Sådan udstiller du dine data

Teknisk vejledning til at sætte Offentlige Data I Spil

Version 1.0, august 2010

IT- og Telestyrelsen
Ministeriet for Videnskab
Teknologi og Udvikling

>

Sådan udstiller du dine data
Teknisk vejledning til at sætte Offentlige Data I Spil

Udgivet af:
IT- & Telestyrelsen
Holsteinsgade 63
2100 København Ø

Telefon: 3545 0000
E-mail: data@itst.dk

ISBN (internet): 978-87-92572-27-1

>

Sådan udstiller du dine data

Teknisk vejledning til at sætte Offentlige Data I Spil

IT- & Telestyrelsen
september 2010

Indhold >

Lidt om offentlige data og det at bringe dem i spil

>

Denne vejledning har til formål at gøre det lidt lettere at få et overblik over, hvad der skal til for, at en offentlig myndighed kan udstille sine data til videre brug af andre myndigheder eller private virksomheder. Vejledningen er skrevet i forbindelse med Videnskabsministeriets initiativ "Offentlige Data I Spil" (ODIS), som samlet har til formål at fremme private virksomheders muligheder for at anvende offentlige data til at skabe nye tjenester og digitale produkter. Denne vejledning berører kun kursorisk juridiske og økonomiske forhold i relation til udstilling af data, idet disse emner behandles mere indgående i en særskilt vejledning.

Udgangspunktet i ODIS er, at offentlige data som udgangspunkt bør stilles til rådighed for andre, som kan bruge dem, og at det ofte kan være svært eller umuligt på forhånd at sige, hvad et givet datasæt kan bruges til. Derfor satses der på at få udstillet så mange data som muligt med så lille en indsats som muligt og lade mere omfattende indsatser følge konkret efterspørgsel. Dette ligger også i tråd med PSI-Loven som på baggrund af et EU-direktiv foreskriver, at offentlige data så vidt muligt skal stilles til rådighed for videreanvendelse på konkret anmodning.

Med offentlige data menes der i denne sammenhæng alle informationer, som anvendes til intern sagsbehandling i den offentlige sektor, og som lagres elektronisk af en offentlig institution eller myndighed. Det er med vilje en meget bred definition som dækker alt fra lister over oldtidsminder til CPRs autoritative liste over alle danske adresser med videre.

Hvad får myndigheden ud af at udstille sine data?

Der er konkrete fordele for de myndigheder, der udstiller deres data.
For det første kan der opstå nye tjenester, som bidrager til at løse myndighedernes opgaver uden ekstra arbejde for myndigheden. Et eksempel kunne være, at oplysninger om ledige pladser på de videregående uddannelser anvendes af private til at udvikle tjenester, som gør det let for de, der ikke kom ind på drømmestudiet, at finde en anden uddannelse. Sådanne løsninger vil ikke koste universiteterne eller Videnskabsministeriet noget, men de vil forbedre udnyttelsen af de tilgængelige studiepladser og dermed fremme institutionernes interesser.

En generel rutine med løbende udstilling af data kan spare ressourcer ved manuelle udtræk, om end der også er en risiko for, at udstilling af data kunne give merarbejde i form af flere henvendelser med spørgsmål. Dialog om data kan vendes til en fordel ved, at data bliver forbedret. For eksempel kan man ved at sammenholde CVRs lister over restauranter med de smileys, der er uddelt af FødevareErhverv, afklare, om nogle restauranter har skiftet navn, adresse, ejer og så videre. Det er endda muligt at give borgere og virksomheder mulighed for at angive rettelser til oplysninger, som de mener er forkerte eller mangelfulde.

Det er også klart, at når mange myndigheder udstiller deres data, vil de ofte kunne genbruge hinandens oplysninger frem for at indsamle på ny. Det giver effektiviseringer, som frigiver ressourcer til andre opgaver.

Hvilke data bør myndighederne udstille?

Som udgangspunkt opfordres offentlige institutioner til at udstille alle data, som ikke er fortrolige eller følsomme.

>

Det kan være kompliceret at vurdere, om et givet datasæt kan offentliggøres, og derfor er den generelle anbefaling at starte med de data, som man er sikker på ikke giver problemer, og derefter overveje at gå videre med dem, hvor der kan være behov for særlige hensyn.

