Sessio 6

Tuula Pääkkönen, Marja-Leena Hynynen, Jukka Kervinen

Kansalliskirjasto

Sanomalehtitietojen rikastaminen – kahdella linjalla: tiedotusta ja tietoa puoluekannoista

Suomalaisen sanomalehdistön historia alkaa vuodesta 1771 ja digitoituna on kaikki Kansalliskokoelmasta löytyvät lehdet aina vuoden 1935 loppuun saakka. Samaan esitysjärjestelmään on vuonna 2019 tuotu myös kirja-aineistot, joista 2000 kirjan avaaminen (Biström, 2019; Laine, 2019; Lehmikoski-Pessa, 2019) tuo aivan uusia aineistoja käyttöön verkossa tai Haka-kirjautumisella yliopistoissa ja ammattikorkeakouluissa. Viestinnällisesti projektin yhteydessä oli

Digitointia on pyritty edistämään etenemällä vuosittain eteenpäin. Joitakin lehtiä on kumppanuusdigitointeina digitoitu koko historiansa ajalta. Koska aineistoa on erilaisia tyyppejä, eri aikakausilta, niin kuinka pidämme vanhemmankin ja haastavaksikin koetun aineiston kiinnostavana nykykäyttäjälle? Tavoitteenamme organisaationa olisi saada aineistoa käyttöön tutkijoille ja kansalaisille organisaation ydintehtävien mukaisesti. Kerromme tässä esityksessä, kuinka nostot sosiaaliseen mediaan, eri aineistoista, esimerkiksi 1800-luvulla julkaistuihin teoksiin, ovat vaikuttaneet aineiston sivukäyttöön. Lisäksi kerromme muutamasta yksittäistapauksesta, joka on luonut aineistojen käyttöön käyttöpiikin.

Kansalaisten ohella, erityisesti tutkijoita, kiinnostaa aineistojen rikastaminen, eli lisätiedot aineistoista, joilla voi päästä aineistojen syntyhistoriaan entistä syvemmin. Vuoden 2019 keväällä Kansalliskirjastossa järjestetyssä sanomalehtisymposiumissa tutkijat kertoivat eri tutkimuksistaan sekä painettujen että digitoitujen sanomalehtien parissa (Lilja & Hakkarainen, 2018). Yksi myös symposiumissa noussut toive useammalta tutkijalta oli selkeä pyyntö lisätä sanomalehtien puoluekantatiedot näkyviin https://digi.kansalliskirjasto.fi – palveluun. Kansalliskirjastossa päätettiin tutkia asiaa sekä sisällöllisenä että tekniseltä kannalta. Niinpä olemme käyneet läpi Päiviö Tommilan ja työryhmän kirjoittamaa ”Suomen sanomalehdistön historia”-kirjasarjaa ja selvittäneet mitä puoluekannoista kerrotaan eri lehdille. Jatkossa näemme kuinka tutkijoiden toiveista tehty rikastaminen vaikuttaa aineistojen tutkimuskäyttöön.

**

Kimmo Kettunen

Kansalliskirjasto

Automaattinen nimien tunnistus 1800-luvun suomenkielisissä digitoiduissa lehtiaineistoissa

Kansalliskirjaston Mikkelin toimipiste on digitoinut Suomessa julkaistuja sanoma- ja aikakauslehtiä vuodesta 1998. Aineisto on käytettävissä verkkopalvelussa digi.kansalliskirjasto.fi. Vapaasti käytettävää lehtiaineistoa on verkkopalvelussa noin 7.5 miljoonaa sivua vuosilta 1771–1929.

