

Rekommendation om beständiga identifierare

Beständiga identifierare bör utformas som URI:er, (Uniform Resource Identifier), enligt det rekommenderade formatet:

[http\(s\)://{domän}/{typ}/{koncept}/{referens}](http(s)://{domän}/{typ}/{koncept}/{referens})

Inom vissa sektorer/områden kan det vara relevant med specifika typer av identifierare.

Beständiga identifierare – en definition

Beständiga identifierare (persistent identifiers/PID:ar) är unika kodsträngar kopplade till både objekt och mer abstrakta konceptuella begrepp. De är en förutsättning för att man ska kunna skapa länkar i ett nätverk.

Det är viktigt att identifierarna är utformade på ett sådant sätt att de är globalt unika och beständiga. Ifall identifierarna endast används internt bör de vara utformade på ett sådant sätt att de kan integreras i formella standarder och format vid extern användning. Beständiga identifierare bör vara enkla och strukturerade efter en verklig uppdelning och inte kopplade till tekniska lösningar som används, då tekniken ändras över tid.

En beständig identifierare kan utgöras av en kombination av institutionens interna identifierare samt information om vilken typ av resurs som identifieras och hur resursen kan nås.

Olika typer av identifierare

URI (Uniform Resource Identifier)

Det finns olika typer av identifierare. På Internet eller i ett datornätverk används ofta en textsträng, URI (Uniform Resource Identifier) för att identifiera en abstrakt eller fysisk resurs. En URI är en identifierare av en resurs på Internet eller i ett datornätverk och kan klassificeras som en platshänvisning (URL), ett namn för resursen (URN) eller både och.

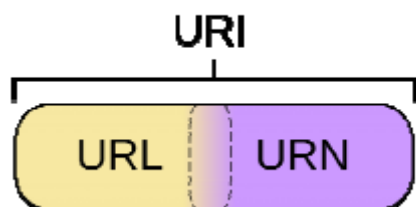


Bild 1. En URI kan vara en URL (Uniform Resource Locator) , en URN (Uniform Resource Name) eller både och. (Källa: "URI Euler Diagram no lone URIs" by David Torres original author derivative work: Qwerty0 (talk) - URI_Venn_Diagram.svg. Licensed under CC BY-SA 3.0 via Wikimedia Commons.)

En typ av URI som de flesta är bekanta med är de adresser vi använder för att peka ut webbsidor på Internet och som läses med hjälp av en webbläsare. Om det är en webbläsare som gör förfrågan till en URI ska det som kommer tillbaka vara en webbsida som kan läsas av en människa. Om det är en RDF-klient som efterfrågar samma URI ska svaret vara en RDF-fil.

Specifika typer av beständiga identifierare

Ibland kan det även vara relevant med specifika typer av identifierare. Inom bibliotekssektorn finns det exempelvis behov av att kunna göra skillnad mellan ett unikt namn och en plats. De flesta bibliotek använder därför idag URN:er (Uniform Resource Name) för publikationer, en standard utvecklad av IETF (Internet Engineering Task Force) som identifierar en resurs. Kungliga biblioteket delar ut URN:er enligt en särskild tillämpning av standarden, URN:NBN. NBN står för National Bibliographic Number. En URL (Uniform Resource Locator) ger information om hur man når resursen och var den finns. URN ska inte bytas ut så länge man inte förändrar innehållet i en resurs och den ska inte heller återanvändas från en annan resurs. Genom URN:NBN kan man göra skillnad mellan ett unikt namn och en plats, i de fall en namngiven resurs eller dess kopior finns på flera platser.

Exempel på URN:NBN (från <https://www.kb.se/om-oss/identifikatorer-och-utgivning/urnnbn/undernamnrymd.html>)

urn:nbn:se:digitalpubl-2005-112

urn:nbn:se - den obligatoriska delen av identifikatorn
:digitalpubl - undernamnrymd reserverad hos KB
-2005-112 - dokumentets nummer tilldelat av organisationen

Inom Life-Science finns också specifika typer av identifierare, LSID:er, Life Science Identifier, <http://www.lsid.info/>, bland annat då det finns behov av att hantera distribuerade resurser och koppla samman metadata och data om samma resurs.

