning Train.
* Min virkende produkt er den primære målbare fremdrift.
* Fokusere på det tekniske og det gode design i produktet.
* Simplicitet, hold det simpelt – KISS, Keep it straight, stupid.   
  Formå at lave lige nok til at få udført opgaven.
* Refleksion over effektiviteten løbende igennem feedback og selv-refleksion.

## 2.4 Scrum

Årsagen til at jeg har valgt Scrum fremfor andre udviklingsmetoder er primært på grund af dets agile tilgang til software udvikling. Et alternativt populært udviklingsform kunne have været vandfaldsmodellen, men problemet med vandfaldsmodellen er at alt skal planlægges på forhånd, det vil sige, kravspecifikationer på næsten alle krav. Med følgende argumenter var det nok til at jeg valgte Scrum fremfor vandfaldsmodellen:

* Jeg forventer ændringer i projektets scope.
* Alle krav er ikke 100% defineret. Det vil sige at Morning Train ikke ved med sikkerhed hvordan slutproduktet skal se ud.
* Jeg ikke har lavet projektet før, eller et lignende projekt, så jeg kan ikke forudse alle krav der kan blive nødvendige.

### 2.4.1 Scrum’s oprindelse

Scrum blev først nævnt af Hirotaka Takeuchi og Ikujiro Nonaka i 1987 i Japan. Det blev beskrevet som en form af hyperaktiv produkt udvikling af produkter. Scrum referer til en strategi anvendt i hold spillet rugby. Strategien går ud på at få en bold tilbage i spillet, når bolden er ude. På grund af ligheder mellem principperne Scrum i spillet rugby og software udviklingsmetoden, blev det fastslået at navnet skulle være Scrum for software udviklingsmetoden.

### 2.4.2 Scrum’s værdier

Scrum er baseret på nogle fundamentale værdier, disse værdier styrker grundlaget for Scrum. Følgende værdier er essentielt for mit Scrum projekt.

* Engagement

Jeg skal være engageret til succes, være villig til at danne realistiske mål. Jeg skal være aktiv, hvor jeg er en del af et team. Min mission er at arbejde sammen med mit team for at realiser mine mål. Jeg skal være ”all-in” på hele projektet.

* Fokus

Jeg skal være fokuseret, men fokusere på få opgaver ad gangen. Jeg skal have en klar vision om min rolle i projektet, og en klar vision om min opgave. Min mission er at bruge min rolle til at fuldføre min opgave.

* Mod

Scrum bifalder ændringer. At kunne udskifte standarder og processer er ikke altid let, men at gøre det for et teams succes, så er det det der skal til. Jeg skal have mod til at kunne engagere, til at kunne lave handlinger, til at være åben, til at forvente respekt.

* Åbenhed

Jeg skal holde alt transparent, alt skal være åbent for feedback og forbedringer.

* Respekt

Jeg skal respektere alle parter der er involveret i dette projekt, og ligeledes skal alle parter respektere, hvad jeg prøver at få udført.

### 2.4.3 Scrum’s roller

Der findes tre kerne roller i Scrum, disse tre roller er ligeværdigt og har hver deres rette ansvarsområde. Rollerne i dette projekt har været følgende:

* Product owner

I dette projekt har Morning Train været den aktive product owner, specifikt Bjarne Bonde. Han har haft til ansvar at stå for visionen af projektet. At opstille relevante krav og herved prioritere disse krav efter hvad der aktuelt, og hvad der er brug for snarest muligt.

* Scrum master

Scrum masteren har til ansvar at håndtere konflikter, blokeringer og forbedringer, som kan hæmme eller fremme min proces for at gennemføre projektet. I dette solo-projekt har jeg ikke specifikt haft en scrum master, jeg har ageret som en scrum master i tider hvor det har været nødvendigt.

* Development team

Development teamet har til ansvar at udvikle produktet. Teamet skal gøre alt for at fuldføre de opgaver som der givet, til at danne et brugbart produkt for kunden. I dette projekt har jeg været ”development teamet”, med relevante kompetencer skabt igennem mit studie har jeg ageret tværfunktionel med produkt udviklingen.

### 2.4.4 Scrum’s artefakter

Følgende tre hovedartefakter fra Scrum har jeg anvendt:

* Product backlog

Produkt backlog er vores liste fyldt med alle opgaver der har en relevans for projektet. Små og store opgaver, som vil til sigt gavne projektet og produktet. Prioritering af opgaverne fastsættes af product owner.

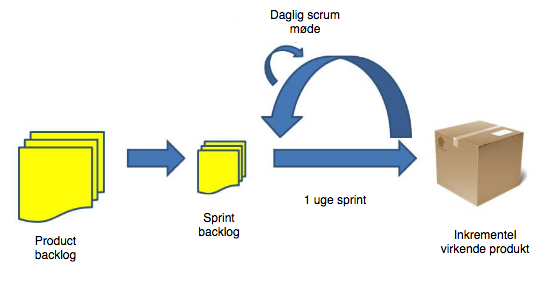
* Sprint backlog

Sprint backloggen er listen over opgaver som jeg skal udføre i min nuværende sprint. Disse opgaver er fastlagt af product owneren i product backloggen, og tilføjet til sprint backlog af development teamet, på baggrund prioritering og relevans med tilsvarende andre opgaver.

* Product increment

Når en sprint er overstået, så har produktet fået tilføjet flere funktioner og herved nået til et stadie hvor produktet danner merværdi til brugeren. Mit mål er at udvikle produktet inkrementel og give brugeren mulighed for at anvende produktet og dets nye funktioner. Sågar hurtigst muligt.

For at opsummere artefakterne: Først danner product owner med relevante udviklere opgaver til en product backlog. Development teamet vælger opgaver fra product backloggen til deres sprint backlog, typisk baseret på prioritering. Herfra begynder udviklingsperioden (sprinten) for development teamet om at fuldføre alle opgaver indenfor sprinten. Figur 3 viser processen for at nå frem til et virkende produkt.



Figur - Scrum flow

# 3. Kravspecifikation

## 3.1 …

## 3.2 …

# 4. Design

## 4.1 …

## 4.2 …

# 5. Implementation og test

## 5.1 …

## 5.2 …

# 6. Evaluering

## 6.1 …

## 6.2 …

# 7. Konklusion

## 7.1 …

## 7.2 …

# 8. Referencer

## 8.1 …

## 8.2 …

# 9. Appendiks

## 9.1 …

## 9.2 …

1. IT-Diplomingeniøruddannelsens afgangsprojekt [↑](#footnote-ref-1)