LIETUVOS PARLAMENTARŲ ŽODYNO DUOMENŲ BAZĖS PROJEKTAS
(PAGAL „LIETUVOS STEIGIAMOJO SEIMO (1920–1922 METŲ) NARIŲ BIOGRAFINĮ ŽODYNĄ“)

Vigintas Stancelis
Vilniaus pedagoginis universitetas
Vilnius Pedagogical University
T. Ševčenkos g. 31, LT-03111 Vilnius
El. paštas: stancelis@vpu.lt

Santrauka
Straipsnyje pristatoma ir aptariama duomenų bazė, sudaroma pagal Lietuvos Respublikos Steigiamojo Seimo (1920–1922 metų) narių biografinį žodyną. Analizuojamas biografinės duomenų bazės pritaikymo tikslingumas ir galimos panaudojimo sferos, aptariami gimininginį užsienio tyrinėtojų projektai. Išsamiai nagrinėjama darbe pritaikyta informacijos atrinkimo ir grupavimo metodika, atskleidžiama įrašo apie parlamentarą struktūra, pagrindžiamas programinės įrangos pasirinkimas. Aptariami pradiniame etape pasiekti rezultatai ir numatomos tolesnės projekto vykdymo gairės.

Esminiai žodžiai: informacijos valdymas; grupinis darbas; duomenų bazė; statistinė informacija; statistinė analizė; Steigiamasis Seimas; biografinis žodynas; biografiniai duomenys; personalija; parlamentaras; kompiuterija; internetas; Microsoft Access.

Sudarant Lietuvos parlamentarų žodyną numatoma sukaupti kiek galima detalesnę medžiagą apie visus XV–XXI a. aukščiausios atstovaujamosios valdžios asmenis, sudaryti jų biogramas, pateikti svarbiausius apibendrinimus. Dar negalutiniais duomenimis, reikės surinkti informaciją apie 3 000–4 000 XVI–XXI a. pradžios parlamentarų. Kiekvieną asmenį apibūdins gausybė parametrų – nuo pavardės, tautiškumo, konfesijos iki politinių pažiūrų ir aktyvumo Seimo kadencijos metu. Daugiatomiu šis leidinys priverstas tapti ne tik dėl nemažo laiko intervalo, kurį užims jo sudarymas, bet ir dėl žymiai buitiškesnės priežasties – su tokia didžiule knyga būtų nepatogu dirbti. O dabar iš to kylantis klausimas – kaip aprėpti, suvokti ir deramai išanalizuoti smegenimis tai, ką net rankomis pakelti sunku?

Rašant straipsnį pirmasis žodyno tomas „Lietuvos Steigiamojo Seimo (1920–1922 m.) narių biografinis žodynas“ jau buvo baigtas ir vyko baigiamieji redagavimo darbai. Nueitas pirmasis etapas leidžia daryti pirmąsias išvadas, siūlyti pataisas ar koreguoti kursą. Virš 500 juodraščio puslapių iškėlė nemažai iššūkių tvarkant informaciją, bandant ją pateikti greitai suvokiamų statistinių lentelių ar diagramų formatu. Šio straipsnio uždavinys – apibendrinti atliktą darbą ir bandyti prognozuoti tolesnę veiksmų programą, kad galutinis leidinys taptų prieinamas ir patogus būsimam vartotojui.

Vienas iš numatomų kelių – aktyvesnis kompiuterinių technologijų panaudojimas, konkrečiai – duomenų bazės sudarymas. Straipsnyje aprašytas šioje srityje atliktas darbas, pristatytas medžiagos atrinkimo klausimynas, grupavimo ir klasifikavimo mechanika, atliktos analizės rezultatai ir kol kas neįveiktos problemos.

Kompiuterinės duomenų bazės reikalingumas

Parlamentarų žodyno išleidimas senuoju, knyginiu formatu neišvengiamai turi tapti reikšmingu pasiekimu. Tokio masto ir apimties leidiniai vėliau tarnauja kaip dirva, kurioje ***ieškoma*** informacijos tolesniems tyrimams ir apibendrinimams. Tačiau spartėjantis mūsų gyvenimo būdas kelia aštrų klausimą: ***kaip greitai surandama***? Ar bus lengva daugiatomyje leidinyje greitai suskaičiuoti tai, kas gali parūpti būsimam tyrinėtojui: kiek buvo bajorų, o kiek stačiatikių, aukštus mokslus baigusių arba kariuomenėje tarnavusių Seimo narių?

Uždavinys skambėtų lyg ir naujai, tačiau sprendimas lieka toks pat kaip ir nuo žilos senovės – sunkiam darbui geriausiai tinka vergai. Šioje epochoje lengviausia įsigyti elektroninį, ir dažnas iš mūsų jau turime net po kelis. Rutininis darbas su vieno tipo, pasikartojančia informacija kaip tik tinkamiausia kompiuterio panaudojimo sfera. Neverta tikėtis, kad mašina už mus atliks analizę ir parašys išvadas, tačiau varginantį paieškos darbą labai palengvins ir paspartins. Lieka tik nukreipti ją tinkamu keliu, pamokyti ištraukti iš nuoseklaus, naratyvinio biogramų teksto istorikui aktualią statistiką.

Dalis tyrinėtojų skaičius ir statistiką laiko nenuginčijamais įrodymais, kita dalis tvirtina priešingai, esą specialiai atrinktais skaičiais galima įrodyti net ir tai, kas neegzistuoja. Nors Markas Tvenas rašė, kad „yra

trys melo rūšys – melas, įžūlus melas ir statistika“, vis dėlto statistika gali pateikti tiek daug ir tiek svarbios tiriamosios medžiagos, kad istorikui jau nebeleistina to ignoruoti.

