KANSALLISKIRJASTO
NATIONALBIBLIOTEKET

Julkaisuarkistot – pysyvän saatavuuden infrastruktuuri?

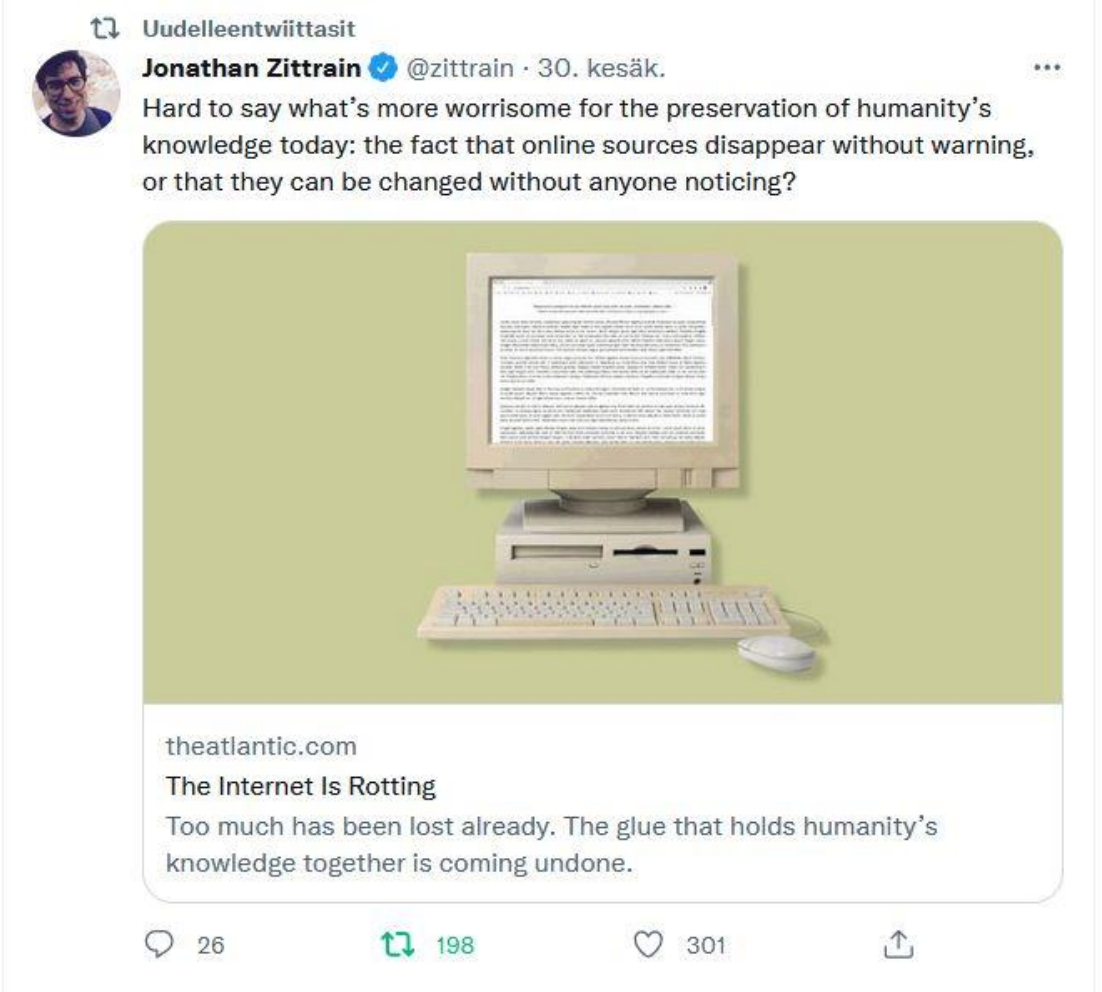
Jyrki Ilva (jyrki.ilva@helsinki.fi)

Missä mennään -webinaari, 16.11.2021

Pysyvä osoite: https://urn.fi/URN:NBN:fi-fe2021111655567

Verkkoaineistojen pysyvyys usein häilyvää

- Verkkoaineistojen pysyvyydessä monenlaisia haasteita
- Aineistoja katoaa saatavilta tai niiden sisältö saattaa muuttua
- Ongelma koskettaa myös tieteellisiä julkaisuja ja muita julkaisuarkistoihin tallennettavia aineistoja



Jonathan Zittrain: The Internet Is Rotting. Atlantic, June 30, 2021.