DOI (Digital Object Identifier) <https://snd.gu.se/sv/om-oss/tjanster/pid> används som identifierare för att tillgängliggöra data inom forskningssektorn, bland annat för att möjliggöra korrekt citering av primärforskaren samt att visa vilken version av data som använts.

Utformning av beständiga identifierare som URI:er

Det rekommenderade formatet för beständiga identifierare följer de riktlinjer som tagits fram inom EU. Formatet är uppbyggt kring ett generellt mönster, <http://{domän}/{typ}/{koncept}/{referens}>. Nedan förklaras beståndsdelarna i formatet:

{domän}

Med domän avses en kombination av värd (ex. data.gov.uk , europa.eu , libris.kb.se) och relevant sektor (ex. utbildning, transport, kultur). Sektorn kan antingen anges som en subdomän eller vara första delen i sökvägen: sektor.värd.se/ eller värd.se/sektor/

{typ}

Med typ menas en av en liten uppsättning termer som definierar vilken typ av resurs som identifieras. Det är följande fyra typer som ska användas: "id" för fysiska objekt, "dok" för dokument, "def" för definition av ett koncept och "set" för en datamängd.

{koncept}

Konceptet kan vara en samling av något, det kan vara typen av det identifierade objektet (ex. skola, väg, stad) eller namnet på ett konceptschema (ex. språk).

Reglerna för val av koncept är att konceptet ska vara ett substantiv, stå i obestämd form singularis, vara den vedertagna termen och att ordet som väljs ska vara det som närmast beskriver konceptet.

{referens}

Referensen pekar ut det specifika objektet, termen eller konceptet: /skola/1212 eller /vaeg/e4 eller /stad/stockholm eller /spraak/swe.

Exempel på beständiga identifierare

Inom projektet har inga exempel identifierats där myndigheterna använder modellen fullt ut, men här är några exempel på tillämpning:

Kungliga Biblioteket:

<http://libris.kb.se/bib/11284499>

sektor = libris
värd = kb.se
domän = libris.kb.se
typ = bib (anger att det är en bokreferens som utpekas)
referens = 11284499 (anger att den specifika boken är "ett urval dikter" av Karin Boye)

Riksantikvarieämbetet:

<https://kulturarvsdata.se/objekt/lamning/7b70c89a-c191-4d73-ac40-4455fd5dd2a7> (ej upplösbar ännu)

domän = kulturarvsdata.se

koncept = objekt

typ = lamning

referens = 7b70c89a-c191-4d73-ac40-4455fd5dd2a7

Teckenstandard och namngivning

Endast gemener från a-z, siffror, bindestreck och understreck ska användas. Övriga skiljetecken, mellanslag, accenter eller specialtecken i redan befintliga identifierare skall tas bort eller ersättas av

bindestreck eller understreck. Om en befintlig identifierare innehåller svenska tecken (ex. väg och språk) ska dessa översättas enligt följande:

Bokstav	Översättning
å	aa
ä	ae
ö	oe

Språkidentifiering

För att ange vilket språk en identifierad resurs följer ska ISO 639-3 3 4 för språkidentifiering användas (ex. swe för svenska och eng för engelska).

Att tänka på vid utformning av identifierare

Tio regler för beständiga identifierare

- ✓ Använd det rekommenderade formatet
`http://{domän}/{typ}/{koncept}/{referens}`
- ✓ Återanvänd identifierare
`http://utbildning.exempel/id/skola/12345`
- ✓ Länka multipla representationer
`http://data.exempel.org/doc/foo/bar.html`
`http://data.exempel.org/doc/foo/bar.rdf`
- ✓ Använd 303-redirect för fysiska ting
`http://example.com/id/alice_brown`
- ✓ Använd en särskild tjänst/system
oberoende av datautgivaren
- ✗ Undvik att ange ägarskap
`http://education.data.gov.uk/ministryofeducation/id/school/1234`
- ✗ Undvik versionsnummer
`http://education.data.gov.uk/doc/school/v01/1234`
- ✗ Undvik automatisk uppräknig
`http://utbildning.exempel/id/skola/12345`
`http://utbildning.exempel/id/skola/12346`
- ✗ Undvik frågesträngar
`http://education.data.gov.uk/doc/school?id=123456`
- ✗ Undvik filändelser
`http://education.data.gov.uk/doc/schools/123456.csv`

Använd det rekommenderade formatet

Det rekommenderade formatet är <http://{{domän}}/{{typ}}/{{koncept}}/{{referens}}> En beständig identifierare bör följa rekommendationen och innehålla information som inte kommer att förändras över tid.

