

4. Uniform Resource Identifier (URI)

Sisältöä päivitetty viimeksi maaliskuussa 2023.

Useimmilla verkkojulkaisuilla ei ole perinteistä tunnusta eikä myöskään PIDiä, vaan tunnukseksi käytettävä verkko-osoite eli Uniform Resource Identifier (URI). Jos URIn tunnusterooli halutaan osoittaa, voidaan käyttää nimitystä Cool URI, koska URista ei näe onko se tunnus vai vain osoite. "Cool" viittaa siihen, ettei URI koskaan muutu. Kun identifioitu objekti siirtyy toiseen osoitteeseen, alkuperäisen URIn pysyvyys taataan uudelleenohjaamalla (redirect) uuteen URI-osoitteeseen. Uudelleenohjausmahdollisuuden ansiosta on mahdollista väittää, että verkossa, toisin kuin perinteisessä kirjastossa, julkaisun sijainti hyllyssä (signum) ja julkaisun tunnus voivat olla sama asia.

Internetin teknisistä standardeista vastaava Internet Engineering Task Force erotti alun perin verkko-osoitteet ja tunnisteet toisistaan. 1994 julkaistu Functional requirements for Uniform Resource Names (RFC 1737^[m]) määritteli tunnisteille joukon toiminnallisia erityisvaatimuksia. Verkko-osoitteet kuvasi RFC 1738 Uniform Resource Locator (URL) -standardi. Verkon tunnistajärjestelmäksi luotiin RFC 2141^[n] -standardissa vuonna 1997 julkistettu URN (Uniform Resource Names). URN-tunnukset täyttivät RFC 1737:n vaatimukset olemalla pysyviä, ainutkertaisia, skaalautuvia, teknologiarippumattomia ja yhteensopivia perinteisten tunnistajärjestelmien kanssa. URL-osoitteiden (Uniform Resource Locators) ei tarvinnut täyttää mitään näistä vaatimuksista, koska ne eivät ole tunnuksia.

URI määriteltiin vuonna 2005 Internet-standardissa RFC 3986^[i]. Se yhdistää URN-tunnukset ja URL-osoitteet URI-sateenvarjon alle ja väittää, että jokainen URI on tunnus. RFC 3986 -standardin mukaan henkilön sähköpostiosoite ja puhelinnumero ovat hänen pysyviä tunnuksiaan, koska niistä voidaan muodostaa URI-tunnukset tel:<puhelinnumero> ja mailto:<email-osoite>. Erikoista tässä on se, että URI rikkoo RFC 1737 -standardin määrittelemiä tunnusteen vaatimuksia. Sähköpostiosoite voi pysyä samana pitkään - esimerkiksi tämän tekstin laatijalla 36 vuotta - mutta se ei ole pysyvä eikä ainutkertainen, koska ennen pitkää saman osoitteen voi saada joku toinen saman niminen henkilö. Sähköpostiosoite ei ole myöskään yhteensopiva perinteisten ISNI-, ORCID- ja HETU-tunnuksen kanssa.

RFC 1737 mainitaan URI-standardin lähdeluettelossa, mutta tekstissä siihen ei ole viittauksia. Sen ymmärtää, koska URI on täydellinen lehmänkäänös IETF:n alkuperäiseen linjaukseen, jossa osoitteet ja tunnisteet erotettiin toisistaan. Valitettavasti RFC 3986 hyväksyy tunnisteeksi myös paljon sellaista, joka ei sellaiseksi sovellu. Jotta URI olisi vakavasti otettava tai edes jollakin tavoin hyväksyttävä tunnistestandardi, sen olisi pitänyt rajata pois tunnuksiksi kelpaamattomat URI-tunnukset kuten puhelinnumerot. Mutta tällöin olisi jouduttu kaltevalle pinnalle, ja ennen pitkää olisi ehkä päädytty alkuperäiseen URN–URL -jaotteluun.

World Wide Web Consortium (W3C) on antanut RFC 3986:n julkaisemisen jälkeen surkuhupaisia ohjeita siitä, miten tunnukseksi käytettävät URIt tulisi muodostaa. Vuonna 2007 julkistetun Cool URIs for the Semantic Web -ohjeen^[v] mukaan niiden pitäisi olla lyhyitä, helposti muistettavia ja helposti hallinnoitavia. Ohje paljastaa W3C:n käsityksen URI-tunnusten elinkaaresta:

Once you set up a URI to identify a certain resource, it should remain this way as long as possible. Think about the next ten years. Maybe twenty.