Webinaari julkaisuarkistoista ja pitkäaikaissäilytyksestä

- Kansalliskirjasto järjestää webinaarin julkaisuarkistoaineistojen pitkäaikaissäilytyksestä 9.12. klo 9-11
  - Ilmoittautuminen auki 7.12. asti
- Tässä esityksessä keskitytään ensisijaisesti aineistojen pysyvään saatavuuteen

Webinaari: Julkaisuarkistojen aineistot pitkäaikaissäilytykseen?

Created by Jyrki Ilva, last modified on Nov 12, 2021

Aika: **Torstai 9.12.2021 klo 9-11**

Paikka: Zoom-etäkokous, osallistumisohjeet

Ilmoittautuminen: viimeistään tiistaina 7.12.

Julkaisuarkistoihin tallennetaan vuosittain kymmeniä tuhansia julkaisuja, sekä opinnäytteitä, rinnakkaistallenteita, sarjajulkaisuja että muitakin aineistoja. Kansalliskirjasto kerää julkaisuarkistoissa julkaistuja aineistoja talteen osana kulttuuriaineistolain mukaista toimintaansa. CSC ylläpitää kansallista pitkäaikaissäilytyspalvelua. Millä tavalla ja missä laajuudessa julkaisuarkistojen aineistot päätyvät pitkäaikaissäilytykseen?

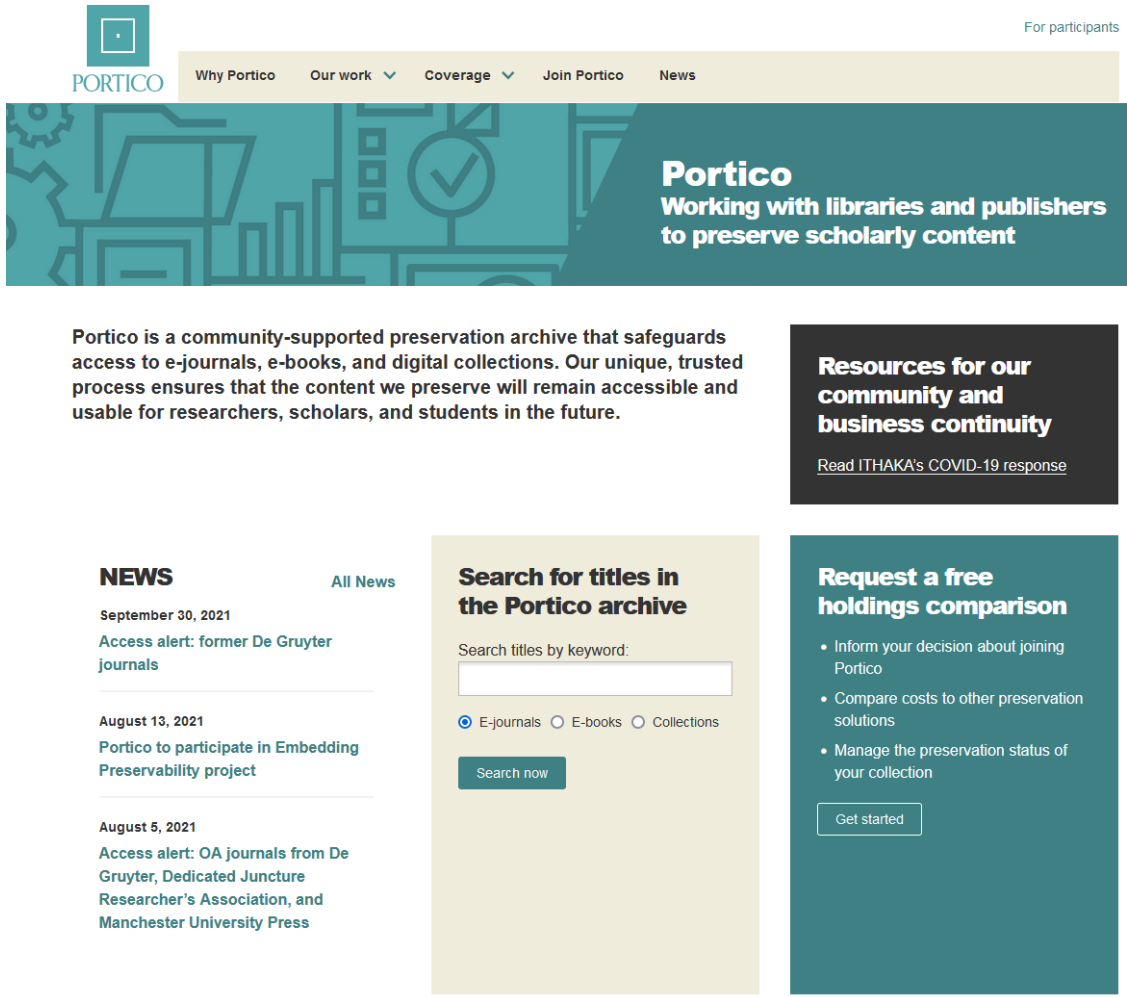
Kansalliskirjaston järjestämässä seminaarissa keskustellaan julkaisuarkistoaineistojen pitkäaikaissäilytykseen (PAS) liittyvistä kysymyksistä. Tilaisuus on suunnattu sekä julkaisuarkistojen parissa työskenteleville että niiden toiminnasta vastaaville henkilöille.

