

Finna ja OAI-PMH

Johdanto

Aineistotiedot haravoidaan Finnaan käyttäen OAI-PMH -protokollaa (versio 2.0). Finna toimii siis OAI-PMH -harvesterina, ja aineistotiedot haravoidaan osallistuvan organisaation OAI-PMH -providerista. OAI-PMH on yksinkertainen HTTP-pohjainen protokolla, joka ei ota kantaa esim. haravoitavan metadatan formaattiin. Ideana OAI-PMH:ssa on se, että ensimmäisellä kerralla haravoidaan kaikki tietueet repositorystä, ja jatkossa vain edellisen haravoinnin jälkeen tapahtuneet lisäykset, muutokset ja poistot. Tämä ns. inkrementaalinen haravointi tapahtuu tietueiden muutosaikaleiman perusteella. Finna siis pitää kirjaa siitä, milloin se on viimeksi haravoinut muutokset, ja pyytää providerilta seuraavalla kerralla vain muuttuneet tiedot.

Jotta kerralla käsiteltävä tietomäärä ei kasvaisi liian suureksi, käytetään OAI-PMH:ssa sivutusta. Jos yhdellä haravointipyynnöllä tulee providerin määrittämää raja-arvoa enemmän tietueita, palautetaan niitä vain tuon raja-arvon verran ja lisätään pyynnön loppuun ns. resumptionToken, jota käyttämällä haravoija voi pyytää seuraavan tietuejoukon. Sopiva raja-arvo riippuu providerin rakenteesta ja tietueiden koosta, mutta vaihtelee tyypillisesti 100 - 5000 välillä. Usein jokainen haravointipyynnö edellyttää providerilta jotain esityötä, jota voidaan vähentää kasvattamalla kerralla palautettavien tietueiden määrää.

Määrittelyt

Openarchives.org:n sivuilta löytyy [OAI-PMH -protokollan määrittely](#). Siellä on myös [Implementation Guidelines -dokumentti](#), ja erityisesti sen [Guidelines for Repository Implementers -osa](#). Näistä löytyvät vastaukset useimpiin kysymyksiin.

Finna käyttää erityisesti seuraavia toimintoja: Identify, ListRecords ja ListIdentifiers. Suosittelemme kuitenkin toteuttamaan pakolliset toiminnot niin, että ne menevät [validaattorista](#) läpi virheett. Esim. ongelmatilanteissa voi olla tarpeen tutkia yksittäiselle tietueelle palautettavia metatietoja GetRecord-toiminnolla.

Jos repositoryn sisältö jakautuu loogisiin kokonaisuuksiin, joita halutaan mahdollisesti käsitellä erillään, voidaan OAI-PMH:ssa käyttää settejä. Tietue voi kuulua useampaankin settiin. Settien tukeminen ei ole välttämätöntä.

Finna käyttää haravointiin [RecordManageria](#), PHP:lla toteutettua avoimen lähdekoodin ohjelmistoa, jota voi vapaasti käyttää vaikka oman repositoryn testaamiseen. RecordManagerissa on myös [OAI-PMH -provider](#), joten sitä voi käyttää tarvittaessa oman repositoryn toteuttamisessa. Lisätietoa löytyy [RecordManagerin wiki-sivuilta](#). RecordManager tukee tällä hetkellä seuraavia metadataformaatteja:

- Dublin Core
- Qualified Dublin Core
- EAD
- ESE
- FORWARD
- LIDO
- MARCXML (MARC21)

OAI-PMH -spesifikaatio edellyttää, että provider pystyy tuottamaan tietueita Dublin Core -formaattissa. Finnassa tätä ei kuitenkaan edellytetä, jos Finnassa käytettävä formaatti on jokin muu. Yhteentoimivuuden näkökulmasta DC:n tukeminen on kuitenkin suositeltavaa. Käyttöä OAI-PMH:lle voi olla muissakin tilanteissa kuin Finnan kanssa.

OAI-PMH -spesifikaatio ei määrittele pääsynvalvontaa. Tyypillisesti OAI-PMH -provider päästää läpi vain tietyistä IP-osoitteista tulevat pyynnöt.

