Big data og records management

”Verdiløse arkiv?”, Norsk Arkivråd - 20. oktober 2015

Tine Weirsøe, Scandinavian Information Audit

www.information-audit.dk

Lidt om denne præsentation…

45 minutter til at fortælle om et meget komplekst emne, som jeg dårligt selv forstår.

Først lidt om mig og derefter har jeg opdelt præsentationen i tre dele

1. Data, data, data overalt…
2. Cases og eksempler på big data og records management/dokumentationsforvaltning
3. Forslag til hvordan vi som profession kan bidrage til data revolutionen

Lidt om mig…

Konsulent og direktør i Scandinavian Information Audit siden 2004

Fagansvarlig og ekstern lektor ved MIR Master i Informatinsforvaltning & Records Management ved Aalborg Universitet og Københavns Universitet/IVA

Før det 20 år som records manager, arkivleder, arkivar, datakonsulent

Opgaver i de nordiske lande, EU, USA og Japan

”Et er et søkort at forstå et andet skib at føre”

Ludvig Holberg, fra Den politiske kandestøber, 1773

Data, data, data overalt…

Data opsamles, sætter spor

- Mobiltelefoner, iPhones, Smartphones
- GPS, PC'ere, iPads
- Digitale transaktioner
- Emails og meget mere

At håndtere så megen data, som for eksempel Google gør, udvikler enorme mængder varme. Processerne kræver derfor store anlæg til nedkøling - som her på Googles datacenter i The Dalles, Oregon. Foto: Connie Zhou/AP Photo/Google

Data, data, data overalt…

Hvor kommer
big data fra?

Data, data, data overalt…

- Er data det 21'ende århundredes guld eller olje?
- Databaserede virksomheder:
  - Google
  - Amazon
  - Facebook
  - LinkedIn
  - Cure4you
  - Boligsiden

- Nogle af verdens mest værdifulde virksomheder i 2015
- Deres råstof er data – data om os alle sammen
- Der værdi ligger i evnen til at forstå data, handle data og kapitalisere på data
- Et nyt begreb er "algoritme økonomi"

Data, data, data overalt…

Ingen fast vedtaget definition på big data

Big data er et begreb indenfor datalogi, der bredt dækker over indsamling, opbevaring, analyse, processering og fortolkning af enorme mængder af data…

Kilde: Wikipedia

The capability to manage a huge volume of disparate data, at the right speed and within the right time frame, to allow real-time analysis and reaction. Big data is typically broken down by three characteristics, including volume (how much data), velocity (how fast that data is processed), and variety (the various types of data).

Kilde: Big Data for Dummies; published by John Wileys and Sons, 2013

Data, data, data overalt…
Big Data vs Little Data

Big Data is what organizations know about people — be they customers, citizens, employees, or voters. Data is aggregated from a large number of sources, assembled into a massive data store, and analyzed for patterns. Big Data is what enables banks to predict credit card fraud by analyzing billions of transactions, marketers to understand customer sentiment by analyzing millions of interactions on social media, and retailers to target promotions and offers by analyzing millions of purchases.

In contrast, **Little Data** is what we know about ourselves. What we buy. Who we know. Where we go. How we spend our time. We've always had a sense for these things — after all, it's our lives. But thanks to the combination of mobile, social, and cloud technologies, it's easier than ever to gain insight into our own behavior.

Kilde: Mark Bonchek; Harvard Business Review, May 03 2013

Data, data, data overalt

Big data karakteriseres ofte med de tre V'er

- ✓ Volume: Enorme mængder data
- ✓ Velocity: Hastighed
- ✓ Variety: Forskellige typer – medier, kilder

»Data er det 21. århundredes olie. Men olie er noget sort, klæbrigt stads, indtil nogen laver det om til brændstof. På sammen måde er data ikke særlig intelligent, medmindre man har algoritmer, der fører til konkrete handlinger med værdi.«