Matematinės statistikos metodai praplečia tradicinio istorinio tyrimo galimybes, jo problematiką. Be to, tai ne tik ekstensyvi temų ir informacijos kiekių plėtra, bet ir savotiška kokybės kontrolė, skelbiamų teiginių pagrįstumo rodiklis. Statistiniai apibendrinimai labai efektyviai atskleidžia ir išryškina neapdairiai paliktas „baltąsias dėmes“, o tai vėliau gali pasitarnauti tyrimų plėtrai ir tobulinimui.

Iš to istorikui kyla du uždaviniai:

- išmokti kelti „teisingus“ klausimus jau sukauptiems statistiniams duomenims;
- išmokti kaupti, skaityti ir analizuoti kiekybinius duomenis, kad iš to būtų galima gauti išsamias ir pagrįstas išvadas.

Duomenų bazė (toliau – DB) – populiarėjantis istorikų aplinkoje terminas, ir nemažai jų daliai tai nebe tik žodis, bet ir įrankis. Vis dėlto pats termino taikymas tiek tarp istorikų, tiek kompiuterijos specialistų šiuo metu yra labai laisvas ir apima didžiulį spektrą informacijos saugojimo ir valdymo būdų. Pagal apibrėžimą duomenų baze laikytinas bet kuris struktūrizuotas susijusių duomenų rinkinys apie vieną ar daugiau informacijos objektų[1], šiuo atveju – konkretų parlamentarą. Duomenis sudaro tam tikras faktų rinkinys, o naudinga informacija jie tampa tik po to, kai sutvarkomi, struktūrizuojami pagal kurį nors tikslingai pasirinktą principą[2].

Įprastai istorikui pritaikytos duomenų bazės kūrimas prasideda vienu iš dviejų kelių:

- imamasi jau turimų duomenų analizės ir galvojama apie tai, į kokius klausimus pavyktų atsakyti išnagrinėjus medžiagą;
- išsikeliami tyrinėtojui aktualūs klausimai ir jau po to ieškoma duomenų, kurie padėtų į juos atsakyti.

Atsižvelgiant į tai, kad parlamentarų žodynas yra tęstinis leidinys, prisideda ir papildomas – laiko – veiksnys. Medžiaga renkama, apdorojama ir skelbiama atskirais etapais, todėl gautos pirmosios išvados galės būti panaudotos vėlesnei darbo plėtrai. Ankstyva empirinė, kad ir ne visai sistemingos informacijos analizė gali tapti pamatu kur kas sklandesnei, teoriškai pagrįstai koncepcijai, pagal kurią vėl būtų ieškoma trūkstamų ar naujų duomenų. Surinkti duomenys galės būti panaudoti ne tik šiame projekte, bet ir kituose socialinio elito tyrimuose.

Prieš pasineriant į pristatomo tyrimo problematiką ir tuo labiau technologinę specifiką, verta pasikartoti tradicinį bet kurios DB projektavimo klausimyną, atsakymai į kurį išsamiau buvo aptarti ankstesniame straipsnyje[3].

1. Kokie duomenys reikalingi būsimam tyrimui?
2. Kokia turės būti surinktos informacijos apimtis?
3. Ar pakaks atitinkamos kompetencijos personalo duomenims kaupti ir apdoroti?
4. Kaip bus užtikrinta informacijos apsauga nuo praradimo?
5. Kaip surinkti duomenys bus panaudoti pagrindinio tyrimo rėmuose?

Pirminė analizė leido pakankamai sėkmingai atsakyti į daugumą čia iškilusių reikalavimų. Steigiamojo Seimo narių biogramos buvo sudaromos remiantis 20 punktų klausimynu, leidžiančiu užtektinai detaliai aprašyti parlamentaro asmenybę ir veiklą. Surinkti duomenys apie atskirus asmenis vėliau buvo apibendrinti statistinėse lentelėse ir pasitarnavo rašant analitinę leidinio dalį. Kol kas santykinai maža informacijos apimtis (150 Seimo narių biogramų) neleido iškilti problemoms, susijusioms su papildomų darbo rankų pasitelkimu ir kitais organizaciniais rūpesčiais. Darbo tikslu buvo prototipinis, bandomasis, metodikai patikrinti skirtas DB modelis, kurio duomenys skirti vidiniam autorių kolektyvo naudojimui, ir šiuos uždavinius pavyko įgyvendinti užtektinai gerai. Kitame etape tikslinga pereiti prie pilnavertės, prieinamos tiek autorių kolektyvui, tiek knygos skaitytojui, DB sudarymo, o tai kartu gerokai pakelia reikalavimų kartelę.

Logiška, kad buvo žvalgomasi analogijų, užsienyje jau įgyvendintų panašių darbų, kuriais remiantis būtų galima sudaryti tinkamiausios struktūros projektą. Ironiška, bet artimiausią savo šaltinio pobūdžiu, klausimynu ir DB atlikimo technologija pavyzdį – Kelno universitete sudarytą duomenų bazių seriją apie Vokietijos XIX a. pabaigos – XX a. pradžios parlamentarus – pavyko rasti jau tik po to, kai buvo atlikti pagrindiniai mūsų DB kūrimo darbai. Vis dėlto tolesniam šios bei galimų kitų biografinio pobūdžio DB tobulinimui vokiečių tyrinėtojų įnašas gali suteikti reikšmingos naudos, todėl verta kiek išsamiau apžvelgti esminius šių elektroninių publikacijų bruožus.

