Retningslinjer for stabile http-URIer

Indledning

De efterfølgende retningslinjer for udformning og anvendelse af stabile http-URIer er rettet mod initiativerne under den fællesoffentlige digitaliseringsstrategi 2016-2020. Retningslinjerne er blevet til på baggrund af tilsvarende arbejde i regi af EU og W3C.

ISAs rapport *"D7.1.3 - Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC"* [1], har været det primære grundlag og fra W3Cs specifikation *"Data on the Web Best Practices"* [2] er der hentet yderligere råd.

Som et resultat af erfaringsindsamlingen til ISAs rapport definerede rapportens forfattere ti regler, *"10 rules for persistent URIs"*, hvis formål er at formidle 'best practice' for design og publicering af stabile URIer. De ti regler er opdelt i fem punkter, der tilrådes, og fem punkter, der frarådes:

ISAs 10 regler er i dansk regi blevet til 14 retningslinjer. Først og fremmest tales der i nærværende skrift ikke om regler, der potentielt kunne blive for rigide, men om retningslinjer, der gerne skulle give mere fleksibilitet og mindst muligt administrativt arbejde. Alle 10 regler er videreført som retningslinjer. Fire yderligere retningslinjer er tilføjet for mere præcist at beskrive, hvad der menes med at "overholde det vedtagne mønster".

Danske retningslinje	ISA-regel
URIer skal overholde det vedtagne mønster	Follow the pattern
En URI skal defineres som en HTTPS-URI	
Lad alle http-URIer tilhøre domænet https://data.gov.dk/ eller et af dettes underdomæner	
URIer skal i videst muligt omfang være menneskeligt læselige ord	
Termer anvendt i URIer skal være engelsksprogede	
Genbrug i videst muligt omfang eksisterende identifikatorer	Re-use existing identifiers

[1] https://joinup.ec.europa.eu/catalogue/distribution/study-persistent-uris-identification-best-practices-and-recommendations-topic
[2] https://www.w3.org/TR/dwbp/

Danske retningslinje	ISA-regel
Link til flere repræsentationer	Link to multiple representations
Anvend 303-omdirigering til ikke-informationsressourcer	Implement 303 redirects for real-world objects
Brug dedikerede services	Use dedicated service
URIer skal ikke udtrykke ejerskab	Avoid stating ownership
URIer må ikke versioneres	Avoid version numbers
URIers referencedel må ikke have fortløbende automatisk opdatering	Avoid using auto-increment
URIer må ikke indeholde søgestrenge	Avoid query strings
En URI må ikke afsluttes med filendelse	Avoid file extensions

Retningslinjerne

Det http-baserede internet, også blot kaldet web'et, er i dag en de facto platform for interoperabilitet mellem it-systemer og vil fremover i stigende grad være den platform, det offentliges data og dokumenter vil blive udstillet og udvekslet ved hjælp af.

Web'et er i sin natur et decentraliseret netværk af data og dokumenter, der er entydigt identificeret ved hjælp af http-URIer.

Web'et er også en platform, hvor ændringer af identifikatorer (http-URIer) i et it-system vil kunne påvirke andre it-systemer negativt. Hvis potentialet og værdien af web'et skal udnyttes, er der derfor behov for at sikre, at http-URIerne holdes stabile og vedvarende, så de, når de anvendes til opslag, altid peger på samme ressource.

Retningslinjerne søger at sikre, at de URIer, der defineres til brug under digitaliseringsstrategien, får den nødvendige stabilitet.

Retningslinjernes fokus er på data, datasæt, webAPIer, datamodeller og modelelementer, mens de ikke anses for nødvendige for adresser for websider i almindelighed.

Brug af https

Hvor ISAs URI-mønster er baseret på brug af HTTP URIer, anbefaler retningslinjerne brug af HTTPS.

§	**En URI skal defineres som en HTTPS-URI** *Opfylder ISA-regel:* Follow the pattern

En http URI identificerer entydigt en ressource. En http URI kan også fungere som en adresse, der kan åbnes for at hente yderligere information om ressourcen.

HTTPS URIer har samme mulighed og giver samme fordele, men når URIen bruges til opslag krypteres kommunikationen, og dermed forøges sikkerheden.

Derfor anvendes HTTPS URIer som URIer.

Brugervenlige URIer

I vid udstrækning er brugen af HTTP URIer og HTTPS URIer overladt til maskiner. En tekststreng som

> http://eksempel.dk/xle4an3v/k48gn2kg/93ng7s34/

kunne derfor godt anvendes som del af en valid URI, men af hensyn til den menneskelige side af anvendelsen, er en sådan URI ikke altid hensigtsmæssig.

