Maisterintutkielma

Tietojenkäsittelytieteen maisteriohjelma

Eduskunnan täysistuntojen pöytäkirjojen muuntaminen semanttiseksi dataksi ja julkaiseminen verkkopalveluna

Laura Sinikallio

22.2.2022

Matemaattis-luonnontieteellinen tiedekunta

Helsingin yliopisto

Ohjaaja

Prof. Eero Hyvönen

Tarkastajat

Prof. Eero Hyvönen, prof. Tomi Männistö

Yhteystiedot

PL 68 (Pietari Kalmin katu 5)
00014 Helsingin yliopisto

Sähköpostiosoite: info@cs.helsinki.fi
URL: https://www.cs.helsinki.fi/

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Study programme
Matemaattis-luonnontieteellinen tiedekunta	Tietojenkäsittelytieteen maisteriohjelma

Tekijä — Författare — Author

Laura Sinikallio

Työn nimi — Arbetets titel — Title

Eduskunnan täysistuntojen pöytäkirjojen muuntaminen semanttiseksi dataksi ja julkaiseminen verkkopalveluna

Ohjaajat — Handledare — Supervisors

Prof. Eero Hyvönen

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Maisterintutkielma	22.2.2022	60 sivua, 11 liitesivua

Tiivistelmä — Referat — Abstract

Parlamentaaristen aineistojen digitointi ja rakenteistaminen tutkimuskäyttöön on nouseva tutkimuksenala, jonka tiimoilta esimerkiksi Euroopassa on tällä hetkellä käynnissä useita kansallisia hankkeita. Tämä tutkielma on osa Semanttinen parlamentti -hanketta, jossa Suomen eduskunnan täysistuntojen puheenvuorot saatetaan ensimmäistä kertaa yhtenäiseksi, harmonisoiduksi aineistoksi koneluettavaan muotoon aina eduskunnan alusta vuodesta 1907 nykypäivään. Puheenvuorot ja niihin liittyvät runsaat kuvailutiedot on julkaistu kahtena versiona, parlamentaaristen aineistojen kuvaamiseen käytetyssä Parla-CLARIN XML -formaatissa sekä linkitetyn avoimen datan tietämysverkkona, joka kytkee aineiston osaksi laajempaa kansallista tietoinfrastruktuuria.

Yhtenäinen puheenvuoroaineisto tarjoaa ennennäkemättömiä mahdollisuuksia tarkastella suomalaista parlamentarismia yli sadan vuoden ajalta monisyisesti ja automatisoidusti. Aineisto sisältää lähes miljoona erillistä puheenvuoroa ja linkittyy tiiviisti eduskunnan toimijoiden biografisiin tietoihin. Tässä tutkielmassa kuvataan puheenvuorojen esittämistä varten kehitetyt tietomallit ja puheenvuoroaineistojen keräys- ja muunnosprosessi sekä tarkastellaan prosessin ja syntyneen aineiston haasteita ja mahdollisuuksia.

Toteutetun aineistojulkaisun hyödyllisyyden arvioimiseksi on Parla-CLARIN-muotoista aineistoa jo hyödynnetty poliittiseen kulttuuriin liittyvässä digitaalisten ihmistieteiden tutkimuksessa. Linkitetyn datan pohjalta on kehitetty semanttinen portaali, Parlamenttisampo, aineistojen julkaisemista ja tutkimista varten verkossa.

ACM Computing Classification System (CCS)
Information systems→ Information retrieval → Document representation → Data encoding and canonicalization
Information systems→ Information retrieval → Document representation → Ontologies
Applied computing → Computers in other domains → Digital libraries and archives

Avainsanat — Nyckelord — Keywords

Linkitetty avoin data, RDF, semanttinen web, Parla-CLARIN

Säilytyspaikka — Förvaringsställe — Where deposited

Helsingin yliopiston kirjasto

Muita tietoja — övriga uppgifter — Additional information

Ohjelmistojen opintosuunta

Sisällys

1 Johdanto

Nykypäivän tietoyhteiskunnassa tavalliselle ihmiselle saatavilla oleva tiedon määrä on ennennäkemätön. Verkossa oleva julkinen, avoin data on loputtomasti kopioitavissa ja jaettavissa. Avoin data on ideaali, mutta se ei määritä, kuinka data on julkaistu [27]. Dataa voidaan julkaista esimerkiksi Excel-taulukkona, tietokantakopiona tai palvelurajapintojen kautta. Avoin data voi olla rakenteellista tai ei-rakenteellista. Siinä missä strukturoidun aivoimen datan määrä verkossa kasvaa jatkuvasti, on merkittävä osa datasta edelleen strukturoimatonta [43], manuaaliseen ihmiskäyttöön tarkoitettua [25]. Haasteeksi jää kuinka avointa dataa voitaisiin moninaisuudessaan käsitellä automatisoidusti, älykkäästi ja kattavasti.

Semanttisen webin (*Semantic Web*) aate pyrki tuomaan saataville enemmän strukturoitua dataa, jota on mahdollista hyödyntää tehokkaammin automatisoidusti [43]. Semanttinen web voidaan nähdä nykyisen *World Wide Webin* parannuksena, jossa data julkaistaan koneymmärrettävässä muodossa [25, 27]. Semanttinen web muodostaa toiminnallisen tiedon verkon (*web of actionable information*), jossa linkitetyt käsitteet ja tiedot niiden välisistä suhteista rakentuvat semanttisen datan verkoksi [59, 27].

Tiede ja tutkimus ovat olleet keskeisiä semanttisen datan aatteen edistäjiä [59]. Esimerkiksi kulttuuriperinnöllisten aineistojen (*cultural heritage*) julkaisu ja tutkimus on noussut erittäin merkittäväksi semanttisen webin -teknologioiden sovelluskohteeksi [28]. Eräs kulttuuriperinnöllisesti merkittävä aineisto ovat kansallisten hallintoelinten asiakirjat, kuten parlamentaaristen keskustelujen pöytäkirjat. Pöytäkirjat läpi historian tarjoavat ainutlaatuista näkökulmaa aikansa yhteiskunnan toiminnasta ja arvoista. Parlamentaariset keskusteluaineistot ovat sisällöltään, rakenteeltaan ja kieleltään uniikkeja ja siksi tärkeä tutkimuskohde digitaalisissa ihmistieteissä [50, 2].

Suomen eduskunnan täysistuntojen pöytäkirjat ovat saatavilla eduskunnan verkkopalveluista [17, 18] avoimena datana aina eduskunnan alun, valtiopäivien 1907, pöytäkirjoista lähtien. Digitoidut pöytäkirjat ovat kuitenkin saatavilla aikakaudesta riippuen eri formaateissa ja eri verkko-osoitteista. Paikoin niiden löytäminen voi olla haastavaa, sillä eduskunnan hakupalveluiden käyttö vaatii asiantuntijuutta, kuten tässä tutkielmassa tullaan myöhemmin havainnollistamaan.

Suomen eduskunnan täysistuntojen pöytäkirjat on nyt ensimmäistä kertaa saatettu yh-

tenäiseen, rakenteiseen muotoon. Pöytäkirjoista kerätyt puheenvuorot ja niihin liittyvät oleelliset kuvailutiedot on muunnettu Eduskunnan puheet -aineistoksi, joka kattaa koko Suomen eduskunnan historian 1907–2021 ja mahdollistaa uudenlaisen kattavan, laajamittaisen ja automatisoidun pöytäkirjojen tutkimisen. Aineisto on toteutettu kahtena versiona: semanttisen webin linkitettynä avoimena datana sekä parlamentaaristen aineistojen julkaisuun kehitetyssä Parla-CLARIN XML -muodossa.

Eduskunnan pöytäkirjojen muuntaminen on osa Semanttinen parlamentti -konsortiohanketta, jossa luodaan eduskunnan tietokannoista datapalvelu ja tutkimusympäristö, joka palvelee niin tutkijoita kuin aiheesta kiinnostuneita kansalaisia ja median edustajia. Hanke alkoi vuonna 2020 ja päättyy vuoden 2022 lopussa. Hankkeen ensituloksista on tehty jo useampia julkaisuja. Hanketta kokonaisuudessaan kuvaavat [31] ja [32][1], pöytäkirjojen muunnosta on kuvattu tiiviisti artikkelissa [63] sekä eduskunnan toimijoiden tietojen keräämistä ja muuntamista [40]. Hankkeessa syntyneeseen aineistoon pohjautuvia ensimmäisiä tutkimustuloksia poliittiseen kulttuuriin ja kieleen liittyen on esitelty julkaisussa [21].

Tässä tutkielmassa esitellään Semanttinen parlamentti -hanke ja ennen kaikkea hanketta varten toteutettu Suomen eduskunnan täysistuntojen pöytäkirjojen muunnosprosessi. Hankkeen lopputuotteena olevia datapalvelua ja tutkimusympäristöä varten pöytäkirjat ja niistä löytyvät puheenvuorot oli saatava yhtenäiseen ja koneluettavaan muotoon. Tässä tutkielmassa esitellään, kuinka tämä toteutettiin ja millaisia työkaluja prosessia varten oli luotava. Luvussa 2 alustetaan hankkeen taustalla olevia semanttisen ja linkitetyn datan periaatteita, luvussa 3 tarkastellaan erilaisia parlamentaaristen aineistojen kuvantamisratkaisuja ja vastaavia hankkeita maailmalta. Luvussa 4 esitellään Semanttinen parlamentti -hanke, aineiston tietomallit sekä julkaisuratkaisut. Luvussa 5 kuvataan pöytäkirjojen muunnoksen konkreettinen toteutus ja luvussa 6 analysoidaan sen tuloksia. Lopuksi esitetään tutkielman yhteenveto luvussa 7.

[1] Hyväksytty julkaistavaksi

2 Linkitetty data

Semanttisen webin kehitys lähti kunnolla vauhtiin 2000-luvun alussa ja sen merkkipaaluna pidetään vuonna 2001 *Scientific American* -julkaisussa ilmestynyttä Berners-Leen ja kumppaneiden artikkelia "*The Semantic Web*" [9], joka kiteytti ajatuksen semanttisesta webistä koneluettavana tiedon verkkona, jota "älykkäät toimijat" (*intelligent agents*) voivat hyödyntää [25, 59]. Semanttisen webin ideaalina on lisätä verkossa tarjolla olevan strukturoidun tiedon määrää tehokkaamman automaation mahdollistamiseksi; rikastaa ihmisluettavaa dataa koneluettavilla semanttisilla merkinnöillä (annotaatiot, *annotations*), jolloin webistä muodostuisi maailman suurin tietokanta [43, 9, 49]. Näin syntyy *World Wide Webin* seuraava edistysaskel *Web 3.0*, semanttinen web (*Semantic Web, Web of Data*) [52].

2.1 RDF ja tietämysverkot

Semanttisen, koneluettavan datan aatetta seuraa tarve yhteiselle tietorakenteelle [52]. [World Wide] Webin standardeja ja suosituksia ylläpitävä *World Wide Web Consortium* (W3C) kehitti jo vuonna 1997 semanttisen webin kuvailukielen ja tietomallin *Resource Description Framework*, **RDF** [59]. Kolmikkorakenteeseen ja universaaleihin tunnisteisiin pohjaavasta RDF-mallista tuli W3C:n suositus semanttisen datan koodaukseen (*encoding*) vuonna 1999 [59, 49]. Sitä voidaan pitää nykypäivän *lingua francana* semanttisen webin ja linkitetyn datan teknologioissa [53].

Hyvönen [27] kuvaa RDF-mallin periaatteita: Sen keskiössä on nimensä mukaisesti resurssi (*resource*), mikä tahansa asia, josta voidaan kuvata tietoa. Jotta resurssia voidaan kuvata on sillä oltava tunniste (*identifier*), jolla siihen voidaan viitata. Resurssien sekaantumisen estämiseksi tunnisteen on oltava yksilöivä ja yksikäsitteinen. RDF-koodatut resurssit on nimetty yleensä URI-tunnisteilla (*Uniform Resource Identifier*) [49]. URI:t ovat tiettyä rakennetta noudattavia merkkijonoja[2]. Esimerkiksi taiteilija Leonardo da Vincin URI Wikipediasta louhitussa DBpedia-järjestelmässä on
<http://dbpedia.org/resource/Leonardo_da_Vinci> [13]. Tunnisteesta käy ilmi muun

[2]Esimerkiksi HTML-sivujen sijainnin kertovat URL-osoitteet (*Uniform Resource Locator*) ovat URI-tunnisteiden erikoistapaus [27].

muassa URI:n skeema (http) sekä resurssin tunnisteen luonnista vastuussa oleva auktoriteetti (dbpedia.org). Tunnisteen rakenteesta voi lukea lisää RFC-3986-julkaisusta [8].

RDF-aineisto koostuu kolmikkomuotoisista väittämistä (*statements*) [49]. Kolmikko (*triple*) on muotoa: **<subjekti predikaatti objekti>** (*subject predicate*[1] *object*). Näillä kolmikoilla kuvataan resursseja ja niiden ominaisuuksia. Kolmikon osat voisi hieman kuvaavammin nimetä myös: *<resurssi, resurssin ominaisuus, ominaisuuden arvo>* [27]. Esimerkiksi halutessamme kuvata aiempaa esimerkkiresurssia, Leonardo Da Vincia, voisimme kertoa tämän olevan 15.4.1452 syntynyt henkilö, joka maalasi Mona Lisan. Kolmikkorakenteella syntyisivät seuraavat englanninkieliset väittämät:

```
<Leonardo da Vinci> <is a> <Person> .
<Leonardo da Vinci> <was born on> <the 15th April 1452> .
<Mona Lisa> <was authored by> <Leonardo da Vinci> .
```

RDF-muodossa ensimmäinen väittämä voisi kuulua esimerkiksi seuraavasti:

```
<http://dbpedia.org/resource/Leonardo_da_Vinci>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://xmlns.com/foaf/0.1/Person> .
```

Jokainen väittämän jäsen on ilmaistu URI-tunnisteella, myös subjektin ja objektin välistä suhdetta kuvaavaa predikaatti. Tunnisteen avulla määritelty suhde on nyt semanttisesti annotoitu ja tunnisteella voidaan löytää vaivattomasti aineistosta kaikki vastaavanlaiset suhteet.

Väittämän objekti voi olla tyypiltään myös literaali. Tiedot, jotka itsessään eivät ole keskeisiä resursseja, kuten lukumäärät, päivämäärät ja kutsumanimet, ilmoitetaan literaalimerkkijonona. Literaalin arvolle määritellään kuitenkin tyyppi URI-tunnisteella. Tunnisteille määritellään usein lyhenteitä, prefiksejä, jolloin niistä tulee ihmissilmin paljon miellyttävämpiä lukea. Seuraavassa esimerkissä *xsd:* on lyhenne URI:n alulle <http://www.w3.org/2001/XMLSchema#>.

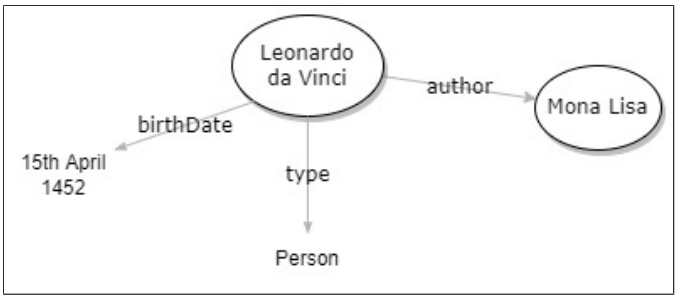
```
<http://dbpedia.org/resource/Leonardo_da_Vinci>
<http://dbpedia.org/ontology/birthDate>
"1452-04-15"^^xsd:date .
```

Seuraavasta väittämästä huomaamme, että resurssi Leonardo Da Vinci onkin tällä kertaa kolmikon objektina. Mona Lisa on itsessään aineiston resurssi, subjekti, josta viitataan toiseen resurssiin, Leonardo Da Vinciin.

[1] Käytetään myös *property*.

```
<http://dbpedia.org/resource/Mona_Lisa>
<http://dbpedia.org/ontology/author>
<http://dbpedia.org/resource/Leonardo_da_Vinci> .
```

Näemme, että RDF-aineiston kolmikoilla aineistosta muodostuu verkko (*graph*), jossa solmut (*nodes*) linkittyvät toisiinsa semanttisesti merkityillä suhteilla. Visualisoituna esimerkkiväittämistämme syntyy seuraava pieni verkko:



RDF-kolmikot muodostivat tiedosta **linkitettyä dataa** (*linked data*). Linkitetty data muodostaa **tietämysverkon** (*knowledge graph*). Yksittäiset tietämysverkot voidaan helposti linkittää toisiinsa. Yksi aineisto voi viitata toiseen, sillä URI:t ovat maailmanlaajuisia, URI:n avulla kuka vain voi viitata tai luoda linkin sen määrittämään resurssiin tai noutaa sen kuvauksen [59]. Linkitetyn datan verkot kutovat yhdessä semanttista webia, jota niin ihmiset kuin koneet voivat selata; linkitetyssä datassa yksittäisen data-elementin pohjalta on mahdollista löytää vaivattomasti muita siihen linkitettyjä resursseja kulkemalla tietämysverkossa solmusta toiseen, resurssista resurssiin [52, 46].

Yhteiskäyttö edellyttää, että käytetyt merkinnät ja niiden semantiikka ovat laajalti ymmärrettävissä. Tätä voidaan edesauttaa esimerkiksi hyödyntämällä ennaltasovittuja, laajasti käytössä olevia predikaattisanastoja tai ontologioita [65]. Esimerkiksi *Dublin Core Metadata Element Set* [14] tarjoaa viisitoista tiedon julkaisulle keskeistä ominaisuutta, kuten *Title*, *Creator* ja *Date* sekä luonnollisen kielen määritelmät niille [49]. Ontologiat (tai skeemat) taas ovat laajempia, rakenteisia kuvauksia esimerkiksi yksittäisen aineiston luonteesta, sen sisältämistä käsitteistä ja niiden välisistä suhteista [25].

2.2 Semanttisen webin standardit

Semanttisen webin kehityksen myötä tiedon julkaisemiselle ja hyödyntämiselle on kehitetty monia standardeja ja julkaistu keskeisiä aineistoja. Jo esitellyn RDF-mallin lisäksi

merkittäviä W3C:n standardeja ovat esimerkiksi ontologiakieli OWL (*Web Ontology Language*, ks. [72]), RDF-aineistojen kyselykieli SPARQL (ks. [69]) sekä SKOS-malli (*Simple Knowledge Organization System*, ks. [45]) [25].

Semanttisen webin ja linkitetyn datan parissa toimivan ja sitä kehittävän yhteisön tavoitteena on alun alkaen ollut tiedon vapaa jakaminen [25]. Tätä periaatetta korostetaan usein puhumalla **linkitetystä avoimesta datasta** (*linked open data*, LOD). Tim Berners-Lee on kehittänyt viiden tähden luokitusmallin datan julkaisulle [1]. Mallin tarkoitus on kannustaa julkaisemaan dataa verkossa mahdollisimman käyttökelpoisessa ja avoimessa muodossa [27]. Kukin tähti määrittelee tavoitteen, joiden avulla edetään kohti laadukasta, viiden tähden julkaisua [1, 27]:

- ⋆ Data on avoimesti verkossa missä vain muodossa (vaikka PDF-tiedostona).
- ⋆⋆ Data on julkaistu rakenteisessa muodossa, kuten Excel-tiedostona, koneellisen käytön helpottamiseksi.
- ⋆⋆⋆ Data on julkaistu avoimia standardeja ja formaatteja käyttäen, mikä edistää sen käyttöä eri järjestelmissä. Esim. CSV-formaatti Excelin sijaan.
- ⋆⋆⋆⋆ URI-tunnisteiden käyttö epätäsmällisten nimikkeiden sijaan, jotta dataan voidaan helposti viitata.
- ⋆⋆⋆⋆⋆ Data on linkitetty URI-tunnisteilla myös ulkopuolisiin aineistoihin, jolloin julkaistulle datalle luodaan konteksti.

Hyvönen ym. [33] ovat ehdottaneet mallin laajentamista seitsemään tähteen. Aineistojen hyödyntämisen ja laadun arvioimisen helpottamiseksi aineiston julkaisuun tulisi lisätä vielä seuraavat askeleet:

- ⋆⋆⋆⋆⋆⋆ Datan ohessa julkaistaan sen skeema ja dokumentaatio, jotta myös ulkopuoliset voisivat vaivattomasti ymmärtää ja hyödyntää sen sisältöä.
- ⋆⋆⋆⋆⋆⋆⋆ Datan laatua on arvioitu ja sen provenienssi (alkuperä, muodostustiedot, jne.) on ilmoitettu, jotta muidenkin on helpompi arvioida datan luotettavuutta.

2.3 Linkitetyn datan aineistot ja sovellukset

Yksi ensimmäisiä ja edelleen käytetyimpiä linkitetyn avoimen datan aineistoja on DBpedia [25]. DBpedian aineisto on koottu muuntamalla Wikipedian sisältöä strukturoituun muotoon, RDF-aineistoksi. Aineisto on kattava niin kooltaan kuin aihealueiltaan ja linkittyy myös muihin avoimiin linkitetyn datan aineistoihin [3]. DBpedia-aineisto on tarkasteltavissa usein eri keinoin, esimerkiksi SPARQL-kyselyillä, RDF-tiedostoja lataamalla sekä syöttämällä http-skeeman mukainen URI-tunniste verkkoselaimen osoiteriville. Vuonna 2020 DBpedian raportoitiin sisältävän yli 21 miljardia RDF-kolmikkoa [26]. Hyvin monet muut aineistot linkittyvät DBpediaan ja siitä onkin muodostunut eräänlainen linkitetyn datan solmupiste [25].

Vuonna 2012 julkaistiin toinen Wikipediaan pohjautuva hanke, Wikidata [75]. Wikimedia (jonka osa Wikipedia on) käynnisti Wikipedian ”sisarhankkeen”, jossa tietoa haluttiin saattaa rakenteiseen muotoon ja näin monipuolisemmin käytettäväksi. Wikimedia seurasi Wikipedian avoimuuden ja joukkoistamisen periaatetta, tieto on kaikkien avoimesti lisättävissä ja muokattavissa. Wikidata-aineisto on saatavilla useassa eri formaatissa, muun muassa RDF-muodossa. Nykyään myös osa Wikipediassa esitettävästä tiedosta saadaan Wikidatasta. Joukkoistuksen ansiosta Wikidata kasvaa jatkuvasti ja sitä hyödynnetäänkin laajasti erilaisissa yhteisöissä ja sovelluksissa [60]. Merkittäviä aineistoja on luonnollisesti monia muitakin, mutta kaksi edelle esiteltyä tulevat usein aiheeseen perehtyjälle ensimmäisinä vastaan.

Linkitetty data ja tietämysverkot ovat käytössä myös teollisuudessa. Noy ym. [46] kuvaavat merkittäviä teollisuuden toteutuksia: Esimerkiksi Microsoft on jo useita vuosia työstänyt laajamittaisia tietämysverkkoja. Niistä muutamia esimerkkejä ovat Bing-hakukoneelle vastauksia tuottava Bing-tietämysverkko sekä henkilöitä, yrityksiä ja työelämän taitoja yhdistävä LinkedIn-tietämysverkko. Googlen *Google Knowledge Graph* muodostuu yli 70 miljardista väittämästä ja yli miljoonasta resurssista. Valtava verkko on yli vuosikymmenen kestäneen tiedonkeruun tulos ja palvelee useita Googlen tuotteita ja ominaisuuksia. Facebook on puolestaan muodostanut maailman suurimman ”sosiaalisen verkon”. Teollisuuden linkitetyn datan aineistot eivät kuitenkaan aina ole kaikille täysin avoimia, avointa dataa, toisin kuin esimerkiksi DBpedia ja Wikidata [25].

Luotu linkitetty avoin data on jo saavutus itsessään, mutta jotta sitä voitaisiin todella hyödyntää, on sen oltava yleisesti saatavilla ja hyödynnettävissä. Esimerkiksi Wikida-

taan pääse käsiksi muuan muassa Wikidata-palvelurajapinnan sekä verkkosivujen kautta[1]. Shadbolt ym. [59] kuvaavat lähivuosien tyypillistä semanttisen webin projektia: Projektissa luodaan aihepiirille uusi ontologia. Aineisto muodostetaan joko jostain jo olemassa olevasta datasta tai kerätään itse. Valmis aineisto sijoitetaan yksittäiseen repositorioon. Aineistoa ja sen verkkoja tulkitaan ja koottu tieto jaetaan kyseisen projektin tarpeisiin kehitetyssä rajapinnassa.

Rajapinnan voi tässä tapauksessa tulkita ohjelmointirajapintaa laajemmin muinakin aineiston tarkastelun työkaluina. Esimerkiksi Suomessa on kehitetty *Linked Data Finland* -alusta *LDF.fi* [56], joka tarjoaa alustan linkitetyn avoimen datan julkaisulle ja dokumentoinnille sekä aineistojen selaamiselle [33]. Palvelun tavoitteena on automatisoida aineistojen julkaisua ja dokumentointia sekä samalla yhtenäistää käytäntöjä ja kehittää niiden laatua. Julkaistavalle aineistolle generoidaan SPARQL-kyselyrajapinnan lisäksi muun muassa kotisivu, jolla aineistojen kuvaukseen ja dokumentaation voi tutustua, ohjeet aineiston hyödyntämiseen SPARQL-kyselyillä sekä koko aineiston latauslinkit. Euroopan tasolla esimerkin datan julkaisemisesta SPARQL-kyselyrajapinnan kautta tarjoaa Europeana [35].

Aineistojen kuvailutietojen selailun lisäksi myös varsinaisen datan selaamiseen on olemassa sovelluksia, jotka mahdollistavat aineiston hyödyntämisen ilman erityisempiä teknisiä valmiuksia. Esimerkiksi British Library on muuntanut osan kansallisesta bibliografiastaan (*British National Bibliography*) linkityksi avoimeksi dataksi [11]. SPARQL-kyselyjen lisäksi aineisto on selattavissa aineiston kotisivuilla olevalla hakutoiminnolla, joka toimii tyypillisen kirjaston hakutoiminnon tavoin hakusanoilla. Aineiston selaus yksittäisillä hakusanoilla ja yksittäisiä tuloksia tarkastelemalla ei kuitenkaan vielä anna kovin kattavaa kokonaiskuvaa aineistosta. Lisäksi käyttäjän on tiedettävä mitä etsiä.

Suomessa kehitetty semanttisten portaalien Sampo-malli pyrkii kehittämään linkitetyn datan julkaisua. Sampo-malli hyödyntää *Sampo-UI*-käyttöliittymäkehystä, jonka avulla käyttäjät voivat monipuolisesti tarkastella aineistoja [34]. Julkaistavalle aineistolle luodaan oma Sampo-portaali, verkkosivusto, joka tarjoaa useita työkaluja aineiston tutkimiseen ilman tarvetta teknisille taidoille. Aineisto on esitetty ihmissilmin selkeässä muodossa ja sitä voi rajata useiden valmiiden hakufasettien avulla. Lisäksi tuloksia voidaan analysoida ja visualisoida esimerkiksi aikajanoin ja tilastoin.