Tilaisuus on ilmainen, ja siihen voi **ilmoittautua viimeistään tiistaina 7.12.** tällä verkkolomakkeella. Etäkokouksen osoite lähetetään ilmoittautuneille tilaisuutta edeltävänä päivänä.

Tervetuloa mukaan!

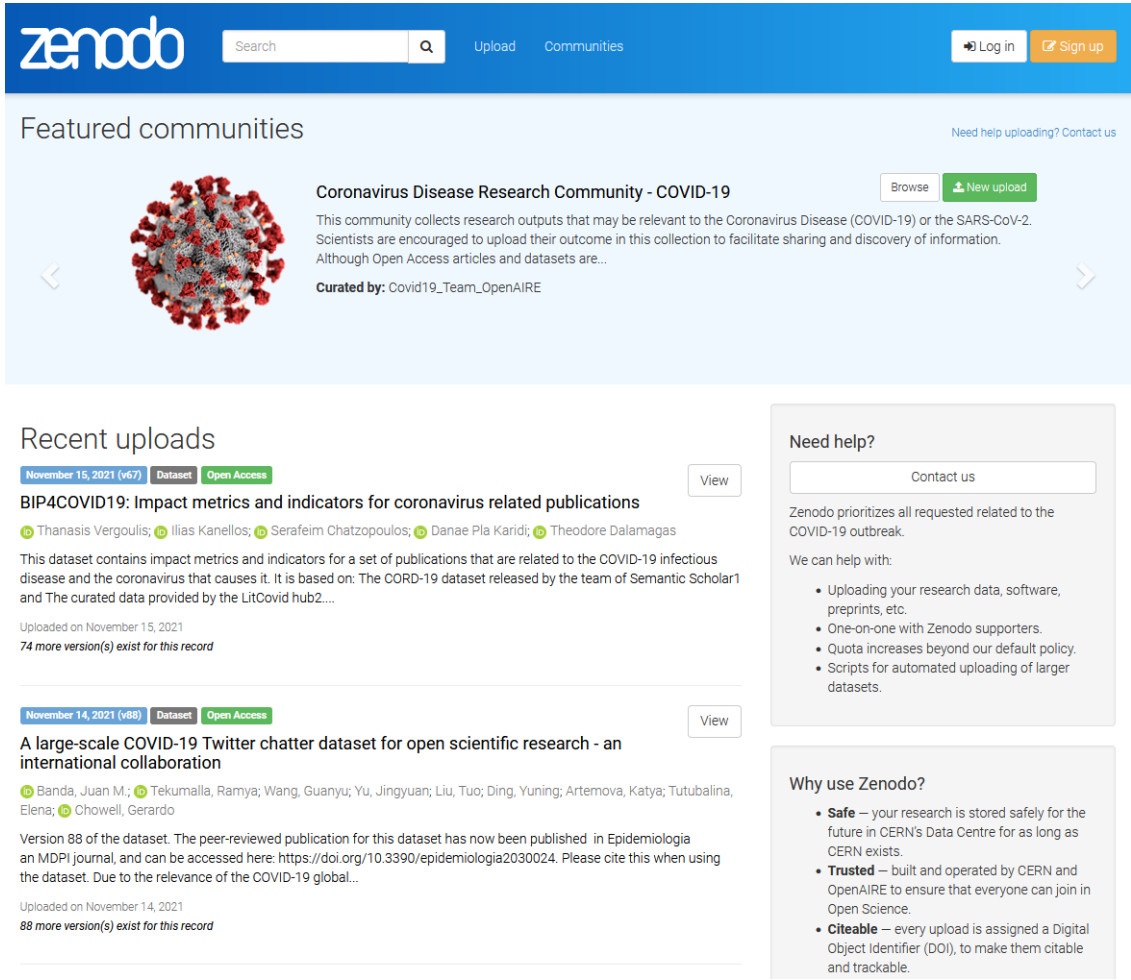
Pysyvä saatavuus ja pitkäaikaissäilytys

- Pysyvä saatavuus ja pitkäaikaissäilytys kytköksissä toisiinsa, mutta kuitenkin eri asioita
  - Pitkäaikaissäilytys ei sinällään tarkoita sitä, että säilytetyt aineistot olisivat laajasti saatavilla
- Pysyvä saatavuus ei sekään ole aivan uusi idea
  - Käytännön esimerkkinä tieteellisten kustantajien käyttämä Portico-palvelu
  - Pysyvyyden pituus ja saatavuuden laajuus enemmän tai vähemmän suhteellisia asioita



Pysyvä saatavuus ja julkaisuarkistot

- Maailmalla julkaisuarkistoilla, pitkäaikaissäilytyksellä ja pysyvällä saatavuudella on usein läheinen yhteys
  - PAS-järjestelmiä saatetaan rakentaa myös organisaatiotasolla
- CERN:in ja OpenAIRE:n Zenodo-julkaisuarkistoon viitataan usein pitkäaikaissäilytyksen ja pysyvän saatavuuden takaavana ratkaisuna
  - ”Tutkimuksesi on tallessa niin pitkään kuin CERN on olemassa”



Entä Suomessa?

- OKM:n rahoittamat ja CSC:n rakentamat kansalliset pitkäaikaissäilytyspalvelut, Kulttuuriperintö-PAS ja Fairdata-PAS
- Pysyvästä saatavuudesta on puhuttu aika ajoin kunnianhimoisesti pitkäaikaissäilytyksen päälle rakennettavana lisäpalveluna
  - Joskus tulevaisuudessa, kenties – mutta miten eletään siihen asti?