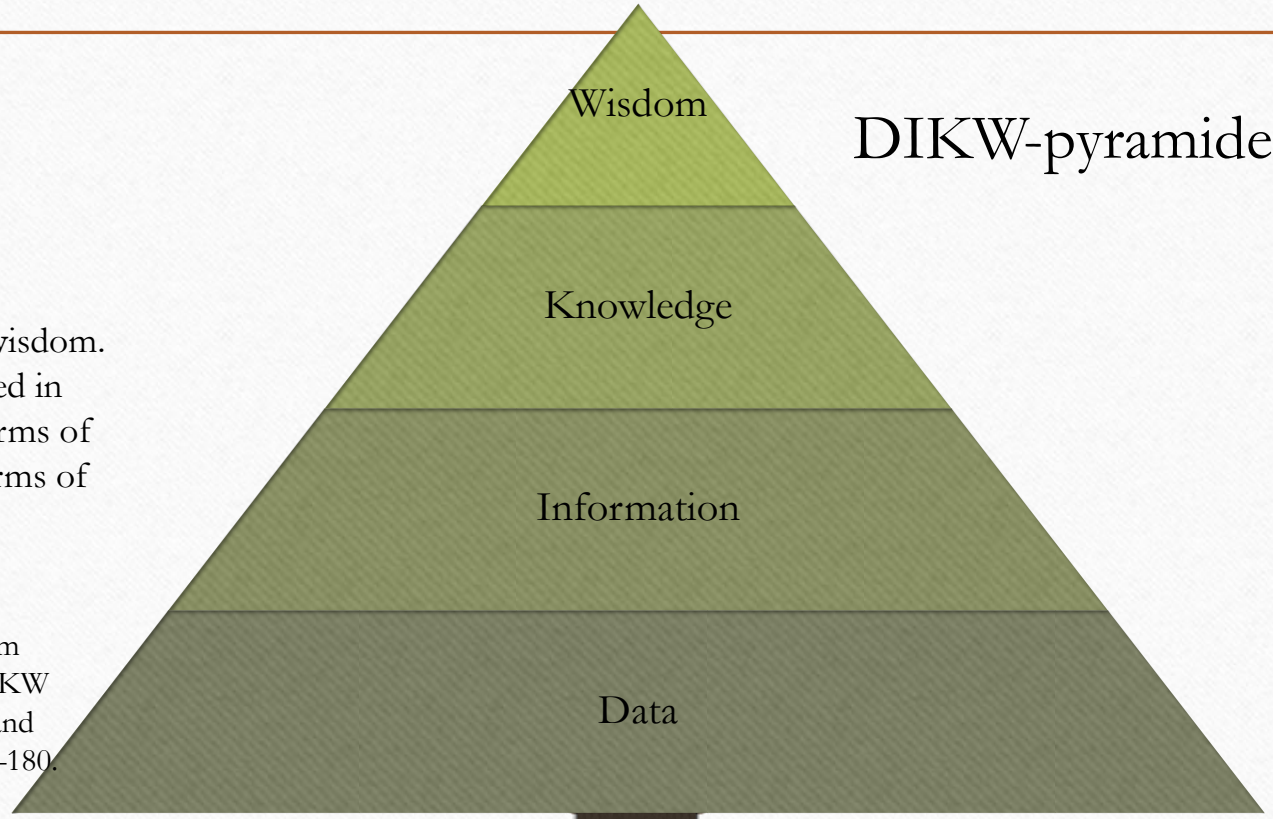
Citat: Peter Søndergaard, en af topfigurerne hos det amerikanske analysefirma Gartner.

Big data og infektionsovervågning

En gruppe it-ingeniører fra Google publicerede i 2009 en artikel i det videnskabelige tidsskrift Nature. Dér forklarede de, at de nu var i stand til at forudse den næste vinterinfluenza i USA. Og ikke bare hvordan den spredte sig nationalt, men også dens spredning helt ned på enkelte regioner og delstater.

