URN ja muut PID-tunnukset

Juha Hakala
Kansalliskirjasto
2017-03-28

Käsitteitä

- Identifier / tunnus
  - unique string or pointer that permanently establishes the identity of a resource, institution or person; alone or in combination with other elements
- Persistent identifier / pysyvä tunnus, PID
  - identifier which may be resolved. Resolution services available depend on the identified resource, resolvers, and digital asset management systems used. Services may also change over time.
- Resolution / resoluutio
  - the act of supplying services related to the identified resource, institution or person such as translating the identifier into one or more current locators for the resource, or delivering metadata about the resource or institution or person in an appropriate format.

Käsitteitä (2)

- URN, Uniform resource name
  - Resurssin pysyvä nimi, jonka tulee olla tekniikkariippumaton ja riippumaton resurssin sijainnista
- URL, Uniform resource locator
  - Resurssin sijainti verkossa (tai verkkoarkistossa)
- URI Uniform resource identifier
  - Mikä tahansa resurssin nimi tai sijainti
  - URI:n ”isä” on Tim Berners-Lee, jonka mielestä URN-tunnusten tarvitsema resoluutio on tarpeeton lisävaihe ja että Webissä sijaintitiedot voivat olla tunnuksia, kunhan niitä ei muuteta
  - ”URLs don’t change, people change them”, ja todellakin, käytännön kokemusten ja tutkimusten perusteella URL:t ovat epäluotettavia

PID on siis…

- Resurssin, organisaation tai henkilön (tai henkilön julkisen identiteetin kuten taiteilijanimen) uniikki ja pysyvä tunnus, joka voidaan muuttaa resurssin tai jonkin siihen liittyvän objektin verkko-osoitteeksi resoluutioksi kutsutun prosessin avulla
- PID-tunnusjärjestelmiä ovat:
  - Archival Resource Key (ARK)
  - Digital Object Identifier (DOI)
  - Handle
  - Persistent URL (PURL)
  - Uniform Resource Name (URN
- URI, joka ei edellytä resoluutiota, ei ole PID, vaan esim. cool URI tai permalink

PID-järjestelmien luotettavuudesta

- PID-järjestelmien luotettavuuden kannalta keskeisiä tekijöitä on ainakin kolme: standardoinnin ja teknisen dokumentaation taso, resoluutio-ohjelmistojen saatavuus, ylläpito-organisaation luotettavuus
- URN on Internet-standardi, DOI ISO-standardi; muita järjestelmiä ei ole standardoitu
- Teknisen dokumentaation taso vaihtelee; esimerkiksi Handle on tässä suhteessa ongelmallinen
  - Handlea voi soveltaa ns. väärin, koska tekninen dokumentaatio on liian väljä
  - PID-tunnuksessa ei esim. pitäisi käyttää koko Unicode-merkkivalikoimaa vaan vain tulostettavissa olevia ASCII-merkkejä sekaannusten välttämiseksi

PID-ohjelmistot

- Handlen ja DOI:n suosion yksi perusta on avoimen lähdekoodin Handle server –ohjelmisto
- ARK:illa ja PURL-tunnuksella on myös valmis ohjelmistoalusta
- URN:lle ei ole yhteistä ohjelmistoalustaa, vaan eri tahot ovat kehittäneet omia järjestelmiään
  - Suomessa käytössä Kansalliskirjaston kehittämä resolveri; Saksassa, Sveitsissä ja Itävallassa käytetään Saksan kansalliskirjaston ohjelmistoa
- Vaikka ohjelmiston saatavuus on merkittävä tekijä, PID-järjestelmää ei kuitenkaan pitäisi valita etupäässä sen perusteella