Der kan være mange forskellige grunde til, at data ikke umiddelbart kan frigives, for eksempel:

- Ophavsret - en del myndigheder anvender data, som ejes af andre, for eksempel billederne i Kunstindeks, som fotografen har ophavsretten til. I sådanne tilfælde skal det afklares om, og på hvilke vilkår, data kan frigives.
- Personhenførbarhed – hvis der er tale om oplysninger om enkeltstående fysiske eller juridiske personer, kan der være udfordringer ved at udstille oplysningerne. IT- og Telestyrelsen er ved at udarbejde en vejledning i, hvordan persondata kan håndteres.
- Fortrolighed – relativt få oplysninger må betragtes som fortrolige af hensyn til rigets sikkerhed, igangværende efterforskninger, forretningshemmeligheder eller andre væsentlige hensyn.
- Økonomi – hvis det vil være meget dyrt at ændre de systemer, hvor data ligger, så de udstiller data, kan det være nødvendigt at vente med at udstille selve datasættene. Det vil dog i disse tilfælde være hensigtsmæssigt at offentliggøre en beskrivelse af datasættet, så eksempelvis andre myndigheder kan se at og hvor data findes.

Hvilket format skal jeg bruge?

>

Der skal ikke herske tvivl om, at nogle udstillingsformater gør det lettere at genbruge data end andre formater. For eksempel kan web services med standardiserede XML-formater meget let tilgås af maskiner og integreres i andre tjenester.

Mindre kan dog ofte også gøre det, og det kan være svært at vurdere, om det vil give værdi at bruge en masse kræfter, tid og penge på at lave lækre web services til alle mulige datasæt.

Derfor lyder den overordnede opfordring:

Udstil data i det bedste format, du har dem i!

Da mange data har en struktur, vil det være optimalt at benytte det format som bedst understøtter den struktur. Det er i mange tilfælde XML, men det kan også blive for tungt i store datasæt med meget ens struktur, og i den slags tilfælde er det bedre at bruge kommaseparerede filer eller andre simplere formater. Summa summarum må det komme an på en konkret vurdering ud fra datasættet og de muligheder, der eksisterer i systemerne.

Det er bedre at udstille mange data i middelmådige formater end ingen (eller få) data i superformater, så lad være med at bruge for mange kræfter før I kender efterspørgslen.

Der er jo altid mulighed for at udvikle bedre formater hen ad vejen, hvis det viser sig, at der er efterspørgsel efter bestemte data i bestemte formater. PSI-loven giver ligefrem mulighed for, at de, der er interesserede i at genbruge data, kan betale marginalomkostningerne ved udvikling af snitflader.

Hvordan bruger jeg et givet format?

>

Når myndigheden skal udstille nye data – altså data som ikke har været udstillet før – bør man vælge det format, der giver den bedste balance mellem omkostninger og egnethed til formålet. For hvert format er der nogle ting, som du bør være opmærksom på, og dette afsnit sigter på at berøre dem.

Dette afsnit handler kun om, hvordan snitfladerne bedst tilrettelægges, så maskiner kan tilgå dem direkte. Råd og vejledning om, hvordan netsteder og webløsninger bør udformes, kan du finde andetsteds.

Web services

For data som ændres ofte, og hvor hvert enkelt udtræk er begrænset i størrelse, er det meget relevant at udstille data via web services.

Der er flere forskellige måder at lave en web service, men nogle af de mest brugte er SOAP og REST. Generelt kan SOAP mere end REST, men REST services er meget nemme at udvikle og anvende, så det er en meget brugt standard.

IT- og Telestyrelsen har lavet en dansk profil af REST services, som kaldes OIOREST, se http://www.oiorest.dk/.

Database

Ligesom web services giver direkte databaseadgang mulighed for at tilgå data dynamisk. Databaser har den fordel, at de kan give brugerne mulighed for selv at sammensætte netop det udtræk, de er interesserede i.

Der er dog nogle sikkerhedsmæssige bekymringer ved at tillade eksterne databaseudtræk, og en databaseadgang er kun nyttig, hvis strukturen i databasen og betydningen af de enkelte tabeller og felter er veldokumenteret.

Ofte kan der relativt enkelt og billigt oprettes web services, som udstiller data fra en database, hvilket kan være en let måde at afhjælpe sikkerhedsbekymringerne.

XML

XML er et meget udbredt format til udveksling af data, fordi det giver gode muligheder for at bevare strukturen i data og den måde filerne opbygges på lader udviklerne skrive dele af dokumentationen ind sammen med data, uden at det forstyrrer læsningen af dem.

Offentlige myndigheder i Danmark er forpligtede til at anvende profilen OIOXML (se http://www.itst.dk/it-arkitektur-og-standarder/standardisering/datastandardisering/oioxml-udvikling/) til systematisk dataudveksling, og alle eksisterende OIOXML-formater kan findes og genbruges på www.digitalisér.dk.