Det er nødvendigt og praktisk, dels at denne type URIer har en form, der gør dem lettere at se hensigten med, dels gør dem lettere at huske. Den ovenstående URI ville både være svær at huske og vil heller ikke ville give nogen form for information om den ressource, den identificerer. Hvis URIens tekststreng derimod ser ud som denne:

> http://example.dk/document/geometry/circles/

er det lettere at anvende den, både for brugere og udviklere. Navngivningen af URIen giver tilmed et fingerpeg om, at URIen identificerer et dokument vedrørende cirkler som geometriske figurer.

Derfor følgende retningslinje:

§	**URIer skal i videst muligt omfang være menneskeligt læselige ord**
	Opfylder ISA-regel: Ingen tilsvarende ISA-regel

Retningslinjen bruger udtrykket ”i videst muligt omfang” i erkendelse af, at anvendelse af ord ikke altid er den mest hensigtsfulde løsning i en URIs referencedel, hvor der eksempelvis kan være behov for at anvende en UUID. Dette uddybes i afsnittet ’Brug af {reference}’.

Bemærk at det at vi, som mennesker, kan aflæse URIen ikke strider mod konventionen om, at ”URIer ikke må være betydningsbærende”. Hensigten med det udsagn har aldrig været, at en URI ikke måtte være forståelig eller let genkendelig. Hensigten har primært været, at en URI skal kunne anvendes maskinelt, som den er. Det skal med andre ord ikke være nødvendigt først at behandle URIens enkelte dele, før URIen kan anvendes.

Engelsksprogede URIer

Af digitaliseringsstrategiens modelregler fremgår det, at alle termer, der indgår i URIer for metadata, skal være engelske. Dette er i overensstemmelse med den eksisterende konvention for vokabularer og services, der i dag primært udtrykkes ved brug af engelske termer. Brugen af et sprog internationalt gør det lettere at dele data og metadata globalt.

Derfor følgende retningslinje:

§	**Termer anvendt i URIer skal være engelsksprogede**
	Opfylder ISA-regel: Ingen tilsvarende ISA-regel

Bemærk at der med udtrykket ’engelsksprogede’ ikke er taget stilling til, om de anvendte termer skal være britisk-engelsk, amerikansk-engelsk eller anden form for engelsk. Det overlades til den URI-definerende aktør at foretage det valg, der giver den bedste formidling.

Brug vedtagne URI-mønstre

En fordel ved at følge fastlagte mønstre ved dannelse af URIer, er alene rettet mod den menneskelige bruger. Velkendte og genkendelige mønstre gør det lettere at se, hvilken type ressource man som bruger er ved at tilgå.

ISAs regel ”Follow the pattern” handler om opbygning af URIer. ISAs forslag er, at URIer skal have følgende mønster:

```
http://{domain}/{type}/{concept}/{reference}
```

I vores regi bliver dette til følgende retningslinje:

§	**URIer skal overholde det vedtagne mønster**
	Opfylder ISA-regel: Follow the pattern

hvor følgende mønster skal følges:

```
https://{domæne}/{type}/{emne}/{reference}/
```

Mønsterets enkelte led behandles i de efterfølgende afsnit.

Bemærk at den afsluttende skråstreg kan undværes, hvis URIens identificerer en model. Det vil sige hvis {type}= ’model’.

Brug af {domæne }

For at sikre størst mulig stabilitet benyttes ”data.gov.dk” som {domæne}, med mulighed for at tilføje en række underdomæner.

Det primære domæne, ”data.gov.dk”, bør forbeholdes til de URIer, der identificerer forekomster, der har bred, fællesoffentlig karakter.

Alle URIer, der følger retningslinjerne, skal altså enten starte med tekststrengen ”https://data.gov.dk/” eller med en tekststreng, hvor et reguleret underdomæne indgår ”https://<*underdomæne*>data.gov.dk/”.

§	**Lad alle http-URIer tilhøre domænet https://data.gov.dk/ eller et af dettes underdomæner** *Opfylder ISA-regel:* Follow the pattern

Bemærk at retningslinjen ikke er ensbetydende med, at alle ressourcer, alle data, skal have en konkret ’fysisk’ placering under domænet. Det er muligt at lade ressourcer være placeret under andre domæner, hvortil der kan omstilles ved hjælp af en 303-omdirigering (’URL redirection’).

Bemærk at omdirigering til andre domæner end det primære domæne ikke fritager det anvendte domæne fra at følge de øvrige retningslinjer i dette dokument. Der bør eksempelvis ikke omdirigeres fra det HTTPS-baserede primære domæne til et mindre sikkert HTTP-baseret domæne.