10–20 vuotta on lyhyt ikä pysyvälle tunnisteelle, mutta pitkä URI:lle. W3C-ohjeen luvussa viisi on kahdeksan Cool URI-esimerkkiä, joista seitsemän tuotti jo 2018 HTTP 404 -virheen. Niiden elinkaari oli siis lyhyempi kuin yllä ohjenuoraksi asetettu 10–20 vuotta. Tätä kirjoitettaessa ainoa edelleen toimiva URI antaa tulokseksi identifioitun henkilön tietojen sijaan vain hänen tuolloisen työnantajaorganisaationsa julkaisuluettelon, joten sekään ei enää toimi oikein.

Kahdeksan URIn perusteella ei voi tehdä yleisiä päätelmiä URI-tunnusten elinajanodotteesta. Mutta Herbert Van de Sompel ja hänen tutkimusryhmänsä ovat analysoineet tieteellisten julkaisujen viitteiden avulla Cool URIn pitkäikäisyyttä, ja osoittaneet että sekä sisältöryjähdys eli viitattun dokumentin muuttuminen^[v] että linkkimätä eli viitattujen dokumenttien katoaminen^[vi] ovat yleisiä ongelmia. Mitä vanhempi linkki, sitä todennäköisempää on, ettei se enää toimi.

Verkon käyttäjien kannalta Cool URI -tunnusten suurin puute on se, että ne eivät ole "cool". Kirjastojen kannalta RFC 3986 -standardin merkittävien käytännön ongelmia on se, että niitä käytetään pysyvinä tunnuksina. Mitä enemmän aikaa kuluu, sitä hankalampaa niitä on pitää toiminnallisena. Jokin ennalta arvaamaton muutos, kuten kehysorganisaation domain-nimen muuttuminen tai kirjastojärjestelmän vaihto, voi muuttaa kaikki URL-osoitteet niin, ettei uudelleenohjaus vanhasta osoitteesta uuteen ole teknisesti mahdollista.

Merkittävin periaatteellinen haaste on, että Cool URIt eivät ota perinteisiä tunnistajärjestelmiä huomioon (niiden käyttöä verkkoaineistojen identifiomisessa ei edes mainita RFC 3986:ssa). Verkossa julkaistun kirjan ISBN saa, kiitos RFC 3986:n, kirjan verkko-osoitteesta kilpailevan tunnuksen. Ja niitä on yhtä monta kuin kirjasta on verkossa kopioita. Kirjastonhoitaja, joka tallentaa e-kirjan kirjastonsa julkaisuarkistoon, tuskin edes ajattelee antaneensa kirjalle samalla uuden tunnuksen, jonka tulisi olla ISBN:n tavoin pysyvä ja ainutkertainen. Mutta lukija, joka viittaa kirjaan, voi käyttää viitteessä ISBN:n tai DOI-tunnuksen sijasta URL-osoitetta. Van de Sompel kumppaneineen löysi 600.000 tapausta, joissa julkaisulla oli DOI, mutta viitteessä käytettiin DOI-tunnuksen asemesta URL-osoitetta. Tämä hankaloittaa sen selvittämistä, mitä julkaisua tai julkaisun versiota on käytetty lähteenä varsinkin jos URL on lakannut toimimasta linkkimädän tai sisältöryjähdysvuoksi.

Keskeisin kaikkia luotettavia tunnistajärjestelmiä yhdistävä kriteeri on hallinnointi. Tunnuksien jakelua pitää rajoittaa, ja kun tunnus on annettu, sen toiminnallisuudesta on pidettävä huolta. Jos kuka tahansa saa antaa tunnuksia mille tahansa ilman että edes tietää mitä tekee, syntyy kaaos. Tutkijan ORCID-tunnuksen korvaaminen hänen sähköpostiosoitteellaan riittää tutkijan tavoittamiseen ehkä ainakin niin kauan kuin hän työskentelee samassa organisaatiossa. Mutta ORCIDin tai ISNI:n kaltaisen pysyvän ja toiminnallisen tunnuksen vaatimuksia sähköpostiosoite ei täytä.