Sampo-portaalissa aineisto ei ole vain helposti selattavissa, vaan portaali voi tukea myös tiedon löytämistä (*knowledge discovery*) [28]. Käyttäjän ei tarvitse lähtökohtaisesti tietää

[1] `https://www.wikidata.org/wiki/Wikidata:Main_Page`

mitä kaikkea aineisto sisältää, vaan erilaiset portaalin työkalut ja tulosten linkitykset ohjaavat käyttäjää tekemään löytöjä. Sampo-mallilla on julkaistu jo useita Sampo-portaaleja, kuten Suomen historian keskeisten henkilöiden elämänkerta-aineiston pohjalta Biografiasampo[1] [30], talvi- ja jatkosodan henkilö-, tapahtuma-, paikka-, yms. tietoja kokoava Sotasampo[2] [37] sekä keskiaikaisten eurooppalaisten käsikirjoitusten liikkeitä kartoittavan aineiston pohjalta MMM-portaali[3], *Mapping Manuscript Migrations* [29].

Euroopan unionin rahoittama *data.europa.eu*-portaali[4] puolestaan tarjoaa käyttäjäystävällisen tavan selata kokonaista aineistojen joukkoa[5]. Portaalin kautta käyttäjä voi löytää yli miljoona erilaista eurooppalaista aineistoa, jotka ovat ladattavissa linkitettynä datana [54]. Aineistoja voi hakea vaivattomasti erilaisten kategorioiden ja hakuehtojen avulla ja aineistot ovat vapaasti ladattavissa useissa eri formaateissa.

[1] `https://biografiasampo.fi/`
[2] `https://www.sotasampo.fi`
[3] `https://mappingmanuscriptmigrations.org/en/`
[4] `https://data.europa.eu/en`
[5] Portaali korvasi aiemmat *Eu Open Data Portal* ja *European Data Portal* -palvelut.

3 Parlamentaariset aineistot

Digitaalisissa ihmistieteissä parlamentaaristen aineistojen on arvioitu olevan sanomalehtiaineistojen ohella tyypillisimpiä digitoitavia aineistoja [2]. Parlamentaariset aineistot tarjoavat monipuolisia tutkimusmahdollisuuksia eri tutkimusaloille: lingvistiikka, sosiologia, politiikantutkimus, tiedonhaku, luonnollisen kielen käsittely ja niin edelleen [66]. Parlamentaariset aineistot kattavat laajoja ajanjaksoja sekä toisin kuin osa sanomalehdistä, ne ovat myös avoimesti saatavilla, mikä tekee niistä niin kansallisesti kuin kansainvälisesti arvokkaita tutkimuskohteita [50, 44].

Parlamentaarisia aineistoja on viime aikoina hyödynnetty monissa digitaalisten ihmistieteiden hankkeissa, joissa niitä on yksinkertaisen digitoinnin lisäksi muunnettu erilaisiin strukturoituihin formaatteihin sekä hyödynnetty ja julkaistu eri tavoin. Yleistyvä tutkimuksen ala on myös synnyttänyt suosituksia ja käytäntöjä parlamentaaristen aineistojen julkaisulle (ks. esim. [44, 48]). Seuraavaksi tarkastellaan näistä yhtä, Parla-CLARINia sekä erilaisia parlamentaaristen aineistojen muunnoshankkeita niin Suomesta kuin muualta maailmasta.

3.1 Parla-CLARIN

Parla-CLARIN on uusi eurooppalainen suositus parlamentaaristen aineistojen koodaukselle [23]. Suositukset muodostavat skeeman, joka lähtökohtaisesti huomioi tyypillisimmät parlamentaaristen aineistojen ominaispiirteet sekä mahdollistaa kansallisten aineistojen kansainvälisen tutkimuksen ja yhteistoimivuuden [50]. Parla-CLARIN pohjautuu TEI-suosituksiin, mikä puolestaan on XML-pohjainen malli aineistojen koodaamisesta erityisesti akateemisiin tarpeisiin [68]. Parla-CLARINia alettiin kehittää vuonna 2019 osana *CLARIN European Research Infrastructure for Language Resources and Technology* -tutkimusinfrastruktuuria (ks. [12]). CLARINin järjestämissä erilaissa hankkeissa ja tapahtumissa oltiin havaittu, että parlamentaarisia aineistoja on koodattu hyvin eri tavoin, mikä vaikeutti niiden käyttöä. Parla-CLARINia ryhdyttiin kehittämään tähän tarpeeseen [47]. Tänä päivänä Parla-CLARIN on laajasti käytössä parlamentaaristen aineistojen koodauksessa [22].

Parla-CLARIN-skeemassa, tai -formaatissa, korostuu tarjolla olevan tiedon maksimaalinen

kirjaus sekä itse koodausprosessin dokumentointi parlamentaaristen aineistojen erityispiirteet huomioiden. Formaatti sisältää lukuisia elementtejä esimerkiksi aineiston käsittelyn, kuten tavuviivoituksen poiston ja muunnoksen tekijöiden dokumentointiin. Seuraavassa on esitetty Parla-CLARINin mukaisen tiedoston karkea rakenne skeeman dokumentaation [23] mukaisesti:

```
<teiCorpus xml:lang="xx" xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader>
<!-- Common corpus metadata -->
 </teiHeader>
 <TEI xml:id="id.1">
  <teiHeader>
<!-- Document metadata -->
  </teiHeader>
  <text>
   <body>
<!-- Document text -->
   </body>
  </text>
 </TEI>
<!-- More TEI elements here -->
</teiCorpus>
```

Parla-CLARIN-muotoisen XML-tiedoston juurielementti on siis <teiCorpus>, jonka sisään varsinaiset aineistot upotetaan <TEI>-elementteinä. Yksittäinen <TEI>-elementti sisältää tyypillisimmin yhden istunnon tai päivän aineistot (*document*). PARLA-Clarin-skeeman tarkemmasta rakenteesta kerrotaan lisää luvussa 4.5.

3.2 Parlamentaaristen aineistojen muunnoshankkeet

Suomen eduskunnan aineistot, kuten pöytä- ja asiakirjat on digitoitu, eli sähköistetty, jo eduskunnan itsensä toimesta, kuten kappaleessa 4.2 tullaan havainnollistamaan. Suomessa parlamentaarisia aineistoja on kerätty ja muunnettu tutkimuskäyttöön muutamaan otteeseen aiemminkin. Aineistot ovat kuitenkin useimmiten olleet vain pieniä palasia kaikesta olemassa olevasta datasta ja lähinnä vain tutkijoille suunnattuja.

Tähänastisista hankkeista kattavuudeltaan poikkeuksellisen laaja on Salla Simolan tutkimusprojekti, jossa hän tutki poliittista puhetta ja jota varten hän keräsi suomenkielisiä eduskunnan täysistuntojen puheenvuoroja vuosilta 1907–2018. Tätä varten hän muunsi kuvantunnistustekniikalla vuosien 1907–2015 pdf-muotoiset pöytäkirjat tekstimuotoon. Kirjoitushetkellä hänen tuloksiaan on julkaistu vasta luonnoksena [62].

Toisaalla esimerkiksi Mansikkaniemi ym. [42] ovat kieliopillisesti annotoineet 2000-luvun täysistuntojen puheenvuoroja ja linkittäneet ne istuntojen videotallenteisiin. Aineisto on julkaistu Kielipankin palveluissa[1]. Suomen eduskunnan puheenvuoroja löytyy puhujatiedoilla täydennettyinä myös Harvardin kansainvälisestä *ParlSpeech*-aineistosta [55], joskin sen kattavuudessa on havaittu puutteita [32].

Myös Andrushchenko ym. [2] keräsivät ja annotoivat automatisoidusti puheenvuoroja vuosilta 1980–2018 sekä eduskunnan toteuttamia ja litteroimia kansanedustajien haastatteluja. Lisäksi he kehittävät työkaluja aineistonsa lingvistiseen tutkimukseen. Vuosien 1999–2014 puheenvuoroja on kerätty ja automaattisesti aihemallinnettu julkaisussa [41].

Maailmalla on jo toteutettu tai parhaillaan toteutetaan useita hankkeita, joissa erilaisia parlamentaarisia aineistoja muunnetaan strukturoituihin ja koneluettaviin muotoihin niiden hyödynnettävyyden parantamiseksi. Eräs merkittävä projekti oli Euroopan parlamentin istuntokeskusteluiden muuntaminen ja julkaiseminen linkitettynä avoimena datana [73]. Tähän *LinkedEP*-aineistoon muunnettiin istuntojen puheenvuorot, jotka linkittyivät tietokantaan parlamentin jäsenistä. Aineistoa rikastettiin myös muun muassa puheiden aihe-, aika- ja paikkatiedoilla. Aineisto kattaa kaikki Euroopan parlamentin istuntojen keskustelut heinäkuusta 1999 heinäkuuhun 2017.

Kansallisen tason hankkeissa linkitetyn datan aineistoja on tuottanut muun muassa alankomaalainen *PoliMedia*-hanke, jossa tutkittiin parlamentaarista keskustelua ja sen käsittelyä mediassa [36]. Hankkeessa muunnettiin Alankomaiden parlamentin keskusteluja linkitetyksi avoimeksi dataksi. Myös Latvian parlamentin keskusteluja on muunnettu linkitetyksi avoimeksi dataksi [10]. Puheista muodostunut *LinkedSaeima*-aineisto mukailee *LinkedEP*-aineistolle kehitettyä rakennetta.

Euroopan ulkopuolella parlamentaarisia keskusteluja on muunnettu linkitetyksi dataksi esimerkiksi Kanadassa [7]. Aineisto on laajasti niin tutkijoiden kuin muuten aiheesta kiinnostuneiden käytettävissä, sillä se on selattavissa *Lipad*-verkkosivuilla[2] myös ilman erityisiä semanttisen webin teknologian valmiuksia. Myös Iso-Britanniassa on kehitetty *The Hansard Viewer* -verkkosivusto[3] paikallisen parlamentaarisen aineiston selaamiseen, joskin kirjoitushetkellä sivusto vaikuttaisi olevan vielä kehitystyön alla.

Slovenian parlamentaarisia keskusteluja on puolestaan muunnettu Parla-CLARIN-muotoiseksi aineistoksi [50]. Kyseinen vuosituhannen vaihteessa valmistunut *siParl*-aineisto oli

[1] `https://www.kielipankki.fi/aineistot/`

[2] `https://lipad.ca/`

[3] `https://shinyviz.smu.edu/shiny/public/hansard-shiny/`

ensimmäinen täysimittainen Parla-CLARIN-suosituksen mukainen aineisto ja sen tarkoitus oli toimia myös esimerkkinä muille formaattia hyödyntäville. Myöhemmin Parla-CLARIN-formaattia on hyödynnetty muun muassa Islannin parlamentaaristen aineistojen muunnoksessa [66].

Muita aineiston rakenteellisia ratkaisuja on toteutettu esimerkiksi Norjassa, jossa Norjan parlamentin keskusteluja on muunnettu ja rikastettu taulukkopohjaiseksi *Talk of Norway*-aineistoksi [39].

Tuore, eurooppalainen *ParlaMint*-hanke puolestaan pyrkii kokoamaan ja harmonisoimaan lukuisia kansallisten hankkeiden tuottamia parlamentaarisia aineistoja [22]. Hanke on Parla-CLARIN-suositusten kehittäjien seuraava askel pyrkimyksessä muodostaa kansainvälinen parlamentaaristen aineistojen korpus, jossa yksittäisten maiden kontribuutiot noudattavat samaa rakennetta ja olisivat näin vaivattomasti hyödynnettävissä monipuolisesti ja kansainvälisesti. Vuonna 2021 hankkeessa oli mukana aineistoja jo 17 eri Euroopan maasta.

4 Semanttinen parlamentti

Kuten edellä todettiin, osia suomalaisista parlamentaarista aineistosta on muunnettu erilaisiin tarkoituksiin aiemminkin. Semanttinen parlamentti -hanke on nyt koonnut ja harmonisoinut Suomen eduskunnan aineistoja koko sen olemassaolojen ajalta, vuodesta 1907 alkaen. Esitellään seuraavaksi tarkemmin hanke ja sen tavoitteet.

4.1 Semanttinen parlamentti -hanke

Semanttinen parlamentti on Suomen Akatemian rahoittama konsortiohanke, joka on osa Digitaaliset ihmistieteet - DIGIHUM 2020-2022 -ohjelmaa [67]. Hanke käynnistyi vuonna 2020 ja päättyy vuoden 2022 lopussa. Hankkeessa ovat mukana Helsingin yliopiston HELDIG -keskuksessa sekä Aalto-yliopiston tietotekniikan laitoksella toimiva Semanttisen laskennan tutkimusryhmä (*Semantic Computing Research Group*, SeCo)[2] sekä Turun yliopiston Eduskuntatutkimuksen keskus[3].

Semanttinen parlamentti -hankkeessa, työnimeltään SemParl, luodaan eduskunnan tietokannoista uudenlainen linkitetyn avoimen datan palvelu ja tutkimusympäristö, Parlamenttisampo [57]. Palvelu rikastaa eduskuntadataa muilla tietolähteillä kuten biografisilla tiedoilla ja lainsäädännöllä. Palvelu rakennetaan semanttisen webin teknologioiden pohjalta ja se on suunnattu niin tutkijoille, kansalaisille, medialle kuin valtionhallinnon edustajille. Hankkeessa tutkitaan parlamentaarista, edustuksellista kulttuuria sekä poliittisen kielen käyttöä pitkällä aikavälillä. Lisäksi tarkastellaan näitä teemoja kansanedustajien, heidän verkostojensa, lainsäädäntötyön ja poliittisen kommunikaation kautta. Hanke selvittää myös uusien digitaalisten viestimien vaikutusta parlamentaariselle kulttuurille.

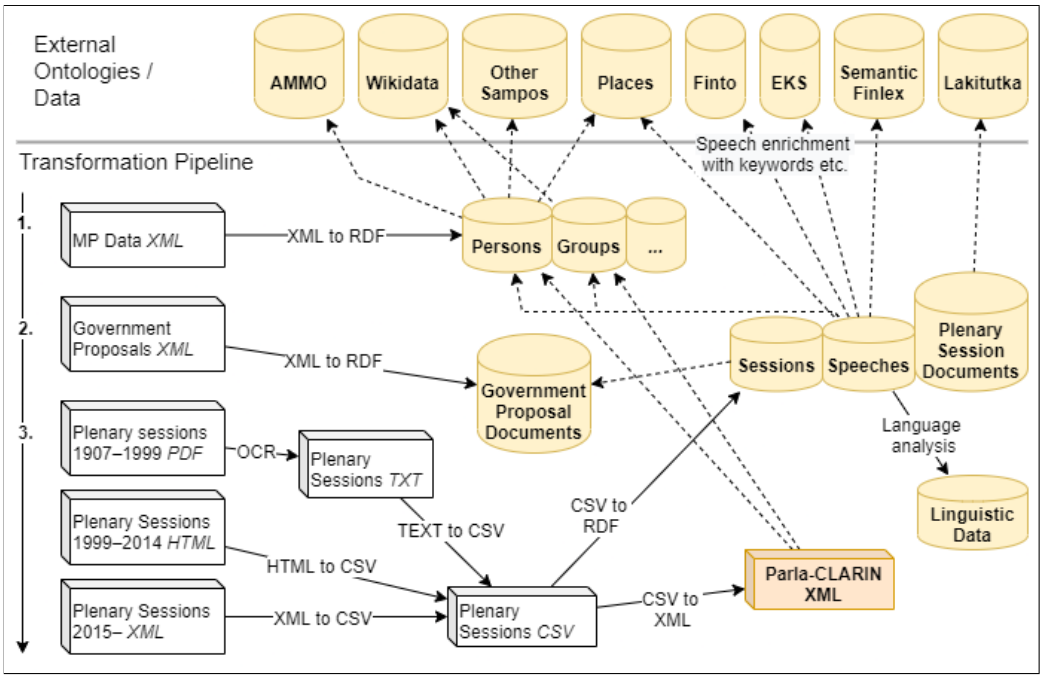
Hankkeen prosessi ja tulokset koostuvat monesta eri osasta ja vaiheesta, joita on kuvattu yleistasolla kuvassa 4.1. Kuvassa vasemmalla ovat listattuna keskeisimmät lähdeaineistot, ylhäällä ulkoiset ontologiat, joihin hankkeen tuottamia aineistoja linkitetään sekä keskellä aineiston tuoton vaiheita ja lopputuloksia. Katkoviivanuolilla havainnollistetaan eri aineistojen linkittymistä toisiinsa, kiinteillä nuolilla aineiston muunnoksen vaiheita.

SemParl-hankkeen keskiössä ovat kaksi toisiinsa linkittyvää aineistokokonaisuutta: Edus-

[2] https://seco.cs.aalto.fi/

[3] https://www.utu.fi/fi/yliopisto/yhteiskuntatieteellinen-tiedekunta/eduskuntatutkimuksen-keskus

Kuva 4.1: Semanttinen parlamentti -hankkeen ylätason prosessikuvaus. Laatija: Mikko Koho.



kunnan toimijoiden biografiset ja prosopograafiset tiedot sekä eduskunnan täysistuntojen puheenvuorot (prosessikuvauksen aineistot 1: MP (*Member of Parliament*) data ja 3: *Plenary sessions*, eduskunnan täysistunnot). Eduskunnan toimijoilla viitataan ensisijaisesti kansanedustajiin, mutta myös muihin eduskunnan pöytäkirjoissa esiintyviin henkilöihin ja virkamiehiin, kuten oikeuskanslereihin tai ministereihin, jotka eivät ole toimineet kansanedustajina. Näistä henkilöistä sekä heidän muodostamistaan ryhmistä kerättiin tietoa eri lähteistä kuten eduskunnan avoimesta kansanedustajat-tietokannasta, Biografiasammosta ja Wikidatasta.

Toinen keskeinen kokonaisuus, eduskunnan täysistuntojen puheenvuorot, kerättiin täysistuntojen pöytäkirjoista aina vuodesta 1907 alkaen. Puheenvuoroista ja niihin liittyvistä lisä- ja kuvailutiedostoista muodostettiin kokonaisuus, joka mahdollistaa laajamittaisen suomalaisen poliittisen keskustelun ja päätöksenteon tarkastelun. Puheenvuoroaineisto linkittyy puhujatietojen kautta yllä esiteltyyn eduskunnan toimijat -aineistoon.

Merkittävän, joskin huomattavasti pienemmän kokonaisuuden muodostavat myös hallituksen esitykset (aineisto 2, *Government proposals*). Hankkeessa kerättiin myös saatavilla olevat tiedot hallituksen esityksistä, jotka ovat keskeinen osa parlamentaarista keskustelua. Näillä tiedoilla rikastettiin puheaineistoa ja sen analysointimahdollisuuksia.

Eduskunnan toimijoiden tiedot kerättiin ja muunnettiin hankkeen aikana laajaksi Edus-

kunnan toimijat -tietämysverkoksi RDF-formaattiin. Prosessin toteutti Petri Leskinen ja sitä on kuvattu tarkemmin artikkelissa [40]. Tämä tutkielma keskittyy toisen kokonaisuuden, puheenvuorojen, kokoamiseen ja muuntamiseen. Hallituksen esitysten keräys ja muuntaminen oli kiinteästi yhteydessä puheaineistojen käsittelyyn ja myös se esitellään tiivisti tutkielman loppupuolella. Seuraavaksi esitellään puheenvuoroaineiston lähdeaineisto, eduskunnan täysistuntojen pöytäkirjat.

4.2 Eduskunnan täysistuntojen pöytäkirjat

Suomen eduskunta on koko historiansa ajan, vuodesta 1907 nykypäivään, laatinut täysistunnoistaan julkisia pöytäkirjoja [51]. Joskin vuosina 1915–1916 eduskunta ei kokoontunut. Pöytäkirjat sisältävät istunnon vaiheiden ja asioiden käsittely- ja päätöstietojen lisäksi myös kaikki käytetyt puheenvuorot [74]. Päätösten yhteydessä käytävän keskustelun merkitys oli Suomessa ymmärretty jo aiemminkin, sillä täysistuntojen puheenvuoroja oli alettu kirjaamaan muistiin jo 1800-luvulla [51, 74].

Lähtökohtaisesti puheenvuorot on pyritty kirjaamaan pöytäkirjoihin sanatarkasti. Pöytäkirjan laatijat korjaavat kuitenkin jossain määrin puheenvuoroja. Eero Voutilainen [74] eduskunnan pöytäkirjatoimistosta kuvaa tyypillisimpiä muutoksia: Muun muassa murteellisia fonologisia piirteitä vaihdetaan yleiskielisemmäksi (*mää, mie* → *minä*) ja empimisiä ja täytesanoja poistetaan (*öö, elikkä*). Näiden korjausten tavoitteena on parantaa puheenvuorojen luettavuutta ja ymmärrettävyyttä. Alusta alkaen pöytäkirjoissa on kuitenkin pyritty pitäytymään puhujan alkuperäisissä sanomisissa, mutta tehtyjen muutosten määrä ja luonne ovat vaihdelleet aikojen saatossa. Nykypäivänä puheenvuorojen kirjaamisprosessia ohjaavat eduskunnan pöytäkirjatoimiston sisäiset, tarkat kirjaamisohjeet [16].

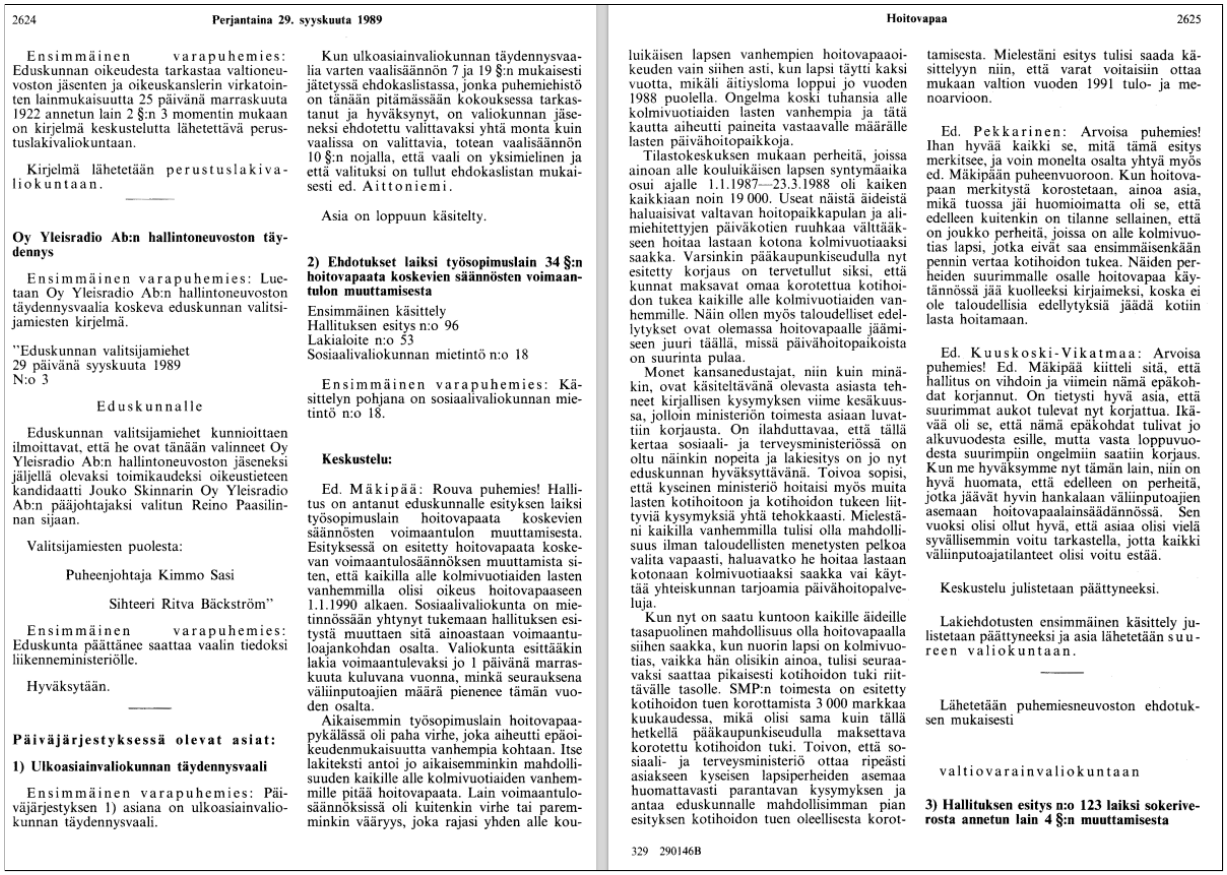
Pöytäkirjojen pääasiallinen kirjauskieli on suomi. Puheenvuorot kirjataan aina puhutulla kielellä, joka voi olla suomi tai ruotsi, eduskunnan viralliset kielet, tai kumpikin. Aikakaudesta riippuen ruotsinkielisen puheen perään on pöytäkirjassa voitu liittää puheen suomennos.

Eduskunnan täysistuntojen, ja näin ollen pöytäkirjojen, rakenne on yli satavuotisen historiansa aikana pysynyt yleisellä tasolla melko samankaltaisena. Kukin istunto alkaa ilmoitusasioilla, joiden jälkeen käsitellään päiväjärjestyksessä olevat asiat ja lopuksi julistetaan istunto päättyneeksi ja kerrotaan seuraavan istunnon ajankohta.

Yksittäinen päiväjärjestyksessä oleva asia, eli **asiakohta**, alkaa asiakohdan otsikolla, jon-

ka jälkeen luetellaan mahdolliset siihen liittyvät eduskunnan asiakirjat, kuten hallituksen esitykset ja valiokuntien mietinnöt. Istuntoa johtava puhemies alustaa aiheen, jonka jälkeen seuraa mahdollinen kansanedustajien keskustelu aiheesta. Asian käsittely päättyy puhemiehen päätöspuheenvuoroon ja/tai muuhun kirjaukseen lopputulemasta.

Kuva 4.2: Esimerkki täysitunnon pöytäkirjan rakenteesta. Tämä ja muut tutkielmassa esitellyt näytteet alkuperäisestä aineistosta ovat saatavilla CC BY 4.0 lisenssin alla.



2624 **Perjantaina 29. syyskuuta 1989**

Ensimmäinen varapuhemies: Eduskunnan oikeudesta tarkastaa valtioneuvoston jäsenten ja oikeuskanslerin virkatointen lainmukaisuutta 25 päivänä marraskuuta 1922 annetun lain 2 §:n 3 momentin mukaan on kirjelmä keskustelutta lähetettävä perustuslakivaliokuntaan.

Kirjelmä lähetetään perustuslakivaliokuntaan.

Oy Yleisradio Ab:n hallintoneuvoston täydennys

Ensimmäinen varapuhemies: Luetaan Oy Yleisradio Ab:n hallintoneuvoston täydennysvaalia koskeva eduskunnan valitsijamiesten kirjelmä.

"Eduskunnan valitsijamiehet
29 päivänä syyskuuta 1989
N:o 3

Eduskunnalle

Eduskunnan valitsijamiehet kunnioittaen ilmoittavat, että he ovat tänään valinneet Oy Yleisradio Ab:n hallintoneuvoston jäseneksi jäljellä olevaksi toimikaudeksi oikeustieteen kandidaatti Jouko Skinnarin Oy Yleisradio Ab:n pääjohtajaksi valitun Reino Paasilinnan sijaan.