- Missä määrin julkaisuarkistot vastaavat – tai voisivat vastata – aineistojen pysyvää saatavuutta koskeviin tarpeisiin?

Organisaatiot ja vastuu

- Pysyvä saatavuus ei toteudu itsestään, vaan vaatii työtä ja riittäviä resursointia
- Edellyttää käytännössä sitä, että taustalla on jokin vakaa ja pysyvä organisaatio
  - Organisaatiolla pitää olla valmiudet ottaa vastuu aineistoista ja palveluista myös pitkällä tähtäimellä
  - Organisaation pitää olla valmis sitoutumaan siihen, että se pitää huolta aineistoista
  - Jos tämä ei ole jossain vaiheessa mahdollista, vastuu pitää välittää eteenpäin jollekin muulle taholle
  - Käytännössä fuusiot, organisaatiomuutokset, rahoitustilanteen vaihtelu ja uudet strategisen tason priorisoinnit saattavat aiheuttaa haasteita vakainakin pidetyille organisaatioille

Palvelut ja tekninen infrastruktuuri

- Tällä hetkellä kaikilla suomalaisilla korkeakouluilla ja monilla tutkimuslaitoksilla on käytössään julkaisuarkisto
  - Palvelukonseptit saattavat ajan myötä elää – esim. julkaisuarkistojen ja tutkimustietojärjestelmien välinen työnjako selkiytynyt vähitellen
  - Miten pitkäikäinen konsepti julkaisuarkisto on?
- Palveluita voidaan tuottaa joko organisaation omin voimin tai yhteistyössä
  - Suurin osa suomalaisista organisaatioista käyttää Kansalliskirjaston tarjoamaa keskitettyä palvelua
  - Pysyvän saatavuuden takaamiseksi teknisen infrastruktuurin ylläpitoon ja kehittämiseen pitää olla riittävät resurssit