Šiuo metu internete viena ar kita forma prieinamos 4 panašios struktūros šio mokslinio kolektyvo sudarytos duomenų bazės[4]:

- BIOSOP – Biographien sozialdemokratischer Parlamentarier in den deutschen Reichs - und Landtagen 1867–1933;

- BIOKAND – Sozialdemokratische Reichstagsabgeordnete und Reichstagskandidaten 1898–1918;
- BIOWEIL – Kollektive Biographie der Landtagsabgeordneten der Weimarer Republik 1918–1933;
- BIORAB – Biographien der Mitglieder deutscher Nationalparlamente (Teil III: 1919–1933).

Dalis jų sudarytos remiantis jau paskelbtu spausdintiniu leidiniu, pvz., BIOSOP – W. H. Schröder'o *Sozialdemokratische Parlamentarier in den deutschen Reichs - und Landtagen 1867–1933*, o BIOKAND – to paties autoriaus *Sozialdemokratische Reichstagsabgeordnete und Reichstagskandidaten 1898–1918*[5].

Skirtingai nuo monografijų, kuriose informacija pateikiama naratyvo forma, grupuojant pagal nagrinėjamą problematiką, elektroniniame variante žinios apie parlamentarus pateikiamos tematinėse lentelėse, skirstomose pagal asmens duomenis, išsilavinimą, partinę ir parlamentinę veiklą. Esant galimybei, pateikiama nuotrauka ar net garso įrašai. BIORAB ir BIOSOP DB pateikiami parlamentarų kalbų 1–2 min. trukmės garso įrašų fragmentai. Įrašo formatas (8 bitų 11 KHz mono PCM) atitinka to meto (1919–1933 m.) garso įrašymo technikos galimybes, be to, aukštesnių standartų panaudojimas negalėtų pastebimai pagerinti šaltinio perteikimo, tik padidintų failų apimtį ir apsunkintų DB prieinamumą internetu. Daugumoje šių DB, greta alfabetinio sąrašo, įdiegta paieška net pagal keletą kriterijų vienu metu.

Detaliau apžvelgus paminėtų DB struktūrą pažymėtina, kad sukaupti duomenys pristatomi gana lakoniškai. Žemiau pateikta BIOKAND DB struktūrinė analizė:

Asmens duomenų lentelė sudaryta iš tokių laukų: *vardas ir pavardė, lytis, gimimo data, gimimo vieta, mirties data, mirties vieta, konfesija, įgyta specialybė, tėvo specialybė*.

Vaizdiniai ir garsiniai dokumentai: *nuotrauka, kalbos fonograma* – vaizdo medžiagos pateikta mažai, parlamentarų bylos su nuotraukomis sudaro iki 10 proc. nuo bendros apimties, be to, yra 5 garso įrašai.

Šeimyninio statuso lentelė: *šeimyninė padėtis, santuokos metai, vaikų skaičius* (šis laukas dažniausiai lieka neužpildytas).

Išsilavinimo lentelė: *išsilavinimo eil. nr., mokymosi įstaigos tipas, mokymosi įstaigos vieta, įstojimo data, mokslų baigimo* arba *išstojimo data.*

Dalyvavimas Reichstage: *faktinė parlamentaro kadencijos pradžia, faktinė parlamentaro kadencijos pabaiga, rinkimų apygarda arba partinis sąrašas, sušaukto Reichstago eilės numeris, sušaukto Reichstago kadencijos pradžios ir pabaigos datos*.

Dalyvavimas Landtage: sudaryta Reichstago lentelės principu.

Likimas nacistinės diktatūros periodu: lentelė detaliau nestruktūrizuota, tiesiog stulpeliu surašomi esminiai biografiniai momentai – emigracija, grįžimas į Vokietiją ir kt.

Likimas po 1945 m.: *įvykio pradžios data, įvykio pabaigos data, vieta, įvykio pobūdis* (pvz., grįžimas į Vokietiją, darbas tam tikrose pareigose, sunkios ligos pradžia).

Gyvenimo aprašymas: biogramos tekstas; gali būti sutrumpintas, dažniausiai apie 500–1 500 spaudos ženklų.

Svarbu pažymėti, kad, greta jau turimų duomenų lentelių, prieinama nemaža dalis spausdintinių publikacijų tekstų, iš kurių galima susidaryti aiškesnį įspūdį apie nagrinėjamą objektą. Ypač aktualu, kad kartu su BIORAB pateikiamas didelės apimties straipsnis apie panaudotus šaltinius ir metodus[6]. Be to, BIOSOP suteikia galimybę atsisiųsti sukauptus duomenis gana paplitusiu SPSS formatu, taigi atveria galimybę kiekvienam besidominčiam asmeniui savarankiškai atlikti tokią statistinę analizę, kokia jam atrodytų tikslingiausia. Siekiama kaupti lankytojų atsiliepimus, susirašinėti su šia problema besidominčiais asmenimis. Visa tai sudaro efektyvią informacinę sistemą, tinkamą ne tik atsiskaityti už atliktą tyrimą, bet ir inicijuoti tolesnę mokslinę veiklą.

Tinkamiausio duomenų bazės tipo pasirinkimas

Labai dažnai istorikai turi reikalų su tekstinėmis DB. Tipiškiausias to pavyzdys – internetinės mokslinių publikacijų bazės. Tai ganėtinai paprastos struktūros ir lygiai taip pat ribotų galimybių sistemos, tačiau ir to pakanka joms keliamam uždaviniui – suteikti galimybę rasti reikalingas publikacijas pagal autorių, antraštę, populiarius raktažodžius ir pan. Tačiau kokios nors statistinės analizės atlikimas saugomų publikacijų viduje čia neprieinamas – ir per didelė apimtis, ir pernelyg skirtinga saugoma informacija. Pavyzdžiui, didžiulės apimties Internet Medieval Sourcebook arba Internet Modern History Sourcebook[7] yra begalė įvairiausių šaltinių tekstų. Šiose DB lengva susiorientuoti, kai teksto ieškoma pagal šalį, epochą ar problemą, tačiau nėra galimybės atlikti kompleksinės čia sukauptos informacijos analizės.