Valitsijamiesten puolesta:

Puheenjohtaja Kimmo Sasi

Sihteeri Ritva Bäckström"

Ensimmäinen varapuhemies: Eduskunta päättänee saattaa vaalin tiedoksi liikenneministeriölle.

Hyväksytään.

Päiväjärjestyksessä olevat asiat:

1) Ulkoasiainvaliokunnan täydennysvaali

Ensimmäinen varapuhemies: Päiväjärjestyksen 1) asiana on ulkoasiainvaliokunnan täydennysvaali.

Kun ulkoasiainvaliokunnan täydennysvaalia varten vaalisäännön 7 ja 19 §:n mukaisesti jätetyssä ehdokaslistassa, jonka puhemiehistö on tänään pitämässään kokouksessa tarkastanut ja hyväksynyt, on valiokunnan jäseneksi ehdotettu valittavaksi yhtä monta kuin vaalissa on valittavia, totean vaalisäännön 10 §:n nojalla, että vaali on yksimielinen ja että valituksi on tullut ehdokaslistan mukaisesti ed. Aittoniemi.

Asia on loppuun käsitelty.

2) Ehdotukset laiksi työsopimuslain 34 §:n hoitovapaata koskevien säännösten voimaantulon muuttamisesta

Ensimmäinen käsittely
Hallituksen esitys n:o 96
Lakialoite n:o 53
Sosiaalivaliokunnan mietintö n:o 18

Ensimmäinen varapuhemies: Käsittelyn pohjana on sosiaalivaliokunnan mietintö n:o 18.

Keskustelu:

Ed. Mäkipää: Rouva puhemies! Hallitus on antanut eduskunnalle esityksen laiksi työsopimuslain hoitovapaata koskevien säännösten voimaantulon muuttamisesta. Esityksessä on esitetty hoitovapaata koskevan voimaantulosäännöksen muuttamista siten, että kaikilla alle kolmivuotiaiden lasten vanhemmilla olisi oikeus hoitovapaaseen 1.1.1990 alkaen. Sosiaalivaliokunta on mietinnössään yhtynyt tukemaan hallituksen esitystä muuttaen sitä ainoastaan voimaantuloajankohdan osalta. Valiokunta esittääkin lakia voimaantulevaksi jo 1 päivänä marraskuuta kuluvana vuonna, minkä seurauksena väliinputoajien määrä pienenee tämän vuoden osalta.

Aikaisemmin työsopimuslain hoitovapaapykälässä oli paha virhe, joka aiheutti epäoikeudenmukaisuutta vanhempia kohtaan. Itse lakiteksti antoi jo aikaisemminkin mahdollisuuden kaikille alle kolmivuotiaiden vanhemmille pitää hoitovapaata. Lain voimaantulosäännöksissä oli kuitenkin virhe tai paremminkin vääryys, joka rajasi yhden alle kouluikäisen lapsen vanhempien hoitovapaaoikeuden vain siihen asti, kun lapsi täytti kaksi vuotta, mikäli äitiysloma loppui jo vuoden 1988 puolella. Ongelma koski tuhansia alle kolmivuotiaiden lasten vanhempia ja tätä kautta aiheutti paineita vastaavalle määrälle lasten päivähoitopaikkoja.

Tilastokeskuksen mukaan perheitä, joissa ainoan alle kouluikäisen lapsen syntymäaika osui ajalle 1.1.1987—23.3.1988 oli kaiken kaikkiaan noin 19 000. Useat näistä äideistä haluaisivat valtavan hoitopaikkapulan ja alimiehitettyjen päiväkotien ruuhkaa välttääkseen hoitaa lastaan kotona kolmivuotiaaksi saakka. Varsinkin pääkaupunkiseudulla nyt esitetty korjaus on tervetullut siksi, että kunnat maksavat omaa korotettua kotihoidon tukea kaikille alle kolmivuotiaiden vanhemmille. Näin ollen myös taloudelliset edellytykset ovat olemassa hoitovapaalle jäämiseen juuri täällä, missä päivähoitopaikoista on suurinta pulaa.

Monet kansanedustajat, niin kuin minäkin, ovat käsiteltävänä olevasta asiasta tehneet kirjallisen kysymyksen viime kesäkuussa, jolloin ministeriön toimesta asiaan luvattiin korjausta. On ilahduttavaa, että tällä kertaa sosiaali- ja terveysministeriössä on oltu näinkin nopeita ja lakiesitys on jo nyt eduskunnan hyväksyttävänä. Toivoa sopisi, että kyseinen ministeriö hoitaisi myös muita lasten kotihoitoon ja kotihoidon tukeen liittyviä kysymyksiä yhtä tehokkaasti. Mielestäni kaikilla vanhemmilla tulisi olla mahdollisuus ilman taloudellisten menetysten pelkoa valita vapaasti, haluavatko he hoitaa lastaan kotonaan kolmivuotiaaksi saakka vai käyttää yhteiskunnan tarjoamia päivähoitopalveluja.

Kun nyt on saatu kuntoon kaikille äideille tasapuolinen mahdollisuus olla hoitovapaalla siihen saakka, kun nuorin lapsi on kolmivuotias, vaikka hän olisikin ainoa, tulisi seuraavaksi saattaa pikaisesti kotihoidon tuki riittävälle tasolle. SMP:n toimesta on esitetty kotihoidon tuen korottamista 3 000 markkaa kuukaudessa, mikä olisi sama kuin tällä hetkellä pääkaupunkiseudulla maksettava korotettu kotihoidon tuki. Toivon, että sosiaali- ja terveysministeriö ottaa ripeästi asiakseen kyseisen lapsiperheiden asemaa huomattavasti parantavan kysymyksen ja antaa eduskunnalle mahdollisimman pian esityksen kotihoidon tuen oleellisesta korottamisesta. Mielestäni esitys tulisi saada käsittelyyn niin, että varat voitaisiin ottaa mukaan valtion vuoden 1991 tulo- ja menoarvioon.

Ed. Pekkarinen: Arvoisa puhemies! Ihan hyvää kaikki se, mitä tämä esitys merkitsee, ja voin monelta osalta yhtyä myös ed. Mäkipään puheenvuoroon. Kun hoitovapaan merkitystä korostetaan, ainoa asia, mikä tuossa jäi huomioimatta oli se, että edelleen kuitenkin on tilanne sellainen, että on joukko perheitä, joissa on alle kolmivuotias lapsi, jotka eivät saa ensimmäisenkään pennin vertaa kotihoidon tukea. Näiden perheiden suurimmalle osalle hoitovapaa käytännössä jää kuolleeksi kirjaimeksi, koska ei ole taloudellisia edellytyksiä jäädä kotiin lasta hoitamaan.

Ed. Kuuskoski-Vikatmaa: Arvoisa puhemies! Ed. Mäkipää kiitteli sitä, että hallitus on vihdoin ja viimein nämä epäkohdat korjannut. On tietysti hyvä asia, että suurimmat aukot tulevat nyt korjattua. Ikävää oli se, että nämä epäkohdat tulivat jo alkuvuodesta esille, mutta vasta loppuvuodesta suurimpiin ongelmiin saatiin korjaus. Kun me hyväksymme nyt tämän lain, niin on hyvä huomata, että edelleen on perheitä, jotka jäävät hyvin hankalaan väliinputoajien asemaan hoitovapaalainsäädännössä. Sen vuoksi olisi ollut hyvä, että asiaa olisi vielä syvällisemmin voitu tarkastella, jotta kaikki väliinputoajatilanteet olisi voitu estää.

Keskustelu julistetaan päättyneeksi.

Lakiehdotusten ensimmäinen käsittely julistetaan päättyneeksi ja asia lähetetään suureen valiokuntaan.

Lähetetään puhemiesneuvoston ehdotuksen mukaisesti

valtiovarainvaliokuntaan

3) Hallituksen esitys n:o 123 laiksi sokeriverosta annetun lain 4 §:n muuttamisesta

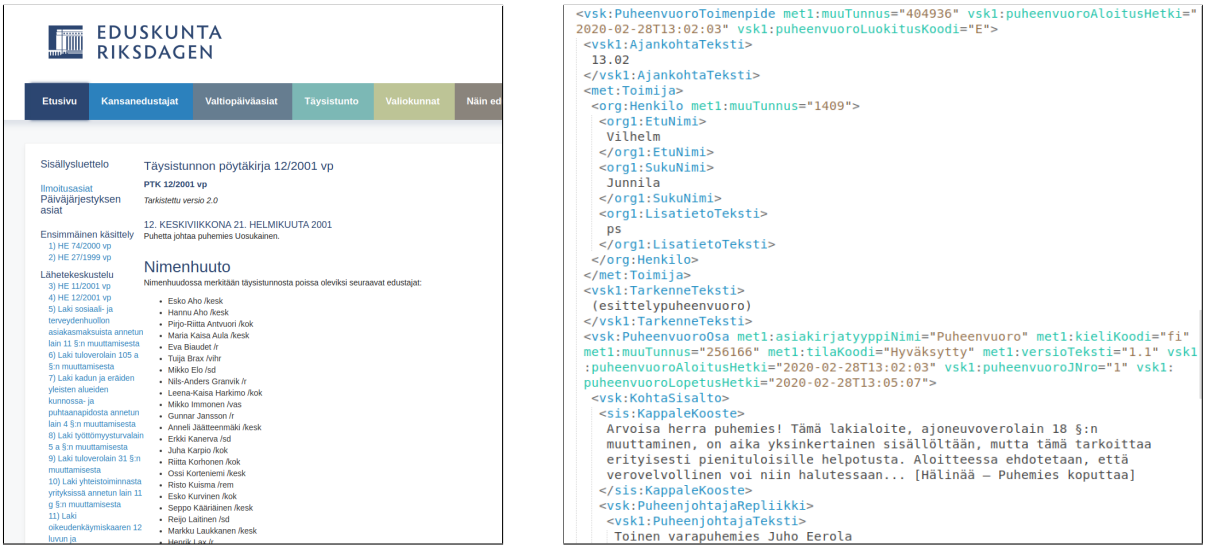
329 290146B

Kuvassa 4.2 näkyy esimerkki pöytäkirjan rakenteesta. Esimerkki on otettu täysistunnon *87/1989 vp* pdf-muotoisesta pöytäkirjasta. Käsiteltävien aiheiden ja asiakohtien otsikot on tummennettu ja keskusteluosioiden alku merkitty selkeästi **Keskustelu:**-rivillä[1]. Kunkin puheenvuoron alkuun on merkitty puhujan nimi ja rooli eduskunnassa, joita seuraa itse puheenvuoro. 1900-luvun pöytäkirjoissa puhemiehistön puheenvuorojen yhteydessä on ilmoitettu vain rooli. Mahdolliset muistiinkirjatut välihuudot ynnä muut merkittävät keskeytykset, kuten puhemiehistön kehotukset ja nuijan koputukset on upotettu puheenvuorojen sekaan sulkeissa.

[1]Paikoin käytetty myös termiä yleiskeskustelu.

Istunnot numeroidaan niin, että uusien valtiopäivien alussa aloitetaan aina alusta. Esimerkiksi pöytäkirja *PTK 15/1975 vp* kattaa vuoden 1975 valtiopäivien 15. täysistunnon. Pöytäkirjoja tutkiessa on hyvä tietää, että valtiopäivät eivät seuraa kalenterivuotta. Edeltävän vuoden valtiopäivät jatkuvat useimmiten pitkälle seuraavan kalenterivuoden puolelle. Lisäksi eduskunnan historiassa on useamman kerran pidetty samana ”vuonna” kahdet valtiopäivät, varsinaiset ja ylimääräiset (tai ensimmäiset ja toiset), mikäli hallitus hajosi kesken valtiopäivien. Siis esimerkiksi *PTK 15/1975 II vp* kattaa eri istunnon kuin *PTK 15/1975 vp*.

Kuva 4.3: Esimerkki täysistuntojen pöytäkirjoista (a) HTML- ja (b) XML-formaatissa.



(a) PTK 12/2001 vp

(b) PTK 15/2020 vp

Kaikki eduskunnan täysistuntojen pöytäkirjat ovat avoimesti saatavilla digitoidussa muodossa erinäisistä eduskunnan verkkopalveluista. Koko eduskunnan historian kattavaa aineistoa ei ole kuitenkaan saatavilla mistään yhdestä paikasta yhdenmukaisessa muodossa. 1900-luvun sekä valtiopäivien 2000 digitoidut pöytäkirjat ovat saatavissa kuvista koostettuina PDF-tiedostoina eduskunnan Avoin data -palvelusta [17]. Yhteen tiedostoon on koottu usean täysistunnon pöytäkirjat peräkkäin ja yhdet valtiopäivät koostuvat 1–8 tällaisesta tiedostosta.

Valtiopäivistä 2001 alkaen PDF-muotoiset pöytäkirjat löytyvät eduskunnan valtiopäiväasioiden ja -asiakirjojen hakupalvelusta [18]. Hakua ei voi kuitenkaan pitää täysin ongelmattomana, sillä jos hakukriteeriksi valitsee *Toimija*-kategoriasta *Täysistunto*, tarjoaa haku pöytäkirjoja vain valtiopäiviltä 2015 alkaen, kun taas ilman tätä kriteeriä löytyvät myös vuosituhannen alun täysistuntojen pöytäkirjat, mutta myös kaikki muut lähivuosien

asiakirjat.

Täysistunnosta 86/1999 vp alkaen pöytäkirjat ovat saatavilla myös HTML-muodossa istuntokohtaisilla verkkosivuilla[1]. Kukin pöytäkirja koostuu pääsivusta sekä mahdollisista erillisistä asiakohtakohtaisen keskusteluosion sisältävistä sivuista. Tuoreimpien istuntojen pöytäkirjoja lukuun ottamatta näiden pöytäkirjojen selailu on haastavaa, pöytäkirjan löytääkseen käyttäjän on käytännössä tiedettävä pöytäkirjan pääsivun tarkka verkko-osoite.

Valtiopäiviltä 2015 alkaen pöytäkirjat ovat saatavissa lisäksi vielä runsaasti metadataa sisältävässä XML-muodossa eduskunnan Avoin data -palvelusta. Pöytäkirjakohtaisia XML-tiedostoja voi tarkastella verkkosivuilta tietokantahakupalvelujen [20] kautta, joskin käyttäjältä edellytetään jälleen tarkkaa oikeiden hakuehtojen tuntemusta. Käyttäjän tulee muun muassa osata valita oikea tietokantataulu, *VaskiData*, sekä noudattaa merkkitarkkuutta (*case sensitivity*). XML-muotoiset pöytäkirjat voi myös ladata ohjelmallisesti suorina kutsuina palvelun rajapinnalle [19]. Kuvassa 4.3 on näytteet pöytäkirjoista sekä HTML- että XML-muodossa.

4.3 Pöytäkirjojen kuvailutiedot

Eri lähdeformaatin pöytäkirjat sisälsivät eriävän määrän kuvailutietoja. Modernimmat pöytäkirjat valmiiksi koneluottavissa muodoissaan tarjosivat enemmän tietoa eritoten puhujasta ja pöytäkirjasta. Taulukkoon 4.1 on koottu mitä tietoja kustakin formaatista on ollut saatavilla. Kaikissa muodoissaan pöytäkirjat ovat sisältäneet täysistunnon perustiedot: päivämäärä, tunnus sekä alku- ja loppuajat. Lisäksi puhujasta on aina kirjattu vähintään sukunimi ja rooli eduskunnassa (kansanedustajuutta ei ole aina erikseen mainittu), lisäksi puheen mahdollinen erityistyyppi, kuten *vastauspuheenvuoro* on aina ilmoitettu. Myös asiakohdat ja niihin liittyvät asiakirjat on aina ilmoitettu.

Viime vuosituhannen alkuperäiset pöytäkirjat on koottu valtiopäivittäin niteiksi, jotka alkavat listalla kyseisten valtiopäivien kansanedustajista vaalipiireittäin. Nämä listaukset sisältävät tiedon myös edustajien etunimistä, mutta koska tämän tiedon poimiminen ja yhdistäminen kunkin puheen puhujaan ei ollut ongelmatonta, laskettiin puhujan etunimen käytännössä puuttuvan tämän lähdeformaatin kuvailutiedoista.

[1] Esimerkiksi PTK 6/2011 vp:
`https://www.eduskunta.fi/FI/Vaski/sivut/trip.aspx?triptype=ValtiopaivaAsiakirjat&docid=ptk+6/2011`

Muut formaatit, HTML ja XML, sisälsivät jo puheen yhteydessä tiedon puhujan koko nimestä ja sen hetkisestä puolueesta. Näistä formaateista löytyi myös tieto pöytäkirjan kirjauksen tilasta (esim. *Hyväksytty*) ja versiosta. HTML-pöytäkirjoista löytyivät myös linkit asiakohtiin liittyvien asiakirjojen verkkoversioihin.

Taulukko 4.1: Pöytäkirjaformaateista saatavilla olleet tiedot. *HTML saatavilla täysistunnosta 86/1999 alkaen.

	PDF	**HMTL**	**XML**
Käytetty lähdeformaatti	1907-1999*	1999*-2014	2015-2021
Lähdemateriaalista löytyvät tiedot			
Puhujan etunimi	-	X	X
Puhujan puolue	-	X	X
Puhujan eduskuntatunnus	-	-	X
URL asiakohdan keskustelusivulle	-	X	-
URL asiakohtaan liittyvään asiakirjaan	-	X	-
Pöytäkirjan status	-	X	X
Pöytäkirjan versio	-	X	X
Puheen status	-	-	X
Puheen versio	-	-	X
Puheen alkuaika	-	-	X
Puheen päättymisaika	-	-	X
Puheen kieli	-	-	X
Kaikista formaateista löytyvät tiedot			
Täysistunnon päivämäärä, tunnus, alku- ja päättymisajat. Puhujan sukunimi ja rooli. Puheen tyyppi, asiakohta sekä asiakohtaan liittyvien asiakirjojen tunnukset.			

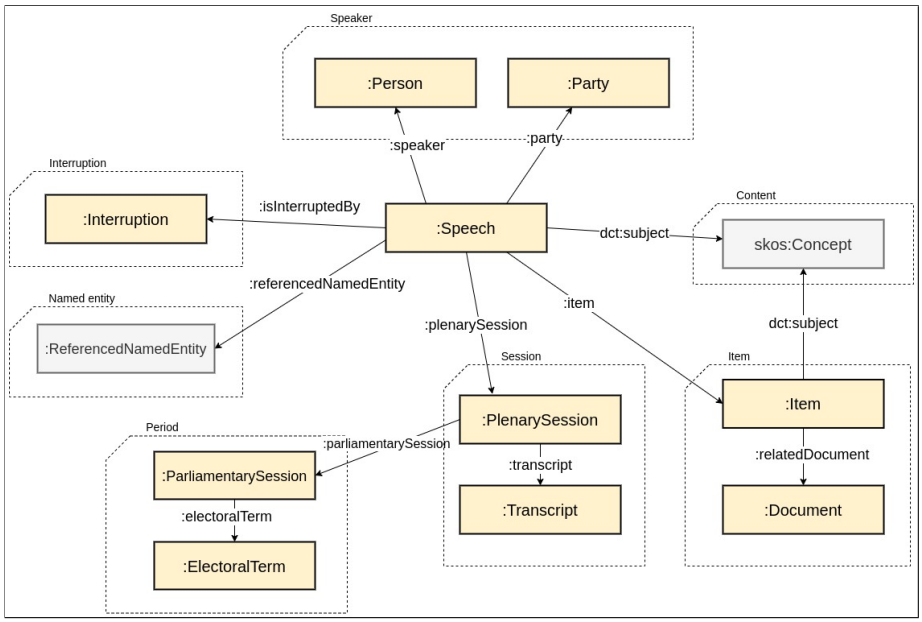
Uusimmat, XML-muotoiset, pöytäkirjat sisälsivät lisäksi vielä tarkempia puhekohtaisia tietoja: yksittäisen puheenvuoron alku- ja päättymisaika, puhekohtainen kirjauksen versio ja tila, puheen kielet sekä puhujan eduskuntatunnus. Näitä tietoja ei kuitenkaan ollut tarjolla aivan kaikille kerätyille puheenvuoroille, kuten puhemiesten kommenteille ja muille edustajien lyhyille kommenteille. XML-pöytäkirjat sisälsivät lisäksi runsaasti muita lisätietoja, kuten erilaisia sisäisiä arkistointitunnuksia, jotka eivät kuitenkaan olleet tämän tutkimuksen kannalta oleellisia.

4.4 Eduskunnan puheet -tietämysverkko

Eduskunnan puheiden kuvaaminen linkitettynä datana vaati uuden skeeman eli aineiston kuvauksen mallin. Skeeman kehityksessä on hyötykäytetty mahdollisuuksien mukaan kansainvälisiä nimiavaruuksia ja tietomalleja, kuten *<skos:prefLabel>* verkon resurssien ensisijaisen nimen predikaattina. Nimiavaruus *skos* viittaa *Simple Knowledge Organization System* (SKOS) -tietomalliin [45]. Valtaosalle verkon suhteista on kuitenkin luotu kyseisen aineiston tarpeita vastaavia uusia predikaatteja ja luokkia. Skeeman yleisin prefiksi *:* on lyhenne puheiden kuvaamista varten luodulle *<http://ldf.fi/schema/semparl/>*-

nimiavaruudelle.

Kuva 4.4: Eduskunnan puheet -tietämysverkon keskeiset luokat ja niiden väliset suhteet.



Kuvassa 4.4 havainnollistetaan skeeman keskeiset luokat sekä niiden väliset suhteet. Yksityiskohtainen skeema löytyy kokonaisuudessaan liitteestä A. Laatikoissa on skeeman keskeisiä luokkia, nuolet osoittavat tavan ja predikaatin, jolla eri luokat linkittyvät toisiinsa. Skeeman keskiössä on puheenvuoro, *:Speech*-luokka. Jokaiseen puheeseen, eli puheluokan yksikköön, on liitetty kaikki keskeinen tieto puheenvuorosta: puheen sisältö merkkijonona, aika- ja istuntotiedot, kieli, puheen tyyppi ja niin edelleen. Puhe linkittyy Eduskunnan toimijat -tietämysverkkoon ennen kaikkea predikaattien *:speaker* ja *:party* kautta. Niillä puhe linkittyy puhujaan sekä puhujan puolueeseen URI-tunnisteiden avulla.

Puheenvuoroihin kirjatut keskeytykset ja välihuudot on jätetty puheenvuorojen sisältöön alkuperäisessä kirjoitusasussaan, mutta poimittu myös omaksi luokakseen (*:Interruption*). Keskeytyksiä voi siis tarkastella omana kokonaisuutenaan, johon on mahdollisuuksien mukaan linkitetty keskeyttäjä. Puhe on aina osa jotain täysistuntoa (*:PlenarySession*), mikä puolestaan kuuluu tiettyihin valtiopäiviin (*:ParliamentarySession*) ja edelleen vaalikauteen (*:ElectoralTerm*). Käytetyn täysistunnon pöytäkirjan kirjaustiedot on tallennettu luokkaan *:Transcript*.

Puheenvuorolla on useimmiten asiakohta, johon se liittyy (*:Item*). Asiakohtaan puolestaan voi liittyä yksi tai useampi eduskunnan asiakirja (*:Document*), kuten valiokunnan mietintö tai hallituksen esitys. Puheiden sisällöt ja asiakohtien otsikot tullaan automatisoidusti asiasanoittamaan (*skos:Concept*) (predikaatin <*dct:subject*> prefiksi *dct* viittaa *Dublin*

Core Terms -kuvailutietosanastoon [14]) sekä puheen sisällöistä tullaan etsimään myös nimetyt entiteetit (*:ReferencedNamedEntity*). Näiden tietomallin luokkien toteutus on vielä kesken ja laadittu eri henkilöiden toimesta, joten ne on kuvassa 4.4 merkitty harmaalla.

Kuva 4.5: Esimerkki puheesta RDF-muodossa.

```
1  @prefix dct: <http://purl.org/dc/terms/> .
2  @prefix : <http://ldf.fi/schema/semparl/> .
3  @prefix skos: <http://www.w3.org/2004/02/skos/core#> .
4  @prefix speeches: <http://ldf.fi/semparl/speeches/> .
5  @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
6
7
8  speeches:s1998_101_8 a :Speech ;
9      :content """Arvoisa rouva puhemies! [...] Tämä pieni tarkasteluni ei ole mitenkään
10     kaiken kattava mutta haluan joihinkin keskeisiin seikkoihin kiinnittää huomiota. (Hälinää)""" ;
11     :diary <https://www.eduskunta.fi/FI/vaski/Poytakirja/Documents/ptk_101+1998.pdf> ;
12     :endDate "1998-09-09"^^xsd:date ;
13     :groupOfSpeaker <http://ldf.fi/semparl/groups/g9651517078802765164> ;
14     :isInterruptedBy speeches:in1998_101_8_1 ;
15     :item <http://ldf.fi/semparl/items/i1998101681> ;
16     :orderNumber 8 ;
17     :page 3968 ;
18     :parliamentaryRole <http://ldf.fi/semparl/groups/Oppositiopuolue> ;
19     :party <http://ldf.fi/semparl/groups/Q1138982> ;
20     :plenarySession <http://ldf.fi/semparl/times/plenary-sessions/ps_101_1998> ;
21     :roleGivenInSource "Kansanedustaja"@fi ;
22     :speaker <http://ldf.fi/semparl/people/p405> ;
23     :speakerAsInSource "Ed. Jääskeläinen" ;
24     :speechType <http://ldf.fi/semparl/speechtypes/Puheenvuoro> ;
25     :yearOfSpeech 1998 ;
26     dct:date "1998-09-09"^^xsd:date ;
27     dct:language <http://id.loc.gov/vocabulary/iso639-2/fin> ;
28     skos:prefLabel "Vp 1998 - istunto 101 - puhe 8 (Jouko Jääskeläinen)"@fi .
29
```

Kuvassa 4.5 esitetään esimerkki skeeman toteutumisesta: eräs puheenvuoro skeeman mukaisena linkitettynä datana (puheen sisältöä lyhennetty). RDF-muotoinen linkitetty data on koodattu Turtle-syntaksilla. RDF-kolmikkojen kirjaamiseen kehitetty Turtle mahdollistaa kolmikkojen kuvaamisen tiivisti ja ihmissilmälle ystävällisesti lyhenteitä eli prefiksejä hyödyntäen [6]. Riveillä 1–5 määritellään aineiston kuvaamisessa käytetyt prefiksit ja *Speech*-luokan yksilö (*instance*) alkaa riviltä kahdeksan. Rivillä kahdeksan määritellään puhe *speeches:s1998_101_8*, joka on tyyppiä *:Speech*. *a* on lyhenne predikaatista *http://www.w3.org/1999/02/22-rdf-syntax-ns#type*, eli se siis määrittelee subjektin *speeches:s1998_101_8* luokan. Täysin aukikirjoitettuna kyseinen rivi, tai kolmikko, kuuluisi:
<http://ldf.fi/semparl/speeches/s1998_101_8>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://ldf.fi/schema/semparl/Speech> ; .
Prefiksien ja Turtle-syntaksin edut käyvät nopeasti ilmeisiksi. Riveillä 9–28 määritellään

kyseisen puheen muut ominaisuudet ja niiden arvot. Turtle-syntaksissa subjektia ei määrittelyn jälkeen tarvitse toistaa joka rivillä, joten sisennetyillä riveillä ilmoitetaan vain kunkin väittämän predikaatti ja objekti. Jokaisen rivin väittämä tulkitaan kuitenkin kokonaisena kolmikkona.