Ohjelmistojen elinkaari

- Avoimen lähdekoodin julkaisuarkisto-ohjelmistot
  - Ohjelmistot vaativat ylläpitoa ja kehitystyötä
  - Koodi ei synny itsestään vaan kehitystyöhön pitää löytää resursseja tavalla tai toisella
- Suomessa käytössä DSpace
  - Julkaistiin alun perin vuonna 2002
  - Käytössä tuhansissa organisaatioissa eri puolilla maailmaa
  - Kesällä 2021 julkaistu versio 7 ajanmukaisti ohjelmiston arkkitehtuuria ja teknisiä ratkaisuja
- Kansainvälisen DSpace-yhteisön toimintaa rahoitetaan nykyään pääosin jäsenmaksuilla
  - Kansalliskirjasto on DSpace-yhteisön jäsen
  - DSpace nyt myös SCOSS:in rahoituskohteena

Pysyvät osoitteet

- Pysyviin tunnisteisiin perustuvat verkko-osoitteet yksi keino helpottaa organisaatioiden, palveluiden ja ohjelmistoversioiden muutoksista aiheutuvia ongelmia
  - Vaikka julkaisun verkko-osoite muuttuu, se on edelleen löydettävissä pysyvän osoitteen avulla
- Suomalaisissa julkaisuarkistoissa käytetään yleisesti URN:eja, joissakin käytössä myös Handle ja DOI
  - Järjestelmästä riippumatta pysyviä osoitteita käyttävät organisaatiot sitoutuvat ylläpitämään tietoa siitä, missä osoitteessa tunnistetta vastaava dokumentti sijaitsee
  - Tiedot toimitetaan resolverille, joka ohjaa käyttäjät oikeaan osoitteeseen
  - DOI:n etuna kansainvälisten palveluntarjoajien (Crossref, DataCite) rakentama tieteellisille aineistoille optimoitu infrastruktuuri
  - Handle- ja DOI-pohjaisten osoitteiden uudelleenohjaus saattaa olla organisaatiomuutoksissa jonkin verran mutkikkaampaa kuin URN:in

Julkaisuoikeudet

- Julkaisun pysyvän saatavuuden kannalta on eduksi, jos siihen liittyvistä oikeuksista on sovittu ja ne on määritelty selkeästi
  - Creative Commons -lisenssit yleinen keino jatkokäyttöoikeuksien ilmaisemiseen
  - Esim. CC BY -lisenssi mahdollistaa aineiston tallentamisen ja julkaisemisen muissa palveluissa, ml. julkaisuarkistot
- Muussa tapauksessa julkaisun avoin saatavuus saattaa rajoittua palveluun johon se on alun perin tallennettu
  - Aiheuttaisi ongelmia myös mahdolliselle kansalliselle pysyvän saatavuuden palvelulle
- Julkaisuarkistoissa CC-lisensoitua aineistoa edelleen suhteellisen vähän
  - Rinnakkaistallenteiden osalta kustantajan kanssa solmitut sopimukset rajoittaneet; tallennus perustunut usein kustantajan politiikkaan, ei lisenssiin
  - Opinnäytteissä lisenssin käyttö kiinni tekijän päätöksestä

Kuvailutiedot ja löydettävyys

- Julkaisuarkistoaineistoilla usein hyvä näkyvyys Googlen ja muiden hakukoneiden tuloksissa
  - DSpace-kehittäjät tehneet yhteistyötä Google Scholarin kanssa indeksoitumisen parantamiseksi
  - Julkaisuihin kohdistuvat linkit ja viittaukset parantavat näkyvyyttä
  - Pysyvien osoitteiden käyttö linkityksissä suositeltavaa, mutta saattaa vaikuttaa hakukoneiden tuloksiin
- Myös näkyvyydellä kirjastojen hakukäyttöliittymissä merkitystä
  - Julkaisuarkistojen Dublin Core -muotoista metadataa haravoidaan jatkuvasti mm. Finnan indeksiin
  - Metadatan jatkokäyttö on joitakin poikkeuksia lukuunottamatta vapaata (CC0)
  - Julkaisuarkistojen metadataa saatetaan kirjastoissa konvertoida myös MARC-formaattiin, missä on omat haasteensa
  - Kansalliskirjaston metadataprosessi ja pitkäaikaissäilytys?