- Basis var de over 3 milliarder daglige søgninger, som brugerne hver dag foretager via Google. Så it-ingeniørerne sad med en overflod af data om folks historiske søgninger. Derved kunne de identificere de 50 millioner mest anvendte søgeord blandt amerikanske internetbrugere, som de derefter holdt op mod de amerikanske sundhedsmyndigheders oplysninger om influenzaspredningen i årene 2003-2008.
- På den måde kunne de så se, hvad folk især søgte efter i områder, som allerede var smittet eller lå tæt på områder med smitte. Det kunne være ord som 'hoste', 'feber' eller kombinationer som 'medicin mod hoste og feber'.

Cases og eksempler på big data og records management/dokumentationsforvaltning



DIKW-pyramiden

Structural and/or functional relationships between **d**ata, **i**nformation, **k**nowledge, and **w**isdom. "Typically information is defined in terms of data, knowledge in terms of information, and wisdom in terms of knowledge".

Kilde:
Rowley, Jennifer (2007). "The wisdom hierarchy: representations of the DIKW hierarchy". Journal of Information and Communication Science 33 (2): 163–180.

Case 1: Big data og records management

En privat virksomhed ønskede at få kommentarer til deres Big data-strategi fra en records mamagement vinkel:

- Virksomheden er en fødevarevirksomhed, der er underlagt mange lovgivningskrav, fx til sporbarhed for ingredienser
- Anvender big data i forbindelse med produktionsplanlægning og markedsføring
- Vil gerne sikre beslutningsgrundlag – hvis big data er beslutningsgrundlag, så skal det bevares.

Udfordring:

- Hvordan bevares big data, så det kan søges, læses og analyseres over tid?

Case 2: Big data og records management

En virksomhed vil have en retention- og kassations politik for alle ustrukturerede data:

- IT har fundet ud af, at kun 9% af data er strukturerede, dvs at 91% er ustrukturerede
- Af de 91% ustrukturede data skønnes 40% at have værdi, idet de indgår i logs, produktionssystemer, er transaktionsdata og lignende. De øvrige er "uønskede"
- Ledelsen stiller krav til datakvalitet og vil være sikker på, at kun valide data bevares og indgår i big data analyser

Udfordringer:

- Hvordan identificeres de ustrukturerede data uden værdi?
- Hvordan sikres principperne om authensitet, integritet, troværdighed og brugbarhed?

(ISO 30300:2011, afsnit 2.3.2.: authentic reliable, integrity, useable)

Case 3: Big data og records management

En virksomhed vil have auditeret deres "big data program" med krav fra ISO 30301 og ISO 15489 samt intern governance:

- Manglende kontrol med kvalitet og brugen af big data er et problem
- Ledelsen ser værdien i big data, men ønsker at få en status på kvalitet og brug af big data i virksomheden

Udfordring:

- Krav og anbefalinger i ISO 30301 og ISO 15489 kan ikke umiddelbart anvendes på ustrukturedere data, uden tilknytning til et system, uden metadata, der er dannet tilfældigt og som et øjebliksbillede…

Case 4: Big data og records management

En statsejet virksomhed ville indarbejde deres records management governance (arkiveringspolitik, procedurer, retentionplan m.m.) i deres big data governance for at få bede kontrol med big data.

- Alle ser store værdier i big data, men kan ikke få det til at spille sammen med arkivloven og anden specifik lovgivning samt en vedtaget IT-strategi, der blandt andet omfatter et mål om oprydning og sletning af ustrukturerede data. Og hvad med persondataloven hvis enkeltpersoner kan spores i big data?

Udfordringer:

- Bevaring af big data
- Persondatalovens krav om sporbarhed relateret til big data
- Kan Rigsarkivet modtage big data?