PID-vastuut

- Koska PID-järjestelmät on tarkoitettu pysyviksi, myös niiden määrityksiä on pidettävä yllä tekniikan kehittyessä
- Jos näin ei tehdä, syntyy ongelmia
  - Handlen määritykset ovat osittain ristiriidassa URI syntax –määrityksen kanssa, mikä on johtanut yhteismitattomiin ratkaisuihin
- Ohjelmistojen vanhentuminen tai keskitetyn palvelun romahtaminen voi kriisiyttää PID-järjestelmän nopeastikin
  - PURL-järjestelmän toiminnassa oli ongelmia kuukausikaupalla vuonna 2016, kunnes järjestelmän vastuu siirtyi Internet Archivelle ja ohjelmistot päivitettiin

Resoluutiosta

- URN:t ja muut PID-tunnukset esitetään yleensä HTTP URI –muodossa siten, että resolverin osoite lisätään tunnukseen
  - http://urn.fi/URN:NBN:fi:hulib-201703131422
- Osoite ei ole osa tunnusta, ja tavoitteena on ettei sitä ennen pitkää enää tarvita
  - Global Resolver Discovery Service - palvelu, jonka avulla oikea resolveri voidaan löytää verkosta, on tarkoitus rakentaa joskus tulevaisuudessa
  - PID-tunnuksessa pitää olla vinkki jonka avulla resolverin voi löytää; yllä olevassa esimerkissä se on ”nbn:fi:”

    -> resolveri löytyy Suomesta, KK:n ylläpitämänä

Resoluutiosta (2)

- Jos halutaan tarjota eri toimijoiden ylläpitämiä palveluja, tarvitaan samalle resurssille useampia PIDejä
  - Kansalliskirjasto käyttää URN-tunnusta, kustantaja DOI:ta ja yliopiston julkaisuarkisto Handlea samalle julkaisulle
- Eri toimijoilla ei ole ollut intressiä luoda yhteisiä järjestelmiä eikä tilanne näytä jatkossa muuttuvan
  - Eri järjestelmien edustajat ovat keskustelleet teknisestä yhteismitallisuudesta, jolloin palvelujen ja niiden parametrien kutsut toteutettaisiin yhteismitallisesti -> helpompi rakentaa resolvereita, jotka tukevat kahta tai useampaa PIDiä
- Kansalliskirjaston URN-palvelut: https://www.kiwi.fi/display/URN

URN ja muut PID-järjestelmät tunnisteina

- URN rakentuu olemassa olevien tunnusjärjestelmien varaan
  - Tunnusjärjestelmälle määritellään nimiavaruus (name space) jossa noudatetaan kyseisen järjestelmän pelisääntöjä sen suhteen, mitä voidaan identifioida ja millä tavoin
    - URN:ISBN -> ISBN-standardin pelisäännöt voimassa
  - Tarjottavat resoluutiopalvelut riippuvat nimiavaruudesta sekä järjestelmistä, joita e-aineistojen hallinnoinnissa käytetään
- Muut PID-järjestelmät sallivat käyttäjän kehittää omia tunnusjärjestelmiä ja identifioida mitä ja miten tarvitaan
  - DOI = digitaalinen objektien tunnus, ei digitaalisten objektien tunnus, eli sitä voi soveltaa mihin vain
- Tunnuksen antajalla on vastuu siitä, että nimeämiskäytännöt ovat asianmukaisia ja resoluutio toimii

URN-tunnusten ja muiden PIDien käytöstä

- Handle, DOI ja URN ovat selvästi suosituimmat järjestelmät
- URN: useat kansalliskirjastot ja niiden yhteistyökumppanit sekä URN-nimialueita rekisteröineet tahot
- DOI: esim. tieteelliset kustantajat, data-arkistot
- Handle: esim. yliopistot, DSpace-julkaisuarkiston käyttäjät, useat muut tahot
- Koko Webin mittakaavassa vain murto-osalla dokumenteista on PID, useimmilla on vain URL-osoite joka ei ole pysyvä
  - Tieteellisten artikkeleiden viittauksissa olevat URL-osoitteet rapautuvat hälyttävän nopeasti
    - Link rot: aineistoa ei enää löydy
    - Content drift: linkin takaa löytyvä dokumentti on muuttunut