Sudaromos ir mažos apimties, specializuotos, konkrečiai tyrimų problematikai skirtos tekstinės DB. Vienas iš pavyzdžių – A. N. Zacharovo vadovaujami projektai „Боярские списки 1706–1710 годов“ ir „Повесточные сказки думных людей конца XVII – начала XVIII века“[8]. Juose publikuojami Rusijos valstybinio senųjų aktų archyvo dokumentų tekstai (РГАДА, ф. 210), atspindintys valdovo dvaro gyvenimą XVI–XVIII a.

pradžios laikotarpiu ir pateikiantys informaciją apie sostinės didikų biografinius faktus ir jų buitį. Akivaizdūs tokio tipo DB privalumai – originalaus šaltinio teksto pateikimas, atviras priėjimas internetu.

Akivaizdus ir trūkumas – silpna struktūrizacija. Šaltinio informacija pateikiama vientisu teksto masyvu, kurioje specifiniam tyrinėjimui reikalingos medžiagos atsirinkimas pareikalaus nemažų darbo laiko sąnaudų. Šį trūkumą lemia tikrai ne autorių požiūris į savo darbą, tačiau pats pasirinktas DB tipas. Anaiptol, sudarytojai atliko didelį ir labai naudingą darbą, sudarydami gausybę klasifikacinių lentelių, pavyzdžiui, abėcėlinį personalijų sąrašą, rangų ir pareigybių kategorijas, jų subkategorijas, santrumpų ir simbolių lenteles. Visa tai iš tikrųjų palengvina „navigaciją" šiame informacijos sraute, tačiau bent kiek sudėtingesnę statistinę analizę vis tiek teks daryti „rankomis".

Internete publikuota DB neabejotinai privalo laikytis HTML (hypertext) formato, tačiau tas pats formatas dažnai taikomas ir su internetu nesusietoms DB, ypač toms, kurios platinamos kompaktinėmis plokštelėmis. **Hipertekstinės DB** modelis irgi turi neabejotinų privalumų ir trūkumų. Skaitytojai jau gerai išmano šios sistemos galimybes, jiems nereikia ieškoti, ką ir kur spausti. Informacinė medžiaga paprastai pateikiama nedidelės apimties, greitai suvokiamų teksto segmentų forma, kur kiekvienas dažniau pasikartojantis informacijos vienetas – asmenybė, vietovė, įvykis – nuorodomis susietas su atskirame puslapyje pateikiamu aprašymu. Gerai sudarytoje hipertekstinėje DB skaitytojas turi galimybę vienu metu atsiversti didelį skaičių susijusių dokumentų ir juos lygiagrečiai analizuoti. Kiek supaprastintas, tačiau užtektinai efektyvus šio tipo DB modelis panaudotas jau minėtoje Kelno universiteto BIOWEIL duomenų bazėje, kurioje parlamentarai suskirstyti pagal partinius, teritorinius ir kitus sąrašus.

Pav. 1. Galimos hipertekstinės parlamentarų žodyno DB eskizas

Deja, šie reikšmingi privalumai turi savo kainą. Sistema, kuri labai draugiška skaitytojui, kur kas mažiau draugiška skaičiuotojui. Visus ryšius, faktų ir įvykių sąsajas reikia rasti sudarinėjant duomenų bazę. Tai, kas patenka į kompiuterio ekraną – iš tiesų galutinis produktas, taip, kaip ir popierinė knyga. Žinoma, lygiagrečiai galima suprogramuoti atskirą paieškos mechanizmą. Tai palengvintų turimos informacijos suradimą, tačiau papildymas naujai atrastąja vis tiek liktų ypač problematiškas.

Lietuvos parlamentarų žodynas reikalauja kitokio informacijos pateikimo ir valdymo, be to, svarbu išlaikyti darbo tęstinumo galimybę – taisyti pastebėtas klaidas, pildyti naujais įrašais. Didelis informacijos kiekis ir ganėtinai šabloniška jos forma (anketiniai duomenys) leidžia išnaudoti šiuo metu populiariausią DB tipą – **reliacinę DB**, tai yra tarpusavyje susietų lentelių visuma su aiškiai struktūrizuota ir kiek įmanoma standartizuota informacija. Būtent šis tipas dominuoja versle, materialinių resursų ir personalo valdymo srityje, nes greitai ir patogiai suranda, apibendrina ir pateikia kiekybinę informaciją. Skirtumai tarp verslo valdymo ir istorinio tyrimo užtektinai ryškūs, kad sukeltų gausybę problemų. Vis dėlto ši technologija yra pakankamai lanksti jas peržengti. Gana detalias metodines pastabas ir patarimus, kaip adaptuoti istorinius duomenis MS Access programai, pateikia K. G. Aliavdinas publikacijoje „База данных: „Трудовые конфликты на текстильных фабриках Центрально-Промышленного Района в 1895–1901 гг."[9]

Pav. 2. Bandomoji parlamentarų žodyno DB, sukurta reliaciniu pagrindu veikiančios Microsoft Access aplinkoje

Duomenų atranka ir grupavimas

Tai, ką žmogus suvokia kaip rišlų tekstą, kompiuterinei programai tėra beprasmių duomenų masyvas. Programa gali tai valdyti, bet negali interpretuoti. Didžiuliame tekste ji akimirksniu suras reikiamą segmentą, pavyzdžiui, „metus studijavo Varšuvoje, Moravskos slaptuose privačiuose moterų kursuose", tačiau tikrai negalės apsispręsti, kokiai išsilavinimo kategorijai tai priskirti. Tai problema, tačiau ji nedidelė. Didelė problema yra tai, kad to dažnai negali padaryti ir pats žmogus, be kito žmogaus ar ištiso kolektyvo paramos. Todėl į pirmą vietą iškyla ne techninė duomenų atranka (apie tai bus rašoma žemiau tekste), bet atrankos ir grupavimo kriterijai, pagal kuriuos turėtų būti atrenkama informacija.