Tunnusten ja niiden toiminnallisuuden pysyvyys riippuu van osittain tunnistajärjestelmästä ja sen teknisestä toteutuksesta. Myös tunnuksia jakavan ja/tai niiden toiminnallisuudesta vastuun kantavan organisaation tulee olla pitkäikäinen ja ihannetapauksessa sillä on pysyvä, sopimukseen tai lakiin perustuva vastuu säilyttää identifioitut julkaisut. Tällöin pysyvien tunnistajärjestelmien toiminnallisuuden takuumiehiksi tulevat julkaisijoiden rinnalle kansalliskirjastojen ja kansallisarkistojen kaltaiset vakaat organisaatiot, joilla on vankea kokemus myös perinteisistä tunnistajärjestelmistä. Kustantajan antamien tunnusteiden pysyvyys ei riipu kustantajan elinkaaresta, jos kustantajalla on yhteistyökumppanina kansalliskirjaston kaltaisen organisaatio. Tällä tavoin on toimittu jo pitkään esimerkiksi kirjojen osalta: suomalaiset kustantajat antavat ISBN-tunnukset julkaisuilleen itse, mutta kirjojen pysyvistä säilyttämisestä ja tarjoamisesta asiakkaille vastaavat Kansalliskirjasto ja muut vapaakappalekirjastot. Jos korkeakoulu antaa väitöskirjalle Cool URI -tunnukset, niiden toiminnallisuutta on vaikea taata silloin, kun väitöskirjan viimeinen säilynyt kopio on Kansalliskirjastossa. Mutta jos väitöskirjan tunnus on esimerkiksi URN, sen toiminnallisuus ei riipu siitä, mikä organisaatio vastaa julkaisun säilytyksestä.

Merkittävä pitkän aikavälin ongelma Cool URI -tunnusten pysyvyyden kannalta on se, että Internet-domaineja ei voi omistaa. Niitä voi ainoastaan vuokrata muutamaksi vuodeksi kerrallaan. Kun domainin vuokra-aika päättyy, sen hallintaoikeus voi siirtyä toiselle toimijalle. Tällöin domainin Cool URI-tunnukset joko lakkaavat toimimasta kokonaan tai linkittyvät eri dokumentteihin.

Verkkoarkistot ratkaisevat tämän ongelman vain osittain. Niiden URI-linkit vievät aina samaan dokumenttiin, mutta vain niin kauan, kun kyseinen arkisto on tuotannossa. Internet Archive täyttää pian 27 vuotta, mutta toimiiko se vielä 270 vuoden päästä? Jokseenkin varmoja voimme olla vain siitä, että Kansalliskirjaston Kulttuuriaineistolain perusteella kokoama verkkoarkisto tulee säilymään käytössä hyvin pitkään.

Valitettavasti verkkoarkistosta löytyvä sivu aina vastaa alkuperäistä. Jos sivu muodostuu useista tiedostoista, ne on voitu haravoida eli noutaa arkistoon WWW-palvelimelta tai -palvelimilta eri aikoina. Tällöin arkiston tarjoama verkkosivu voi olla yhteensopimattomista palasista koottu tilkkutäkki, kuten sääennuste, jonka sääkartta ja teksti kuvaavat eri päiviä.

Keskustelu Cool URI -tunnusten ja PID-tunnusten keskinäisestä paremmuudesta on vellone jo vuodesta 1998, jolloin Tim Berners-Lee esitteli idean artikkelissaan Cool URIs don't change. Internetistä löytyy toki esimerkkejä siitä, että Cool URI -tunnuksia voidaan käyttää menestyksekkäästi rajatuissa kohteissa ainakin jonkin aikaa. Esimerkiksi paikkatietojen tunnisteena sitä on käytetty jo vuosia, paikkatiedon yhteiskäyttöä tukevan INSPIRE-direktiivin^[vii] nojalla ja määräyksestä. Huolellisesti suunnitellen ja tunnuksia hallinnoiden nämä Cool URI -tunnukset saadaan elämään vielä tästä eteenkinpäin vuosikymmeniä. Mutta Cool URIt ovat teknologiarippuvaisia ja toimivat enintään siihen asti, kunnes verkon perusteknologiat kuten HTTP-protokolla muuttuvat. Tähän voi mennä kymmeniä vuosia tai pidempäänkin, ehkä jopa satoja vuosia. Mutta riittävän pitkällä aikavälillä mitään varmuutta edes HTTP:n kaltaisen perusteknologian säilymisestä ylöspäin ei ole. Jos voimme valita teknologiarippuvaisen ja teknologiarippumattoman tunnistejärjestelmän välillä, miksi ottaa riski ja valita edellinen?