URN-standardit: nykytilanne

- Nykyiset RFC:t ovat vuosien takaa; osin 90-luvulta
- RFC 1737 Functional Requirements for URNs
- RFC 2141 URN Syntax
- RFC 2483 URI Resolution Services Necessary for URN Resolution
- RFC 3401 Dynamic Delegation Discovery System (DDDS). Part One: The Comprehensive DDDS
- RFC3402-3405 DDDS:n tekniset määritykset
- RFC 3406 URN Namespace Definition Mechanisms
- Vain RFC 2141 on ”oikea” Internet-standardi, muut ovat ”Informational”, ”Experimental” tai ”Best current practice”

Vanhojen URN-standardien ongelmia

- URN-syntaksin epäajantasaisuss
  - URI-syntaksin (RFC 3986) uusia piirteitä ei voi käyttää
- Palvelujen määrittelymekanismi on kömpelö
  - Yhdenkin resoluutiopalvelun lisääminen vaatii uuden RFC:n
  - RFC 2483 -palveluihin ei liity parametreja, joten esimerkiksi metadataa pyydettäessä ei voi määritellä formaattia
- URN-nimialueiden rekisteröinti on hankalaa
  - Vaatii erillisen RFC-julkaisun -> korkeahko kynnys, mutta URN-nimialueita silti yli 40
- Yksikään URN-resolveri ei tue palvelupyyntöjen välittämistä RFC-julkaisuissa 3401-3405 määritellyllä, DNS-pohjaisella kokeellisella tekniikalla

URN-standardien päivittäminen

- URN-syntaksin uudistaminen
  - Työ valmistui helmikuussa 2017, kuusi vuotta kestettyään
  - Uusi RFC julkaistaan lähitulevaisuudessa
  - URI-syntaksin fragment ja query otettu huomioon
- Nimialuiden rekisteröintikäytäntöjen uudistaminen
  - RFC2141bis kuvaa aiempaa kevyemmän menettelytavan
- Resoluutiopalvelut ja niiden parametrit määritellään jatkossa IANA:n ylläpitämässä rekisterissä URN-nimialueiden tapaan -> uusien palvelujen lisääminen helppoa
- Uuden URN-syntaksin mukaan palvelupyynnöt voidaan liittää URN-tunnukseen ja ne voidaan suunnata joko resolverille tai suoraan kohdejärjestelmään

Uusittu URN-syntaksi

- URN-tunnus rakentuu edelleen kolmesta osasta:
  - Merkkijonosta ”urn:”
  - Nimiavaruuden tunnuksesta eli NID:stä (esim. ”ISBN:”)
  - Varsinaisesta tunnuksesta eli NSS:stä (Namespace specific string), joka voi olla esim. ISBN-tunnus
- Uutta: tunnukseen voidaan lisätä f-, r- tai q-komponentti; f-komponentti vastaa URI-fragmenttia ja q-komponentti querya, mutta niillä ei ole roolia identifioinnissa
  - r-komponentti on URN-syntaksin erikoisuus; sen avulla voidaan suunnata pyyntö resolverille
- Muissa PID-järjestelmissä on vain rajalliset mahdollisuudet lisätä tunnukseen resoluutiota koskevia pyyntöjä

Esimerkkejä

- ISBN- ja NBN-nimialueiden URN-tunnukset:
  - http://urn.fi/URN:ISBN:978-951-51-2899-7
  - http://urn.fi/URN:NBN:fi:hulib-201703081380
- URN ja f-komponentti
  - http://urn.fi/URN:ISBN:978-951-51-2899-7#Luku2
  - Voi käyttää vain tietyillä nimialueilla (ei esim. ISSN:ssä) ja jos tiedostoformaatti sen sallii
- URN ja r-komponentti (Dublin Core-tietue)
  - http://urn.fi/URN:ISBN:978-951-51-2899-7?=s=URC&p=dc
  - R-komponentin syntaksia ei ole vielä sovittu, yllä on vain esimerkki siitä millaiselta se voisi näyttää kun halutaan julkaisun kuvaileva metatieto Dublin Core –muodossa