Puheet, asiakohdat ja istunnot muodostavat Eduskunnan puheet -tietämysverkon selkärangan. Verkko linkittyy erilaisten rikastamistoimenpiteiden ja luokkien ominaisuuksien kautta laajasti muihin tietolähteisiin, kuten Eduskunnan toimijat -tietämysverkkoon, eduskunnan verkkopalveluihin ja niin edelleen. Aineiston rikastamis- ja linkitystoimenpiteistä tarkemmin luvussa 5.

4.5 Eduskunnan puheet Parla-CLARIN-formaatissa

Eduskunnan puheet -aineisto on tuotettu kahtena eri versiona, juuri esiteltynä linkitetyn avoimen datan RDF-tietämysverkkona sekä XML-muotoisena aineistona, joka on koodattu Parla-CLARIN suositusten mukaan. Parla-CLARIN-muotoinen aineisto on jossain määrin itsenäisempi kokonaisuus kuin RDF-muotoinen aineisto, joka linkittyy laajasti erilaisiin ulkoisiin lähteisiin. Parla-CLARIN-aineistossa on myös hyödynnetty Eduskunnan toimijat -aineiston henkilötietoja, mutta tietämysverkon URI-tunnisteiden sijaan aineistosta on poimittu vain tekstuaaliset tiedot. Puhujille luodaan sen sijaan tiedostokohtaiset *xml:id*-tunnisteet. Tunnisteet rakentuvat henkilöiden nimien pohjalta, joten mikäli nimi ei muutu, tunnisteet toistuvat samanlaisina tiedostosta toiseen, joten yksittäisiä henkilöitä voi hakea koko aineistosta yhdellä tunnisteella.

Jokainen Parla-CLARIN-tiedosto muodostaa itsenäisen kokonaisuuden, yhdet valtiopäivät puheineen ja kuvailutietoineen. Tässä formaatissa kuvailutiedot ovat suppeammat, sillä kaikkia RDF-aineistoon liitettyjä rikastetietoja ei aineistoversion keskeisten käyttäjien kesken koettu tarpeellisiksi, esimerkiksi valtiopäivien alku- ja päättymispäivämäärät, erilliset puheen järjestysnumero ja pitovuosi sekä linkit asiakirjojen verkkoversioihin. Aineistosta löytyvät kuitenkin kaikki pöytäkirjoissa esiintyneet tiedot: ajankohta, istunto, asiakohta, asiakirjat, puheen tyyppi, puhujan nimi ja rooli. Kuitenkin myös Parla-CLARIN versiossa puheisiin on liitetty verkko-osoite alkuperäiseen pöytäkirjaan sekä automaattisesti tunnistettu puheen kieli.

Parla-CLARIN XML -tiedoston juurielementti on <teiCorpus>. Elementti sisältää ensimmäisenä <teiheader>-elementin, joka sisältää tiedoston kuvailutiedot: otsikon ja tietoa sisällöstä, esimerkiksi tiedoston puheenvuorojen lukumäärän. Elementti sisältää myös kes-

keiset <listPerson>- ja <listOrg>-elementit, joissa listataan kaikki tiedostossa esiintyvät puhujat sekä heidän puolueensa sekä näiden *xml:id*-tunnisteet. Puheen yhteydessä puhuja määritellään näillä tunnisteilla puheen *who*-atribuutin kautta. Seuraavassa typistetty esimerkki:

```
<teiCorpus xml:id="Speeches_2019" >
  <teiheader>
    ...
    <profileDesc>...</profileDesc>
      <settingDesc>...</settingDesc>
      <particDesc>
        <listPerson>
          <person xml:id="Satu_Hassi">
            <persName>
              <surname>Hassi</surname>
              <forename>Satu</forename>
              <roleName>Kansanedustaja</roleName>
            </persName>
            <sex value="F">Female</sex>
            <birth when="1951-01-01"/>
            <affliation ref="#party.VIHR"/>
          </person>
        </listPerson>
        <listOrg>...</listOrg>
    </profileDesc>
  </teiheader>
  <TEI xml:id="ptk_10_2019">
    <teiHeader>...<teiHeader>
    <text>
      <body>
        <div>
          <head>...<head>
          <div>
            <note speechType="Puheenvuoro" xml:id="2019_10_10"/>
            <u next="2019_10_10.2" who="#Satu_Hassi" xml:id="2019_10_10.1">Arvoisa
            puhemies! Tässä nyt keskusteltiin Rinteen todennäköisesti tulevan hallituksen
            ohjelmasta.[...] tutustumaan siihen varsinaiseen tekstiin, joka on julkaistu.
            </u>
            <vocal>
              <desc>Välihuutoja</desc>
            </vocal>
            <u prev="2019_10_10.1" who="#Satu_Hassi" xml:id="2019_10_10.2"> Enää ei [...]
            perusteellisen ruotimisen aika on sitten ensi viikolla.
            </u>
            <vocal who="Jukka_Gustafsson">
              <desc>Jukka Gustafsson: Hyvä puhe!</desc>
            </vocal>
          </div>
          <div> <!-- Lisää puheita !--> </div>
```

<teiHeader>-elementin jälkeen tulevat varsinaiset puheenvuorot <TEI>-elementtien sisällä. Yksi <TEI> pitää sisällään yhden täysistunnon puheenvuorot. <TEI>-elementti

alkaa jälleen <teiheader>-elementillä, jossa kerrotaan istuntokohtaisia kuvailutietoja. Puheenvuorot löytyvät <text>-elementtiin kääritytyn <body>-elementin sisältä asiakohdittain -elementteinä. Varsinainen yksittäinen puheenvuoro muodostaa vielä oman -elementtinsä, jonka sisällä <note> sisältää puheen kuvailutietoja, <u> itse puheen sisällön (*utterance*) ja <vocal> mahdolliset puheen keskeytykset. Puheet on pilkottu niin, että <u> sisältää vain varsinaisen puhujan puheen sisältöä. Mikäli puheenvuoron sisään on kirjattu esimerkiksi välihuuto, esitetään puheenvuoro useamassa <u>-elementissä, joiden väliin on merkitty keskeytys <vocal>-elementtinä. Konsortion sisäisessä käytössä ollut Parla-CLARIN-tiedostojen rakenteen epävirallinen dokumentaatio on nähtävillä kokonaisuudessaan liitteenä B.

4.6 Aineiston julkaisu

Semanttinen parlamentti -aineisto, niin puheet kuin eduskunnan toimijoiden tiedot, on julkaistu kokonaisuudessaan hankkeen sisäiseen käyttöön vuonna 2021. Julkiseen käyttöön aineisto avataan vuoden 2023 alussa CC BY 4.0 -lisenssillä. Aineisto on saatavilla FinnParla Linked Open Data -palveluna jo esitellyllä Linked Data Finland -alustalla. Tästä palvelusta linkitetty data on saatavissa SPARQL-kyselyinä palvelun rajapintaan, mikä on keskeinen aineiston käyttötapa analyysejä varten [31]. Lisäksi aineisto voidaan ladata omalle laitteelle niin linkitettynä datana kuin Parla-CLARIN XML -muodossa.

LDF.fi-alusta noudattaa aineiston julkaisun seitsemän tähden mallia [33], joten aineiston käytön ja arvioinnin tueksi julkaisulle on laadittua käyttöä tukevaa dokumentaatiota. Julkaisun kotisivua[1] varten on laadittu aineiston skeema sekä kuvaus SPARQL-palvelusta (*SPARQL Service Description*). Kuvaus määrittelee tarkasti aineiston muodostavat tietämysverkot sekä muun muassa aineiston luojat, lisenssit, oikeuksienhaltijat, rajapinnan osoitteen, nimiavaruuden sekä julkaisun arvosanan seitsemän tähden asteikolla. Yksittäisistä verkoista ilmoitetaan myös esimerkiksi lyhyt kuvaus sisällöstä, nimiavaruus, käytetyt sanastot sekä käytetyt lähteet.

SPARQL [69] on kyselykieli RDF-muotoisille aineistoille. SPARQL-kyselyitä voi suorittaa esimerkiksi verkkoselaimessa YASGUI-alustalla[2], joka mahdollistaa kyselyiden helpon editoinnin, mutta myös tulosten tarkastelun ja tallennuksen omalle laitteelle [71]. Kyselyt onnistuvat myös ohjelmallisesti, esimerkiksi Python-ohjelmalla SPARQLWrapper-kirjastoa

[1] `http://www.ldf.fi/dataset/semparl`

[2] `https://yasgui.triply.cc/`

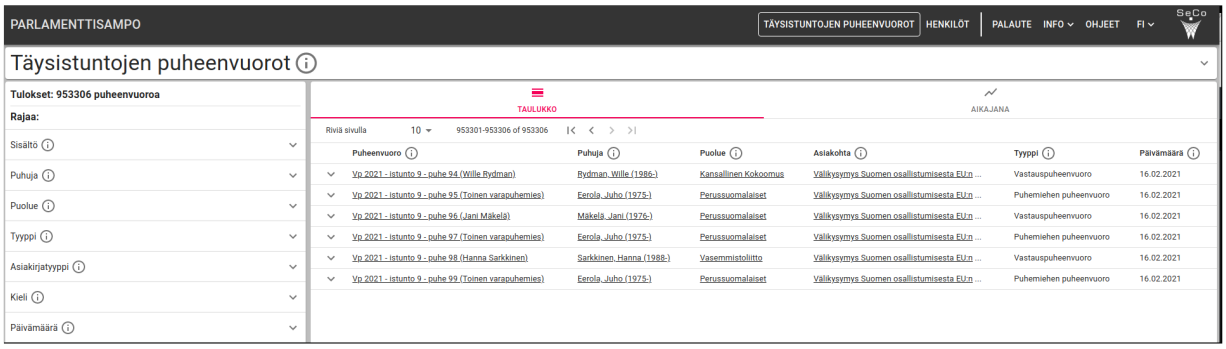
(ks. [64]) hyödyntäen. SPARQL-kielellä voidaan muodostaa kyselyjä, joilla voidaan tutkia aineistoa hyvinkin kattavasti. Alla olevalla kyselyllä aineistosta voidaan esimerkiksi löytää kaikkien tiettyyn puolueeseen kuuluneiden naiskansanedustajien puheet tietyltä ajanjaksolta.

```
PREFIX bioc: <http://ldf.fi/schema/bioc/>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX semparl: <http://ldf.fi/schema/semparl/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT DISTINCT ?speech
WHERE {
  VALUES ?party { <http://ldf.fi/semparl/groups/Q1138982> }
  VALUES ?gender { <http://schema.org/Female> }
  ?speech
        semparl:party ?party ;
        semparl:speaker ?speaker ;
        dct:date ?date .
  ?speaker bioc:has_gender ?gender .
  FILTER(?date > "1979-01-01"^^xsd:date && ?date <= "1999-12-31"^^xsd:date)
}
ORDER BY ?date
```

Datapalvelun pohjalta on rakennettu myös semanttinen portaali, Parlamenttisampo (kuva 4.6) [32]. Portaali on rakennettu semanttisen webin periaatteiden pohjalta ja on osa luvussa 2 esiteltyä semanttisten Sampo-portaalien sarjaa. Parlamenttisammossa käytetään aineiston linkitetty data -versiota. Parlamenttisampo-verkkosivulla kuka tahansa voi vaivattomasti tarkastella aineistoa, niin puheita kuin henkilötietoja, ilman erityisiä teknisiä taitoja. Portaali tarjoaa käyttäjälle useita aineiston rajauksen ja analysoinnin työkaluja, kuten fasettihaun puhujan, puolueen ja päivämäärän perusteella sekä sanahaun puheen sisällöistä (kuvassa portaalin vasemmassa reunassa).

Kuva 4.6: Parlamenttisammon puheenvuorot -näkymä. Sivustoa kehitetään edelleen, joten näkymä voi poiketa lopullisesta versiosta.



Hakua vastaavat puheet listataan käyttäjälle selkeänä listana, yksi puhe per rivi (kuvassa keskellä). Yksittäistä puhetta klikkaamalla käyttäjä pääsee puheen kotisivulle, jossa listataan puheen sisällön lisäksi kaikki siihen liittyvät kuvailutiedot. Aineiston linkitetyn luonteen mukaisesti esimerkiksi puheen puhujan nimeä klikkaamalla käyttäjä pääsee suoraan puhujan kotisivulle tarkastelemaan tämän henkilökohtaisia tietoja. Henkilösivulta löytyvät niin ikään linkit kaikkiin kyseisen henkilön puheenvuoroihin. Käyttäjä voi lisäksi tarkastella aineistoja aikajanalla sekä muiden kuvaajien avulla. Myös Parlamenttisampo avataan yleiseen käyttöön vuoden 2023 alussa.

Julkaisunsa myötä aineisto yhdistyy ontologiainfrastruktuuriltaan osaksi laajempaa Linked Open Data Infrastructure for Digital Humanities in Finland (LODI4DH) -hanketta rakentaen kansallista digitaalisten ihmistieteiden linkitetyn datan tietoinfrastruktuuria [31]. Myös hankkeessa tuotettu lähdekoodi tullaan julkaisemaan avoimeen käyttöön aineiston julkaisun yhteydessä.

5 Eduskunnan täysistuntojen puheenvuorojen muuntaminen

Eduskunnan puheet -aineiston keskeisin yksikkö on **puhe** (*speech*). Puhuttaessa eduskunnan puheista ja puheenvuoroista keskitytään usein nimenomaan pöytäkirjojen *Keskustelu*-osioiden puheenvuoroihin, jotka monesti ovatkin sisällöltään antoisimpia ja tutkimuksen kohteena. Tässä hankkeessa pyrittiin kuitenkin keräämään täysistuntojen puheet laajimmassa merkityksessään. Pöytäkirjoista kerättiin kaikki niihin merkityt lausunnat, niin varsinaiset keskustelupuheenvuorot kuin puhemiehistön istunnon kulkua ohjaavat kommentit sekä edustajien lyhyet esimerkiksi äänestykseen liittyvät huomautukset. Kaikki nämä erilaiset puheenvuorot kulkevat yksikön *puhe* alla. Tällä ratkaisulla haluttiin varmistaa, että kaikki mahdollinen tieto istuntojen etenemisestä saadaan talteen. Yksittäisten puheiden lisäksi myös niihin upotetut välihuudot on poimittu vielä omiksi kokonaisuuksikseen.

Täysistuntojen puheenvuorojen kerääminen koko Suomen eduskunnan historian ajalta vaati monia erilaisia lähestymistapoja. Kuten aiemmassa luvussa kuvattiin, lähdeaineisto koostui erilaisista formaateista ja lähteistä. Prosessin helpottamiseksi ja tulosten laadun parantamiseksi lähdeaineistona käytettiin valmiiksi koneluettavaa versiota (HTML/XML) aina kun mahdollista. Näistä kahdesta hyödynnettiin mieluiten XML-muotoista aineistoa, joka oli jo valmiiksi pitkälle käsiteltyä ja sisälsi formaateista eniten valmiita kuvailutietoja. Seuraavaksi käydään läpi kuinka eri aikakausien puheet koottiin ja muunnettiin.

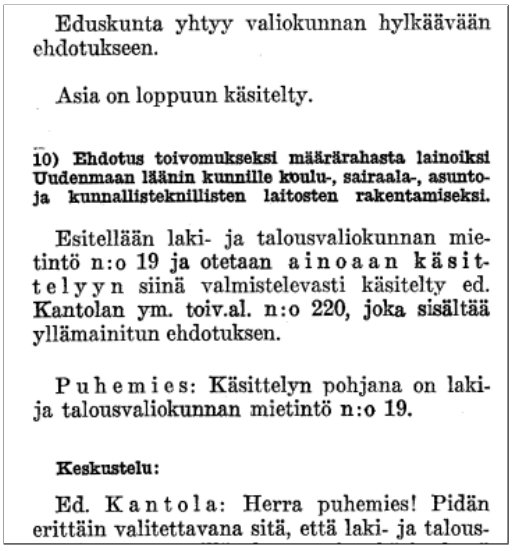
5.1 PDF-pohjaisten puheiden kerääminen

Alkuperäisten pöytäkirjojen skannatuista kuvista useamman pöytäkirjan PDF-tiedostoiksi koottu lähdeaineisto kattoi selkeästi suurimman ajanjakson kerättävästä aineistosta, valtiopäiviltä 1907 aina valtiopäivien 1999 puoliväliin. Lisäksi sen keräys- ja muunnosprosessi oli kaikista kolmesta lähtöaineistosta haastavin ja monimutkaisin.

Puheiden kerääminen

PDF-muotoinen aineisto tuli ensimmäiseksi muuntaa koneluettavaan muotoon, joten kaikki PDF-tiedostot muunnettiin ensin tekstintunnistuksella (OCR, *optical character recognition*). Muunnoksen toteutti Senka Drobac Tesseract 5 -työkalua hyödyntäen, prosessi on kuvattu tarkemmin artikkelissa [63]. Muunnettu pöytäkirja-aineisto oli tarjolla kahtena eri versiona: tekstintunnistuksen tulosten standardoidussa kuvausformaatissa hOCR-tiedostoina [4] sekä tekstitiedostoina. Näistä käyttöön otettiin tekstitiedostot niiden yksinkertaisuuden ja (ihmissilmin) luettavuuden takia. Tämä helpotti tulosten manuaalista evaluointia sekä tietojen poimintaan käytettävien säännöllisten lausekkeiden muotoilua. Toisaalta valinnalla menetettiin mahdollisia alkuperäisen pöytäkirjan muotoilutietoja.

Kuva 5.1: Esimerkki pöytäkirjojen tekstintunnistusprosessin tuloksista (PTK 49/1967, sivu 885).



Eduskunta yhtyy valiokunnan hylkäävään ehdotukseen.

Asia on loppuun käsitelty.

10) Ehdotus toivomukseksi määrärahasta lainoiksi Uudenmaan läänin kunnille koulu-, sairaala-, asunto- ja kunnallisteknillisten laitosten rakentamiseksi.

Esitellään laki- ja talousvaliokunnan mietintö n:o 19 ja otetaan ainoaan käsittelyyn siinä valmistelevasti käsitelty ed. Kantolan ym. toiv.al. n:o 220, joka sisältää yllämainitun ehdotuksen.

Puhemies: Käsittelyn pohjana on laki- ja talousvaliokunnan mietintö n:o 19.

Keskustelu:

Ed. Kantola: Herra puhemies! Pidän erittäin valitettavana sitä, että laki- ja talous-

(a) Alkuperäinen PDF

```
Eduskunta yhtyy valiokunnan hylkäävään
ehdotukseen.

– Asia on loppuun käsitelty.

10) Fhdotus toivomukseksi määrärahasta lainoiksi
Uudenmaan läänin kunnille koulu-, sairaala-, asunto-
ja kunnallisteknillisten laitosten takentamiseksi.

Esitellään laki- ja talousvaliokunnan mie-
tintö n:o 19 ja otetaan ainoaan käsit-
telyyn siinä valmistelevasti käsitelty ed.
Kantolan ym. toiv.al, n:o 220, joka sisältää
yllämainitun ehdotuksen.

Puhemies: Käsittelyn pohjana on laki.
ja talousvaliokunnan mietintö n:o 19.

Keskustelu:

Ed. Kantola: Herra puhemies! Pidän
erittäin valitettavana sitä, että laki- ja talous-
```

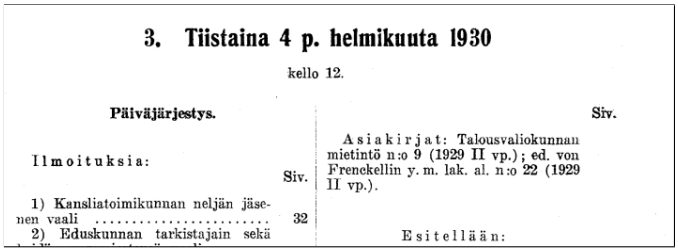
(b) Tekstintunnistuksella tuotettu tekstitiedosto

Tekstintunnistusprosessin tulokset olivat pääosin varsin hyvälaatuisia. Pöytäkirjoissa olleet sisällysluettelot, taulukot ja muut rakenteeltaan poikkeukselliset elementit sisälsivät paljon virheitä, mutta varsinaisten keskusteluosioiden muunnokset olivat keskimäärin hyvin luotettavia, joskaan eivät täydellisiä. Kuvassa 5.1 näkyy melko hyvin tulosten keskiarvoa edustava esimerkki. Kuvassa vasemmalla on kuvakaappaus alkuperäisestä pöytäkirjan PDF-versiosta ja oikealla sama kohta muunnettuna tekstitiedostona. Muunnos on pääosin erittäin tarkka, mutta sisältää pieniä virheitä, kuten hyvin tyypillisen E-merkin korvautumisen F-merkillä asiakohdan otsikossa. Lisäksi kolmannen tekstirivin alla ollut pieni painotahra on tulkittu ajatusviivaksi rivin alkuun ja puhemiehen kommentissa rivin

päättävä tavuviiva on muuttunut pisteeksi. Muunnosvirheistä kiitettävä osa oli niin säännöllisiä, että ne olivat kohtuullisella vaivalla ennakoitavissa puheita poimittaessa, eritoten merkkivaihdokset (esim. E → F, å → ä, : → ; jne.).

Muunnostulosten laatu vaihteli vuosikymmeneltä toiselle. 1990-luvun aineistojen tulokset olivat erittäin hyvälaatuisia, kun taas vuosisadan alun pöytäkirjojen muunnokset sisälsivät selkeästi enemmän virheitä. Merkittävät tulosten laatuun vaikuttavat tekijät olivat alkuperäisten kuvien laatu sekä pöytäkirjojen painamiseen käytetty paperi, joka on ajoittain ollut hyvin heikkolaatuista. La Mela [38] on tehnyt samankaltaisia havaintoja omissa tutkimuksissaan eduskunta-aineistojen parissa.

Kuva 5.2: Pöytäkirjan otsikkorivi.



3. Tiistaina 4 p. helmikuuta 1930

kello 12.

Päiväjärjestys. Siv.

Ilmoituksia: Siv.

1) Kansliatoimikunnan neljän jäsenen vaali 32

2) Eduskunnan tarkistajain sekä

Asiakirjat: Talousvaliokunnan mietintö n:o 9 (1929 II vp.); ed. von Frenckellin y. m. lak. al. n:o 22 (1929 II vp.).

Esitellään:

Yksittäinen merkittävä haaste tekstintunnistusmuunnoksessa olivat pöytäkirjojen otsikkorivit. Jokaisen täysistunnon pöytäkirja alkaa selkeästi strukturoidulla rivillä, joka sisältää keskeistä tietoa istunnosta: sen järjestysnumeron ja päiväyksen. Kuten kuvasta 5.2 näkyy, siinä missä muu asiakirja on aseteltu kahteen sarakkeeseen, venyy otsikkorivi kummankin sarakkeen puolelle. Tämän seurauksena otsikkorivit olivat hyvin usein leikkaantuneet muunnoksessa kahdelle eri riville tai muuten korruptoituneet. Useita peräkkäisiä pöytäkirjoja sisältävässä tiedostossa näiden otsikkorivien paikantaminen oli keskeistä, jotta niitä seuraavat puheenvuorot liitettäisiin oikeaan istuntoon. Jotta lopullisen puheaineen tuloksia voitaisiin istuntojen osalta pitää luotettavana, päätettiin otsikkorivit korjata manuaalisesti. Tekstitiedostot käytiin läpi otsikkorivejä tunnistamaan kirjoitetun Python-ohjelman avulla ja kaikki puuttuvat otsikot paikannettiin ja korjattiin.

Otsikkorivien korjauksien yhteydessä havaittiin, että valtiopäivien 1908 pöytäkirjatiedostoissa oli joukko istuntoja kahteen kertaan kahdessa eri tiedostossa. Nämä kaksoiskappaleet poistettiin aineistosta. Lisäksi ylimääräisten valtiopäivien 1932 ja 1935 pöytäkirjat on eduskunnan palveluissa arkistoitu muista pöytäkirjoista poiketen 'Asiakirjat'-kategorian alle 'Pöytäkirjat'-kategorian sijaan. Näiden kaksien valtiopäivien pöytäkirjat oli koottu tiedostoiksi, jotka poikkeuksellisesti sisälsivät myös kyseisten valtiopäivien muita asiakirjoja. Näistä tiedostoista käytettiin vain varsinaiset pöytäkirjat kattavat osiot. Vuoden 1917 varsinaisten ja ylimääräisten valtiopäivien välissä pidettiin kaksi epävirallista, mutta historiallisesti merkittävää hallinnon merkkihenkilöiden istuntoa, joiden pöytäkirjat on

liitetty valtiopäivien koontitiedostoon. Nämä kaksi alunperin numeroimatonta istuntoa numeroitiin, jotta niille voitaisiin luoda muita istuntoja vastaavat tunnisteet.

Kun tekstintunnistuksella tuotettu tekstiaineisto oli valmisteltu, voitiin aloittaa varsinainen tietojen kerääminen. Koska tekstitiedostot eivät sisältäneet muita alkuperäisiä muotoiluvihjeitä kuin kappalevälit, perustui aineiston kerääminen pöytäkirjojen rakenteen tuntemukseen ja säännöllisiin lausekkeisiin. Aineiston keruuta varten luotiin Python-ohjelma, joka käy tiedostot läpi rivi riviltä ja erilaisia ehtoja vasten erottelee aineistosta aika- ja puhujatiedot, asiakohtien otsikot, mainitut asiakirjat, itse puheenvuorot sekä sivunumerot.

Koska puheita keräävän ohjelman tuli käsitellä hyvin erilaisia rakenteita ja tietoja sekä sietää erilaisia poikkeuksia, kasvoi ohjelman toimintalogiikka varsin nopeasti laajaksi ja monimutkaiseksi. Sen seikkaperäinen ymmärrys vaatii selkeyteen pyrkivästä nimeämisestä ja koodin kommentoinnista huolimatta lukijaltaan lähdemateriaalin tuntemusta, kuten seuraavasta ohjelman otteesta on havaittavissa.

```
for i in range(len(rows)-1):
    if end_compendium(rows[i]):
        break
    if page_header(rows[i]):
        if not document_start(rows[i]):
            if page == -1:
                page = handle_page_num(rows[i], parliament_year)
                if page != -1:
                    page -= 1
        if page != -1:
            page += rows[i].count('\f')
    if document_start(rows[i]):
        if current_speech:
            clean_content = edit_content(current_speech)
            speaker, content = get_speaker('␣'.join(clean_content))
    ...
```

Sama päti erilaisiin säännöllisiin lausekkeisiin, joita vasten ohjelman toimintalogiikan eri ehdot tarkistetaan. Jotta aineistosta saataisiin poimittua halutut tiedot mahdollisimman kattavasti, pyrittiin lausekkeissa huomioimaan tyypillisimmät tekstin muunnoksessa tai alkuperäisessä kirjauksessa tapahtuneet virheet ja poikkeamat. Esimerkiksi kansanedustajan puheen aloitusrivin ”perustapauksenkin” tunnistaminen vaatii melkoista taiteilua:

```
def speech_starters(row, row2, row3):
    speech_start = re.compile("^[E|F]d[\.,]␣?(af|von|v\.)?␣?[A–ZÅÄÖ].*[:;]")
    speech_start2 = re.compile(
        "^[A–ZÅÄÖ][^(:]*iniste[rt]i␣(af|von)?␣?[A–ZÅÄÖ].*[:;]")
    long_title = re.compile(
        '^[A–ZÅÄÖ][^(:]*iniste[rt]i␣(af|von)?␣?[A–ZÅÄÖ].*–$')
    ...
    if speech_start.match(row) or speech_start2.match(row) \
        ...
        return True
    return False
```

Tekstipohjainen aineisto kattoi pöytäkirjat lähes sadan vuoden ajalta. Luonnollisesti niin pitkän aikajakson aikana pöytäkirjojen käytännöt eivät ole pysyneet identtisinä vaikka yleistason rakenne onkin melko vakaa. Esimerkiksi asiakohtaan liittyvät asiakirjat kirjattiin 1900-luvun alussa lausemaisena kokonaisuutena, kun taas vuosisadan lopulla ne kirjattiin selkeänä listana.