Tilladte underdomæner publiceres på http://data.gov.dk/catalogue/subdomains/ sammen med oplysning om, hvilken organisation der har ansvaret for underdomænet. Før dette er sket, kan underdomænet ikke anvendes.

Underdomæner skal repræsentere offentlige forretningsområder og have et abstraktionsniveau, der er tilstrækkelig højt til at gøre underdomænet langtidsholdbart, således at de definerede URIer er stabile og uafhængige af fx ressortomlægninger.

Et eksempel på et muligt underdomæne kunne være ’geo’ – for geodata. URIer udformet med dette underdomæne kunne derfor starte med følgende tekststreng:

”https://geo.data.gov.dk/”

Den organisation, som er ansvarlig for et underdomæne, skal sikre, at al relevant brug af underdomænet er tilgængelig for alle relevante organisationer inden for forretningsområdet.

Brug af {type}

Tekststrengen {type} er et enkelt ord, der skal vælges fra et fast sæt af mulige udtryk. De foreløbigt accepterede udtryk er:

Type	Forklaring
'id'	'id' angiver, at URIen er en identifikator for et konkret objekt, der ikke er en informationsressource. Objektet selv kan med andre ord ikke tilgås direkte via internettet. Bemærk at der dermed ikke umiddelbart er skelnet mellem reelle og fiktive forekomster. 'id' kan efterfølges, på pladsen for {begreb}, enten af det anonyme 'thing' eller af en mere specifik typeangivelse, eksempelvis 'person', 'place', 'organization' eller 'event' (se efterfølgende afsnit "Brug af {emne}").
'doc'	{doc} er dokumenter, der beskriver konkrete objekter, der ikke er informationsressourcer.
'model'	{model} er alle former for metadatadefinitioner, ofte udtrykt som databeskrivende modeller. Under modeller hører blandt andet, men ikke udelukkende begrebsmodeller, kernemodeller, vokabularer, anvendelsesmodeller og -profiler.
'dataset'	{dataset} bruges, når URIen skal anvendes til identifikation af et sæt eller en samling af data. Et datasæts dataobjekter eller dataforekomster identificeres ved at benytte datasættets http-URI som basisURI og tilføje relevant identifikator.
'api'	{api} bruges når URIen anvendes til at identificere et http-baseret webapi (eller webservice).

Brug af {emne}

Tekststrengen {emne} skal entydigt beskrive det emneområde, ressourcen definerer eller er defineret under. For URIer, der indeholder et underdomæne, forventes det, at emneområdet er en delmængde af det ressortområde, der er tilknyttet underdomænet.

Eksempelvis vil en model for emneområdet familiestruktur (family structure) kunne tildeles følgende URI,

`https://data.gov.dk/model/`**`familystructure`**

hvor **`familystructure`** udgør {emne}-delen af URIen.

Tekststrengen {emne} kan være 'flerledet'; eksempelvis som disse fire emneområder:

```
/infrastructure/roads/
/infrastructure/rails/
/infrastructure/roads/construction
/infrastructure/roads/maps
```

For URIer dannet på basis af det primære domæne, "http://data.com.dk/", gælder det, at når {type} er ***'id'*** eller ***'doc'*** skal {emne} være et af følgende emneområder:

Emne	Forklaring
'thing'	er overbegrebet for alle typer af forekomster i den virkelige verden. Ved at anvende det anonyme 'thing' gives der ingen oplysninger i URIen om det refererede objekts natur. Ved at bruge mere specifik URI-del end 'thing', eksempelvis 'organization', *kan* udvalgte URI-mønstre reserveres til myndigheder hvis ressortområde dækker det pågældende emne.
'person'	Angiver at URIen identificerer en person.
'place'	Angiver at URIen identificerer et sted eller en lokation, eksempelvis "Samsø".
'organization'	Angiver at URIen identificerer en organisation i bred forstand, eksempelvis "Røde Kors".
'event'	Angiver at URIen identificerer en hændelse eller en begivenhed; eksempelvis "Folketingsvalget 2015".
'core'	Kun hvis {type}='model' – angiver at modellen er en kernemodel/et vokabular – en model som beskriver et snævert emneområde
'profile'	Kun hvis {type}='model' – angiver at modellen er en anvendelsesmodel/application profile – en model som sammensætter kernemodeller/vokabularer til en specifik anvendelse

Bemærk at ovenstående liste vil kunne udvides, men ikke reduceres efter retningslinjernes offentliggørelse.