Pirmiausia duomenų bazėje bus nemažai duomenų, kurie yra unikalūs ir nereikalauja priskyrimo jokioms grupėms. Vardas, pavardė, gimimo ar mirties datos yra vertingos būtent savo individualumu. Tačiau dažniau pasikartojanti informacija, pavyzdžiui, socialinė kilmė, išsilavinimas, profesija ir pan., savaime prašosi suskirstymo į kategorijas. Žinoma, originalaus biogramų teksto įrašai kur kas labiau atskleidžia laiko dvasią, pavyzdžiui: „baigė dviklasę, paskui mokėsi privačiai", „iki 15 metų amžiaus lankė Šmidto fabriko ir Garliavos pražios m-klą", „lankė rusų šventadieninius kursus suaugusiems", bet tokiu atveju jie nepasiduoda statistinei apskaitai. Todėl duomenų grupavimo, priskyrimo vienai ar kitai kategorijai fazė labai svarbus etapas sudarant bet kokią duomenų bazę. Kuo tiksliau duomenys bus suklasifikuoti, tuo į daugiau klausimų bus pajėgi atsakyti galutinė sistema.

Socialinė statistika, o tai, kam bus skirta projektuojamoji DB, iš esmės ir yra socialinės statistikos tyrimo objektas, apibrėžia tokius duomenų grupavimo principus: statistinių tyrimų kintamieji (duomenys) skirstomi į ***kokybinius*** ir ***kiekybinius***. Savo ruožtu, kokybiniai skirstomi į ***pavadinimų*** ir ***rangų*** skales, o kiekybiniai – į ***intervalų*** ir ***santykinę*** skales.

Struktūrinė schema, sudaryta remiantis V. Rudzkienės tipologija[10]

Kiekybinių duomenų biogramų tekste praktiškai nėra – nekalbama apie pajamų dydį, valdomos žemės apimtis ir pan. Planuojant biogramas, nebuvo keliamas uždavinys kiekybiškai apibendrinti seiminę veiklą, pateikiant tikslius pasisakymų, apeliacijų, lankytų posėdžių skaičius, be to, galimybė tai padaryti labai abejotina, net su pačių naujausių laikų duomenimis. Vieninteliai dažnai sutinkami kiekybiniai kintamieji – vaikų skaičius Seimo nario šeimoje ir kiek kartų jis buvo perrinktas į vėlesnius Seimus. Žinoma, pirmasis rodiklis atrodo labai išskirtinis bendroje leidinio koncepcijoje, tačiau gali praversti tiems, kurie domisi elito problematika. Technine prasme abu punktai gražiai dera su intervalų skale ir dėl mažo verčių diapazono nereikalauja jokio apibendrinimo, tarkime, užtenka nurodyti bendrą vaikų skaičių, be to, jį galima diferencijuoti pagal lytis.

Visi kiti duomenys – ***kokybiniai*** kintamieji. Vieni iš jų, pavyzdžiui, tautybė, konfesija, kilmė, priklauso pavadinimų, kitaip vadinamai nominaliajai skalei. Šiuo atveju esminis atrankos kriterijus glūdi pačiame duomenų segmento pavadinime – nereikalingi papildomi duomenys, kad atskirtume lenką nuo lietuvio, judėją nuo liuterono.

Žymiai sudėtingesnė padėtis su rangų skale. Išsilavinimo, profesinės veiklos, politinio aktyvumo Seime gradacija turi būti vykdoma pagal iš anksto nustatytus prioritetus, išrikiuojant juos tam tikra nuoseklia eile, kur paprastai pirmesnis įrašas pranašesnis už einantį po jo. Įgyvendinant šią procedūrą, tenka susidurti su nemažai ir objektyvių, ir subjektyvių sunkumų.

Ypač marga XX a. pradžios parlamentarų išsilavinimų ir profesinio užimtumo įvairovė netelpa į galiojančias klasifikacijas. Vienas ir tas pats teiginys „privatus išsilavinimas" talpina platų išsilavinimų spektrą: nuo paprasto raštingumo iki universitetui artimo lygmens. Neramiais valstybės atsikūrimo metais greitai keitėsi darbinė parlamentarų veikla ir ją taip pat sunku sutalpinti griežtuose DB rėmuose. Akivaizdu, kad šiam, ir, žinoma, ne tik šiam darbui reikalingas lietuviškas HISCO[11]. Tarptautinis istorikų projektas (veikia nuo 1968 m.), skirtas istorinei profesijų klasifikacijai, apima maždaug 300 metų chronologines ribas (1690–1970 m.). Juo siekiama aprašyti, susisteminti ir kodifikuoti darbinės veiklos sritis, daugiausia apimant Vakarų Europos šalis. Taigi panašios sistemos reikalingumas Lietuvos istorijos tyrimams abejonių nekelia.

Kaip minėta, Steigiamojo Seimo narių biogramų struktūrą nulėmė iš anksto apsibrėžtas klausimynas, iš pradžių gana trumpas, vėliau išsiplėtęs iki 20 pozicijų. DB buvo pildoma remiantis autorių jau pateiktu galutiniu biogramos tekstu, kuris, kaip ir bet koks rišlus tekstas, daug kur turėjo nukrypti nuo nustatyto formalaus šablono. Vieni iš tų nukrypimų buvo tikrai teigiami, pateikiantys daugiau medžiagos tyrimui, kitais atvejais teko susidurti su didelėmis informacijos spragomis, kurias kėlė tiek šaltinių trūkumas, tiek ir subjektyvi autorių pozicija neskirti dėmesio „nereikšmingiems" aspektams. Dėl to DB struktūra iš esmės seka parlamentaro biogramos struktūrą, tačiau, esant būtinybei arba atsiveriant naujoms galimybėms, interpretuoja ją palyginti laisvai.