Kuva 5.3: Esimerkki pöytäkirjojen muutoksesta asiakohtaan liittyvien asiakirjojen merkitsemisessä.

```
Päiväjärjestyksessä olevat asiat:
1) Ehdotuksen asetukseksi työnvälitystoimesta

sisältävä suuren valiokunnan mietintö n:o 3,
joka tämän kuun 2 päivänä pantiin pöydälle
tähän täysi-istuntoon, esitetään ja otetaan toi-
seen käsittelyyn siinä sekä työväenasiain-
valiokunnan mietinnössä n:o 1 valmistelevasti
käsitelty edustaja Hjeltin y. m. eduskuntaesitys
n:o 32, joka sisältää ehdotuksen työnvälitystoi-
mistoja koskevaksi asetukseksi.

Puhemies: Toiseen käsittelyyn esitetään
itse asetusehdotus, ja suuren valiokunnan ehdo-
tus tulee olemaan käsittelyn pohjana. Ensin
sallitaan yleiskeskustelu, sitten ruvetaan pykä-
littäin tarkastamaan asetusta.
```

(a) Valtiopäivät 1911

```
Päiväjärjestyksessä olevat asiat:
1) Hallituksen esitys televisio- ja radiotoimintaa
koskevaksi lainsäädännöksi

Toinen käsittely

Hallituksen esitys 34/1998 vp

Suuren valiokunnan mietintö 3/1998 vp
Liikennevaliokunnan mietintö 6/1998 vp

Ensimmäinen varapuhemies:
Käsittelyn pohjana on suuren valiokunnan mie-
tintö n:o 3. Ensin sallitaan asiasta yleiskeskuste-
lu, sen jälkeen ryhdytään lakiehdotusten yksi-
tyiskohtaiseen käsittelyyn.
```

(b) Valtiopäivät 1998

Kuvaan 5.3 on poimittu esimerkki aineistosta valtiopäiviltä 1911 ja 1998. Siinä missä vuosisadan lopun asiakirjalistaus on melko yksinkertaista paikantaa ja käsitellä, vaatii vuosisadan alun asiakirjojen koonti ja erittely selvästi enemmän vaivaa. Esimerkinkaltaisten rakennemuutoksien takia oli puheidenkerääjän useimmista ratkaisuista tehtävä erilliset versiot eri aikakausille.

Aikakaudesta huolimatta aineistosta poimitut tiedot kerättiin ”raakaversiona” CSV-tiedostoihin, yksi puhe per rivi ja yhdet valtiopäivät per tiedosto. Tiedostoon kootuille tiedoille tehtiin keräyksen yhteydessä ensimmäisiä siistimis- ja erotteluoperaatioita; puhujatiedot poimittiin omaan sarakkeeseensa, poistettavat tavuviivat merkittiin ja esimerkiksi

mahdollisia asiakirjalyhenteitä täydennettiin (*rah.al. n:o 66* → *raha-aloite n:o 66*). Jokaiseen puheeseen liitettiin myös pöytäkirjasta poimittuja aika- ja istuntotietoja.

Puheiden korjaus ja ensimmäinen rikastus

Karkeamuotoinen aineisto kulki seuraavaksi läpi useista automatisoiduista siistintä- ja rikastusoperaatioista. Ensinnäkin puhujista tiedettiin parhaassakin tapauksessa vain sukunimi, joten näiden vaiheiden keskeinen tehtävä oli tunnistaa puhevuorojen pitäjät. Tässä apuna käytettiin jo aiemmin esiteltyä Eduskunnan toimijat -tietämysverkkoa, joka koostuu eduskunnan toimijoiden biografisista tiedoista.

Puheenvuoron päiväyksen, tiedetyn sukunimen ja mahdollisen sukunimikaimat erottelevan etukirjaimen perusteella pyrittiin löytämään vastaava henkilö Eduskunnan toimijat -aineistosta. Mikäli sopivia osumia ei löytynyt, yritettiin pöytäkirjasta poimittua sukunimeä korjata yleisimpien löydettyjen muunnosvirheiden varalta, kuten *Törn**g**vist* → *Törn**q**vist*. Jos tämäkään ei toiminut, haettiin läheisin osuma kaikista tiedetyistä sukunimivaihtoehdoista samankaltaisuuden perusteella. Kun sopiva henkilö löytyi henkilöaineistosta, kerättiin puheen yhteyteen tiedot hänen etunimestään, Eduskunnan toimijat -tietoverkon URI-tunnisteestaan, sen hetkisestä puolueesta ja muusta ensisijaisesta ryhmäjäsenyydestä, sekä niiden URI-tunnisteista, syntymävuodesta sekä sukupuolesta. Puhemiehistön jäsenille, joista puheen yhteydessä ei ollut merkitty edes sukunimeä, nämä tiedot jouduttiin etsimään pelkän ajankohdan ja roolin perusteella.

Henkilötietojen täydentämisen lisäksi aineistoa pyrittiin siivoamaan turhista välimerkeistä, aikatietoja siivottiin sekä haettiin automatisoidusti arvio puheen kielestä LAS-työkalulla (ks. [58]). Tieto mahdollisesta puheenvuoron erikoistyypistä poimittiin talteen sekä mahdollisia virheellisesti kahdeksi eri puheeksi tulkittuja yksittäisiä puheenvuoroja pyrittiin paikantamaan ja jälkikorjaamaan. Siistitty ja täydennetty aineisto tallennettiin jälleen CSV-muotoon, yhdet valtiopäivät per tiedosto.

Kuva 5.4 havainnollistaa näitä vaiheita. Aineiston raakaversio (a) sisältää vasemmalta katsoen muun muassa seuraavia tietoja: 1. Puhujan nimi ja mahdollinen mainittu puhevuoron erityistyyppi sellaisena, kuin ne lähdeaineistossa olivat. 2. Asiakohta, johon puhe liittyy sekä asiakohdan alla mainitut eduskunnan asiakirjat. Asiakirjoihin liitetty myöhempää käsittelyä helpottava tunniste '⋙'. 3. Puheen sisältö. Rivityksestä johtuneet tavuviivat on merkitty poistettavaksi '<REMOVE>'-tunnisteella. Kuvassa (b) samat rivit siistinnän jälkeen: 1. Puhujan päätelty etunimi, 2. Sukunimi, 3. Rooli, 4. Päätelty puolue, 5.

Kuva 5.4: Aineiston siistintä ja rikastus. Esimerkkiote valtiopäivien 1993 aineistosta.

Ensimmäinen varapuhemies	7) Ehdotus laiksi Pohjois-Suomessa annettavista sähkön hinnanalennuksista >>>Hallituksen esitys n:o 358/1992 vp >>>Talousvaliokunnan mietintö n:o 9	Käsittelyn pohjana on talousvaliokunnan mie-<REMO
Fd.Louekoski	7) Ehdotus laiksi Pohjois-Suomessa annettavista sähkön hinnanalennuksista >>>Hallituksen esitys n:o 358/1992 vp >>>Talousvaliokunnan mietintö n:o 9	Rouva puhemies! Käsi-<REMOVE> teltävänä on halli Kun hallitus esittää alennusta jatkettavaksi muodossa Tämä erityinen alennuksen järjestämistapa näyttää ku Toinen seikka, johon hallituksen esityksessä valiokun Jossain määrin erityislaatuiselta tuntuu sellai-<REMO Toistan vielä, että valiokunnan mietintö oli yksimieline
Ed. Ukkola (vastauspuheenvuoro)	7) Ehdotus laiksi Pohjois-Suomessa annettavista sähkön hinnanalennuksista >>>Hallituksen esitys n:o 358/1992 vp >>>Talousvaliokunnan mietintö n:o 9	Arvoi-<REMOVE> sa puhemies! En ymmärrä, jos vali Minun mielestäni on kohtuutonta ryhtyä sit-<REMOVE Sitä paitsi Kemijoki Oy on ostanut sähkönsä sieltä riist
Oikeusministeri Pokka	7) Ehdotus laiksi Pohjois-Suomessa annettavista sähkön hinnanalennuksista >>>Hallituksen esitys n:o 358/1992 vp >>>Talousvaliokunnan mietintö n:o 9	Arvoisa rouva puhemies! Ed. Ukkola on ihan oikeassa Tätä sähköalennusta on sopimusten pohjal-<REMOV Tässä ed. Ukkola myöskin oli oikeassa, että lakiesitys

(a) Aineiston raakaversio lähdetiedostosta keräämisen jälkeen.

Saara-Maria	Paakkinen	Ensimmäinen varapuhemies	SDP	7) Ehdotus laiksi Pohjois-Suomessa annettavista sähkön hinnanalennuksista >>>Hallituksen esitys n:o 358/1992 vp >>>Talousvaliokunnan mietintö n:o 9	Käsittelyn pohjana on talousvaliok
Matti	Louekoski	Kansanedustaja	SDP	7) Ehdotus laiksi Pohjois-Suomessa annettavista sähkön hinnanalennuksista >>>Hallituksen esitys n:o 358/1992 vp >>>Talousvaliokunnan mietintö n:o 9	Rouva puhemies! Käsiteltävänä o Kun hallitus esittää alennusta jatke Tämä erityinen alennuksen järjest Toinen seikka, johon hallituksen e Jossain määrin erityislaatuiselta tu Toistan vielä, että valiokunnan mie
Tuulikki	Ukkola	Kansanedustaja	LKP	7) Ehdotus laiksi Pohjois-Suomessa annettavista sähkön hinnanalennuksista >>>Hallituksen esitys n:o 358/1992 vp >>>Talousvaliokunnan mietintö n:o 9	Arvoisa puhemies! En ymmärrä, jo Minun mielestäni on kohtuutonta r Sitä paitsi Kemijoki Oy on ostanut
Hannele	Pokka	Oikeusministeri	KESK	7) Ehdotus laiksi Pohjois-Suomessa annettavista sähkön hinnanalennuksista >>>Hallituksen esitys n:o 358/1992 vp >>>Talousvaliokunnan mietintö n:o 9	Arvoisa rouva puhemies! Ed. Ukk Tätä sähköalennusta on sopimus Tässä ed. Ukkola myöskin oli oike

(b) Ensimmäisten siistimisien ja rikastusten jälkeen

Asiakohta ja asiakirjat, kuten kuvassa (a), 4. Puheen sisältö, josta siistitty ylimääräiset välimerkit.

Edellä kuvattujen toimien lisäksi jokaiselle puheelle luotiin uniikki tunniste muotoon:

valtiopäivät_täysistunto_puheenJärjestysnumeroIstunnossa

Esimerkiksi puhe *1963_11_2* on vuoden 1963 valtiopäivien yhdennentoista täysistunnon toinen puheenvuoro (puhemiehen puheenvuorot mukaan laskien). Vastaava rakenne pätee kaikille puheille luoduille tunnisteille lähtöaineistosta riippumatta.

Varsinaiset ja ylimääräiset valtiopäivät täytyi huomioida tunnisteissa. Ylimääräisillä valtiopäivillä istuntojen numerointi aloitettiin alusta, joten pelkkää valtiopäivien vuosilukua käyttämällä aineistoon olisi päätynyt paikoin kaksi eri puhetta samalle tunnisteelle. Tästä syystä ylimääräisten valtiopäivien puheiden tunnisteisiin lisätiin valtiopäivien perään *_II*-merkintä (”toinen”). Näin esimerkiksi varsinaisten valtiopäivien puhe *1975_25_2* erottuu järjestysnumeroinniltaan vastaavasta ylimääräisten valtiopäivien puheesta *1975_II_25_2*. Vuoden 1917 poikkeusistunnot eroteltiin vastaavasti mielivaltaisesti valitulla *_XX*-merkin-

nällä.

Tässä vaiheessa PDF-tiedostoista lähtöisin oleva puheaineisto oli valmis muunnettavaksi loppuformaatteihin.

5.2 HTML-pohjaisten puheiden kerääminen

Aineistonkeruun aloitusaikoina, kesällä 2020, täysistuntojen pöytäkirjoista oli saatavilla myös HTML-versiot täysistunnosta 86/1999 alkaen. Pöytäkirjat olivat saatavilla eduskunnan verkkosivuilla. Valmiiksi koneluettavaan muotoon käsin kirjatut pöytäkirjat olivat luonnollisesti huomattavasti helpompia hyödyntää kuin tekstitiedostot, joten niitä käytettiin lähdeaineistona heti pöytäkirjasta 86/1999 lähtien. Valtiopäivien 1999 aineisto koottiin siis kahdesta eri lähteestä ja sen ensimmäinen puolisko sisältää siksi vähemmän kuvailutietoja.

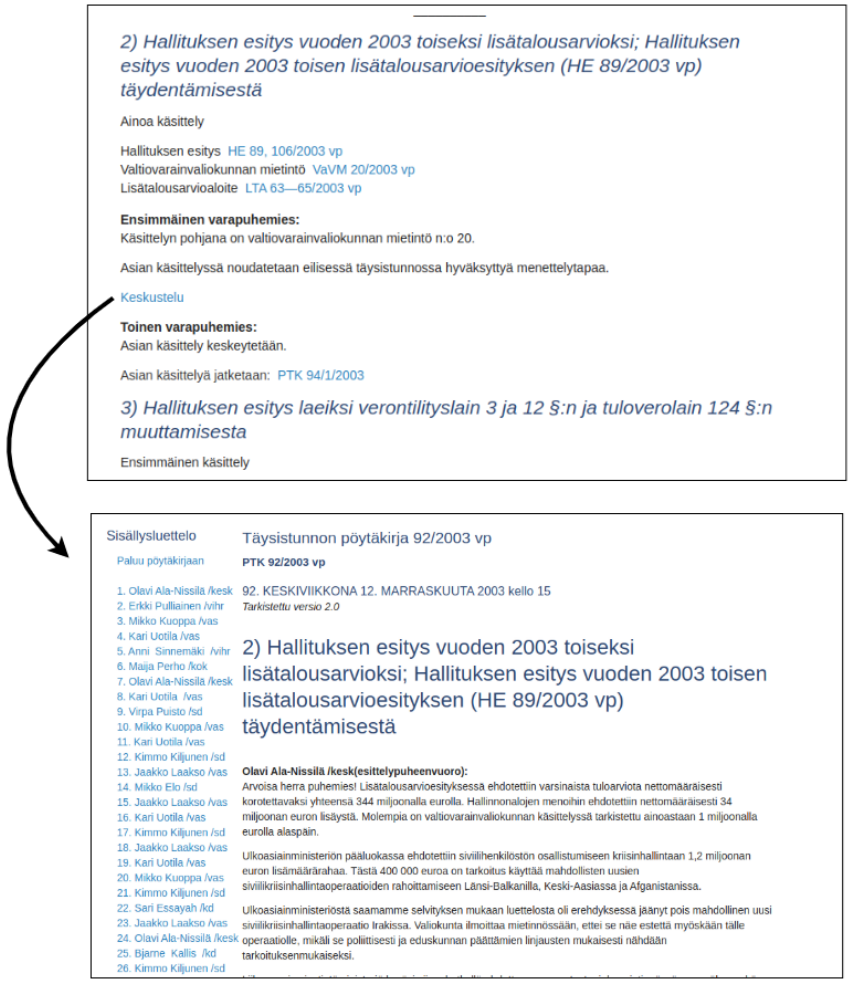
HTML-muotoiset pöytäkirjat rakentuivat pöytäkirjan pääsivusta sekä mahdollisista erillisistä asiakohtakohtaisista keskusteluosion sisältävistä sivuista. Pääsivu sisälsi pöytäkirjan rungon: osallistujatiedot, asiakohtien otsikot ja niihin liittyvät eduskunnan asiakirjat, puhemiesten asiakohdan avaus- ja päätöspuheenvuorot sekä linkin asiakohdan keskustelusivulle. Keskustelusivu sisälsi joitain aikatietoja, asiakohdan otsikon sekä varsinaiset puheenvuorot.

Toisin kuin tekstipohjaisissa pöytäkirjoissa, HTML-versiossa puhujasta on kerrottu suoraan myös etunimi sekä puolue. Kuvassa 5.5 havainnollistetaan HTML-pöytäkirjojen rakennetta. Kuvassa ylempänä kuvakaappaus *PTK 92/2003 vp* pääsivulta, jossa alustetaan lisätalousarviota käsittelevä asiakohta, *Keskustelu*-linkkiä painamalla pääsee tämän asiakohdan keskustelusivulle, jonka alusta on kuvakaappaus kuvan 5.5 alareunassa. Keskustelusivun vasemmassa laidassa on sisällysluettelo kyseessä olevan keskusteluosion käytetyistä puheenvuoroista.

Aineistonkeruuta varten tuli siis ensin kerätä pöytäkirjojen pääsivut, joiden kautta pääsi käsiksi myös itse keskusteluihin. Aineistokeruuta varten luotiin Python-ohjelma, joka latasi verkosta pöytäkirjojen pääsivujen HTML-lähdekoodit. Pääsivujen verkko-osoitteet seurasivat tiettyä kaavaa, joten ohjelma teki verkkokyselyitä istunnon numeroita kasvattaen, kunnes uusia pöytäkirjoja ei enää löytynyt kyseisille valtiopäiville.

Varsinaisten puheiden ja istuntotietojen keräämiseen laadittiin taas erillinen Python-ohjelma, joka BeautifulSoup-kirjastoa (ks. [5]) hyödyntäen poimi ja pilkkoi halutut tiedot

Kuva 5.5: Esimerkki pääsivun asiakohdasta, jonka *Keskustelu*-linkki johtaa varsinaisen keskustelun sisältävälle verkkosivulle.



HTML-aineistosta. Pääsivuilta kerättiin CSV-tiedostoon talteen tietoja istunnosta (päiväys, aikatietoja jne.), pääsivun puheenvuorot sekä kirjatut asiakohdat ja niihin liittyvät keskustelu- ja asiakirjalinkit.

Löydettyjen linkkien perusteella kerättiin keskustelusivujen HTLM-lähdekoodi. Keskustelusivuilta koottiin keskustelupuheenvuorot omaan CSV-tiedostoonsa. Tämän jälkeen pääja keskustelusivujen puheenvuorot piti yhdistää yhteen CVS-tiedostoon niin, että oikeaa pääsivun puheenvuoroa seurasivat istunnossa sitä seuranneet keskustelupuheenvuorot. Myös etusivulla olleita kuvailutietoja, joita ei ollut keskustelusivulla saatavilla (kuten asiakohtaan liittyvät asiakirjat) liitettiin keskustelusivujen puheenvuorojen tietoihin.

HTML-pohjaiset puheenvuorot vaativat melko vähän tietojen siistimis- tai korjaustoimenpiteitä, mutta myös niitä rikastettiin Eduskunnan toimijat -ontologialla liittämällä sieltä

puheisiin tietoja puhujan, puolueen ja puhujan ryhmän URI-tunnisteista, puhujan syntymävuodesta ja sukupuolesta. Myös näille puheille luotiin uniikki tunniste sekä haettiin tieto puheen kielestä. Lopputulos oli vastaava kuin PDF-lähtöisten aineiston siistinnän ja rikastuksen jälkeen: Yksi CSV-tiedosto per valtiopäivät, yksi puhe per rivi. Erona oli vain saatujen kuvailutietojen määrä.

5.3 XML-pohjaisten puheiden kerääminen

Valtiopäivistä 2015 alkaen eduskunta on julkaissut pöytäkirjoja PDF- ja HTML-versioiden lisäksi vielä pitkälle prosessoituina, kuvailutiedoiltaan rikkaina XML-versioina. Nämä pöytäkirjat ovat yksittäin saatavilla eduskunnan Avoin data -palvelun rajapinnasta pöytäkirjojen eduskuntatunnuksen pohjalta (esim. *PTK 1/2018* vp). Rajapintakyselyllä noudetut pöytäkirjat lähetään käärittynä JSON-formaattiin. Itse XML-muotoinen pöytäkirja löytyy vastauksen *rowData*-kohdasta.

Pöytäkirjakysely palautti useimmiten kaksi eri XML-asiakirjaa: Täysistunnon pöytäkirjan pääsivun sekä pöytäkirjan. Näistä ensimmäinen on tiivistelmä täysistunnosta, se listaa käydyt asiakohdat ja niistä tehdyt päätökset, mutta ei itse puheenvuoroja. Koska tavoitteena oli kerätä nimenomaan puheenvuorot, näitä asiakirjoja ei käytetty. Jälkimmäinen, eli itse pöytäkirja sisältää kaikki istuntoon liittyvät tiedot sekä asiakohdat puheenvuoroineen. Puheenvuorot kerättiin näistä asiakirjoista.

Puheenvuorojen ja niihin liittyvien tietojen kerääminen XML-aineistosta oli melko suoraviivaista. Tehtävää varten luotu Python-ohjelma keräsi pöytäkirjat rajapintakutsuina ja poimi BeautifulSoup-kirjaston työkaluja hyödyntäen halutut tiedot pöytäkirjoista CSV-tiedostoon, yhdet valtiopäivät per tiedosto, yksi puhe per rivi. Pöytäkirjat olivat jo valmiiksi hyvin koneluottavassa muodossa ja esimerkiksi puhujatietoja oli varsin paljon ja ne oli valmiiksi pilkottu kuten on nähtävissä kuvassa 5.6. Kuitenkin satunnaiset puutteet, virheet ja poikkeukset pöytäkirjassa oli huomioitava.

Kun XML-pohjaista aineistoa alettiin alun perin kerätä kesällä 2020, havaittiin, että eduskunnan rajapinnan aineistossa oli puutteita. Valtiopäiviltä 2016, 2018 ja 2020 löytyi yhteensä seitsemän pöytäkirjaa, joiden pyyntö rajapintaan ei palauttanut mitään, ne siis puuttuivat XML-muotoisina tietokannasta. Asia otettiin esille eduskunnan tietohallinnon kanssa ja syksyllä 2021 kaikki puuttuvat pöytäkirjat yhtä lukuun ottamatta on saatu tuotua tietokantaan. Yhden edelleen puuttuvan pöytäkirjojen kohdalla on hyödynnetty pöytäkirjan HTML-versiota paikkaamaan puutetta.

Kuva 5.6: Esimerkki XML-muotoisten pöytäkirjojen puhujatiedoista (PTK 22/2015 vp.) Puhujan nimi-, rooli- ja puoluetiedot ovat valmiiksi merkitty ja eritelty. Puheesta on jopa tiedossa sen tarkka ajanhetki.

```
<vsk:PuheenvuoroToimenpide met1:muuTunnus="6854" vsk1:puheenvuoroAloitusHetki="
2015-06-17T14:10:52" vsk1:puheenvuoroLuokitusKoodi="E">
 <vsk1:AjankohtaTeksti>
  14.10
 </vsk1:AjankohtaTeksti>
 <met:Toimija>
  <org:Henkilo met1:muuTunnus="1031">
   <org1:AsemaTeksti>
    Valtiovarainministeri
   </org1:AsemaTeksti>
   <org1:EtuNimi>
    Alexander
   </org1:EtuNimi>
   <org1:SukuNimi>
    Stubb
   </org1:SukuNimi>
  </org:Henkilo>
 </met:Toimija>
 <vsk1:TarkenneTeksti>
  (esittelypuheenvuoro)
 </vsk1:TarkenneTeksti>
 <vsk:PuheenvuoroOsa met1:asiakirjatyyppiNimi="Puheenvuoro" met1:kieliKoodi="fi"
 met1:muuTunnus="3192" met1:tilaKoodi="Tarkistettu" met1:versioTeksti="1.1" vsk1:
 puheenvuoroAloitusHetki="2015-06-17T14:10:52" vsk1:puheenvuoroJNro="1" vsk1:
 puheenvuoroLopetusHetki="2015-06-17T14:18:25">
  <vsk:KohtaSisalto>
   <sis:KappaleKooste>
    Arvoisa puhemies! Jaottelisin avauspuheenvuoroni kolmeen osaan. Sanon ensin
```

Nyt 2020-luvulla pöytäkirjojen kirjaustarkkuutta on lisätty huomattavasti. Siinä missä esimerkiksi puhemiehen vuoronantokommentteja (”*Edustaja Pirttilahti.*”) puheenvuorojen välissä ei ole aiemmin merkitty pöytäkirjoihin, on ne tuoreimpiin pöytäkirjoihin liitetty. XML-versiossa nämä puhemiehen vuorokommentit on liitetty edellisen puhevuoroelementin loppuun *PuheenjohtajaRepliikki*-tunnisteella. Tarkasteltaessa kuitenkin PDF-versioita samoista pöytäkirjoista, on nämä kommentit aseteltu selkeästi omiksi puheenvuoroikseen eikä esimerkiksi sulkumerkeillä keskeytyksiksi. Tästä syystä kerättäessä puheita XML-versioista nämä perään liitetyt puhemiesten kommentit eroteltiin omiksi itsenäisiksi puheikseen.

Myös jo kuvailutiedoiltaan valmiiksi rikkaat XML-aineistosta kerätyt puheet kävivät läpi rikastamisprosesseja. Puheille haettiin tieto puhujan, puolueen ja ryhmän URI-tunnisteista, syntymävuodesta ja sukupuolesta sekä puheen kielestä[1]. Lopputulos oli jälleen vastaava kuin aiemmin kuvailluilla lähdeaineistoilla: Yksi CSV-tiedosto per valtiopäivät, yksi puhe per rivi. Erona saatujen kuvailutietojen määrä.

[1]Teknisesti ottaen XML-pöytäkirjoissa isoon osaan puheista on jo liitetty tieto puhekielestä. Merkintätapa ei ruotsin- ja monikielisten puheiden kohdalla ollut kuitenkaan kovin intuitiivinen. Lisäksi osalta lyhyemmistä puheenvuoroista tieto puuttui. Tästä syystä myös kaikille XML-pohjaisille puheille käytettiin kielitiedon osalta LAS-kielityökalua [58].

5.4 RDF-muunnos

Keräämisen ja valmistelun jälkeen puheenvuorot muunnettiin lopuksi loppuformaatteihin, RDF- sekä Parla-CLARIN XML -muotoihin. RDF-muunnoksen tavoitteena oli siirtää aineisto tekstiavaruudesta käsiteavaruuteen. Aineisto tuli siis semanttisesti annotoida ja muuntaa kolmikoiksi. Keskeisimmille resursseille, eli puheille, tuli ensinnäkin luoda URI-tunnisteet. Niissä hyödynnettiin jo aiemmin luotuja puheiden yksilöllisiä tunnisteita ja koko projektin aineistolle yhteistä alkua *<http://ldf.fi/semparl/>*. Puhe-resurssien hierarkiaa selvennettiin vielä *speeches/*-polulla, jolloin puheiden tunnisteista muotoutui seuraavanlaisia:

<http://ldf.fi/semparl/speeches/s2015_49_79>

missä lopun numerosarja on puhekohtainen.