PID-resolvereista

- Nykyiset resolverit ovat ”tyhmiä”; ne kykenevät yleensä vain linkittämään yhden PID:in yhteen URL-osoitteeseen
  - Ainoa (mutta merkittävä) etu käyttäjille on PIDin ja sen avulla muodostetun linkin pysyvyys
- Resolvereista halutaan ”älykkäämpiä”, koska silloin käyttäjä (ihminen tai sovellus) voi pyytää erilaisia asioita, ja resolveri tietää minne ja miten linkitetään
- Resoluutiopalvelujen pitää kehittyä samassa tahdissa kuin e-aineistojen hallintasovellustenkin
  - Silloin e-julkaisun asemesta voidaan tarjota esim. siihen liittyvä oikeuksien hallinnan metadata tai tekninen metadata
  - Semanttisesta Webistä voidaan saada luotettavampi ja fiksumpi

URN ja DOI - vertailu

- DOI on maksullinen (mutta varsin halpa), URN maksuton
- URN-nimialueiden hallinnointi vaihtelee, DOI:ssa taso on sama riippumatta siitä kuka tunnuksen on antanut
- URN-nimialueiden palvelut ja luotettavuus voivat vaihdella, DOI:ssa periaatteessa aina samat
- DOI:ssa alkuperäistä tunnusta voidaan muokata, URN:ssä se säilyy paitsi jos URI-syntaksi siihen pakottaa
  - ISBN-A (DOI): http://doi.org/10.978.12345/99990
  - URN http://urn.fi/urn:isbn:978-1-2345-9999-0
    - Alkuperäinen ISBN: 978-1-2345-9999-0
- URN-tunnukseen voidaan lisätä resoluutiota ohjaavaa lisäinformaatiota, DOI:ssa se ei toistaiseksi ole mahdollista

URN ja Handle - vertailu

- Molemmat tunnukset ovat maksuttomia
- Sekä URN-tunnuksen nimialueiden että Handle-prefixien hallinnoinnin tasot vaihtelevat
  - Kummallakaan PID-järjestelmällä ei ole keskitettyä valvontaa
- Molemmissa järjestelmissä tunnukseen voidaan lisätä resoluutiota ohjaavaa lisäinformaatiota
  - URN-tunnuksessa lisäys noudattaa URI syntax –määritystä, Handlessa ei
- Handle-järjestelmässä tunnusjärjestelmän ja sen käyttösäännöt voi keksiä itse, URN-järjestelmässä kullakin nimialueella on omat, sen taustalla olevasta standardista tulevat pelisäännöt

Lopuksi

- URL-osoitteet ovat tutkitusti epäluotettava yleinen perusta esimerkiksi linkitetylle datalle
  - URL-osoitteita voi käyttää rajatuissa ympäristöissä (ns. siilot), jos aineistoa ei tarvitse siirtää niistä pois
- PID-tunnuksilla voi rakentaa pitkäikäisiä linkkejä, mutta linkin toimivuus on myös hallinnollinen, ei vain tekninen asia
  - Eri PID-järjestelmien valvonnan tasot vaihtelevat
  - Onko PID:istä mitään hyötyä jos itse dokumentti katoaa?
    - Mahdolliset korvaavat palvelut, kuten uudet versiot dokumentista
- PID-tunnuksia käyttävät lähinnä tahot, joilla on kokemusta julkaisutoiminnasta ja / tai pysyvän säilyttämisen velvoite
  - Pitääkö PID-käyttäjäkuntaa laajentaa, ja jos, niin miten?

Linkkejä

- ARK https://en.wikipedia.org/wiki/Archival_Resource_Key
- DOI http://www.doi.org/
- Handle http://www.handle.net/
- PURL https://archive.org/services/purl/help
- URN https://datatracker.ietf.org/wg/urnbis/about/
  http://openscience.fi/ri-pid