Informaatiotutkimuksen päivät 2010
21. - 22. lokakuuta, Tampere

ABSTRAKTI

Kansalliskirjaston digitointiprosessien kehittäminen ontologialähtöiseen annotointiin: Acerbin historiallinen matkakuvaus prototyyppinä

Yhteystiedot: Tiina Ison, Kansalliskirjasto, tiina.ison@kansallikirjasto.fi.

Kansalliskirjasto kehittää kokonaisvaltaista digitointituotantoaan osana kansallista infrastruktuurikehitystä (Kansallinen Digitaalinen Kirjasto, KDK ja Pitkäaikaisjärjestelmä, PAS). Kansalliskirjaston Konservointi- ja Digitointikeskus toimii taustajärjestelmänä tuottaen kattavia digitaalisia aineistoja ja kokoelmia käyttöön. Kansalliskirjaston digitointipolitiikka 2010 raamittaa kokonaisvaltaista digitointitoimintaa. Samalla digitointituotannossa pyritään jatkuvasti kehittämään digitointiprosesseja ja menetelmiä kohti sisältöjen haettavuutta ja käytettävyyttä.

Osana Travel Europeana -hanketta Kansalliskirjasto on digitoinut italialaisen Skandinavian- ja Lapinmatkaajan Giuseppe Acerbin (1773–1846) kaksiosaisen teoksen ”Travels through Sweden, Finland and Lapland to the North Cape in the years 1789 and 1799”. Digitoitu teksti on erillisessä ESR-Ephemera pilottihankkeessa muunnettu ohjelmallisesti muotoon, joka mahdollistaa sen jakamisen internetissä. Sen lisäksi on luotu menetelmä, jonka avulla erilaiset käyttäjät kuten erikoisalojen tutkijat tai muut tekstin lukijat voivat liittää tekstiin omia merkintöjään eli annotaatioitaan.

Osa annotaatioprosessista on mahdollista toteuttaa automaattisesti hyödyntämällä kieliteknologiaa siten, että esimerkiksi aikamääreet sekä henkilöiden, paikkojen ja instituutioiden nimet (Named Entities) eristetään tekstistä. Osa annotaatioprosessista, kuten erilaisten teemojen tunnistaminen tapahtuu kuitenkin ihmisvoimin. Teemojen tunnistamisessa on käytetty runkona Outline of Cultural Materials -käsitteistöä (OCM), joka on osa kulttuuriantropologian ja arkeologian alan Human Relations Area Files -tietokantaa.

OCM:n lisäksi tekstin kuvailussa on hyödynnetty TEI-kuvailukieltä. TEI mahdollistaa tekstin summittaisen rakenteellisen kuvailun, mutta tarvittaessa kuvailu ja erittely voidaan viedä halutulle tarkkuustasolle aina kieliopillisiin rakenteisiin ja yksittäisiin kirjainmerkkeihin asti.

Tekstin kuvailussa voidaan käyttää hyväksi ulkopuolisia auktoriteettilähteitä. Samalla historiallista nimistöä ja nimiin liittyviä tunnisteita voidaan kerätä tietokantaan, joka on liitettävissä erilliseen, historiallista nimistöä sisältävään paikkatietopalveluun tai muihin jaettuihin palveluihin.

Tavoitteena on, että myöhemmässä vaiheessa nimistöä ja tunnisteita sisältävä auktoriteettipalvelu luo edellytyksiä monikielisten hakujen toteuttamiselle. Acerbin matkakuvaus on käännetty usealle kielelle: englanniksi, ranskaksi, italiaksi, saksaksi, ruotsiksi, suomeksi ja hollanniksi. Teoksen alkuperäinen kieli on englanti, ja hankkeen tässä vaiheessa työskennellään ainoastaan englanninkielisen tekstin parissa.

TEI on XML-formaatti, jossa tekstin analyysi ja annotointi tyypillisesti sulautetaan osaksi tekstiä. TEI mahdollistaa myös tekstin ulkopuolelle sijoitettavat annotaatiot. Toteutetussa ratkaisussa on päädytty malliin, jossa tekstiin koodataan ainoastaan tekstin rakenteelliset elementit kuten sivut, rivit ja sanat. Koko tekstin jokaiselle sanalle on annettu oma tunniste, esimerkiksi P70_ST00058 tarkoittaa 70. sivun 58. sanaa.