Puheeseen liittyvät tiedot, eli CSV-taulukossa puheen rivillä olleet kuvailutiedot liitettiin puheyksilön ominaisuuksiksi. Ominaisuuksien predikaatit olivat suurimmilta osin aineistoa varten luotuja ja *<http://ldf.fi/schema/semparl/>*-nimiavaruuden alla. Tästä käytetään tässä prefiksiä ':'. Kun mahdollista, predikaateissa hyödynnettiin myös muun muassa valmiita SKOS- ja DCT-sanastoja. Puhujan sekä tämän puolueen ja ryhmän URI-tunnisteet oli kerätty jo aiemmin valmiiksi, joten ne voitiin suoraan asettaa asianmukaisten predikaattien objekteiksi.

Puheluokan (*Speech*) lisäksi kerätyistä tiedoista luotiin myös muita resurssiluokkia kuten luvussa 4.4 esiteltiin. Puheenvuoroihin kirjatut mahdolliset keskeytykset poimittiin omiksi *Interruption*-luokan instasseikseen *<http://ldf.fi/semparl/speeches/>*-nimiavaruuden alle. Keskeytys linkitettiin puheeseen puheen *:isInterruptedBy*-predikaatilla. Mikäli alkuperäisen välihuudon kirjaukseen oli selkeästi merkitty huutajan nimi, haluttiin keskeytys linkittää myös kyseiseen henkilöön. Mikäli keskeyttäjä oli pöytäkirjaan selkeästi nimetty (vs. geneerinen suunta, ”Keskustan eduskuntaryhmästä”) tästä oli aikakaudesta riippuen ilmoitettu joko koko nimi tai vain sukunimi. Heidät tunnistettiin, kuten varsinaiset puhujat Eduskunnan toimijat -aineistosta, josta saatiin heidän URI-tunnisteensa.

Puheiden tiedoista löytyvät asiakohdat ja -kirjat ja niiden tiedot eriytettiin omiksi *Item*- ja *Document*-luokikseen. Puheet linkittyvät asiakohtiin *:item*-predikaatilla ja asiakohdat asiakirjoihin *:relatedDocument*-predikaatilla. Molempia rikastettiin mahdollisuuksien mukaan löydetyillä tai automatisoidusti luoduilla linkeillä verkkoversioihin sekä tuoreimpien

hallituksen esityksien osalta päätöstiedoilla ja asiasanoilla. Asiakohdat kuuluvat *<http://ldf.fi/semparl/items/>*-nimiavaruuteen (esim. *<http://ldf.fi/semparl/items/i2018121>*) ja asiakirjat *<http://ldf.fi/semparl/documents/>*-nimiavaruuteen (esim. *<http://ldf.fi/semparl/documents/TAMINO2_30_1909_II>*).

Esiintyvistä täysistunnoista luotiin erillisiä *PlenarySession*-luokan yksilöitä, johon koottiin niihin liittyviä aikatietoja. Näitäkin tietoja rikastettiin ulkoisilla verkkolähteillä, eli linkeillä alkuperäisiin verkkoversioihin. Kunkin valtiopäivien aikatiedot oli myös koottu eduskunnan verkkosivuilta ja ne niitä hyödynnettiin luomalla *ParliamentarySession* eli valtiopäivät-luokka, johon kaikki kyseisien valtiopäivien täysistunnot linkittyivät. Valtiopäivien ja täysistuntojen tunnisteet luotiin *<http://ldf.fi/semparl/times/>* -nimiavaruuden alle (esimerkiksi *<http://ldf.fi/semparl/times/plenary-sessions/ps_26_2015>*).

Myös pöytäkirjat itsessään nostettiin istuntoihin linkittyviksi resursseiksi, *Transcript*, joiden ominaisuuksina olivat muun muassa linkit alkuperäiseen käytettyyn pöytäkirjaan sekä mahdolliset versiotiedot. Pöytäkirjojen tunnisteet ovat saman nimiavaruuden alla kuin muut asiakirjat, eli esimerkiksi *<http://ldf.fi/semparl/documents/ptk_85_1931>*.

Tätä muunnosta varten luotu Python-ohjelma koodasi yllä esitellyt resurssit yhdet valtiopäivät kerrallaan kolmeen eri verkkoon. Ensimmäiseen tulivat puheet ja keskeytykset, toiseen asiakohdat ja -kirjat, kolmanteen istunnot, valtiopäivät sekä pöytäkirjat. Tällä ratkaisulla data kirjautui kolmeen eri Turtle -tiedostoon (.ttl). Tällöin tulosten tarkastelu oli helpompaa, ryhmittely looginen ja tiedostokoko hallitumpi. Kullekin valtiopäiville syntyi siis linkitettynä datana kolme erillistä tiedostoa ja verkkoa. Ratkaisu on kuitenkin lähinnä logistinen ja kukin verkko linkittyy niin tiivisti toisiinsa, että käytännössä puhumme yhdestä Eduskunnan puheet -tietämysverkosta.

Aineistoa täydennettiin vielä yhdellä RDF-tiedostolla, jossa määriteltiin aineiston luokkien *prefLabel*-ominaisuuksia, eli ”nimiä”. Tämä mahdollisti sen, että esimerkiksi Parlamenttisampo voi näyttää puheen tyypistä ihmissilmin miellyttävämpää nimeä ”vastauspuheenvuoro”, puhetyypin tunnisteen *<http://ldf.fi/semparl/speechtypes/Vastauspuheenvuoro>* sijaan.

Tämän vaiheen seurauksena puheaineisto oli muunnettu linkitetyksi dataksi ja valmis testattavaksi ja sen jälkeen julkaistavaksi LDF.fi-palvelussa.

5.5 Parla-CLARIN XML -muunnos

Lopullisen puheaineiston toinen versio toteutettiin Parla-CLARIN XML -formaatissa. Samoin kuin RDF-aineiston luomisessa, XML-aineistossa käytettiin lähtökohtana rikastettuja ja siistittyjä CSV-tiedostoja. Tuloksena oli yksi XML-tiedosto per valtiopäivät.

Muunnos XML-muotoon oli hyvin suoraviivainen, suurin osa XML-versioon tallentuvasta tiedosta oli jo valmisteltu. Muunnosta varten luotiin Python-ohjelma, jolla BeautifulSoup-kirjastoa hyödyntäen rakennettiin Parla-CLARIN-suositusten mukainen XML-tiedosto. Parla-CLARIN-tiedostojen rakenne esiteltiin luvussa 3.1. Ohjelma muunsi CSV-tiedostoon kerätyt puheet ja kuvailutiedot kronologisesti eteneväksi elementtipuuksi.

Muunnoksen työläin vaihe oli tunnistaa puheista keskeytykset ja välihuudot ja erotella ne. Formaattiin kuuluu, että puhe-elementit sisältävät vain varsinaisen puhujan puhetta, joten puheet ja keskeytykset tuli pilkkoa erillisiin osiin. Muunnostyökalu tarkasti ensin jokaisen puheen kohdalla sisälsikö se keskeytyksiä. Jos sisälsi, puheet pilkottiin erillisiin puhe-elementteihin ja kuhunkin niistä lisättiin attribuutiksi tieto siitä, onko kyseisellä puheenosalla edeltäviä tai seuraavia puheenosia ja niiden tunnisteet. Keskeytykset tallennettiin näiden osien väliin omina elementteinään.

Mikäli alkuperäiseen välihuudon kirjaukseen oli selkeästi merkitty huutajan nimi, liitettiin myös keskeytykseen henkilön tunniste. Näin myös kunkin henkilön keskeytykset olivat löydettävissä kyseisen henkilön *xml:id*-tunnisteella. Myös tässä versiossa keskeyttäjät, joista oli kirjattu vain sukunimi tuli tunnistaa Eduskunnan toimijat -aineiston avulla. Kun keskeyttäjän koko nimi oli tiedossa, voitiin hänelle muodostaa *xml:id*-tunniste ja tarvittaessa lisätä hänet kuvailutietojen puhujalistaan, mikäli hän ei muuten kyseisillä valtiopäivillä ollut äänessä.

Lopputuloksena oli itsenäinen aineisto Suomen eduskunnan täysistuntojen puheenvuoroista vuodesta 1907 vuoteen 2021. Kukin tiedosto sisältää kyseisten valtiopäivien puheenvuorot, pöytäkirjoista löytyneet kuvailutiedot ja rikastetut tiedot pöytäkirjoissa esiintyvistä eduskunnan toimijoista. Tiedot on kirjattu helppokäyttöiseen ja -lukuiseen rakenteeseen, joka mahdollistaa aineiston sujuvan automatisoidun käsittelyn.

5.6 Aineiston rikastaminen

Aineiston ensimmäisten keräys- ja muunnosvaiheiden aikana puheiden tietoja rikastettiin jo puhuja- ja puoluetiedoilla. Samalla toteutettiin myös aineiston luotettavuudenkin kannalta ensisijainen rikastustoimenpide eli puheiden linkittäminen alkuperäisiin aineistoihin. Puheen tietoihin lisättiin linkki kunkin lähdedokumentin verkkoversioon, jotta käyttäjä pääsee helposti tarkastelemaan myös alkuperäistä puhetta ja sen kontekstia alkuperäisestä, lähteenä käytetystä päiväkirjasta. Ainoa poikkeus on XML-pohjainen aineisto, jossa saavutettavuuden takia tämä verkkolinkki osoittaa kyseisen pöytäkirjan verkkosivuversioon.

Linkitetyn datan aineistoa, eli aineiston RDF-versiota, pyrittiin rikastamaan ja linkittämään edelleen muihin lähteisiin. Asiakohtiin liittyvien eduskunnan asiakirjoille on joko poimittu tai automatisoidusti pyritty generoimaan verkko-osoite avoimeen sähköiseen versioon, mikäli sellainen on löytynyt. Lisäksi puheiden sisällöistä on poimittu erikseen keskeytykset ja mikäli kyseinen keskeyttäjä on nimeltä mainittu, on keskeytys linkitetty kyseiseen henkilöön.

Aineistoa haluttiin laajentaa kattamaan hallituksen esitykset mahdollisuuksien mukaan. Samoin kuin pöytäkirjat, ovat hallituksen esitykset saatavissa XML-muotoisina asiakirjoina eduskunnan palveluiden rajapinnasta valtiopäivistä 2015 alkaen. Nämä asiakirjat käytiin läpi Python-ohjelmalla ja niistä kerättiin talteen tiedot esityksen otsikosta, mahdollisesta päätöksestä ja sen päivämäärästä sekä pöytäkirjatoimiston laatimat YSO-asiasanat. Näistä muodostettiin hallituksen esitysten RDF-tietämysverkko. Tämän verkon esitykset linkittyivät suoraan jo puheiden keräysvaiheessa luotuihin asiakirjoihin, sillä niiden ja näin jälkikäteen eri lähteestä kerättyjen vastaavien hallitusten esitysten URI:t luotiin identtisiksi.

Puheaineistoa halutaan syventää myös kielellisen analyysin osalta. Puheiden sisällöstä tullaan automatisoidusti tunnistamaan nimetyt entiteetit, jotka puolestaan voidaan linkittää eduskunta-aineiston sisäisiin tai ulkoisiin ontologioihin, kuten paikat GeoNames-ontologiaan (ks. [24]). Lisäksi puheet tullaan automatisoidusti asiasanoittamaan. Edellä mainitut rikastustoimenpiteet ovat vielä kesken ja tullaan toteuttamaan muiden hankkeen jäsenten toimesta, joten niitä ei tarkastellessa tässä tutkielmassa tarkemmin.

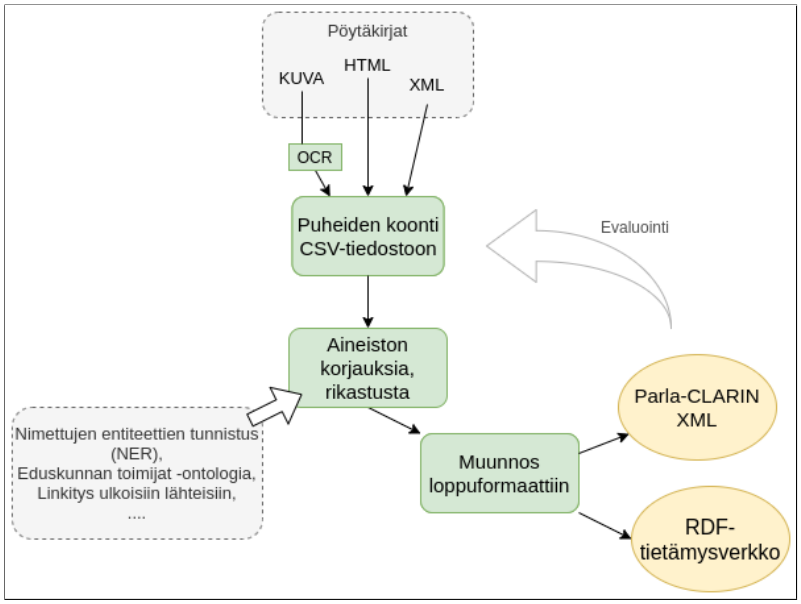
6 Tulosten arviointi

Muunnettu puheaineisto valtiopäivistä 1907 valtiopäiviin 2021 on varsin kookas. Vuoden 2021 lopussa se kattoi lukuina:

- 13 474 täysistuntoa
- 138 valtiopäivät
- 959 130 erillistä puheenvuoroa, joista
  - 662 766 on kansanedustajien pitämiä [2]
- 5 GB (kummatkin aineistoversiot yhteensä)

Aineisto on käynyt läpi jo useita iteraatioita ja on käytössä kaikilla konsortion jäsenillä, niin aineiston sisällön, eli puheiden tutkimuksessa kuin kielellisessä analyysissä sekä julkaisualustojen kehitystyössä. Aineiston muunnostyökalujen ja -prosessin kehitys oli vahvasti evaluaatiojohtoinen ja työkalut ja muunnettu aineisto kävivät läpi lukemattomia muutoksia uusien aineistoversioiden myötä. Kuva 6.1 tiivistää tässä tutkielmassa esitellyn puheiden muunnosprosessin rakennetta. Käytännössä evaluaation nuolen voisi vetää mistä vain vaiheesta mihin vain sitä edeltävään vaiheeseen. Muunnosprosessin modulaarisuus (pilkottu useampiin välivaiheisiin

Kuva 6.1: Puheiden muunnosprosessin luominen oli kokonaisuudessaan syklinen prosessi, jossa muunnettujen puheiden tarkastelu johti muutoksiin muunnosprosessissa.



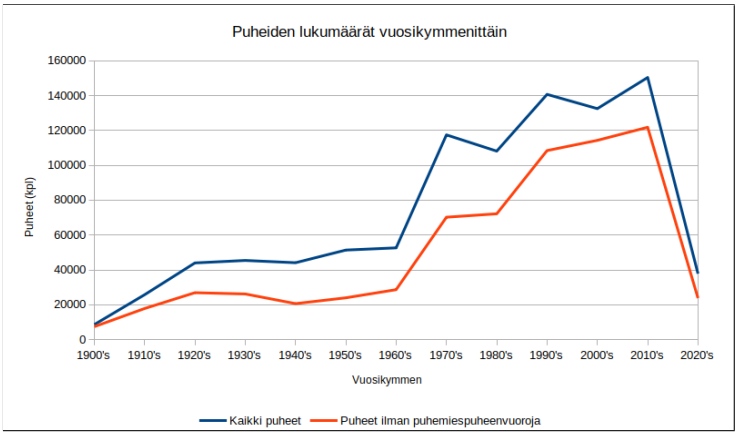
[2] Tarkkaan ottaen oikea määritelmä olisi: ”muun kuin puhemiehen pitämiä”, sillä aineisto sisältää puheenvuoroja ministereiltä, jotka eivät olleet puheen ajanhetkellä kansanedustajia sekä kourallisen muiden viranomaisten, kuten eduskunnan oikeusasiamiesten puheenvuoroja.

ja ohjelmiin) edesauttoi ketterämpää kehitystyötä. Massiivista aineistoa ei aina tarvinnut muuntaa alusta loppuun uusia muutoksia testatakseen.

6.1 Eduskunnan puheet

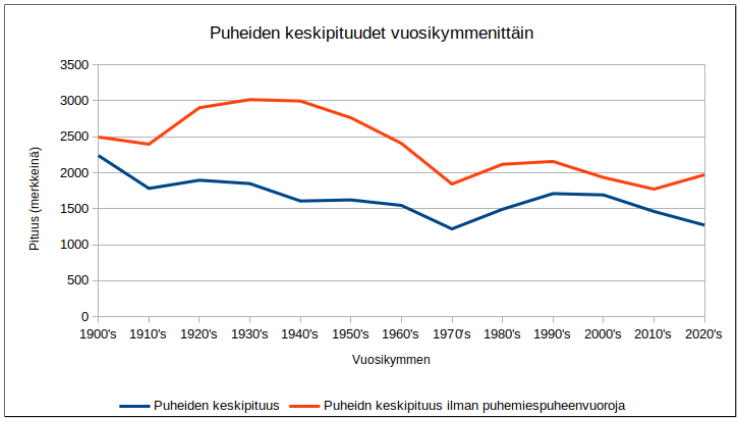
Muunnettu aineisto mahdollistaa aivan uusia suomalaisen parlamentaarisen puheen historian analyysejä. Jo pelkästään aineistoa luonnehtimalla voidaan tehdä esimerkiksi mielenkiintoisia havaintoja puheiden lukumäärän ja merkkipituuksien muutoksista läpi vuosien. Kuvassa 6.2 ovat aineiston puheenvuorojen lukumäärät vuosikymmenittäin. Yksittäisten puheiden lukumäärät ovat y-akselilla ja vuosikymmenet x-akselilla. Sininen käyrä kuvaa kaikkia aineistossa olevia puheenvuoroja, punaisesta on poistettu puhemiesten puheenvuorot. Määrät ovat nousujohteisia ja varsinkin 80-luvulta alkaen puheenvuoroja on pidetty moninkertaisesti. Syytä kasvulle voidaan hakea esimerkiksi muuttuneista keskustelu- ja kirjauskäytänteistä.

Kuva 6.2: Kerättyjen puheiden lukumääriä valtiopäiviltä 1907–2021 vuosikymmenittäin ryhmiteltynä.



Samat käytännemuutokset voivat selittää myös puheiden keskipituuksien muutosta. Kuva 6.3 kuvaa aineiston puheenvuorojen merkkipituuksia. Asetelma on sama kuin kuvassa 6.2, mutta tällä kertaa y-akselilla on puheiden pituus merkkeinä. Tällä kertaa trendi on kuitenkin ollut paikoin laskeva. Kaikkien puheiden ja eipuhemiesten puheenvuorojen keskipituuden isoa eroa selittää se,

Kuva 6.3: Valtiopäiviltä 1907–2021 kerättyjen puheiden keskimääräisiä merkkipituuksia vuosikymmenittäin ryhmiteltynä.



että puhemiesten puheenvuorot ovat kauttaaltaan hyvin lyhyitä. Kansanedustajien puheenvuorot vaikuttaisivat lyhentyneen vuosituhannen taitetta lähestyessä. Tässä voi syynä hyvinkin olla esimerkiksi käytännemuutokset sallituissa puheenvuorojen kestoissa, mutta oletettavasti myös kirjaustarkkuuden lisääntyminen. Moderneihin pöytäkirjoihin on yhä enenevissä määrin merkitty myös pieniä, epämuodollisempia edustajien kommentteja esimerkiksi äänestysasioihin liittyen. Nämä lisääntyneet, lyhyemmät puheenvuorot ovat voineet osaltaan laskea kansanedustajien puheiden keskipituuksia.

Edellä olevat kaaviot havainnollistavat kuinka jo hyvin alkeellisilla aineiston visualisoinneilla saadaan konkreettista ja mielenkiintoista tietoa Suomen eduskunnan toiminnasta läpi sen historian. Aineisto mahdollistaa monenlaisen analyysin nopeasti ja vaivattomasti sekä oleellisten puheiden automatisoidun löytämisen esimerkiksi lähilukua ja tarkempaa analyysiä varten. Muita ensimmäisiä aineistoon pohjautuvia analyysejä on esitelty julkaisuissa [21] ja [32].

6.2 Muunnoksen laatu ja kattavuus

Muunnostulosten ensisijainen arviointimetodi oli manuaalinen tarkistus. Tulosten manuaalinen tarkistus ohjasi muunnosputken kehitystyötä ja auttoi ymmärtämään virheiden luonnetta ja alkuperää. Lähes miljoonan puheen aineiston manuaalinen tarkistus kokonaisuudessaan ei kuitenkaan ole kestävä vaihtoehto, joten muunnetun aineiston laadun arvioinnin tueksi tarvittiin muitakin keinoja.

Aineiston RDF-version tarkistukseen otettiin käyttöön ShEx - *Shape Expressions* -työkalun Python-implementaatio *PyShEx* [61]. *Shape Expressions* on skeeman kuvantamiskieli RDF-verkoille [53]. Työkalua varten kirjoitettiin skeematiedosto, johon RDF-formaattia mukaillen määritellään aineiston oletettu rakenne sekä arvo-, tyyppi- ja määrävaatimukset[1]. Kaksoispisteen, ':', merkitessä prefiksiä *<http://ldf.fi/schema/semparl/>*, puheen (tiivistettyä) skeemaa kuvattiin seuraavanlaisesti:

```
:Speech {
    a [:Speech] ;
    :content xsd:string ;
    :speaker [ <http://ldf.fi/semparl/people/>~ ]?
}
```

Skeema määrittää, että tyyppiä *:Speech* olevalla verkon solmulla eli subjektilla on oltava predikaatti *:content* ja sillä tasan yksi objekti, joka on tyypiltään merkkijono. Li-

[1]ShEx toimii erinomaisesti myös RDF-aineiston suunnittelun ja dokumentaation työkaluna, ks.[70].

säksi kyseisellä subjektilla voi olla predikaatti *:speaker*, johon liittyy yksi tai ei yhtään *<http://ldf.fi/semparl/people/>*-alkuista URI:a. Käytännössä siis puheesta voi olla tiedossa puhujan URI-tunniste, mikäli puhuja on tunnistettu ja voitu linkittää Eduskunnan toimijat -aineistoon ja näitä puhujia tulee olla korkeintaan yksi. Tämän ehdollisuuden määrittää rivin perässä oleva '?'-merkki, joka toimii tässä kuten säännöllisissä lausekkeissa, nolla tai yksi kertaa. ShEx-työkalu tarkistaa annetun aineiston skeematiedostoa vasten ja ilmoittaa kaikki mahdolliset poikkeamat.

```
Focus: http://ldf.fi/semparl/documents/HE_18_2002_vp
Start: http://ldf.fi/schema/semparl/GovernmentProposal
Reason:    Testing <http://ldf.fi/semparl/documents/HE_18_2002_vp> against
shape http://ldf.fi/schema/semparl/GovernmentProposal
  Triples:
    <http://ldf.fi/semparl/documents/HE_18_2002_vp> skos:prefLabel "1) Ehdotus laiksi
    Ahvenanmaan itsehallintolain muuttamisesta  HE 18/2002 vp"@fi .
    <http://ldf.fi/semparl/documents/HE_18_2002_vp> skos:prefLabel "Hallituksen esitys
    HE 18/2002 vp"@fi .
 2 triples exceeds max {1,1}
```

Yllä olevassa esimerkissä subjektia *<http://ldf.fi/semparl/documents/HE_18_2002_vp>* on testattu vasten sen tyypin (*<http://ldf.fi/schema/semparl/GovernmentProposal>*) skeemaa. Subjektilla kuuluisi olla tasan yksi predikaatin *skos:prefLabel* objekti, mutta niitä on löytynyt kaksi kappaletta.

ShEx on erinomainen työkalu koko aineiston perusteelliseen läpikäyntiin, mutta se ei kuitenkaan kykene löytämään kaikkia aineiston ongelmia, eritoten tapauksia, jossa predikaattien ja objektien määrät ja tyypit ovat oikein, mutta objektin arvo on virheellinen. Esimerkiksi alla olevassa tapauksessa kaksi puheenvuoroa on liimaantunut yhteen.

```
speeches:s1907_27_36 a semparls:Speech ;
    semparls:content """Tahdon vaan ed. Rosengvistin huomautukseen
    mainita, että valiokunnassa kyllä on jo ollut tästä puhetta.[...]
    myöskin huomautettiin että matkustajaliikenne verrattain vähän
    lisää työtä rautatiehenkilökunnalle.
    Ed I. Hoikka: Minä kannatan ed. Vuorimaan ehdotusta.""" ;
```

Edustaja Hoikan puheenvuoron alkua ei ole tunnistettu, vaan se on jäänyt osaksi edeltävää puheenvuoroa. Puheenvuoron sisältö on kuitenkin tyypiltään oikein, eikä ShEx kykene tunnistamaan ongelmaa, vaan se vaatisi ihmissilmää tai muita ratkaisuja.

Myös Parlamenttisampo-portaali toimii eräänlaisena laaduntarkkailun työkaluna. Koko RDF-muotoista aineistoa pyörittävä sovellus havainnollistaa ainestoa hyvin kootussa muodossa sekä osoittaa virhetilanteisiin johtavat poikkeukset ja puutteet. Parlamenttisampo

on jo hankkeen sisäisessä käytössä ja aineiston on todettu palvelevan sitä hyvin. Manuaalinen laaduntarkkailu säilyy silti keskeisessä roolissa.

Aineiston kattavuuden arviointi automatisoidusti on myös haastavaa. Aivan tuoreimpia valtiopäiviä lukuun ottamatta ei ole olemassa minkäänlaista valmista tilastoa pöytäkirjoissa olevista puhemääristä. Myös pelkkä istuntojen lukumäärä on täytynyt selvittää itse pöytäkirjoista. Kaikkien istuntojen löytymisen muunnetusta aineistosta voi melko helposti tarkistaa kirjaamalla ylös istuntoja per valtiopäivät -lukemat ja verrata niitä aineistossa olevien istuntojen lukumääriin. Tämäkään tarkistus ei ole aivan mutkatonta, sillä historiasta löytyy useampiakin pöytäkirjoja, jossa istunnolle ei ole kirjattu yhtään puheenvuoroa. Tällöin näiden istuntojen kuuluukin puuttua aineistosta, mutta näiden poikkeuksien huomioiminen vaati manuaalisen tarkistuksen.

Valtiopäivistä 2015 alkaen on eduskunnan tietopalveluista fasettihaulla löydettävissä listauksia kunkin valtiopäivien puheenvuoroista. Näiltä osin löydettyjen puheenvuorojen lukumääriä on vertailtu toisiinsa ja mahdollisien erojen syitä tutkittu. Nämä vertailuluvut eivät ole kuitenkaan aivan ongelmattomia, sillä eduskunnan tietojärjestelmissä puheenvuorojen laskutavat ovat paikoin hieman erilaisia. Esimerkiksi valtiopäivien ensimmäisen istunnon puheenvuoroja ei lasketa varsinaisiksi puheenvuoroiksi, sillä ne käsittelevät vain eduskunnan järjestäytymistä. Välillä tulosten tarkastelu ja lukujen vertailu on kuitenkin kantanut hedelmää myös päinvastaiseen suuntaan, kun eduskunnan tietopalveluissa on havaittu puutteita ja niistä on ilmoitettu eteenpäin.