Reunaehdot

OAI-PMH -protokolla ja Finna asettavat repositorylle joitakin vaatimuksia:

- Jokaisella tietueella on oltava pysyvä yksilöivä tunniste.
- Tietueilla on oltava muutosaikaleima, joka päivittyy jokaisen muutoksen yhteydessä. Suositellaan sekunnin tarkkuutta.
- OAI-PMH -providerin ei tule palauttaa ei-julkista tietoa, vaan se pitää suodattaa pois tietueista.
- Tieto poistetuista tietueista tulee säilyttää, jotta poistettujen tietueiden tunnisteet voidaan raportoida haravojalle. (*
- resumptionTokenia on tuettava. Riippuu tietueiden koosta ja repositoryn toteutustavasta, mikä on järkevä määrä tietueita kerralla palautettavaksi. Tyypillisesti 100 - 5000.
- resumptionToken on haravojan näkökulmasta "läpinäkymätön" merkkijono, eli siihen voi vaikka upottaa kaikki ensimmäisen pyynnön parametrit ja kursorin sijainnin, tai siinä voi käyttää tunnistetta, jolla tiedot haetaan muualta, esim. tietokannasta.

(* Poistettujen tietueiden tuesta voidaan joustaa siinä tapauksessa, että repositoryn tietuemäärä on kohtuullisen pieni, ja repository pystyy vastaamaan ListIdentifiers-kutsuun nopeasti. Tällöin voidaan Finnan puolella ajaa esim. viikoittain ListIdentifiers-kutsut ja poistaa tietueet, joita ei enää näkynyt vastauksissa.

Tyypillisiä kompastuskiviä

- Tietueen metadatan tulee OAI-PMH:n <metadata> -elementin sisään. Tietueella pitää kuitenkin olla oma juurielementti, eli tietue pitää pystyä irrottamaan ehjänä OAI-PMH -kehyksestä niin, että se säilyy kelvollisena XML:nä.
- Kaikkien aikaleimojen tulee noudattaa standardia, eli muoto on 2016-03-21T15:46:31Z ja aikavyöhyke UTC.
- Poistetut tietueet palautetaan muuten samalla tavalla kuin normaalit tietueet, mutta niille ei palauteta metadataa, ja <header> -elementissä on attribuutti status="deleted":

```
<header status="deleted">
  <identifier>oai:unique.repository.id:123</identifier>
  <datestamp>2019-03-01T12:00:00Z</datestamp>
</header>
```

- identifier:n tulee olla maailmanlaajuisesti yksilöllinen. Tietueen tunnisteeseen etuliitteenä voi käyttää esim. palvelun domain-nimeä.
- resumptionTokenia ei saa palauttaa, jos ei sillä löydy mitään. Ts. kun viimeinen tietue on palautettu, ei vastauksessa saa enää olla resumptionTokenia. Providerin puolella tyypillisesti helpoin tapa tarkistaa tilanne on noutaa tietueita yksi enemmän kuin kerralla lähetetään ja lähettää resumptionToken tuon ylimääräisen tietueen kohdalla.
- Dublin Core -formaatin metadataPrefix OAI-PMH:ssa on oai_dc.
- MARC 21:n metadataPrefix on marc21, vaikka muotona käytetäänkin MARCXML:ää.
- Kerralla kannattaa palauttaa tarpeeksi tietueita. Esim. 20 tietueen joukoilla pyyntöjen aiheuttama overhead yleensä hidastaa haravointia merkittävästi. Lähteestä riippuen sopiva määrä voi olla 100 - 3000 tietuetta.

Esimerkkejä

Esimerkeissä käytetään osittain kuvitteellisia tietoja.

Identify - palvelun tunnistetiedot

Pyyntö:

<http://provider.finna.fi/oaipmh?verb=Identify>

Vastaus:

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
    http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2016-11-04T12:26:26Z</responseDate>
  <request verb="Identify">http://provider.finna.fi/oaipmh</request>
  <Identify>
    <repositoryName>Finna - Public Interface</repositoryName>
    <baseUrl>http://provider.finna.fi/oaipmh</baseUrl>
    <protocolVersion>2.0</protocolVersion>
    <adminEmail>nobody@finna.fi</adminEmail>
    <earliestDatestamp>2001-07-08T22:00:00Z</earliestDatestamp>
    <deletedRecord>persistent</deletedRecord>
    <granularity>YYYY-MM-DDThh:mm:ssZ</granularity>
    <compression></compression>
    <description>
      <oai-identifier
        xmlns="http://www.openarchives.org/OAI/2.0/oai-identifier"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai-identifier
          http://www.openarchives.org/OAI/2.0/oai-identifier.xsd">
        <scheme>oai</scheme>
        <repositoryIdentifier>finna.fi</repositoryIdentifier>
        <delimiter>:</delimiter>
        <sampleIdentifier>oai:finna.fi:123</sampleIdentifier>
      </oai-identifier>
    </description>
  </Identify>
</OAI-PMH>
```

ListRecords - tietueet

Pyyntö:

<http://provider.finna.fi/oaipmh?verb=ListRecords&metadataPrefix=marc21>

Vastaus:

```

<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
    http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2016-11-04T12:31:35Z</responseDate>
  <request verb="ListRecords" metadataPrefix="marc21">http://provider.finna.fi/oaipmh</request>
  <ListRecords>
    <record>
      <header status="deleted">
        <identifier>oai:finna.fi:1</identifier>
        <datestamp>2016-03-12T11:59:12Z</datestamp>
      </header>
    </record>
    <record>
      <header>
        <identifier>oai:finna.fi:2</identifier>
        <datestamp>2016-03-21T15:46:31Z</datestamp>
      </header>
      <metadata>
        <record xmlns="http://www.loc.gov/MARC21/slim"
          xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
          xsi:schemaLocation="http://www.loc.gov/MARC21/slim
            http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd"
          type="Bibliographic">
          <leader>00762cam a2200265zi 4500</leader>
          <controlfield tag="001">2</controlfield>
          <datafield tag="035" ind1=" " ind2=" ">
            <subfield code="a">2</subfield>
          </datafield>
          <datafield tag="035" ind1=" " ind2=" ">
            <subfield code="a"> 2</subfield>
          </datafield>
          [paljon lisää metadataa ja tietueita]
        </record>
      </metadata>
    </record>
    <resumptionToken cursor="0">resxyz</resumptionToken>
  </ListRecords>
</OAI-PMH>

```

Jatkopyyntö resumptionTokenilla:

<http://provider.finna.fi/oaipmh?verb=ListRecords&resumptionToken=resxyz>

Vastaus:

```

<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
    http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2016-11-04T12:34:12Z</responseDate>
  <request verb="ListRecords" resumptionToken="resxyz">http://provider.finna.fi/oaipmh</request>
  <ListRecords>
    <record>
      <header>
        <identifier>oai:finna.fi:1201343</identifier>
        <datestamp>2016-11-04T10:46:31Z</datestamp>
      </header>
      <metadata>
        <record xmlns="http://www.loc.gov/MARC21/slim"
          xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
          xsi:schemaLocation="http://www.loc.gov/MARC21/slim
            http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd"
          type="Bibliographic">
          <leader>00872nam a2200265zi 4500</leader>
          <controlfield tag="001">1201343</controlfield>
          <datafield tag="035" ind1=" " ind2=" ">
            <subfield code="a">2</subfield>
          </datafield>
          [viimeisen tietueen loppu metadata]
        </record>
      </metadata>
    </record>
  </ListRecords>
</OAI-PMH>

```

ListIdentifiers - Vain tietueiden headerit

Pyyntö:

<http://provider.finna.fi/oaipmh?verb=ListIdentifiers&metadataPrefix=marc21>

Vastaus:

```

<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
    http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2016-11-04T12:31:35Z</responseDate>
  <request verb="ListIdentifiers" metadataPrefix="marc21">http://provider.finna.fi/oaipmh</request>
  <ListIdentifiers>
    <record>
      <header status="deleted">
        <identifier>oai:finna.fi:1</identifier>
        <datestamp>2016-03-12T11:59:12Z</datestamp>
      </header>
    </record>
    <record>
      <header>
        <identifier>oai:finna.fi:2</identifier>
        <datestamp>2016-03-21T15:46:31Z</datestamp>
      </header>
    </record>
    <resumptionToken cursor="0">resxyz</resumptionToken>
  </ListIdentifiers>
</OAI-PMH>

```