Återanvänd identifierare

Om en resurs redan är unikt refererad ska dess identifierare återanvändas. Om det till exempel redan finns ett nummersystem för att identifiera skolor kan skolan 12345 refereras till som:

<http://utbildning.exempel/id/skola/12345>

Det är dock viktigt att komma ihåg att endast identifierare som själva är beständiga ska användas.

Länka multipla representationer

Olika representationer av samma resurs ska alla länka till varandra med lämplig metod och teknik. I HTML ska till exempel 'link'-elementet användas med 'rel'-värdet satt till 'alternate', och i RDF ska 'dcterms:hasFormat' användas.

Använd 303-redirect för fysiska ting

Om en URI-id refererar till ett fysiskt objekt ska URI:n svara med http-svarskoden 303 och skicka vidare till ett webbläsarvänligt dokument som beskriver objektet.

Använd en särskild tjänst/system

I alla lyckade implementeringar av beständiga URI:er har en oberoende särskild tjänst som är fristående från dataägarna använts. Genom att lyfta ut hanteringen av URI:erna kan driften lättare flyttas runt om något skulle hända med organisationen.

Undvik att ange ägarskap

Namnet på den organisation eller det projekt som skapar en URI ska inte vara en del av URI:n då URI:n blir obsolet om det sker en organisationsförändring eller om projektet läggs ner eller byter namn, exempelvis förvaltare/ägare, titel, ämne m.m. Eftersom offentliga organisationer byter namn och ansvarsområden med jämna mellanrum får inte URI:erna vara beroende av en struktur där till exempel informationsägare finns med.

Undvik versionsnummer

Även om konceptschema, ontologier, taxonomier och vokabulärer ofta går igenom många iterationscykler ska versionsnummer och statusinformation inte vara en del av URI:n.

Undvik automatisk uppräknings

När det skapas URI:er för stora datamängder kommer processen att automatiseras. Ett sätt att göra detta på är att räkna upp ett tal som blir id-numret på resursen. Om detta förfarande används skulle URI:erna för två skolor kunna bli:

<http://utbildning.exempel/id/skola/12345>

<http://utbildning.exempel/id/skola/12346>

Enda tillfällena då automatisk uppräknings får användas är när processen aldrig kommer att upprepas eller om det kan garanteras att exakt samma URI:er skapas om processen måste upprepas.

Undvik frågesträngar

Frågesträngar av typen ?param=value används ofta i URI:er för att skicka med uppslagsvärden till en databas. Detta sätt är dock inte stabilt då det ofta förutsätter att en specifik teknik används. Använd istället det formatet som rekommenderas och använd server-konfigurationen för att tolka URI:erna.

Undvik filändelser

Filändelser är ofta beroende av den teknik som används för att representera informationen och detta är en lösning som inte är stabil då tekniken med stor sannolikhet kommer att förändras. Låt istället serverkonfigurationen avgöra vilken typ av fil som ska returneras och på vilket sätt.