Tallennusformaatit

- Tallennusformaattien elinkaari yksi mahdollinen pullonkaula pysyvän saatavuuden toteutumisessa
  - Pysyvätkö dokumentit luettavina teknisen ympäristön muutoksista huolimatta?
  - Pitkäaikaissäilytysjärjestelmissä tehdään tarvittaessa migraatioita formaatista toiseen; julkaisuarkistoissa ei tällaiseen pääsääntöisesti mahdollisuuksia
- Julkaisuarkistojen aineistot pääosin suhteellisen yksinkertaisia PDF-formaattiin tallennettuja tekstimuotoisia dokumentteja
  - Yleisesti käytetty ja suhteellisen hyvin dokumentoitu formaatti
  - Kasvava osa aineistoista tallennettu jo valmiiksi PDF/A-formaatissa
  - XML-pohjaiset tallennusformaatit (esim. tieteellisten kustantajien käyttämä JATS-XML) eivät ole toistaiseksi yleistyneet prosesseihin ja kustannuksiin liittyvistä syistä johtuen
  - Audiovisuaalisten aineistojen kanssa saattaa olla enemmän ongelmia

Sisällölliset muutokset ja versiointi

- Käytännön elämässä aineistoja saatetaan joutua korjaamaan ja päivittämään
  - Julkaisu on saatettu tallentaa vääränä versiona tai väärässä muodossa
  - Julkaisussa saattaa olla jokin korjaamista vaativa virhe
  - Tieteellisestä artikkelista saatetaan tallentaa ensin tekijän oma versio ja sitten kustantajan versio
- Päivitysprosessien käytännöissä vielä parannettavaa
  - DSpace-julkaisuarkistoissa mahdollisuus tallentaa julkaisuille perustason provenienssitietoja, joihin rekisteröidään myös tiedostojen tarkistussummat
  - Julkaisujen versiointi parantaisi tilannetta mm. viittaamisen kannalta
  - Esim. Zenodossa tähän jo valmiimpia välineitä
  - Kysymyksiä: Miten käyttäjät ohjataan julkaisun uusimpaan versioon? Miten pysyvien tunnisteiden kanssa toimitaan, kun julkaisusta on erilaisia versioita?

Julkaisujen poistaminen saatavilta

- Yksittäisiä aineistoja saatetaan kaikesta huolimatta välillä joutua poistamaan saatavilla
  - Esim. opinnäytteen tekijä saattaa peruuttaa julkaisuluvan
  - Julkaisu saattaa sisältää vanhentunutta tietoa, josta voi olla haittaa
- Poistamiseen liittyvissä käytännöissä vielä monin paikoin kehittämistä
  - Poistetulle julkaisulle on suositeltavaa jättää julkaisuarkistoon ”hautakivisivu”
  - Poistoprosessi on hyvä dokumentoida ja samalla on syytä pitää huolta, että poistettu julkaisu on arkistoitu

Julkaisuarkistot ja pysyvä saatavuus?

- Julkaisuarkistot takaavat kohtuullisen hyvin aineistojen saatavuuden lyhyellä ja keskipitkällä aikavälillä
  - Pitkän aikavälin pysyvyyden osalta ennusteiden tekeminen on vaikeampaa
- Mitä jos pysyvyys ei toteudukaan?
  - Julkaisun tallentaminen moneen eri paikkaan parantaa sen säilymismahdollisuuksia, vaikka jokin yksittäinen palvelu poistuisi käytöstä
  - Vaikka aineisto ei enää olisi saatavilla julkaisuarkistossa, se saattaa löytyä verkkoarkistoista (mm. Internet Archive ja Suomalainen verkkoarkisto) tai vapaakappaleena Kansalliskirjastosta ja muista vapaakappalekirjastoista

KANSALLISKIRJASTO
NATIONALBIBLIOTEKET

www.kansalliskirjasto.fi