Kansalliskirjastossa on pyritty viime vuosina paitsi kartuttamaan lehtiaineistoa myös lisäämään sen

käytettävyyttä eri tavoin. Tätä tarkoitusta varten olemme muun muassa tutkineet automaattista nimien tunnistamista ja merkitsemistä (named entity recognition, NER) aineistosta. Olemme luoneet 1800-luvun lopun ja 1900-luvun alun suomenkielisistä lehtiaineistoista merkityn opetus- ja evaluaatioaineiston, jolla olemme opettaneet Stanford NER –ohjelmiston tunnistamaan ihmisten ja paikkojen nimiä. Lopputuloksena syntynyt malli toimii evaluaatioaineistossamme kohtuullisen hyvin. Ihmisten nimissä malli saavuttaa F-arvon 0.72 ja paikannimissä arvon 0.79 optisesti luetulla kohtuulaatuisella tekstillä. Hyvälaatuisella tekstillä nimiä tunnistetaan jonkin verran paremmin.

Olemme toistaiseksi liittäneet toimivan nimimallin digi.kansalliskirjasto.fi –palvelussa Uuden Suomettaren parannetun optisen luvun aineistoon vuosilta 1869-1918. Nimimallista on myös tekeillä toimiva sovellus, jota voidaan käyttää Kielipankissa. Merkitty opetus- ja evaluaatioaineisto ovat saatavilla Kansalliskirjaston avoimen datan sivustolla https://digi.kansalliskirjasto.fi/opendata. Aineistoa voi käyttää oman nimientunnistusohjelmiston opettamiseen ja evaluoimiseen.

Kirjallisuutta

La Mela, Matti, Tamper, Minna, Kettunen, Kimmo, 2019. Finding Nineteenth-century Berry Spots: Recognizing and Linking Place Names in a Historical Newspaper Berry-picking Corpus. DHN2019.

Ruokolainen, Teemu, Kettunen, Kimmo, 2018. À la recherche du nom perdu – searching for named entities with Stanford NER in a Finnish historical newspaper and journal collection. DAS, 13th IAPR International Workshop on Document Analysis Systems.

Ruokolainen, Teemu, Kettunen, Kimmo (2020). Name the Name - Named Entity Recognition in OCRed 19th and Early 20th Century Finnish Newspaper and Journal Collection Data (ilmestyy).

**

FT Heikki Kokko

Suomen Akatemian Kokemuksen historian huippuyksikkö HEX

Tampereen yliopisto

Miksi median synnyn yhteiskunnallinen merkitys jäi Suomessa historiantutkimuksen katveeseen?

Nykyinen digitaalinen aikakausi tuo esille medioiden syvällisen yhteiskunnallisen merkityksen. Media on yhä selvemmin vallankäytön areena, jossa ratkotaan kehityksen suuntaa niin yksittäisten yhteiskuntien kuin globaalilla tasolla. Lehdistö oli ensimmäinen vaikutuksiltaan väestöä läpileikkaava media. Suomessa sen yhteiskunnallinen läpimurto tapahtui 1800-luvun puolivälissä, kun suomenkielisen julkisuuden ensimmäinen nousu alkoi.

Suomenkielisen julkisuuden ensimmäinen nousukausi on kuitenkin jäänyt Suomen historian kaanonissa varjoon. Sitä on tutkittu lähinnä osana lehdistöinstituution ja journalismin historiaa. Kansalaisyhteiskuntatutkimuksessa suomenkielisen lehdistön nousu on milteipä sivuutettu ja sille on liiennyt oma lukunsa vain harvoissa Suomen historian yleisesityksissä. Näin on, vaikka monet suomenkielisen julkisuuden syntymisen kokeneet aikalaiset tunnistivat sen yhteiskuntaa muuttaneen luonteen.

Tarkastelen esityksessäni sitä, miksi median syntymisen yhteiskunnallinen merkitys 1800-luvun puolivälissä on jäänyt Suomessa unohduksiin. Historiografinen esitykseni kulkee 1800-luvun suurmiesmyytistä kielikysymyksen kautta kansalliseen eheytymiseen maailmansotien välissä ja siitä 1960-luvun uusiin liikkeisiin päätyen lopulta digitaaliseen aikaan.