Checklista – beständiga identifierare

- Undersök om det finns en policy för identifierare vid din institution eller inom din sektor
- Kontrollera om organisationens identifierare är unika internt
- Kontrollera om organisationens identifierare är unika globalt
- Kontrollera om organisationens identifierare är beständiga internt
- Kontrollera om organisationens identifierare är beständiga globalt
- Om identifierarna endast används internt, undersök om de är utformade på ett sådant vis att de kan integreras för extern användning på ett unikt och beständigt sätt, i enlighet med rådande rekommendation
- Kontrollera om en policy för identifierare redan är implementerad i organisationens befintliga system
- Om de beständiga identifierarna uttrycks med URI:er, kontrollera att följande format tillämpas "http://{domän}/{typ}/{koncept}/{referens}"
- Kontrollera att olika representationer av samma resurs länkar till varandra
- Kontrollera att redan befintliga unika identifierare återanvänds så att nya identifierare inte skapas i onödan
- Kontrollera att det inte skapas nya identifierare för olika versioner av samma resurs
- Kontrollera att automatisk numrering vid skapandet av identifierare inte tillämpas eller endast tillämpas i de fall där processen aldrig kommer att upprepas, eller om det kan garanteras att exakt samma URI:er skapas om processen måste upprepas
- Kontrollera att filändelser inte används, utan att det är serverkonfigurationen som avgör vilken typ av fil som ska returneras och på vilket sätt
- Kontrollera att frågesträngar inte förekommer i URI:erna
- Kontrollera att URI:n svarar med http-svarskoden 303 och hänvisar vidare till ett dokument som beskriver objektet när en URI refererar till ett fysiskt objekt
- Namnet på den organisation eller det projekt som skapar en URI ska inte vara en del av URI:n
- Kontrollera att återsökbarheten fungerar, t ex genom att det vid uppdatering av länkar skapas en länk som pekar mot originalinformationen

Tjänster och system för beständiga identifierare

Det finns även ett antal tjänster som gör det möjligt att länka mellan de identifierare som används internt i institutionernas samlingssystem och de identifierare som skapas när informationen t.ex. tillgängliggörs online. Med en uppslagstjänst (resolution service) kan man skapa beständiga identifierare av exempelvis interna identifierare vid institutionen, som identifierar en digital fil, men också referera till flera filer och digitala objekt eller lagra metadata om en resurs. Som ett stöd i att identifiera behov av en tjänst/system för identifierare vid din organisation kan du använda dig av checklistan nedan.

Checklista – tjänster och system för identifierare

- Definiera behoven för en extern uppslagstjänst eller system för identifierare vid din organisation, t ex vad gäller behov kring:
 - att kunna separera namnet för resursen och platsen för resursen
 - att koppla samman metadata och data
 - automatisk versionshantering

- Undersök om tjänsten/systemet behöver användas endast internt eller också externt
- Undersök om tjänsten/systemet behöver integreras med andra system inom organisationen
- Planera för en introduktion till systemet för alla som använder identifierare
- Ta fram en plan för förvaltning och långsiktigt bevarande av systemet
- Kontrollera om systemet är tekniskt tillförlitligt och i enlighet med organisationens policy, t ex vad gäller:
 - säkerheten
 - säkerhetskopiering med redundant teknik
 - automatisk uppdatering av registret över identifierare
- Om det är ett externt system, undersök om systemet är förvaltat av en auktoritativ och trovärdig organisation som tilldelar identifierare/resolvers till resurser
- Kontrollera om systemet är flexibel och kan exponera data utan att informationen förloras, dvs att exponering sker kvalitativt
- Kontrollera om systemet är interoperabelt

Exempel på tjänster och system för beständiga identifierare

Kungliga bibliotekets resolutjänst

<https://www.kb.se/om-oss/identifikatorer-och-utgivning/urnnbn/kallfil-och-uppslagstjanst.html>

EPICHandle

http://www.pidconsortium.eu/pid_demo/

Tjänst för beständiga identifierare genom DOI (Digital Object Identifier) från Svensk Nationell Datajänst, SND

<https://snd.gu.se/en/about-us/services/pid-service>

Några fördjupningstips

"Study on persistent URIs, with identification of best practices and recommendations on the topic for the Member States and the European Commission ", Europeiska kommissionen, 2012: <https://joinup.ec.europa.eu/sites/default/files/document/2013-02/D7.1.3%20-%20Study%20on%20persistent%20URIs.pdf>

Några best practices från W3C-community:

<http://www.w3.org/TR/2015/WD-dwbp-20150224/#dataIdentification>

<http://www.w3.org/TR/ld-bp/#HTTP-URIS> (t ex "URI construction")

Rekommendationen är utarbetad av Riksarkivet i samverkan med Lantmäteriet och samråd med flera experter. Den grundar sig på tidigare "Råd om beständiga länkar" och Digisams checklista för beständiga identifierare.

Version 1.0. Licens CC-0.