Juk vienas ir tas pats pirminio teksto įrašas gali pasitarnauti sudarant daugumą duomenų bazės laukų, kurie atspindėtų skirtingus informacijos aspektus ir leistų pateikti skirtingus apibendrinimus. Pavyzdžiui, mokymosi įstaigos pavadinimas gali būti pagrindu tokioms skirtingoms kategorijoms kaip „mokymo įstaiga", „mokymo įstaigos profilis", „išsilavinimo lygis", „mokymosi vieta" ir pan.

Maža to, kur kas rimtesnis iššūkis laukia, kai toje pačioje duomenų bazėje teks derinti XX–XXI a. pradžios medžiagą su feodalizmo epochos duomenimis. Straipsnyje „Lietuvos Didžiosios kunigaikštystės parlamentarų (XV–XVIII a.) biografinis žodynas: problemos iškėlimas"[12] (autoriai Domininkas Burba, Robertas Jurgaitis, Deimantas Karvelis, Žydrūnas Mačiukas, Raimonda Ragauskienė, Aivas Ragauskas) vien šis periodas skirstomas į 5 iš esmės skirtingus etapus, kurie neabejotinai lems ir statistinių duomenų struktūrą.

Galima alternatyva yra dviejų skirtingų duomenų bazių – feodalizmo ir moderniųjų laikų – sudarymas. Viena vertus, lengva nuspėti, kad tai gali nemažai palengvinti tiek projektavimo, tiek techninio išpildymo darbą, tačiau pažeidžiamas leidinio vientisumo principas. Kita vertus, XVI ir XXI a. problematika paprastai yra tiek skirtinga, kad mažai tikėtina, jog konkrečiam tyrinėtojui rūpėtų visas prieinamų duomenų masyvas. Kol kas projektuojant ir bandant DB, laikomasi bendro, neperiodizuoto parlamentarų sąrašo koncepcijos, o galutinį sprendimą teks priimti, kai bus pradėti rašyti XVI–XVIII a. apimantys tomai.

Įrašo apie parlamentą struktūra

Bendras duomenų apie parlamentarą masyvas buvo sąlygiškai padalytas į tris stambius tematinius blokus: asmens duomenis, išsilavinimas ir profesinė veikla, visuomeninė politinė veikla. DB sudarymo teorija reikalauja kiek galima detalesnio informacijos išskirstymo atskiromis lentelėmis, nes tai palengvina pasikartojančių duomenų įvedimą ir kontrolę, turi įtakos didesniam sistemos našumui. Tačiau atsižvelgiant į konkrečią specifiką – mažą duomenų kiekį (tradicinės DB sąvokos prasme), nutarta neskubėti dėstyti duomenis į labai smulkias lenteles.

Pirmasis blokas – pagrindiniai, enciklopediniai personalijos duomenys. Atrodytų, visa tai galima perkelti į DB beveik mašinaliai, nesusiduriant su didesnėmis problemomis. Tai yra tiesa, tačiau aklas sekimas biogramos tekstu reikštų nepakankamą kompiuterio galimybių išnaudojimą.

1	Alternatyvios pavardės/vardo formos

	Lytis
2	Gimimo metai
3	Gimimo regionas
4	
5	
6	
7	Vaikų skaičius
8	Mirties metai
9	
10	Gimimo ir mirties/palaidojimo regiono sutapimas (Taip/Ne)

Kairiajame lentelės stulpelyje išdėstyti laukai, sudaryti remiantis pirminiu klausimynu, ir didesnio paaiškinimo jie nereikalauja. Dešinėje – papildomi laukai, vienus iš jų lėmė kompiuterinių technologijų ribotumas, kitus – priešingai, jų teikiamos papildomos galimybės. Gana žaismingas pavyzdys, kad informacijos galima išpešti net iš tokio, atrodytų, beviltiškai trivialaus dalyko kaip vardai, priklauso Ukrainos istorikams. Prozopografinėje DB „КанДеп" (Кандидат в депутаты) tarp įvairių statistinių pjūvių buvo atlikta ir semantinė vardų analizė. Ieškota ryšio tarp vardo prasmės ir polinkio į lyderiavimą. Vladimirai sudarė 12,84 proc., Nikolajai (*gr.* – tautų nugalėtojas) – 10,93 proc., Aleksandrai (*gr.* – žmonių gynėjas) – 8,47 proc., Viktorai – 7,65 proc. Iš viso tarp vyrų kandidatų į deputatus asmenys su „valdingais" vardais užėmė 39,98 proc., o tarp išrinktų deputatų tokie vardai sudarė beveik 50 proc.[13] Lietuviškame kontekste tokio tyrimo sėkmė abejotina, tačiau svarbus pats požiūrio netradiciškumas.

1b. Tarp Steigiamojo Seimo narių pasitaikė gana didelis skaičius asmenų, turinčių alternatyvias vardų ir pavardžių formas. Akivaizdu, kad šis reiškinys bus dar aktualesnis apdorojant ankstesnio periodo įrašus.

1c. Lyties parametras nėra aktualus spausdintiniam variantui, tačiau gali žymiai paspartinti kompiuterinę duomenų atranką.

2b, 8b. Ne apie visus parlamentarus pavyks surasti tikslias gimimo arba mirties datas. Kadangi MS Access tikslias datas (***date***, 1945/05/09) ir sveikus skaičius (***integer***, 1945) traktuoja visiškai skirtingai, būtinai rezervuoti atskirus laukus nepakankamai tikslioms datoms.

