Olli Hallamaa

TEI – järjestystä ja metatietoja tekstidokumentteihin

Jokainen kirjoitus on rakenteinen. Yksinkertaisimmillaan tekstin rakenne voi tarkoittaa tekstin asettelua kirjoitusalustalle, siis sitä, miten teksti jakautuu riveille ja sivuille. Usein tekstin rakenne on kuitenkin huomattavasti monisyisempi: se voi jakaantua lukuihin ja alalukuihin, siinä on kappaleita, siihen voi kuulua suoria lainauksia muista teksteistä ja sen osana voi olla taulukoita, kaavioita sekä kuvia. Lukijalle tekstin rakenne on helppo ilmaista typografisten tehokeinojen avulla, mutta tietokoneelle typografisin keinoin esitetty tieto ei välity. Rakenteelliselta tekstiltä näyttävä dokumentti on koneelle pelkkä numeerinen sarja eikä se juuri kykene tulkitsemaan lukijalle ilmeisiä typografisia vihjeitä. Se kyllä osaa käskettäessä vaihtaa kirjasinleikkauksen kokoa ja tyyliä, mutta se ei hahmota näin merkittyä tekstikappaletta otsikoksi kuten kirjan lukemiseen harjaantunut ihmismieli tekee.

Jotta tekstiä voitaisiin tehokkaasti työstää koneellisesti, tietokoneelle tulee yksiselitteisesti kuvata tekstin rakenne sekä kirjoittaa auki riittävä määrä semanttista informaatiota. Rakenteen koodaaminen on tarpeen esimerkiksi silloin, kun sama emotiedosto halutaan julkaista useammalla eri tavalla, esimerkiksi perinteisenä paperijulkaisuna sekä digitaalisena versiona verkossa; yhä useammin saman aineiston pitää myös kääntyä eri tiedostoformaatteja käyttäville sähköisille lukulaitteille.

Tiedostoihin upotettua semanttista informaatiota tarvitaan myös tekstiin kohdistettavien tarkkojen hakujen mahdollistamiseksi. Jos esimerkiksi haluaisimme löytää kuvitteellisesta 1800-luvun saksalaisten kirjekokoelmien kokotekstitietokannasta vain ja ainoastaan Friedrich Hölderlinin kotiopettajana Bordeaux'sta vuosina 1801–1802 kirjoittamat kirjeet, tietokannan pitäisi tekstin ohella sisältää yksiselitteinen informaatio kirjeiden kirjoittajista, kirjoituspaikoista ja -ajoista. Kaikki nämä tiedot eivät aina käy ilmi itse kirjeestä, minkä vuoksi toimittajan pitää lisätä ne kuhunkin kirjeeseen metatiedoiksi. Metatiedot ovat tarpeen silloinkin, kun tiedot käyvät ilmi kirjeestä, sillä tietokone ei varmuudella tunnista tekstissä esiintyvää nimeä kirjoittajaksi, maantieteellistä nimeä kirjoituspaikaksi tai numeerista informaatiota kirjeen päiväykseksi.

Tutkimuksessa tekstejä käytetään monin tavoin. Yhteiskuntatieteilijä saattaa purkaa tekstiksi videoitua keskustelutilannetta, jossa verbaalisen informaation lisäksi merkitystä on muun muassa puheen jaksotuksella, puhujan tunnetilalla tai äänen voimakkuudella. Miten tekstiin merkitään tauon pituus, aggressiivi tai äänen vaimentuminen kuiskaukseksi? Tekstikriittisen edition toimittaja puolestaan joutuu merkitsemään tiedostoihinsa tekstin eri versioiden poikkeamat, kirjoittajan käyttämien lähteiden viitetiedot, tekstin sisäiset viitteet sekä ehkä myös tekstiä selittäviä huomautuksia. Tekstintutkimuksen

tietojenkäsittelyllisiin haasteisiin kuuluu, miten yllä esimerkein kuvattu tekstin rakenteen ja mitä moninaisimman semanttisen informaation sisällyttäminen tiedostoihin on mahdollista toteuttaa standardisti, käyttöjärjestelmistä ja sovellusohjelmista riippumattomasti.

Ratkaisu näihin kaikkiin sekä lukuisiin muihin tekstintutkijan eteen tuleviin haasteisiin on XML (eXtensible Markup Language, suom. laajennettavissa oleva merkkauskieli), jonka kehittämisestä vastaa World Wide Web Consortium (ks. www.w3.org/XML/). Tekstien tutkimuksen käyttöön XML:stä on kehitetty erityinen laajennus TEI (Text Encoding Initiative), jonka ylläpidosta vastaa oma erillinen konsortio (ks. www.tei-c.org).