XML-muotoisen aineiston laadun arviointi on tällä hetkellä vielä pitkälle manuaalista. Myös tämä aineistoversio on kuitenkin luonnollisesti hyötynyt RDF-aineiston tarkistustyökaluilla havaittujen virheiden korjaamisesta, aineiston sisällölliset ongelmat korjaantuvat silloin kummastakin versiosta. Lisäksi XML-versiota aineistosta on jo hyödynnetty Semanttinen parlamentti -hankkeen sisäisessä tutkimuksessa, joten sen käytettävyys on todennettu ja joitain tutkimustyön yhteydessä havaittuja ongelmia jo ratkaistu. Tältä aineistoversiolta puuttuvat kuitenkin vielä automatisoidun, rakenteellisen arvioinnin työkalut. Tähän tullaan panostamaan hankkeen edetessä.

6.3 Haasteet

Edellä on tuotu esille jo joitain muunnosprosessin sekä tuotetun aineiston arvioinnin haasteita. Näin valtavassa projektissa haasteet eivät kuitenkaan lopu tähän.

Tuotettu aineisto sisältää virheitä, niin XML- kuin RDF-muodossa. Tyypillisimpiä virheitä ovat puheen sisällön ongelmat: tekstintunnistusvirheet, pilkkoutuneet puheet tai puheen loppuun liimautuneet puheeseen kuulumattomat pöytäkirjan sisällöt. Lisäksi puheenvuoron puhujia ei ole aina kyetty tunnistamaan, jolloin puhujan tiedot ovat vajavaiset. Tähän on syynä yleisemmin pöytäkirjasta poimitun sukunimimerkkijonon virheellisyys. Vuoden 2021 lopussa noin 1,4 prosentille puheenvuoroista ei ollut automatisoidusti tunnistettu ja linkitetty puhujaa Eduskunnan toimijat -aineistoon. Lisäksi puutteita on puheeseen liittyvien asiakohtien ja asiakirjojen sekä esimerkiksi istunnon alku- ja päättymisaikojen löytämisessä. Kaikki edellä mainitut ongelmat koskevat lähinnä 1900-luvun, eli PDF-lähtöistä aineistoa.

Käsiteltäessä näin laajaa ja monimuotoista aineistoa automatisoidusti, on virheiltä vaikea välttyä. On lähes mahdotonta varautua ohjelmallisesti kaikkiin mahdollisiin aineistossa esiintyviin poikkeamiin. Tiettyä määrää virheitä voidaankin pitää automatisoinnin ja siten myös toistettavuuden hintana. Tällöin kysymys kuuluukin, mikä määrä virheitä on hyväksyttävää? Missä vaiheessa on järkevämpää siirtyä projektissa eteenpäin ainaisten poikkeustapausten käsittelyn ohjelmoinnin sijaan?

Valmiissa aineistossa on myös luonteeltaan hieman erilaisia tietoja. Osa tiedoista on poimittu suoraan käytetyistä lähdeaineistoista, mutta osa on muunnosprosessin aikana päätelty. Eritoten 1900-luvun puheet, joista puhujasta oli alun perin tiedossa vain rooli ja sukunimi. Kaikki muu lopputuloksessa puhujasta tarjolla oleva tieto on päätelty Eduskunnan toimijat -aineistosta: etunimi, sen hetkinen puolue, syntymäaika ja niin edelleen. Mikäli puhuja on tunnistettu väärin, kreditoituu puhe väärälle puhujalle. Myös 1900-luvun kaikki puhemiehistön tiedot, sukunimestä alkaen, on jouduttu päättelemään roolin ja päivämäärän perusteella Eduskunnan toimijat -aineiston puhemiestiedoista.

Äärimmäisissä tapauksissa kaikki puhujaan liittyvät tiedot on jouduttu päättelemään. 1900-luvun pöytäkirjoissa on ajoittain viitattu puhujaan roolin ja sukunimen sijaan pelkällä 'Puhuja'-termillä. Tämä tapahtuu tilanteissa, joissa puhuja on aloittanut puheensa, mutta puhemies keskeyttää hänet. Mikäli puhemiehen keskeytys on eriytetty omaksi puheenvuorokseen (suluissa puheen sekaan upottamisen sijaan), katkeaa puhujan puhe. Puhemiehen kommentin jälkeen alkuperäisen puhujan puhevuoro jatkuu ikään kuin uutena puheenvuorona, mutta puhujatietona on tällä kertaa vain sana *Puhuja*. Tämä jälkimmäinen puheenpätkä poimitaan omaksi puheenvuorokseen, mutta sen puhuja on kopioitava sitä edeltävältä, ei-puhemiehen puheenvuorolta. Mikäli puheen alkuperäistä osaa ei jostain syystä ole saatu poimittua pöytäkirjoista oikein, voidaan puheen jälkiosa liittää väärään

puhujaan. Vastaava käytäntö ei ole erityisen yleinen, muttei myöskään yksittäistapaus, joten virheiden mahdollisuutta ei voida täysin poissulkea.

Edellä kuvatun kaltaisten päättelytoimien tarkoituksena on rikastaa aineistoa ja tarjota käyttäjälle lisää resursseja ja linkityksiä muihin keskeisiin aineistoihin. On kuitenkin keskeistä, että käyttäjälle välitetään selkeä kuva tarjolla olevan tiedon luonteesta ja siitä mistä mikäkin tieto on peräisin. Näin käyttäjä voi tehdä perusteltuja ja oikeassa suhteessa kriittisiä tulkintoja aineistosta tekemistään havainnoista. Tällaisten aineistojen *provenienssin*, tiedon datan alkuperästä ja kontekstista, rooli on nykytieteessä keskeinen kriteeri [59].

Puheenvuoroaineiston yhteydessä tietojen taustat raportoidaan tarkasti, mutta niiden omatoimisesta tarkistamisesta on myös tehty mahdollisimman helppoa. Jokaisen puheen yhteydessä on verkko-osoite kyseisen puheenvuoron sisältävään pöytäkirjan verkkoversioon. Puheenvuoron osoite osoittaa joko suoraan oikeaan asiakohtaan tai tiedoissa on kerrottu puheenvuoron sivunumero, jolloin käyttäjä voi vaivattomasti löytää alkuperäisen puheen tarkastaakseen sen tietojen paikkansapitävyyden. Lisäksi puhujan nimen alkuperäinen pöytäkirjasta poimittu kirjoitusasu on tallennettu RDF-versiossa puheen tietoihin, jolloin käyttäjä voi nopealla vilkaisulla tarkistaa, vaikuttaisiko linkitetty puhuja olevan oikea.

Oman haasteensa muodostaa myös hankkeessa luodun palvelun ja muunnosprosessin ylläpito. Työkalut ja prosessit on nyt luotu ja niillä halutaan tuottaa aineistoa myös hankkeen päätyttyä vuoden 2022 lopussa. Mutta kuka huolehtisi näiden työkalujen käytöstä jatkossa? Eduskunnan tietopalvelut itse, tutkimusryhmä vai joku erillinen taho? Entä kun seuraavan kerran pöytäkirjojen rakenteisiin tulee muutoksia; Kuka kykenee muokkaamaan tai jopa luomaan uusia työkaluja muuttuneisiin tarpeisiin, mikäli tämän hankkeen jäsenillä ei ole siihen resursseja? Näihin kysymyksiin tulisi löytää vastauksia ennen hankkeen päättymistä ja niistä onkin käyty keskusteluja muun muassa eduskunnan edustajien kanssa.

6.4 Jatkokehitys

Tässä tutkielmassa kuvatut työkalut ja prosessit on kontitettu Docker-työkalulla (ks. [15]) helppokäyttöiseksi prosessiksi. Konttiin pakattu prosessi on itsenäinen työkalu, jota voi käyttää erilaisilla alustoilla, eikä se Docker-ohjelman lisäksi vaadi muiden työkalujen asentelua. Käyttäjä pystyy yhdellä komennolla joko muuntamaan uudelleen koko aineiston hankkeessa valmisteltuja lähdeaineistoja hyödyntäen tai vain päivittämään sen kuluvien

valtiopäivien osalta. Puheaineisto on kuitenkin osa isompaa Semanttinen parlamentti -kokonaisuutta ja vaatii muun muassa ajantasaisen Eduskunnan toimijat -aineiston saatavuutta. Tätä varten tullaan hankkeessa vielä luomaan yksi yhtenäinen konttikokonaisuus, jolla voidaan suorittaa niin puhe- kuin henkilöaineiston muunnos ja julkaisu vaivatta ja täysin automatisoidusti.

Aineiston mittavuus tekee tällä hetkellä sen selailusta Parlamenttisampo-portaalissa paikoin hyvin hidasta. Portaalin kyselyt ovat usein hyvin raskaita. Portaalia ja RDF-muotoisen aineiston esittelyä tullaan vielä kehittämään sujuvamman käyttökokemuksen varmistamiseksi. Mahdollisia vaihtoehtoja ovat esimerkiksi vähemmän oleellisten puhemiesten puheenvuorojen ulossulkeminen oletuksena listauksista ja kyselyistä tai aineiston jakaminen eri aikakausien näkymiin, jolloin kerrallaan pyöritettäisiin pienempää osaa aineistosta.

Aineiston laatua tullaan vielä yleisestikin kehittämään ja esimerkiksi tekstintunnistustuloksia tullaan järjestelmällisemmin analysoimaan ja mahdollisia ongelmia voidaan jatkossa korjata esimerkiksi erilaisia sanakirjoja vasten. Tällä pyritään parantamaan perusteellisen kielellisen analyysiin mahdollisuuksia korjaamalla puheiden sisältöjen mahdollisia tekstintunnistusvirheitä. Muutenkin aineiston täydennysmahdollisuuksia tullaan tarkastelemaan lisää. Pohdinnassa on jo muun muassa aineiston kytkeminen sanomalehtiteksteihin, eduskunnan istuntojen videomateriaaleihin ja sosiaalisen median, kuten Twitterin keskusteluihin.

Semanttinen parlamentti halutaan kytkeä myös osaksi laajempaa eurooppalaista parlamentaaristen aineistojen verkostoa. Semanttisen laskennan -tutkimusryhmä on mukana jo aiemmin esitellyn ParlaMint-hankkeen toisessa vaiheessa. ParlaMint 2 käynnistyy vuoden 2022 alussa ja tähän toiseen vaiheeseen on liittynyt kymmenkunta uutta Euroopan maata, joissa joko on käynnissä tai juuri käynnistellään erilaisia parlamentaaristen aineistojen muunnoshankkeita. Suomesta, Semanttinen parlamentti -aineistosta tullaan toimittamaan ParlaMint-korpukseen valtiopäivät 2015–2021.

7 Yhteenveto

Tässä tutkielmassa esiteltiin Semanttinen parlamentti -konsortiohanke ja erityisesti yksi sen keskeisistä prosesseista: Suomen eduskunnan täysistuntojen puheenvuorojen muuntaminen linkitetyksi avoimeksi dataksi sekä parlamentaaristen aineistojen koodaukseen suunnitellun Parla-CLARIN-suositusten mukaiseen XML-muotoon. Tutkielmassa esiteltiin kuinka vuodesta 1907 alkaen pidetyt, lähes miljoona puheenvuoroa kerättiin, yhtenäistettiin, rikastettiin ja muunnettiin Semanttinen parlamentti -skeemojen mukaiseksi. Prosessin tarkoituksena oli luoda rikas aineisto, jota voidaan hyödyntää ohjelmallisesti niin rajapintojen ja verkkosovellusten kautta kuin myös muilla ladatulle aineistolle laadituilla työkaluilla.

Toisin kuin useimmissa suomalaisissa eduskunnan keskustelujen muunnosprojekteissa, Eduskunnan puheenvuorot -aineisto kattaa täysistunnot eduskunnan alusta, vuodesta 1907, aina nykypäivään. Prosessissa on myöskin muunnettu kaikki löydetyt puheenvuorot, myös ruotsinkieliset sekä *Keskustelu*-osioiden ulkopuoliset puheenvuorot. Puhetietoja on rikastettu muun muassa asiakohta ja -kirjatiedoilla, jolloin puheita voi hakea myös esimerkiksi tietyn hallituksen esityksen pohjalta. Tällöin olennaisten puheiden löytyminen ei ole esimerkiksi tiettyjen hakusanojen varassa. Puheet on myös linkitetty laajaan Eduskunnan toimijat -aineistoon, mikä puolestaan avaa uusia mahdollisuuksia puheiden tarkasteluun henkilö- ja puoluelähtöisesti. Muun muassa edellä mainittujen tekijöiden summana tässä tutkielmassa esitelty prosessi on Suomessa ainutkertainen.

Semanttinen parlamentti -hankkeen ollessa edelleen käynnissä on aineistokin edelleen, kuten todettu, kehitystyön alla. Prosessi ja sen tulokset ovat kuitenkin jo niin pitkälle, että on mielekästä vielä lopuksi arvioida luodun linkitetyn datan aineiston tasoa luvussa 2 esiteltyyn seitsemän tähden mallin mukaisesti. Vaikka aineisto on teknisesti ottaen jo julkaistu LDF.fi-alustalla ja Parlamenttisampo on toiminnassa, on pääsy niihin vielä rajattu. Seuraava tarkastelu tehdään siis siltä osin vuoden 2023 alkua, eli aineiston täysin vapaasti julkaisua ajatellen.

Tähdet 1–4 eittämättä toteutuvat: aineisto tulee olemaan verkossa kaikkien saatavilla (1), se on rakenteisessa muodossa (2), joka noudattaa avoimia standardeja (3) sekä käytössä ovat URI-tunnisteet (4). Viides tähti edellyttää aineiston linkittyvän URI-tunnistein myös muihin aineistoihin. Puheet linkittyvät eritoten Eduskunnan toimijat -aineistoon,

mutta myös hankkeen ulkopuolisiin kieli-, paikka- ja asiasanastoihin, joten viideskin (5) tähti toteutuu. Kuudennen (6) tähden lunastavat aineistosta laaditut skeema ja palveludokumentaatio, jotka julkaistaan aineiston yhteydessä. Aineiston provenienssi, eli tieto alkuperästä ja muunnoksista, on myös pyritty tekemään läpinäkyväksi dokumentaatiolla, tietojen selitteillä (esimerkiksi Parlamenttisammon infoboksit) sekä tarjoamalla helpon pääsyn alkuperäisiin lähteisiin vertailua varten. Aineistoa on myös arvioitu, joskin kuten todettu, arvioinnin parissa tullaan vielä työskentelemään. Jo toteutetuilla toimilla voitaneen jo kuitenkin alkaa vähintään värittämään seitsemännen tähden sakaraa (🌟).

Aineiston voitaneen siis nähdä olevan hyvällä mallilla, joskin tehtävää on vielä ja monia haasteita ratkaistavana. Eduskunnan puheet -aineisto, osana laajempaa Semanttinen parlamentti -aineistoa liittyy osaksi niin suomalaista linkitetyn avoimen datan kokonaisuutta ja semanttista webiä kuin myös eurooppalaista parlamentaaristen aineistojen korpusta.

Lähteet

[1] *5-Star Open Data*. URL: http://5stardata.info/en/ (viitattu 12. 01. 2022).

[2] Andrushchenko, M., Sandberg, K., Turunen, R., Marjanen, J., Hatavara, M., Kurunmäki, J., Nummenmaa, T., Hyvärinen, M., Teräs, K., Peltonen, J. & Nummenmaa, J. "Using Parsed and Annotated Corpora to Analyze Parliamentarians' Talk in Finland". *Journal of the Association for Information Science and Technology* (kesäkuu 2021). DOI: 10.1002/asi.24500.

[3] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. & Ives, Z. "DBpedia: A Nucleus for a Web of Open Data". Teoksessa: *The Semantic Web. ISWC 2007, ASWC 2007*. Lecture Notes in Computer Science. 2007, s. 722–735. DOI: 10.1007/978-3-540-76298-0_52.

[4] K. Baierer & T. Breuel, toim. *hOCR - OCR Workflow and Output embedded in HTML*. Helmikuu 2020. URL: http://kba.cloud/hocr-spec/1.2/ (viitattu 18. 12. 2021).

[5] *Beautiful Soup Documentation*. https://www.crummy.com/software/BeautifulSoup/bs4/doc/. (Viitattu 18. 12. 2021).

[6] Beckett, D., Berners-Lee, T., Prud'hommeaux, E. & Carothers, G. *RDF 1.1 Turtle. Terse RDF Triple Language*. Toim. E. Prud'hommeaux & G. Carothers. Helmikuu 2014. URL: https://www.w3.org/TR/turtle/ (viitattu 18. 12. 2021).

[7] Beelen, K., Thijm, T. A., Cochrane, C., Halvemaan, K., Hirst, G., Kimmins, M., Lijbrink, S., Marx, M., Naderi, N., Rheault, L., Polyanovsky, R. & Whyte, T. "Digitization of the Canadian Parliamentary Debates". *Canadian Journal of Political Science* 50.3 (2017), s. 849–864. DOI: doi:10.1017/S0008423916001165.

[8] Berners-Lee, T., Fielding, R. & Masinter, L. *RFC 3986. URI Generic Syntax*. Tammikuu 2005. URL: https://www.ietf.org/rfc/rfc3986.txt (viitattu 08. 01. 2022).

[9] Berners-Lee, T., Hendler, J. & Lassila, O. "The Semantic Web". *Scientific American* 284.5 (2001), s. 34–43. URL: http://www.jstor.org/stable/26059207.

[10] Bojārs, U., Darģis, R., Lavrinovičs, U. & Paikens, P. "LinkedSaeima: A Linked Open Dataset of Latvia's Parliamentary Debates". Teoksessa: *Semantic Systems. The Power of AI and Knowledge Graphs 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, September 9–12, 2019, Proceedings*. 2019, s. 50–56. DOI: 10.1007/978-3-030-33220-4.

[11] British Library. *The British National Bibliography as Linked Open Data*. https://bnb.data.bl.uk/. (Viitattu 21. 01. 2022).

[12] *CLARIN*. URL: https://www.clarin.eu/ (viitattu 18. 12. 2021).

[13] DBpedia. *About: Leonardo da Vinci*. URL: https://dbpedia.org/page/Leonardo_da_Vinci (viitattu 11. 02. 2022).

[14] DCMI Usage Board. *DCMI Metadata Terms*. Tammikuu 2020. URL: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ (viitattu 18. 12. 2021).

[15] Docker. *Use containers to Build, Share and Run your applications*. URL: https://www.docker.com/resources/what-container (viitattu 20. 12. 2021).

[16] Eduskunnan pöytäkirjatoimisto. *Kirjo - Kirjaamisohjeet*. Helsinki, Suomi: Eduskunnan kanslia, 2021.

[17] Eduskunta. *Digitoidut valtiopäiväasiakirjat 1907-2000*. URL: https://avoindata.eduskunta.fi/#/fi/digitoidut/download (viitattu 11. 02. 2022).

[18] Eduskunta. *Haku: Valtiopäiväasiat ja -asiakirjat*. URL: https://www.eduskunta.fi/FI/search/Sivut/Vaskiresults.aspx (viitattu 11. 02. 2022).

[19] Eduskunta. *Ohjeet*. URL: https://avoindata.eduskunta.fi/#/fi/about (viitattu 11. 02. 2022).

[20] Eduskunta. *Tietokantahaku*. URL: https://avoindata.eduskunta.fi/#/fi/dbsearch (viitattu 11. 02. 2022).

[21] Elo, K. & Karimäki, J. "Luonnonsuojelusta ilmastopolitiikkaan: Ympäristöpoliittisen käsitteistön muutos parlamenttipuheessa 1960–2020". *Politiikka* 63.4 (marraskuu 2021). DOI: 10.37452/politiikka.109690.

[22] Erjavec, T., Ogrodniczuk, M., Osenova, P., Pančur, A., Ljubešič, N., Agnoloni, T., Barkarson, S., Pérez, M. C., Çöltekin, Ç., Coole, M., Dargis, R., de Macedo, L. D., de Does, J., Depuydt, K., Diwersy, S., Hansen, D. H., Kopp, M., Krilavičius, T., Luxardo, G., Marx, M., Morkevičius, V., Navarretta, C., Rayson, P., Ring, O., Rudolf, M., Simov, K., Steingrímsson, S., Üveges, I., van Heusden, R. & Venturi, G. "ParlaMint: Comparable Corpora of European Parliamentary Data". Teoksessa: *Proceedings of CLARIN annual conference 2021, 27-29 September, 2021, virtual edition*. 2021, s. 20–25. URL: https://epubl.ktu.edu/object/elaba:108748986/.

[23] Erjavec, T. & Pančur, A. *Parla-CLARIN. A TEI Schema for Corpora of Parliamentary Proceedings v0.2*. Joulukuu 2020. URL: https://clarin-eric.github.io/parla-clarin/ (viitattu 18. 12. 2021).

[24] GeoNames. *GeoNames Ontology*. URL: http://www.geonames.org/ontology/documentation.html (viitattu 25. 01. 2022).

[25] Hitzler, P. "A review of the semantic web field". *Communications of the ACM* 64.2 (2021), s. 76–83. DOI: 10.1145/3397512.

[26] Hofer, M., Hellmann, S., Dojchinovski, M. & Frey, J. "The New DBpedia Release Cycle: Increasing Agility and Efficiency in Knowledge Extraction Workflows". Teoksessa: *Semantic Systems. In the Era of Knowledge Graphs. SEMANTICS 2020 Proceedings*. Lecture Notes in Computer Science. Syyskuu 2020. DOI: 10.1007/978-3-030-59833-4_1.

[27] Hyvönen, E. *Semanttinen web: linkitetyn avoimen datan käsikirja*. Helsinki: Gaudeamus, 2018.

[28] Hyvönen, E. "Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery". *Semantic Web* 11.1 (tammikuu 2020), s. 187–193. DOI: 10.3233/SW-190386.

[29] Hyvönen, E., Ikkala, E., Koho, M., Burrows, T., Ransom, L., Tuominen, J. & Wijsman, H. "Mapping Manuscript Migrations on the Semantic Web: A Semantic Portal and Linked Open Data Service for Premodern Manuscript Research". Teoksessa: *Proceedings of the 20th International Semantic Web Conference (ISWC 2021)*. Lokakuu 2021. DOI: 10.1007/978-3-030-88361-4_36.

[30] Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J. & Keravuori, K. "BiographySampo - Publishing and Enriching Biographies on the

Semantic Web for Digital Humanities Research". Teoksessa: *The Semantic Web. ESWC 2019*. Kesäkuu 2019, s. 574–589. DOI: 10.1007/978-3-030-21348-0_37.

[31] Hyvönen, E., Sinikallio, L., Leskinen, P., Drobac, S., Tuominen, J., Elo, K., La Mela, M., Koho, M., Ikkala, E., Tamper, M., Leal, R. & Kesäniemi, J. "Parlamenttisampo: eduskunnan aineistojen linkitetyn avoimen datan palvelu ja sen käyttömahdollisuudet". *Informaatiotutkimus* 40.3 (marraskuu 2021), s. 216–244. DOI: 10.23978/inf.107899.

[32] Hyvönen, E., Sinikallio, L., Leskinen, P., Drobac, S., Tuominen, J., Elo, K., La Mela, M., Koho, M., Ikkala, E., Tamper, M., Leal, R. & Kesäniemi, J. "Parliament of Finland on the Semantic Web: ParliamentSampo Data Service and Portal for Digital Humanities Research". Teoksessa: *6th Digital Humanities in Nordic and Baltic Countries Conference*. Hyväksytty julkaistavaksi. Maaliskuu 2022.

[33] Hyvönen, E., Tuominen, J., Alonen, M. & Mäkelä, E. "Linked Data Finland: A 7-Star Model and Platform for Publishing and Re-Using Linked Datasets". Teoksessa: *The Semantic Web: ESWC 2014 Satellite Events. ESWC 2014*. Lecture Notes in Computer Science. Lokakuu 2014, s. 226–230. DOI: 10.1007/978-3-319-11955-7_24.

[34] Ikkala, E., Hyvönen, E., Rantala, H. & Koho, M. "Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces". *Semantic Web* 13.1 (2022), s. 69–84. DOI: 10.3233/SW-210428.

[35] Isaac, A. & Haslhofer, B. "Europeana Linked Open Data - Data.Europeana.Eu". *Semantic Web* 4.3 (2013), s. 291–297. DOI: 10.3233/SW-120092.

[36] Juric, D., Hollink, L. & Houben, G.-J. "Bringing Parliamentary Debates to the Semantic Web." Teoksessa: *CEUR Workshop Proceedings 2012*. Vol. 902. 2012, s. 51–60.

[37] Koho, M., Ikkala, E., Leskinen, P., Tamper, M., Tuominen, J. & Hyvönen, E. "WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data". *Semantic Web* 12.2 (tammikuu 2021), s. 265–278. DOI: 10.3233/SW-200392.

[38] La Mela, M. "Tracing the Emergence of Nordic Allemansrätten through Digitised Parliamentary Sources". Teoksessa: *Digital histories: Emergent approaches within the new digital history*. Toim. M. Fridlund, M. Oiva & P. Paju. Helsinki University Press, 2020, s. 181–197. DOI: 10.33134/HUP-5-11.

[39] Lapponi, E., Søyland, M. G., Velldal, E. & Oepen, S. "The Talk of Norway: a richly annotated corpus of the Norwegian parliament, 1998–2016". *Language Resources and Evaluation* 52.3 (syyskuu 2018), s. 873–893. DOI: 10.1007/s10579-018-9411-5.

[40] Leskinen, P., Hyvönen, E. & Tuominen, J. "Members of Parliament in Finland Knowledge Graph and Its Linked Open Data Service". Teoksessa: *Further with Knowledge Graphs. Proceedings of the 17th International Conference on Semantic Systems, 6-9 September 2021, Amsterdam, The Netherlands*. Vol. 53. 2021, s. 255–269. DOI: 10.3233/SSW210049.

[41] Makkonen, K. & Loukasmäki, P. "Eduskunnan täysistunnon puheenaiheet 1999-–2014: Miten käsitellä LDA-aihemalleja?" *Politiikka* 61.2 (2019), s. 127–159.

[42] Mansikkaniemi, A., Smit, P. & Kurimo, M. "Automatic Construction of the Finnish Parliament Speech Corpus". Teoksessa: *Proc. Interspeech 2017*. 2017, s. 3762–3766. DOI: 10.21437/Interspeech.2017-1115.

[43] Martínez-Rodríguez, J.-L., Hogan, A. & López-Arévalo, I. "Information extraction meets the Semantic Web: A survey". *Semantic Web* 11.2 (2020), s. 255–335. DOI: 10.3233/SW-180333.

[44] Marx, M. "Advanced information access to parliamentary debates". *Journal of Digital Information* 10.6 (2009), s. 1–11.

[45] Miles, A. & Bechhofer, S. *SKOS Simple Knowledge Organization System Namespace Document - HTML Variant*. Elokuu 2009. URL: https://www.w3.org/2009/08/skos-reference/skos.html/ (viitattu 18. 12. 2021).