3b. Visos gimimo vietos aprašymas yra neabejotinai vertingas, tačiau nepasiduodantis statistiniam apibendrinimui. Papildomo lauko, skirto gimimo regionui, įvedimas gali sudaryti galimybes daryti statistines išvadas.

7b. Apdorotose biogramose beveik visada greta šeimyninės padėties nurodomi vaikų vardai, arba bent jų skaičius ir lytis. Iš to labai nedidelėmis pastangomis galima surinkti demografinę statistiką.

10b. Ryšys tarp gimimo ir mirties (palaidojimo) regionų gali suteikti papildomų duomenų apie aukštųjų visuomenės sluoksnių teritorinį mobilumą.

Antrasis blokas – išsilavinimas ir profesinė (darbinė) veikla. Jeigu pirmasis išlaikė savo pradinę struktūrą ir reikalavo tik papildymų, tai šiuo atveju padėtis labiau komplikuota. Teko atsisakyti pernelyg aptakių mašininiam apdorojimui biogramos laukų, juos suskaldant į didesnį ir detalesnį klausimyną. Be to, labai sunkūs laikai laukia šios lentelės, kai teks susidurti su feodalizmo epochos medžiaga.

11	Išsilavinimas iki išrenkant
	11c
	11d
12	Pagrindinė profesija
	12c
	12d

Būtent čia į pirmąjį planą iškyla bendro istorinio išsilavinimų ir profesijų registro poreikis. Vienas dalykas grupuoti nagrinėjamus asmenis remiantis tik subjektyvia DB pildančio darbuotojo nuojauta, ir kita – kolegialiai apsvarstytu ir patvirtintu rubrikatoriumi. Vis dėlto nemažai šios lentelės laukų taip ir liko aprašomojo pobūdžio, nes juose turėtų būti talpinamos kelios mokymosi įstaigos ar daugybė darbinės veiklos sričių. Vis dėlto tokie laukai kaip 11d ir 12b, jei pavyks parengti pakankamai tikslią klasifikaciją, gali suteikti įdomios statistikos.

Trečiasis blokas – visuomeninė ir politinė veikla iki ir po Seimo. Kaip ir ankstesnėje, daug aprašomojo pobūdžio laukų, kurie nepatogūs naudojant kiekybinius tyrimo metodus. Jeigu išrinkimo apygarda nereikalauja jokio papildomo išskaidymo, visus kitus laukus tikslinga suskaidyti į smulkesnes lenteles, kuriose būtų pavieniai ir, esant galimybei, gretimame lauke datomis papildyti visuomeninių, politinių organizacijų, Seimo pareigybių ir kiti įrašai.

13	Organizacija ar susivienijimas
	Organizacija ar susivienijimas...
14	
15	Pareigos Nr. 1
	Pareigos Nr. 2...
16	Pareigos Nr. 1
	Pareigos Nr. 2...
17	Bendras vėlesnių seimų skaičius
	Pilnas seimų sąrašas

Informacijos suvedimas į DB, duomenų tikslumo kontrolė

Sudarant Steigiamojo Seimo DB, pirmajame etape visa informacija (išskyrus nuotraukas) buvo atrenkama iš biogramų autorių pateikto teksto. Ateityje tikėtina ir siektina, kad dalis medžiagos papildys DB ir iš kitų šaltinių, tačiau dominuojančiu įvedimo būdu vis tiek išliks autorinio teksto analizė. Žinoma, būtų idealu, kad duomenis suvedinėtų pats autorius, tačiau, turint omenyje didelį autorių kolektyvą, ryškiai skirsis tiek jų kompiuterio naudojimo įgūdžiai, tiek techninė ir programinė įranga. Vadinasi, informacijos įvedimą teks pavesti tretiems asmenims, todėl dažnai teks susidurti su situacija, kai DB pildantis darbuotojas bus priverstas priiminėti sprendimus, kuriuos derėtų patikėti tik pačiam biogramos autoriui. Neretai aptakias socialinės kilmės, išsilavinimo, profesinio užimtumo ar kitas formuluotes teks sprausti į griežtus kompiuterinės programos rėmus. Neišvengiamai iškyla informacijos patikimumo ir vientisumo problema, jau neminint galimų klaidų dėl nuovargio ar išsiblaškymo.

Vienas iš tokiu atveju siūlomų kokybės garantijos būdų, ypač originalaus istorinio šaltinio perspausdinimo atveju, – kad tą pačią informaciją lygiagrečiai suvedinėtų du kompiuterio operatoriai, o gauti duomenys vėliau būtų tikrinami trečio asmens[14]. Žinoma, tai galėtų būti puikus vaistas tiek nuo spausdinimo klaidų, tiek nuo neobjektyvių sprendimų klasifikuojant duomenis. Tačiau kaip įveikti žymiai padidėsiančias personalo ir darbo laiko sąnaudas?

Kitas kontrolės būdas, bent iš dalies sugrąžinantis atsakomybę patiems autoriams, – popierinės anketos sudarymas, kuri, skirtingai nei laisvos struktūros klausimynas, reikalautų tikslių atsakymų į tikslias formuluotes – išsilavinimo cenzas, profesinės veiklos kategorija ir pan. Deja, ir šiuo atveju papildomų darbo laiko sąnaudų problema taip pat išlieka aktuali.