Brug af {reference}

Tekststrengen {reference} er det led i URIen, der er givet til den enkelte model, det enkelte datasæt eller den konkrete forekomst.

En model for hybridbiler kunne eksempelvis have referencen `hybridcars`. En samlet URI for modellen kunne dermed være `https://data.gov.dk/model/transport/hybridcar` .

Tilsvarende kan en organisation med referencen `bb16` være identificeret med URIen `https://data.gov.dk/id/organization/bb16/`.

Bemærk at det for URIer hvor {type}='model' er tilladt at identificere enkeltelementer i den pågældende model ved brug af fragmenttegn (#). Indgår der i modellen for hybridbilen et element `Engine`, så identificeres det ved følgende URI:

`https://data.gov.dk/model/transport/hybridcar#Engine`

Selvom URIen som sådan entydigt identificerer det specifikke element, `Engine`, i modellen så vil et opslag af URIen hente modellen i sin helhed. Da modeller er datamæssigt relativt små og da der ofte er brug for flere relaterede oplysninger fra modellen, er brug af fragmenttegnet i denne sammenhæng praktisk og anbefalet.

Brug af fragmenttegn frarådes i alle øvrige tilfælde!

Genbrug universelt unikke identifikatorer i URI-form

I de tilfælde hvor der eksisterer en fælles offentlig identifikator, bør denne gives en http-URI-form.

§ **Genbrug i videst muligt omfang eksisterende identifikatorer**

Opfylder ISA-regel: Re-use existing identifiers

Eksempelvis kunne et cvr-nummer som ’34051178’ (Digitaliseringsstyrelsen) repræsenteres som denne http-uri:

```
https://data.gov.dk/id/organization/34051178/
```

En URN i form af en UUID, der eksempelvis har været anvendt til at identificerer fakturaer kan omdannes fra denne URN:

```
urn:uuid:91aa87da-9f06-11e7-abc4-cec278b6b50a
```

til følgende http-URI:

```
https://data.gov.dk/id/thing/91aa87da-9f06-11e7-abc4-cec278b6b50a/
```

Hvor en eksisterende identifikator ikke ønskes anvendt direkte i den nuværende form, kan URI-repræsentationen eksempelvis anvende en hash-værdi af identifikatoren. Eksempelvis kunne et CPR-nummer som ’2310450637’ gives følgende URI:

```
http://data.gov.dk/id/person/5985e9a05ca2d667b8a3f1b53609f16feccea23c1
262172ddc192c8f/
```

hvor CPR-nummeret er konverteret ved brug af hash-funktionen SHA-3 (Secure Hash Algorithm 3), hvormed entydigheden er bevares samtidigt med at anonymiteten er sikret.

Brug af SHA-3 (og tilsvarende) gør det muligt for et it-system at gentage konverteringen fra, som her et givet CPR-nummer, og altid få samme krypterede kode, der så kan anvendes til forespørgsel i andre systemer. Det er ikke muligt at gå den modsatte vej, det vil sige, der kan ikke konverteres tilbage fra SHA-3-koden til det oprindelige CPR-nummer. Den dannede http-URI kan altså bruges uden frygt for at afsløre selve CPR-nummeret.

En ressource kan have mange repræsentationer

Persistente URIer anvendes til at identificere både informationsressourcer og ’konkrete objekter’. En informationsressource er i denne forbindelse noget der kan transmitteres som en strøm af bytes. Et konkret objekt er et objekt, der i sig selv ikke kan transmittes som en byte-strøm.

§ **Link til flere repræsentationer**

Opfylder ISA-regel: Link to multiple representations

Uanset hvilken type af ressource en URI repræsenterer, så vil forskellige brugsscenarier have behov for forskellige formater af den modtagne information. Afhængigt af om brugeren er et menneske, der skal have en visuel fremstilling, eller om brugeren er et maskinafviklet program, skal den returnerede information være i et format der, kan anvendes af brugeren. Hvis vi, som eksempel, har en URI som denne:

```
http://data.gov.dk/doc/foo/bar/
```

så kan der være behov for, afhængigt af brugeren, at få returneret enten HTML eller et af RDFs serialiseringsformater (eksempelvis JSON-LD).

For begge de (i dette tilfælde) to mulige repræsentationer gælder det, at de skal have deres egen URI, overholdende retningslinjerne. Den enkleste måde at opnå det på vil være at benytte den originale URI tilføjet henholdsvis `.html` og `.jsonld`

De to URIer bliver altså henholdsvis

```
http://data.gov.dk/doc/foo/bar.html
```

og

```
http://data.gov.dk/doc/foo/bar.jsonld
```

En repræsentation af en ressource skal have reference til alle de øvrige repræsentationer af samme ressource.