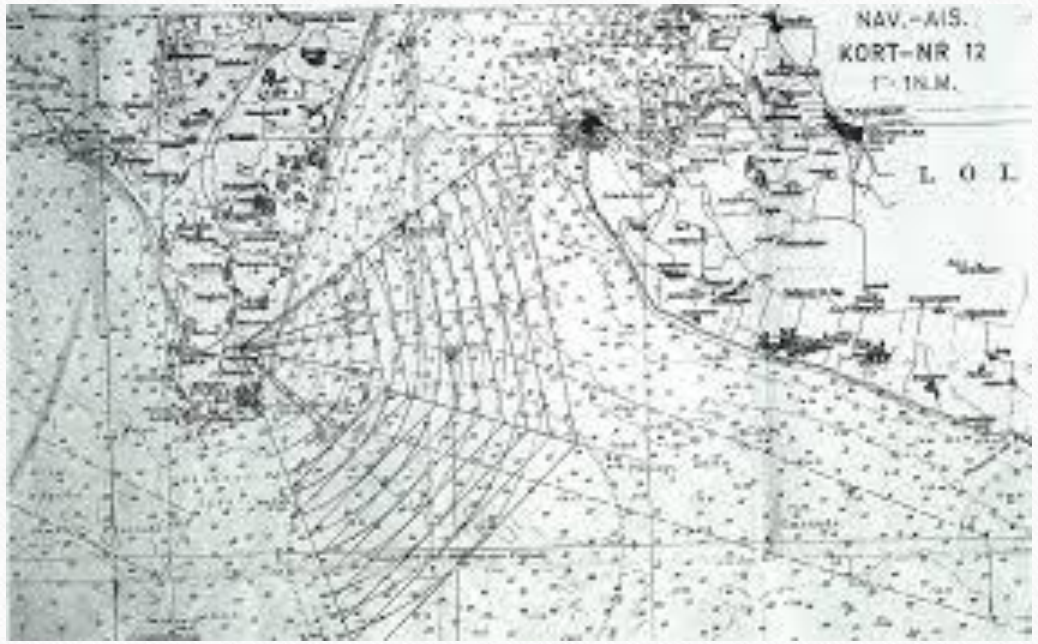
Forslag til hvordan vi som profession kan bidrage til big data og data revolutionen

Arkivarer, records managers og informationsspecialister kan bidrage med

- Struktur – ved at skabe en form for struktur i kaos af data
- Retention og kassations krav
- At finde nye veje for bevaring af flygtige og ustrukturerede data
- At hjælpe med at afklare lovgivning, fx persondatalovgivning i relation til big data
- Datavask og datakvalitet
- Data mining
- At bruge viden om vores organisationer, dens IT-systemer og governance til at at skabe overblik, rammer og interne standarder.

”Et er et søkort at forstå et andet skib at føre”

Vi skal lære at læse søkortet og finde vores rolle på skibet



Afsluttende bemærkninger

- Ingen tvivl om at data har enorm værdi for samfundet, for virksomhederne og for den enkelte
- Udfordringer med styring og kontrol med personfølsomme data – hvordan sikres det, at personer ikke kan spores?
- Personligt får jeg oftere og oftere opgaver, hvor elemeter af big data indgår
- Fremtiden vil bringe mange muligheder for arkivarer, records managers og informationsspecialisters arbejde med big data – vi skal som profession finde ud af, hvordan vi kan biddrage og skabe værdi
- Udfordringer for big data er mange, blandt andet
  - Manglende rammer og standarder
  - Brug af big data til forskning er uafklaret
  - Afklaring af grænseflader mellem IT, Informationssikkerhed, Legal og forretningsfunktioner som produktion og marketing
  - Manglende lovgivning – hvordan må man bruge big data?
  - Hvordan kan big data bevares og sikres, i de tilfælde det er relevanr

Kontaktinformation til Tine Weirsøe: Email tine@weirsoe.com - telefon (+45) 7023 14040
Tilmelding til nyhedsbrev og kursusoversigt på www.information-audit.dk
Diskussioner, jobmuligheder m.m. i LinkedIn gruppen "Forum for Records Management & Documentation