Didžiąją darbo laiko dalį turėtų sudaryti nuoseklus parlamentarų biogramų įvedimas, kai imamas vieną konkretų asmenį aprašantis tekstas ir nuo pradžios iki galo išskaidomas į smulkius informacijos vienetus. Tai įprasta, patogu ir tradiciška. Tačiau gali tekti susidurti su situacija, kai, tarkime, bus priimtas sprendimas papildyti DB kokia nors nauja tema. Pavyzdžiui, iškeltas klausimas, kaip tarybinė represinė sistema palietė Steigiamojo Seimo narių likimus. Vadinasi, DB pildantis darbuotojas, ieškodamas specifinės informacijos, turės skubiai peržiūrinėti didžiulės apimties tekstą. Kaip vienas iš darbą palengvinančių metodų galėtų būti tarp lingvistų bei istorikų populiarėjančios ir kol kas nemokamos ConcApp programos[15] panaudojimas, kuri padeda greitai surasti ir suvesti į vieną vietą rūpimą informaciją.

ConCapp programos langas, atliekant ištremtų Seimo narių paiešką

Išvados

Lietuvos parlamentarų žodyno duomenų bazės tikslas – ne dubliuoti arba pakeisti spausdintinį variantą, tačiau papildyti ir praplėsti tradicinio leidinio galimybes naudojantis naujomis technologijomis. Įgyvendintas, galutinis parlamentarų duomenų bazės modelis leistų:

1. turėti istorinės informacijos šaltinį, pritaikytą greitai kompiuterinei paieškai bei paieškos rezultatų spausdinimui;

2. sukauptus duomenis padaryti lengvai prieinamus mašininiam apdorojimui, kaip kad rūšiavimas, grupavimas, pritaikyti perkelti į skaičiuokles ir statistinės analizės programas tolesniam tyrimui;

3. užtikrinti tyrimo proceso tęstinumą – po spausdintinio leidinio pasirodymo atsiradusi papildoma informacija, klaidų pataisymai būtų registruojami numatomoje DB;

4. sukauptą metodinį ir technologinį patyrimą pritaikyti kituose Lietuvos istorijos duomenų kaupimo ir analizės projektuose;

5. parengta DB būtų mažas, bet konkretus indėlis laipsniškai kaupiant elektroninį Lietuvos istorinių duomenų banką – kad ir laisvos struktūros, tačiau šiuolaikinėmis informacijos technologijos pagrįstą istorinių žinių registrą, kuris sudarytų prielaidas palengvinti ir paspartinti visos tyrinėtojų bendruomenės darbą.

[1] Andriušaitienė, R., Denisovas, V. ir kt. *Duomenų bazės.* Vilnius, 2002, p. 7.

[2] Хэлворсон, М., Янг, М. *Эффективная работа с Microsoft Office97.* Спб, 1997, с. 792.

[3] Stancelis, V. Kompiuterinės duomenų bazės panaudojimas Lietuvos parlamentarizmo istorijos tyrimams. *Parlamentarizmo studijos*. Vilnius, 2004, t. 2, p. 174–189.

[4] *Sozialdemokratische Reichstagsabgeordnete und Reichstagskandidaten 1898–1918* – http://www.zhsf.uni-koeln.de/biokand/; *Biographien sozialdemokratischer Parlamentarier in den deutschen Reichs - und Landtagen 1867–1933* – http://biosop.zhsf.uni-koeln.de/; *Kollektive Biographie der Landtagsabgeordneten der Weimarer Republik 1918–1933* – http://hsr-trans.zhsf.uni-koeln.de/quantum/bioweil/index.html; *Biographien der Mitglieder deutscher Nationalparlamente (Teil III: 1919–1933)* – http://www.zhsf.uni-koeln.de/biorab.

[5] Schröder, Wilhelm Heinz. *Sozialdemokratische Parlamentarier in den deutschen Reichs - und Landtagen 1867–1933: biographien – chronik – wahldokumentation: ein Handbuch*. Düsseldorf, 1995, 1098 s.; Schröder, Wilhelm Heinz. *Sozialdemokratische Reichstagsabgeordnete und Reichstagskandidaten 1898–1918: biographisch–statistisches handbuch*. Düsseldorf, 1986, 355 s.

[6] *Quellen und Methoden* – http://www.zhsf.uni-koeln.de/biokand/texte/biokandquellen.html

[7] *Internet Medieval Sourcebook* – http://www.fordham.edu/halsall/sbook2.html; *Internet Modern History Sourcebook* – http://www.fordham.edu/halsall/mod/modsbook.html

[8] *Боярские списки 1706–1710 годов* – http://zaharov.csu.ru/bspisok.pl; *Повесточные сказки думных людей конца XVII - начала XVIII века* – http://zaharov.csu.ru/povestki.pl

[9] Алявдин, К.Г. *База данных „Трудовые конфликты на текстильных фабриках Центрально - Промышленного Района в 1895–1901 гг.*" – http://www.hist.msu.ru/Labs/Ecohist/DBASES/STRIKES/index.html

[10] Rudzkienė, V. *Socialinė statistika.* Vilnius, 2005, p. 30.

[11] *Historical international standard classification of occupations* – http://hisco.antenna.nl/

[12] *„Lietuvos Didžiosios kunigaikštystės parlamentarų (XV–XVIII a.) biografinis žodynas"*: *problemos iškėlimas* – http://www.parlamentostudijos.lt/Trecias/Istorija_Ragauskas.htm

[13] Белозерский, С.В., Пелех, И.Р., Святец, Ю.А. (Днепропетровск). Политический образ кандидата в депутаты Верховного Совета Украины от Днепропетровской области. Статистический анализ просопографической базы данных "КанДеп". *Информационный бюллетень Ассоциации "История и компьютер", № 23* – http://kleio.asu.ru/aik/bullet/23/49.html

[14] Cohen, D.J., Rosenzweig, R. *Digital history.* 2005 – www.chnm.gmu.edu\digitalhistory\digitizing\4.php.htm

[15] *ConcApp Corcordance and Word Profiler* – http://www.edict.com.hk/pub/concapp