Omdirigering fra en ressource til en anden ressource

Når en URI, der repræsenterer et konkret objekt, en ikke-informationsressource, skal åbnes, så skal der omdirigeres til et dokument, der beskriver objektet. Omdirigeringen sker med en 303 http- svarkode.

§	**Anvend 303-omdirigering til ikke-informationsressourcer** *Opfylder ISA-regel:* Implement 303 redirects for real-world objects

Dette bør gøres på en måde, der er konsistent og overholdende retningslinjerne. Det anbefales, at gøre dette ved at ændre `{type}`-delen af URIen fra `'id'` til `'doc'`. Eksempelvis således:

Fra URIen for det konkrete objekt:

```
https://data.gov.dk/id/organization/bb16/
```

til URIen for dokumentet om objektet:

```
https://data.gov.dk/doc/organization/bb16/
```

Service til håndtering af de persistente URIer

Retningslinjerne anbefaler at benytte et mønster, hvor domænet er fastlagt til `https://data.gov.dk/`. Hvis alle de URIer, der bliver dannet efter retningslinjerne, derefter skulle håndteres af en enkelt service, ville der være skabt et 'single point of failure'; fejler den ene service ville alle opslag på URIerne fejle.

§	**Brug dedikerede services** *Opfylder ISA-regel:* Use dedicated service

Flere distribuerede services skal derfor oprettes til håndtering af URI-opslag.

URI-mønstre der skal undgås

URIer der udtrykker specifikt ejerskab

Mange eksisterende http-URIer er dannet med et domæne, hvori et foranderligt navn indgår. Ofte er det navnet på eller akronymet for den organisation, der ejer websitet. Ændring af organisationsnavn medfører normalt også, at domænets navn bliver udskiftet. De URIer der er dannet med det gamle domænenavn skal derfor også ændres og er dermed ikke længere stabile.

Derfor bør URIer ikke indeholde dele, der angiver et specifikt ejerforhold.

§	**URIer skal ikke udtrykke ejerskab** *Opfylder ISA-regel:* Avoid stating ownership

Ved at bruge retningslinjerne undgås dette.

URIer med versionsnumre

Mange ressourcer gennemløber et livsforløb. Dette sker eksempelvis for dokumenter og for modeller som vokabularer og anvendelsesprofiler. For at holde alle URIer persistente skal information om versionsnummer eller livscyklusstatus holdes ude af URIerne.

§	**URIer må ikke versioneres** *Opfylder ISA-regel:* Avoid version numbers

For ressourcer, der har en livscyklus, eksempelvis en model, kan løsningen være at have en fast URI, der altid omdirigeres til den seneste version af modellen ved brug af dennes URI. Eksempelvis således:

Som fast URI:

`https://data.gov.dk/model/`**`familystructure/`**

URIer for modeller, der repræsenterer modellen på forskellige tidspunkter, henholdsvis 2/3 2017 og 25/4 2017:

`https://data.gov.dk/model/familystructure-02032017/`

`https://data.gov.dk/model/familystructure-25042017/`

Når den persistente URI `https://data.gov.dk/model/familystructure/` anvendes, bliver opslaget omdirigeret til den seneste model, altså til `https://data.gov.dk/model/familystructure-25042017/` .

Automatisk opdatering af fortløbende referencedel

§	**URIers referencedel må ikke have fortløbende automatisk opdatering i distribuerede miljøer** *Opfylder ISA-regel:* Avoid using auto-increment

Dannelse af nye URIer for store datasæt kræver automatiserede processer, der skal garantere, at de producerede URIer er unikke.

Brug af fortløbende numre og automatiseret opdatering af disse er en mulighed hvis, og kun hvis, der er absolut sikkerhed for, at samme URI ikke bliver dannet eller vil blive dannet et andet sted eller på et andet tidspunkt.

URIer må ikke indeholde søgestrenge

Brug af søgestrenge, eksempelvis '`?para alue`', giver ikke persistente URIer; URIens ressource kan principielt variere fra søgning til søgning.

§	**URIer må ikke indeholde søgestrenge** *Opfylder ISA-regel:* Avoid query strings

Persistente URIer må ikke afsluttes med filtypeekstension

De stabile URIer holdes fri af filtypeangivelser. Data eller dokumenter hvor filtypeekstension ønskes anvendt skal gives URIer, der henvises til som beskrevet under afsnittet "En ressource kan have mange repræsentationer".

§	**En URI må ikke afsluttes med filendelse** *Opfylder ISA-regel:* Avoid file extensions