IT- og Telestyrelsen har udviklet et værktøj, OIOCheck, som gør det lettere at kontrollere, at ens XML overholder kravene til OIOXML og at uploade mange XML-skemaer til www.digitalisér.dk på én gang.

RDF

Et relativt nyt format kaldet RDF gør det muligt at udstille data i en form, så maskiner delvist kan forstå og fortolke data. Dette format anvendes blandt andet af den britiske regerings Open Linked Data-projekt (se http://www.data.gov.uk/).

RDF er endnu ikke så udbredt i Danmark, og der er begrænset værktøjsunderstøttelser for det, men det giver særdeles gode muligheder for automatisk viderebehandling af data.

Da formatet er meget frit – på linje med XML – er der behov for profileringer, men der er endnu ikke udviklet en fælles dansk profil af RDF. Dog er der internationale profiler, som kan anvendes som udgangspunkt for en sådan dansk standard.

Regneark

Mange myndigheder har oplysninger liggende i regneark, for eksempel Microsoft Excel. Disse data kan ofte anvendes umiddelbart med de rette beskrivelser af, hvad de forskellige kolonner betyder.

Dog vil der i en del tilfælde være makroer og formler i regnearkene, som kan være noget mere besværlige at have med at gøre. Det er derfor tilrådeligt at dokumentere sådanne beregninger ved siden af regnearket, da det generelt er mere tilgængeligt for brugerne at læse.

Kommaseparerede filer

CSV-filer kan være et meget nyttigt format, da det er kompakt og dermed velegnet til at overføre store sæt af data med samme struktur.

Dog er formatet så spartansk, at data ofte er ubrugelige uden dokumentation, idet det kan være nærmest umuligt at gætte, hvilken betydning de forskellige kolonner har. Det er derfor særligt vigtigt for kommaseparerede formater, at dokumentationen af de enkelte felter er præcis.

Desuden er det afgørende, at strukturen i filen overholdes, idet en enkelt udeladelse af et felt kan forrykke læsningen af alle resterende data i filen uden reel mulighed for at rette op på det, da det ikke vil kunne afgøres, hvordan de resterende data skal fortolkes.

Tekstdokument

Klassiske dokumenter i formater som Word, ODF, OOXML eller PDF kan være tilstrækkelige til udstilling af visse former for data – for eksempel relativt stabile adresselister eller tilsvarende. Det kan være billigt at udstille i, da det ofte er det format, data er født i.

Formatet giver ikke støtte til at holde strukturen stringent, hvilket ofte medfører, at det er svært at indlæse data maskinelt. Sørg derfor for at bruge skabeloner som basis for dokumenter, der skal udstille data til genbrug, så det kræver mindst muligt at hive informationen ud af dokumenterne.

Det kan også støtte videreanvendelse af data at bruge typografiopmærkning så meget som muligt, så det bliver lettere for en maskine at skelne overskrifter (helst typeangivne) fra indhold og så videre.

Generelt anbefales det ikke at udstille i tekstbehandlingsformat, hvis data forefindes i et andet format.

Scannet billede

Vel nok den mindst velegnede form for de fleste data, men både TIFF og JPG-2000 kan i det mindste opmærkes med dokumentation af, hvad der er på billedet – helt op til at opmærke et billede af et dokument med hele tekstindholdet af dokumentet.
Det kan være relevant at udstille data som billeder, hvis data ikke er født elektronisk – et oplagt eksempel er gamle kirkebøger og andre arkivalier - og et billede er bedre end ingenting.

Proprietære formater

En del fagsystemer med videre har egne dataformater, som de kan lagre eller eksportere data i.

Det kan i visse tilfælde være tilstrækkeligt at udstille data i et sådant format – især hvis det forventes, at videreanvendelse vil ske i et tilsvarende system som det, de kommer fra. Det bør altid oplyses, hvor man kan finde mere information om sådanne proprietære formater, eksempelvis ved at oplyse et link til leverandørens hjemmeside.

Generelt anbefales det dog at udstille data i ikke-proprietære formater, hvor det kan lade sig gøre.

HTML

Ganske mange data ligger i dag tilgængelige i HTML-format på forskellige netsteder. Et eksempel på det er den autoritative liste over universiteter i Danmark, som kun udstilles som en HTML-side på Universitets- og Bygningsstyrelsens hjemmeside (http://www.ubst.dk/institutioner-og-okonomi/institutionsoversigt). Dette kan sagtens være tilstrækkeligt, da de pågældende oplysninger er meget stabile og begrænsede i omfang.