[46] Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A. & Taylor, J. "Industry-Scale Knowledge Graphs: Lessons and Challenges". *Communications of the ACM* 62.8 (heinäkuu 2019), s. 36–43. DOI: 10.1145/3331166.

[47] Odjik, J. *CLARIN ParlaFormat workshop*. Toukokuu 2019. URL: https://www.clarin.eu/blog/clarin-parlaformat-workshop (viitattu 03. 01. 2022).

[48] Palmirani, M. & Vitali, F. "Akoma-Ntoso for Legal Documents". Teoksessa: *Legislative XML for the Semantic Web: Principles, Models, Standards for Document Management*. Toim. G. Sartor, M. Palmirani, E. Francesconi & M. A. Biasiotti. Law, Governance and Technology Series. Dordrecht: Springer Netherlands, 2011, s. 75–100. DOI: 10.1007/978-94-007-1887-6_6.

[49] Pan, J. Z. ”Resource Description Framework”. Teoksessa: *Handbook on Ontologies*. Toim. S. Staab & R. Studer. International Handbooks on Information Systems. Berlin, Heidelberg: Springer, 2009, s. 71–90. DOI: 10.1007/978-3-540-92673-3_3.

[50] Pancur, A. & Erjavec, T. ”The siParl corpus of Slovene parliamentary proceedings”. Teoksessa: *Proceedings of the Second ParlaCLARIN Workshop*. Toukokuu 2020, s. 28–34.

[51] Pekonen, O. ”Debating "the ABCs of parliamentary life": the learning of parliamentary rules and practices in the late nineteenth-century Finnish Diet and the early Eduskunta”. Tohtorinväitöskirja. Jyväskylä: University of Jyväskylä, 2014.

[52] Po, L., Bikakis, N., Desimoni, F. & Papastefanatos, G. ”Linked Data Visualization: Techniques, Tools, and Big Data”. Synthesis Lectures on the Semantic Web: Theory and Technology 19 (maaliskuu 2020). DOI: 10.2200/S00967ED1V01Y201911WBE019.

[53] Prud'hommeaux, E., Labra Gayo, J. E. & Solbrig, H. ”Shape Expressions: An RDF Validation and Transformation Language”. Teoksessa: *Proceedings of the 10th International Conference on Semantic Systems*. SEM '14. Association for Computing Machinery, syyskuu 2014, s. 32–40. DOI: 10.1145/2660517.2660523.

[54] Publications Office of the European Union. *About data.europa.eu*. URL: https://data.europa.eu/en/about/about-dataeuropaeu (viitattu 21. 01. 2022).

[55] Rauh, C., De Wilde, P. & Schwalbach, J. *The ParlSpeech data set: Annotated full-text vectors of 3.9 million plenary speeches in the key legislative chambers of seven European states*. Versio V1. 2017. DOI: 10.7910/DVN/E4RSP9.

[56] Semantic Computing Research Group. *Linked Data Finland. Living Laboratory Data Service for the Semantic Web*. URL: https://www.ldf.fi/ (viitattu 18. 12. 2021).

[57] Semantic Computing Research Group. *Parlamenttisampo - eduskunta semanttisessa webissä*. URL: https://seco.cs.aalto.fi/projects/semparl/ (viitattu 24. 09. 2021).

[58] Semantic Computing Research Group. *SeCo Language Analysis Services*. URL: http://demo.seco.tkk.fi/las/# (viitattu 18. 12. 2021).

[59] Shadbolt, N., Berners-Lee, T. & Hall, W. ”The Semantic Web Revisited”. *IEEE Intelligent Systems* 21.3 (toukokuu 2006), s. 96–101. DOI: 10.1109/MIS.2006.62.

[60] Shenoy, K., Ilievski, F., Garijo, D., Schwabe, D. & Szekely, P. ”A Study of the Quality of Wikidata”. *Journal of Web Semantics* 72 (huhtikuu 2022). DOI: 10.1016/j.websem.2021.100679.

[61] *ShEx - Shape Exprssions*. URL: https://shex.io/ (viitattu 20. 12. 2021).

[62] Simola, S. *A Century of Partisanship in Finnish Political Speech*. Luonnos. 2020. URL: https://sites.google.com/site/sallasimolaecon/home/research (viitattu 11. 11. 2021).

[63] Sinikallio, L., Drobac, S., Tamper, M., Leal, R., Koho, M., Tuominen, J., La Mela, M. & Hyvönen, E. "Plenary Debates of the Parliament of Finland as Linked Open Data and in Parla-CLARIN Markup". Teoksessa: *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Vol. 93. 2021, 8:1–8:17. DOI: 10.4230/OASIcs.LDK.2021.8.

[64] *SPARQL Endpoint interface to Python*. https://sparqlwrapper.readthedocs.io/en/latest/main.html. (Viitattu 18. 12. 2021).

[65] Steffen, S. & Studer, R., toim. *Handbook on Ontologies*. 2nd ed. 2009. International Handbooks on Information Systems. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

[66] Steingrímsson, S., Barkarson, S. & Örnólfsson, G. T. "IGC-Parl: Icelandic Corpus of Parliamentary Proceedings". Teoksessa: *Proceedings of the Second ParlaCLARIN Workshop*. Toukokuu 2020, s. 11–17.

[67] Suomen Akatemia. *Digitaaliset ihmistieteet - DIGIHUM (2016-2022)*. https://www.aka.fi/tutkimusrahoitus/ohjelmat-ja-muut-rahoitusmuodot/akatemiaohjelmat/digitaaliset-ihmistieteet---digihum-2016-2022/. (Viitattu 18. 12. 2021).

[68] TEI Consortium. *P5: Guidelines for Electronic Text Encoding and Interchange*. Elokuu 2021. URL: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html (viitattu 18. 12. 2021).

[69] The W3C SPARQL Working Group, toim. *SPARQL 1.1 Overview*. Maaliskuu 2013. URL: https://www.w3.org/TR/sparql11-overview/ (viitattu 18. 12. 2021).

[70] Thornton, K., Solbrig, H., Stupp, G., Labra Gayo, J., Mietchen, D., Prud'hommeaux, E. & Waagmeester, A. "Using Shape Expressions (ShEx) to Share RDF Data Models and to Guide Curation with Rigorous Validation". Teoksessa: *The Semantic Web. ESWC 2019.* Vol. 11503. Lecture Notes in Computer Science. 2019, s. 606–620. DOI: 10.1007/978-3-030-21348-0_39.

[71] Triply. *Yasgui*. URL: https://triply.cc/docs/yasgui (viitattu 18. 12. 2021).

[72] W3C OWL Working Group, toim. *OWL 2 Web Ontology Language*. Joulukuu 2012. URL: https://www.w3.org/TR/2012/REC-owl2-primer-20121211/ (viitattu 24. 01. 2022).

[73] van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M. & Beunders, H. "The debates of the European Parliament as Linked Open Data". *Semantic Web* 8.2 (2017), s. 271–281. DOI: 10.3233/SW-160227.

[74] Voutilainen, E. "Tekstilajitietoista kielenhuoltoa: puheen esittäminen kirjoitettuna eduskunnan täysistuntopöytäkirjoissa". Teoksessa: *Puheesta tekstiksi – Puheen kirjallisen esittämisen alueita, keinoja ja rajoja*. Toim. L. Tiittula & P. Nuolijärvi. Suomalaisen Kirjallisuuden Seura, 2016, s. 162–191.

[75] Vrandečić, D. & Krötzsch, M. "Wikidata: A Free Collaborative Knowledgebase". *Communications of the ACM* 57.10 (2014), s. 78–85. DOI: 10.1145/2629489.

Liitteet

Liite A Eduskunnan puheet -tietoverkon skeema

Tästä skeemassa on kuvattu puheiden keräys- ja muutostyön yhteydessä luotu eduskunnan puheiden tietomalli. Luvussa 4.4 mainittujen, puheaineistoon linkittyvien luokkien *Speaker*, *Party*, *ReferencedNamedEntity*, *Concept* ja *ElectoralTerm* tarkkaa kuvausta ei ole tässä yhteydessä, sillä ne on toteutettu toisaalla.

Element type	Element (=predicate)	Cardinality	Range (=object)	Value examples	Notes	Might exist in final data:	Readily available in source formats:

Namespaces:

:	**http://ldf.fi/schema/semparl/**
semparl_linguistics:	**http://ldf.fi/schema/semparl/linguistics/**
bioc:	**http://ldf.fi/schema/bioc/**
rdf:	**http://www.w3.org/1999/02/22-rdf-syntax-ns#**
rdfs:	**http://www.w3.org/2000/01/rdf-schema#**
xsd:	**http://www.w3.org/2001/XMLSchema#**
dct:	**http://purl.org/dc/terms/**
speeches:	**http://ldf.fi/schema/semparl/speeches/**

Speech <http://ldf.fi/semparl/speeches/s1988_100_1>

Speaker	:speaker	0..1	bioc:Person	<http://ldf.fi/semparl/people/p910985>	Person URI interpreted from speaker name. *Can be empty for chairmen	All*	
	:party	0..1	:Party	<http://ldf.fi/semparl/groups/Q506591>	Party URI interpreted from speaker name. *Can be empty for chairmen	All*	
	:roleGivenInSource	1	rdfs:LangString	Opetusministeri	Role of the speaker that was mentioned	All	All
	:speakerAsInSource	1	rdfs:Literal	Ed. Turunen	Speaker's name as written in the source	All	All
	:parliamentaryRole	0..1	:ParliamentaryRole	<http://ldf.fi/semparl/groups/Oppositiopuolue>	Parliamentary role of the speaker's party (Government, opposition, garetaker government, government official)	All	
	:groupOfSpeaker	0..1	bioc:Group	<http://ldf.fi/semparl/groups/g3367533099705406466>	Group the speaker is part of (parliamentary group or body, committee, etc)	All	
Speech	skos:prefLabel	1	rdf:langString	"Vp 2021 - istunto 2 - puhe 3 (Veikko Vennamo)"@fi	Unique string label for speech	All	
	:orderNumber	1	xsd:integer	1	Order number of the speech in a session	All	
	:content	1	rdfs:Literal	Herra puhemies! Minä puhun nyt!	Original speech content, without end-of-line hyphenation	All	All
	dct:language	0..*	rdfs:Resource	<http://id.loc.gov/vocabulary/iso639-2/fin>	Automatically recognized languages used in the speech (fi and/or sv). (XML source has language info, some issues there, currently not used)	All	
	:speechType	1	:SpeechType	<http://ldf.fi/semparl/speechtypes/Puheenvuoro>	Type of of the speech.	All	All
	:isInterruptedBy	0..*	:Interruption	<http://ldf.fi/semparl/speeches/in1908_II_10_12_1>	URI of interruption found in the speech content	All	All
Named entities	semparl_linguistics: referencedNamedEntity	0..*	semparl_linguistics:NamedEntity	semparl_linguistics:ne_1970_1	List of named entities referenced in a speech	All	WIP
Temporal info	dct:date	1	xsd:date	YYYY-MM-DD	Date of session start	All	All
	:endDate	1	xsd:date	YYYY-MM-DD	Date of session end (session migth go over midnight)	All	
	:startTime	0..1	xsd:time	hh:mm:ss	Speech specific start time	XML	XML
	:endTime	0..1	xsd:time	hh:mm:ss	Speech specific end time	XML	XML
	:yearOfSpeech	1	xsd:integer	1975	Calendar year the speech was held on	All	All
Item	:item	0..1	:Item	<http://ldf.fi/semparl/items/i1935100751>	URI of related topic/item in minutes (asiakohta)	All	
Session	:plenarySession	1	:PlenarySession	<http://ldf.fi/semparl/times/plenary-sessions/ps_100_1935>	URI to plenary session where speech was held	All	
Reference	:diary	1	rdfs:Resource	<https://s3-eu-west-1.amazonaws.com/eduskunta-asiakirja-origir	URL of the online book	All	
	:page	0..1	xsd:integer	357	Page number where speech is found in original pdf	PDF	PDF
Status	:status	0..1	:Status	<http://ldf.fi/semparl/status/Ready>	Speech specific status of the speech transcript	XML	XML
Version	:version	0..1	xsd:decimal	1.1	Speech specific version of the speech transcript	XML	XML
	dct:subject	0..n	skos:Concept		Extracted subject matter	NOT YET IMPLEMENTED	

Element type	Element (=predicate)	Cardinality	Range (=object)	Value examples	Notes	Might exist in final data:	Readily available in source formats:
Interruption				<http://ldf.fi/semparl/speeches/in1908_II_10_12_1>			
	:content	1	rdfs:Literal	Hyvä puhe! Naurua Puhemies koputtaa	Interruption content. If a clear interruptor was named (e.g. in 'Ed. Kaikkonen: Hyvä puhe!'), it has been removed from the content	All	All
	:interrupter	0..1	rdfs:Literal	"Vasemmalta"	The source of interruption if mentioned; a named entity, be it group, general direction or person	All	All
	:speaker	0..1	bioc:Person	<http://ldf.fi/semparl/people/p1274>	URI to person interrupting, if they have been clearly named	All	All
	skos:prefLabel	1	rdf:langString	"Vp 1935 - istunto 78 - puhe 86 - Keskeytys - 9"@fi	Unique interruption label	All	All
	:yearOfSpeech	1	xsd:integer	1975	Calendar year of the interruption	All	All
	dct:date	1	xsd:date	YYYY-MM-DD	Date of interruption	All	All
Item				<http://ldf.fi/semparl/items/i1908_II1034>			
	:plenarySession	1	:PlenarySession	<http://ldf.fi/semparl/times/plenary-sessions/ps_100_1935>	URI to plenary session where the item was on the agenda	All	All
	skos:prefLabel	1	rdf:langString	"Toisen varapuhemiehen vaali"@fi	Title of the item	All	All
	:relatedDocument	0..*	:Document	<http://ldf.fi/semparl/documents/M_5_2019_vp>	URI to a document related to the item	All	
	:diary	1..2	rdfs:Resource	<https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTk	URL of the online book	All	HTML
Document (sub-classes: GovernmentProposal, LegislativeMotion, Interpellation, Account, Debate				<http://ldf.fi/semparl/documents/TAMINO2_30_1909_II>			
	skos:prefLabel	1	rdf:langString	"Valiokunnan mietintö HaVM 6/2015 vp"@fi .	Title of the document as in source	All	All
	skos:altLabel	0..*	rdf:langString	"Valiokunnan mietintö HaVM 6/2015"@fi .	Possible alternative wording of the document title when referred to on multiple occasions	All	All
	:id	0..1	xsd:string	HaVM 6/2015 vp	Official id of the document	HTML, XML	HTML, XML
	:url	0..1	rdfs:Resource	<https://www.eduskunta.fi/valtiopaivaasiakirjat/TaVM+1/2012>	URL to online version of the document	All	HTML
GovermentProposal (subclass of Document, also properties from Document)							
	dct:title	1	rdf:langString	"Hallituksen esitys eduskunnalle laeiksi lääkelain sekä Lääkealan turvallisuus- ja kehittämiskeskuksesta annetun lain muuttamisesta"@fi	Title of the document as in source	XML	XML
	dct:subject	0..*	rdf:langString	"apteekit"@fi,	YSO keywords	XML	XML
	:decision	0..1	rdf:langString	"Hyväksytty"@fi	Decision made for the proposal	XML	XML
	:dateDecided	0..1	xsd:date	YYYY-MM-DD	Date of the decision made for the proposal	XML	XML
PlenarySession (=täysistunto)				<http://ldf.fi/semparl/times/plenary-sessions/ps_10_1908_II>			
	skos:preflabel	1	rdf:langString	“3/1975 II vp”@fi	Id of the plenary session	All	All
	:parliamentarySession	1	:ParliamentarySession	<http://ldf.fi/semparl/times/parliamentary-sessions/parl_ses_1935_v>	Parliamentary session the plenary session is part of	All	

Element type	Element (=predicate)	Cardinality	Range (=object)	Value examples	Notes	Might exist in final data:	Readily available in source formats:
Temporal information	dct:date	1	xsd:date	YYYY-MM-DD	Start date of the session	All	All
	:orderNumber	1	xsd:integer	1	Order number of the plenary session in that parliamentary session	All	All
	:endDate	1	xsd:date	YYYY-MM-DD	End date of the session (might run over midnight)	All	
	:startTime	0..1	xsd:time	hh:mm:ss	Start time of the session	All	All
	:endTime	0..1	xsd:time	hh:mm:ss	End time of the session	All	All
Transcript	:transcript	1	:Transcript	<http://ldf.fi/semparl/documents/ptk_13_1931>	URI of minutes transcript	All	All
ParliamentarySession (=valtiopäivät)				<http://ldf.fi/semparl/times/parliamentary-sessions/parl_ses_2021>			
	skos:preflabel	1	rdf:langString	"Valtiopäivät 1975 II"@fi	Id of the parliamentary session	All	All
	:electoralTerm	0..1	:ElectoralTerm	<http://ldf.fi/semparl/times/electoral-terms/e_1975-09-27-1979-03-23>	The electoral term the parliamentary session is part of	All	
Temporal information	:startDate	1	xsd:date	YYYY-MM-DD	Start date of the session	All	
	:endDate	0..1	xsd:date	YYYY-MM-DD	End date of the session (missing for current)	All	
Transcript				<http://ldf.fi/semparl/documents/ptk_85_1931>			
	skos:prefLabel	1	rdf:langString	"PTK 3/1975 II vp"@fi	Id of the minutes transcript, second parliamentary session marked by 'II'	All	
	:status	0..1	:Status	<http://ldf.fi/semparl/status/Validated>	Status of the transcript	XML	XML
	:version	0..1	xsd:decimal	1.1	Version of the transcript	XML	XML
	:url	1	rdfs:Resource	<https://www.eduskunta.fi/FI/Vaski/sivut/trip.aspx?triptype=Valtio	URL to used online transcript	All	
SpeechType							
	skos:prefLabel	1	rdf:langString	"Varsinainen puheenvuoro"@fi		All	All
Status							
	skos:prefLabel	1	rdf:langString	"Tarkistettu"@fi		XML, HTML	XML
Language							
	skos:prefLabel	1	rdf:langString	"suomi"@fi, "finska"@sv, "Finnish"@en		All	
ParliamentaryRole							
	skos:prefLabel	1	rdf:langString	"Oppositiopuolue"@fi		All	

Liite B Parla-CLARIN-muotoisen aineiston rakenne

Tässä liitteessä on hahmoteltu Parla-CLARIN-formaatissa olevien Eduskunnan puheet -aineiston tiedostojen rakenne. Yksi tiedosto sisältää yksien valtiopäivien puheenvuorot ja kuvailutiedot. Rakennetta on avattu luvussa 4.5.

Rough Structure of the XML format

<teiCorpus> All speeches from one Valtiopävät
  <teiheader>
    … metadata... see picture 1
      <particDesc> All speakers and parties of the whole document
        … metadata... see pictures 2 and 3
  <TEI> One session
    <teiheader>
      ...metadata... see picture 4
    <text> container for body
      <body> Speeches from that one session
         Speeches related to to the same topic (or topicless subsequent speeches)
          <head> Topic (if there is one)
            <listBibl> Possible listed related documents
           One speech
            <note> speech metadata
            <u> actual speech
            <vocal> Possible interruptions during/ right after speech

A **<note>**-element has attributes:

- type: value is 'speaker'
- speechType: default is 'Puheenvuoro'
- link: link to the original source/Internet resource
- xml:id: unique ID for the entire speech (see *)

It *may* have attributes:

- page: page number of the start of the speech in the original pdf (if such was used as the source)
- start:** starting datetime of that speech (exists only Valtiopäivät 2015→)
- end:** ending datetime of that speech (exists only Valtiopäivät 2015→)

An **<u>**-element *has* attributes:

- who: reference to speaker metadata
- xml:id: unique ID for that speech/speech part

It *may have* attributes:

- prev:* ID of the previous part of the speech
- next:* ID of the next part of the speech

See examples in picture 4

A **<vocal>**-element *may have* attribute:

- who: reference to speaker metadata if the interrupter has been named in the source material

More information:

* Possible interjections and reactions from others have been separated from each speech by splitting the original speech transcription into actual speech and interruption segments. The original speech's parts share the same start of an ID plus an extra digit to order the separated parts.

Eg. Original ID and speech:

2010_5_12 "I would like to start (Chairman: Silence, please!) by telling a tale."

becomes:

2010_5_12.1 "I would like to start"
"Chairman: Silence, please!"
2010_5_12.2 "by telling a tale."

** Time references to the whole speech, not just an individual part (see * above)

Examples of data structure:
Picture 1:

```
<?xml version="1.0" ?>
<teiCorpus xml:id="Speeches_2020" xml:lang="eng">
   <teiheader>
      <fileDesc>
         <titleStmt>
            <title>Finnish Parliament plenary session speeches during Diet 2020</title>
            <title>Suomen eduskunnan täysistuntojen puheet Valtiopäivillä 2020</title>
         </titleStmt>
         <extent>
            <measure quantity="5735" unit="speeches">5735 speeches</measure>
         </extent>
      </fileDesc>
      <encodingDesc>
         <classDecl>
            <taxonomy>
               <desc>Types of chairmen</desc>
               <category xml:id="elderMember">
                  <catDesc>
                     <term>Ikäpuhemies</term>
                  </catDesc>
               </category>
               <category xml:id="chairman">
                  <catDesc>
                     <term>Puhemies</term>
                  </catDesc>
               </category>
               <category xml:id="secondViceChair">
                  <catDesc>
                     <term>Toinen varapuhemies</term>
                  </catDesc>
               </category>
               <category xml:id="viceChair">
                  <catDesc>
                     <term>Ensimmäinen varapuhemies</term>
                  </catDesc>
               </category>
            </taxonomy>
         </classDecl>
      </encodingDesc>
      <profileDesc>
         <settingDesc>
            <setting>
               <name key="FIN" type="country">Finland</name>
            </setting>
         </settingDesc>
         <particDesc>
            <listPerson>…
            </listPerson>
            <listOrg>…
            </listOrg>
         </particDesc>
      </profileDesc>
   </teiheader>
```

Picture 2:

```
<particDesc>
  <listPerson>
    <person xml:id="Kauko_Tuupainen">
      <persName>
        <surname>Tuupainen</surname>
        <forename>Kauko</forename>
        <roleName>Ikäpuhemies</roleName>
      </persName>
      <sex value="M">Male</sex>
      <birth when="1940-01-01"/>
      <affliation ref="#party.PS"/>
    </person>
    <person xml:id="Ben_Zyskowicz">
      <persName>
        <surname>Zyskowicz</surname>
        <forename>Ben</forename>
        <roleName>Puhemies</roleName>
      </persName>
      <sex value="M">Male</sex>
      <birth when="1954-01-01"/>
      <affliation ref="#party.KOK"/>
    </person>
    <person xml:id="Jutta_Urpilainen">
      <persName>
        <surname>Urpilainen</surname>
        <forename>Jutta</forename>
        <roleName>Ensimmäinen varapuhemies</roleName>
      </persName>
      <sex value="F">Female</sex>
      <birth when="1975-01-01"/>
      <affliation ref="#party.SDP"/>
    </person>
    <person xml:id="Anssi_Joutsenlahti">
      <persName>
```

Picture 3:

```
<particDesc>
  <listPerson>…
  </listPerson>
  <listOrg>
    <org role="Oppositiopuolue" xml:id="SDP"/>
    <org role="Hallituspuolue" xml:id="PS"/>
    <org role="Hallituspuolue" xml:id="KOK"/>
    <org role="Oppositiopuolue" xml:id="VIHR"/>
    <org role="Hallituspuolue" xml:id="KESK"/>
    <org role="Oppositiopuolue" xml:id="VAS"/>
    <org role="Oppositiopuolue" xml:id="SD"/>
    <org role="Oppositiopuolue" xml:id="R"/>
    <org role="Oppositiopuolue" xml:id="KD"/>
    <org role="other" xml:id="SININEN_TULEVAISUUS"/>
    <org role="Oppositiopuolue" xml:id="LIIK"/>
    <org role="Hallituspuolue" xml:id="RKP"/>
  </listOrg>
</particDesc>
```

Picture 4:

```
<TEI xml:id="ptk_10_2019">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>PTK 10/2019</title>
      </titleStmt>
      <sourceDesc>
        <bibl>
          <edition n="5.0">Valmis</edition>
        </bibl>
      </sourceDesc>
    </fileDesc>
    <profileDesc>
      <settingDesc>
        <setting>
          <date when="2019-06-04">2019-06-04</date>
          <time from="14.00" to="14.46"/>
        </setting>
      </settingDesc>
    </profileDesc>
  </teiHeader>
  <text>
    <body>
      <div>
        <head>Ilmoituksia - Kertomukset</head>
        <div>
          <note link="https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_10+2019+2.2.aspx" speechType="PuhemiesPuheenvuoro" type="speaker" x
          <u ana="#secondViceChair" who="#Paula_Risikko" xml:id="2019_10_1">Eduskunnalle on annettu Valtiontalouden tarkastusviraston erilliskertomus edu
        </div>
      </div>
      <div>
        <head>
          Ilmoituksia - Muut ilmoitukset
          <listBibl>
            <head>Related documents:</head>
            <bibl>Pöytäkirjan liite</bibl>
          </listBibl>
        </head>
        <div>
          <note link="https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_10+2019+2.4.aspx" speechType="PuhemiesPuheenvuoro" type="speaker" x
          <u ana="#secondViceChair" who="#Paula_Risikko" xml:id="2019_10_2">Tasavallan presidentin avoin kirje valtioneuvostossa toimeenpannusta muutokse
        </div>
      </div>
      <div>
        <head>
          Valtiontalouden tarkastusviraston  erilliskertomus  eduskunnalle valtion vuoden 2018 tilinpäätöksen ja hallituksen vuosikertomuksen tarkastukse
          <listBibl>
            <head>Related documents:</head>
            <bibl>Kertomus  K 13/2019 vp</bibl>
          </listBibl>
        </head>
        <div>
          <note link="https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_10+2019+3.aspx" speechType="PuhemiesPuheenvuoro" type="speaker" xml
          <u ana="#secondViceChair" who="#Paula_Risikko" xml:id="2019_10_3">Lähetekeskustelua varten esitellään päiväjärjestyksen 3. asia. Puhemiesneuvos
        </div>
        <div>
          <note end="2019-06-04T14:05:38" link="https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_10+2019+3.aspx" speechType="Puheenvuoro"
          <u who="#Anna-Kaisa_Ikonen" xml:id="2019_10_4">Arvoisa puhemies! Kiitos Valtiontalouden tarkastusvirastolle tästä erilliskertomuksesta. Teen si
        </div>
        <div>
          <note end="2019-06-04T14:08:40" link="https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_10+2019+3.aspx" speechType="Puheenvuoro"
          <u next="2019_10_5.2" who="#Timo_Heinonen" xml:id="2019_10_5.1">Arvoisa rouva puhemies! Edustaja Ikonen piti kokeneena kuntien ja kaupunkien ja
          <vocal who="Paavo_Arhinmäki">
            <desc>Paavo Arhinmäki: Eduskunta sen kumoaa eikä hallitus!</desc>
          </vocal>
          <u prev="2019_10_5.1" who="#Timo_Heinonen" xml:id="2019_10_5.2">Hallitusohjelmaa kun lukee, niin näyttää, että sitä ei ole kumottu, vaan aktiiv
        </div>
```