XML on tiedon esittämiseen ja käsittelemiseen kehitetty avoin standardinomainen suositus, joka on tarkoitettu tekstimuotoisen rakenteisen tiedon kuvauskieleksi. Luonteeltaan XML-standardi on elementtiperustaisen koodauksen kielioppi ja tyyppikuvaus. Se kertoo, minkälainen on hyvin muodostettu XML-dokumentti, ja määrittelee XML-dokumentin loogisen rakenteen. Dokumentti on hyvin rakennettu, kun siinä on vähintään yksi elementti, ainoastaan yksi juurielementti ja elementtien alku- ja lopputagit menevät dokumentin hierarkkisen elementtirakenteen mukaisesti aidosti sisäkkäin.

Käytännössä hierarkkinen rakenne syntyy siten, että XML-dokumentteihin kirjoitettava tieto merkataan tageilla. Tagi eli tunniste on kulmasulkeisiin kirjoitettu kuvaus, joka kuuluu elementtiin. Elementin muodostavat alkutagi ja sitä vastaava lopputagi sekä niiden välissä oleva teksti. Esimerkiksi

<otsikko>Johdanto</otsikko>

muodostavat elementin, joka määrittää kirjan yhden rakenteellisen osan ja sen sisällön. Koska XML on puhdas metakieli, se ei sisällä elementtien määrittelyä, vaan tagit ovat käyttäjän vapaasti valittavissa. Tämä tekee XML:stä helposti laajennettavan. Toisaalta laajennettavuudesta seuraa, että XML-dokumenttiin pitää liittää metatietoa, joka määrittelee dokumentissa käytetyt elementit ja niille mahdollisesti kuuluvat attribuutit. Attribuuteilla voidaan antaa elementeille tarkempia arvoja. Esimerkiksi kuvitteellisen taiteilijatietokannan elementissä

<kirjailija ala=”lastenkirjailija”>Anni Swan</kirjailija>

tagipari <kirjailija></kirjailija> osoittaa Anni Swanin taiteenalan, ja alkutagiin liitetty attribuutti ala=”lastenkirjailija” antaa kirjailijalle tarkemman lisämääreen.

Metatiedot XML-dokumentin rakenteesta ja määreistä annetaan niin sanotussa skeemassa. Sen avulla voidaan selvittää, onko dokumentti validi. Käytännössä tiedoston validiteetin tarkastamisen tekee XML-tiedostojen kirjoittamiseen suunniteltu ohjelma, XML-editori. Skeeman perusteella editori tunnistaa, mitä elementtejä dokumentti

käyttää ja mitkä attribuutit ovat sallittuja.

TEI on luonteeltaan ehdotus siitä, kuinka XML-standardin mukaisia tiedostoja tulisi kirjoittaa tekstientutkimuksessa. TEI-suositus sisältää valtaisan joukon valmiiksi määriteltyjä elementtejä ja attribuutteja, joilla voidaan merkata tekstiin muun muassa kappaleiden tyypit, runon rakenne, päiväykset, nimet, lyhenteet ja niiden tulkinta, mittayksiköt, tekstin pääkielestä poikkeavien sanojen kieli, osoitteet, tekstin sisäiset viitteet, bibliografiset tiedot, sanakirjan rakenne, editorin huomautukset, vaihtoehtoiset lukutavat, kriittisen apparaatin merkinnät, grafiikka ja tekstiin liittyvä non-verbaalinen informaatio. Jos TEI:n valmiit elementit eivät riitä, käyttäjä voi vapaasti lisätä dokumenttiin omia elementtejä kunhan vain noudattaa XML-kielioppia ja dokumentoi luomansa muodosteet metatietoihin.

TEI:n viimeisin versio P5 (proposal 5) tarjoaa tekstintutkijalle huolella suunnitellun ja laajasti käytetyn XML-ratkaisun. TEI:n ongelmana on, ettei XML-dokumentti sellaisenaan sovellu käytettäväksi juuri mihinkään; sen hyödyntäminen edellyttää lähes aina jatkojalostusta. Tämän tekee tietokone, kunhan sille on osattu kirjoittaa tarkoitukseen sopiva ohjelma. Muunnosohjelman tuottamat tiedostot pitää edelleen kyetä jalostamaan julkaisuksi, mieluiten automaattisesti. Suomessa TEI:tä käytetään muun muassa Kotimaisten kielten tutkimuskeskuksen tekstikorpuksissa, SKS:n Edith-yksikössä, SLS:n Topelius-editioissa sekä Helsingin yliopiston englantilaisen filologian korpushankkeissa.