Sanojen tunnisteet syntyvät automaattisesti osana tekstin digitointia. Digitoinnin yhteydessä jokaisesta sivusta muodostuu yksi Alto XML -muotoinen tiedosto. Digitointiprosessin osana Alto-tiedostot liitetään METS-paketteihin pitkäaikaistallennusta silmällä pitäen. METS/Alto-tiedostoja ei kuitenkaan ole optimoitu internetiä varten. Siitä syystä Alto-tiedostoista tuotetaan riisuttu versio, jossa kuitenkin säilytetään osa Alton rakenteista ja tunnisteista, mm. sivujen ja sanojen tunnisteet.

Antamalla jokaiselle sanalle oma tunniste tekstin annotoinnissa voidaan viitata mihin kohtaan tahansa tekstissä. Koska digitointiprosessissa automaattisesti syntyneet tunnisteet ovat hankalia käsitellä, tunnisteiden poimimista varten on otettu käyttöön tekninen ratkaisu (ns. tooltip), jossa tunniste poimitaan talteen hiirellä klikkaamalla.

Käyttämällä kahta tunnistetta voidaan viitata yhtä sanaa laajempaan tekstisegmenttiin. Esimerkiksi P70_ST00058:P70_ST00089 tarkoittaa segmenttiä, joka alkaa 70. sivun 58. sanasta ja päättyy saman sivun 89. sanaan. Tähän segmenttiin voidaan liittää haluttu, tekstin ulkopuolelle sijoittuvaan tiedostoon tai tietojärjestelmään tallennettava kommentti. Koska annotaatio sijoittuu tekstin ulkopuolelle, segmentit voivat olla sisäkkäisiä tai ristikkäisiä. Esimerkiksi yllä mainittuun segmenttiin sisältyy sijaintipaikassa P70_ST00059 vuosiluku 1365 ja sijaintipaikassa P70_ST00060:P70_ST00063 ilmaisu ”Magnus, king of Sweden”.

TEI on harmonisoitu kulttuuriperintöalan CIDOC CRM -käsitemallin kanssa. CIDOC CRM -puolestaan on yhdistetty bibliografisen kuvailun käsitemalleihin (kansainvälisen kirjasto-organisaation IFLAn kehittämät FRBR, FRAD ja FRSAD). Viimeksi mainitut ovat toimineet lähtökohtana luotaessa uutta RDA-luettelointisääntöä.

CIDOC CRM:n ja FR-maailman välille on rakennettu kaksi siltaa: oliomallina ilmaistu ontologia FRBRoo ja sanastojen yhdistelyä palveleva Vocabulary Mapping Framework, joka on toteutettu semanttisessa webissä käytettävinä tripletteinä. Ontologiat voidaan kuvailla oliomalleilla ja ne voidaan ilmaista tripletteinä, joten näiden mallinnustasojen ja toteutustapojen välillä ei ole ristiriitaa.

Kaikki edellä mainituista käsitemalleista sisältävät paitsi käsiteluokkien myös käsitteiden välisten suhteiden (relaatioiden) kuvauksia. Teemojen tunnistamisen lisäksi tekstistä voidaan TEI:n avulla tunnistaa myös tekstiin sisältyviä, esimerkiksi tapahtumia kuvaavia moniulotteisia relaatioita, jotka yhdistävät toisiinsa esimerkiksi henkilön, toiminnan, ajan, paikan, välineen ja

aiheen.

FR-mallit on ensisijaisesti luotu ohjaamaan bibliografista luettelointia ja kuvailua. FRSAD muodostaa linkin teoksesta niihin teemoihin ja niitä edustaviin termeihin (FRSADissa ”nomen”), joita teoksessa ilmenee. Tässä tapauksessa teemoja on kuvailtu OCM-käsitteistön avulla, mutta esimerkiksi tapahtumien temaattisessa kuvailussa on mahdollista hyödyntää myös CIDOC CRM -käsitemallia.

Toteutetussa tekstin annotoinnissa ei ole kyse vain hakuelementtien liittämisestä tekstiin. Hakuelementit johtavat tekstin äärelle. TEI ja temaattisessa kuvailussa käytetyt käsitemallit ja käsitteistöt johtavat sekä tekstin sisään, sen rakenteisiin ja teemoihin, että sen ulkoisiin kytköksiin.

Toteuttajat:
Tiina Ison, projektipäällikkö
Eeva Murtomaa, temaattinen koodaus
Mika Nyman, tekninen asiantuntija, TEI, CIDOC CRM

Kansalliskirjaston digitointipolitiikka, 2010
http://www.kansalliskirjasto.fi/attachments/5v5daJ8e3/5smlIebLV/Files/CurrentFile/KK_Digitointipolitiikka.pdf