I en del tilfælde kunne det dog være ønskeligt at have data i en form, som lettere lader sig downloade og bearbejde, men da det er let og billigt at henvise til en side på et netsted, kan det være et godt udgangspunkt i udstilling af data.

Typisk vil det være mest hensigtsmæssigt at anvende tabeller i HTML-dokumenterne til at holde data, og her er det vigtigt, at de forskellige datafelter, der udstilles, får angivet id'er, som gør det let at finde og bearbejde data. Yahoo har udviklet et værktøj (http://developer.yahoo.com/yql/), som kan udtrække struktureret information fra en hjemmeside, og sådanne værktøjer kan meget mere med data, hvis de er omhyggeligt opmærket.

Hvordan skal det dokumenteres?

>

Overordnet er rådet om dokumentation at gøre det så godt som muligt. Dokumentationen er langt vigtigere end formatet, da det er relativt enkelt at konvertere veldokumenterede data mellem formater, men meget sværere at anvende uforståelige data selv i de bedste formater.

Det ultimative ideal er, at data kobles til en informationsmodel opbygget af semantikdefinitioner på www.digitalisér.dk, men det vil i mange tilfælde være at overdrive - en PDF med beskrivelse kan ofte være nok

Dog er det altid væsentligt at sikre, at dokumentationen har det rette indhold for at gøre det muligt og let at genbruge data. Beskriv gerne mindst:

- Hvilket format data er udstillet i
- Hvor data kan findes. Angiv gerne en URL, IP-adresse, eller hvad der nu måtte være relevant. For data, som endnu ikke udstilles, kan det give mening at offentliggøre en beskrivelse af datasættet enten på Digitalisér.dk eller myndighedens egen hjemmeside med angivelse af et telefonnummer eller en e-mail-adresse, man kan kontakte for at bede om at få udstillet data.
- Formål med datasættet. En beskrivelse af hvilke data der indgår og hvorfor er ofte meget værdifuld for at gøre det lettere at vurdere, om et datasæt er relevant til en given anvendelse.
- Hvordan datasættet er opbygget – herunder specifik dokumentation af, hvilke felter der betyder hvad.
- Hvilke vilkår gælder for tilgang og anvendelse. Hvis der kræves betaling, godkendelse, hensyn til fortrolighed eller andet, er det vigtigt at angive dette.
- Kvalitet af data – hvor korrekte mener du selv, dine data er, hvor kommer de fra, og hvor tit opdateres de?
- Kontaktoplysninger, så brugerne kan give feedback på datakvaliteten, formatet eller andet.

Husk at oprette en datakilderegistrering på www.digitalisér.dk, så interesserede kan se, at data findes, og hvordan de kan tilgås.

Checkliste

>

1. Få et overblik over, hvilke data du har
   Det er nyttigt for dig selv at vide og bydende nødvendigt for at kunne udstille dem
2. Vurdér om der er fortrolighedshensyn
   Hvis der er, så far med lempe, ellers er det bare om at få udstillet.
3. Opret datakilderegistreringer for det hele - også for data du ikke umiddelbart udstiller
   Det giver faktisk værdi at beskrive data, også selvom du ikke udstiller dem lige nu. Det kan være, at data har så meget værdi for nogen, at de vil hjælpe med at få dem udstillet.
4. Udstil data som du har dem
   Du kan tids nok lave forbedringer, hvis nogen får nok nytte af det.
5. Dokumentér dine data
   Dokumentation er afgørende for brugbarheden - og det giver tit dig et bedre overblik over dine data

Læs mere og få hjælp

>

Der er mange steder at finde mere information og få konkret hjælp.

Der er information på IT- og Telestyrelsens hjemmeside:
http://www.itst.dk/digitale-losninger/offentlige-data.

Du bør naturligvis også klikke lidt rundt i Datakildekataloget på Digitalisér.dk:
http://digitaliser.dk/ressourcer?tabContainerResources=tabDatakildeResources

Der er også mange tilsvarende aktiviteter i udlandet. Et sted, hvor du kan få meget information – på engelsk – er hjemmesiden for det britiske projekt om Open Linked Data:
http://data.gov.uk/

Standardiseringsorganisationen W3C har også udgivet flere vejledninger på engelsk:

- http://www.w3.org/DesignIssues/GovData
- http://www.w3.org/TR/gov-data/

Du kan også sende en e-mail til IT- og Telestyrelsen på data@itst.dk eller ringe på 3545 0000, hvis du er i tvivl om noget, eller du har brug for konkret hjælp med at få sat dine data i spil.

<

Overskrift
Bagside kursiv tekst

Bagside brødtekst