Joissakin järjestelmissä URI-tunnuksien sijaan on lupa tai suoranainen vaatimus käyttää Internationalized Resource Identifier – eli IRI-tunnuksia. IRI-määrittäjä RFC 3987^[viii] julkaistiin heti RFC 3986:n jälkeen, ja se laajentaa URI-tunnuksissa käytettävissä olevan suppean merkkivalikoiman Unicodeksi. Jokainen URI on siis myös IRI, mutta jos tunnuksessa käytetään RFC 3986:ssa kiellettyjä merkkejä, se on IRI.

IETF yritti päivittää RFC 3987:n ajan tasalle 2010-luvun alussa, mutta yritys epäonnistui, koska IRI-tunnuksiin liittyviä teknisiä ongelmia ei pystytty ratkaisemaan. Sen vuoksi pysyvyissä tunnisteissa pitäisi käyttää vain URI-tunnuksissa hyväksyttyjä merkkejä, vaikka IRI-tunnusten käyttö olisi periaatteessa mahdollista. Tunnuksien merkkivalikoiman rankka rajoittaminen on järkevää myös käyttäjien kannalta. Esimerkiksi emoji-ja sisältävä IRI olisi tunnuksena hankala, koska emoji-symbolien ulkonäkö vaihtelee käyttöjärjestelmästä riippuen. Vielä tärkeämpi peruste merkkivalikoiman rajoituksille on "Unicode confusables", eli merkit, joiden ulkoasu on sama tai lähes sama. Jos esimerkiksi Handle-tunnuksen takaliitteessä oleva latinalainen c korvataan kyrillisellä c-kirjaimella, tunnus näyttää käyttäjälle samalta kuin ennen, mutta muokattu tunnus resoluutuu eri dokumenttiin kuin alkuperäinen^[ix]. Verkkosoitteissa näitä huijauksia torjutaan siten, että selaimet eivät hyväksy verkko-osoitteita joissa sekoitetaan eri kirjaimistoja. Handle.Net-resolveriin on teknisesti mahdollista rakentaa vastaava ominaisuus, mutta emme tiedä onko sitä tehty. Handle-standardi ei aseta mitään rajoituksia eri merkistöjen rinnakkaiskäytölle samassa tunnuksessa.

Cool URI -tunnukset eivät siis täytä kirjastojen tunnisteille asettamia pysyvyyden, ainutkertaisuuden ja hallinnoitavuuden vaatimuksia. Siksi niitä käsitellään tässä ohjeessa jatkossa vain kursorisesti.

^[i] <https://tools.ietf.org/html/rfc3986>

^[ii] <https://tools.ietf.org/html/rfc2141>

^[iii] <https://tools.ietf.org/html/rfc1737>

^[iv] <http://www.w3.org/2001/sw/sweo/public/2007/cooluris/>

^[v] <https://doi.org/10.1371/journal.pone.0167475>

^[vi] <https://doi.org/10.1371/journal.pone.0115253>

^[vii] INSPIRE mahdollistaa paikkatietojen yhteentoimivuuden yli organisaatioiden ja valtioiden rajojen. Esimerkiksi ilmanlaatuun ja vesistöihin liittyvät ympäristövaikutukset eivät tunne rajoja, joten tietojen tulee olla yhteentoimivia, laadukkaita ja helposti saatavilla, jotta päätöksiä voidaan perustaa niiden varaan. Kyseessä on EU-direktiivi, jonka tavoitteena on Euroopan yhteisön paikkatietoinfrastruktuuri. INSPIRE tulee sanoista Infrastructure for Spatial Information in the European Community. (Katsottu 4.1.2021, <https://www.maanmittauslaitos.fi/kartat-ja-paikkatieto/paikkatietojen-yhteentoimivuuus/inspire/mika-inspire>)

^[viii] <https://tools.ietf.org/html/rfc3987>

^[ix] <http://websec.github.io/unicode-security-guide/visual-spoofing/>