



Kokybinių tyrimų duomenų archyvavimas ir sklaida: technologijos, galimybės ir realūs sprendimai

Daiva Vitkutė-Adžgauskienė
Andrius Utkā



UNIVERSITAS
VYTAUTI MAGNI
MCMXXII



MOKSLAS • EKONOMIKA • SANGLAUDA



EUROPOS SĄJUNGA
EUROPOS SOCIALINIS FONDAS

Kuriame Lietuvos ateitį



ŠVIETIMO IR MOKSLŲ MINISTERIJA



K A U N O
TECHNOLOGIJOS
UNIVERSITETAS

KOKYBINIŲ HSM TYRIMŲ DUOMENŲ ĮGIJIMO, AUTORINIŲ TEISIŲ UŽTIKRINIMO, ARCHYVAVIMO, DOKUMENTAVIMO IR SKLAIDOS SISTEMOS SUKŪRIMO GALIMYBIŲ STUDIJA

Kristina Juraitė, Jūratė Kavaliauskaitė, Rimvydas
Laužikas, Aušra Rimaitė, Irena Šutinienė, Andrius
Utkas, Daiva Vitkutė-Adžgauskienė

Įžvalgos



Kokybinių duomenų tipai pagal tyrimo metodą

Kokybinių duomenų archyvavimo sistema turi būti pritaikyta saugoti:

- Struktūruotų, pusiau struktūruotų ir nestruktūruotų interviu duomenis;
- Fokus grupių diskusijų užrašus;
- Stebėjimų protokolus (lauko užrašai) ir įrašus;
- Struktūruotus arba pusiau struktūruotus dienoraščius;
- Analizės dokumentus;
- Asmeninius dokumentus;
- Spaudos ištraukas;
- Fotografijas;
- Natūralios kalbos/pokalbių įrašus.



Kokybinių duomenų tipai pagal duomenų formatą

Kokybinių duomenų archyvavimo sistema turi būti pritaikyta saugoti įvairių formatų duomenis:

- Tekstinius duomenis;
- Statiško vaizdo duomenis;
- Vaizdo įrašų duomenis;
- Garso duomenis;
- Daugialypės terpės duomenis.



Įvairaus išsamumo žymėjimo (metaduomenų) lygiai

- Tyrimo (duomenų kolekcijos) aprašai. Jie apima bendrąjį duomenų kolekcijos aprašą, tyrimo metodologijos aprašą, su tyrimu susijusių duomenų sąrašą (katalogą).
- Duomenų failo aprašai.
- Duomenų failo turinio aprašai (pvz., interviu struktūra).
- Ryšių tarp failų aprašai.
- Kintamųjų lygio metaduomenys (susijusių kiekybinių duomenų aprašai).
- Duomenų valdymo taisyklių aprašai (objektai, įvykiai, teisės).



Metaduomenų standartai

Kokybinių tyrimų duomenų archyvai dažniausiai naudoja šiuos metaduomenų standartus:

- **Data Documentation Initiative (DDI)** – socialinių mokslų duomenims skirtas metaduomenų standartas, labiau orientuotas į kiekybinių duomenų aprašymą.
- **Dublin Core Metadata Initiative (DC)** – paprastos struktūros dokumento aprašo standartas, skirtas aprašyti bendrąsias skaitmeninio dokumento savybes – pavadinimą, autorių, temą, tipą, šaltinį, kalbą ir t.t.
- **Text Encoding Initiative (TEI)** – standartas, daugiausia orientuotas į tekstynų aprašų ir ontologijų kūrimą, skirtas dokumentų savybių ir jų turinio aprašymui.
- **Metadata Encoding and Transmission Standard (METS)** – skaitmeninės bibliotekos objektų aprašomųjų, administracinių ir struktūrinių metaduomenų standartas.
- **Preservation Metadata: Implementation Strategies (PREMIS)** – metaduomenų rinkinys, orientuotas į duomenų išsaugojimui skaitmeniniame archyve reikalingus aprašus, apimančius objektų, įvykių, agentų, intelektualios nuosavybės teisių apibūdinimą.



Kokybinių duomenų archyvavimo iniciatyvų pavyzdžiai

- ESDS Qualidata Archive, Jungtinė Karalystė (kartu su SQUAD ir QUADS projektais).
- Australian Qualitative Archive (AQuA, ADA Qualitative), Australija.
- Archive for Live Course Research (ALLF), Vokietija, Bremeno universitetas.
- Irish Qualitative Data Archive (IQDA), Airija.
- Finish Social Science Data Archive (FDS), Suomija.
- Swiss Centre of Expertise in the Social Sciences (FORS), Šveicarija.
- Prancūzijos socialinių mokslų kokybinių duomenų bankas (galimybių studija).
- National Archive of Criminal Justice Data (NACJD), JAV, Mičigano universitetas.



Svarbiausios problemos

- Duomenų įgijimo proceso organizavimas.
- Sistemoje naudojami duomenų tipai ir formatai.
- Archyvuose naudojami metaduomenų standartai.
- Duomenų kokybės užtikrinimas.
- Duomenų anonimizavimas (nuasmeninimas).
- Autorinių teisių apsauga.
- Sutartys su turinio teikėjais ir naudotojais.
- Išsamaus tyrimo aprašo parengimas.
- Projekto duomenų failų katalogas.
- paieškos sistemos funkcionalumas.
- Duomenų pateikimo vartotojui formatai.
- Prieiga, tiksliniai vartotojai, vartotojų teisės.



Kokybinių duomenų archyvo formavimo procesai (1)

Formuojant kokybinių HSM duomenų archyvą, vykdomi šie pagrindiniai su tyrimo duomenų įgijimu, apdorojimu ir sklaida susiję procesai:

- Tyrimų identifikavimas ir atranka;
- Duomenų įgijimas (skaitmeninimas; įkėlimas į sistemą; duomenų pilnumo patikrinimas, duomenų skaitmeninimas);
- Duomenų apdorojimas (duomenų failų konvertavimas; anonimizavimas; duomenų aprašymas metaduomenimis; kokybės patikrinimas);
- Archyvo įrašo suformavimas;
- Sklaida (duomenų paieškos ir naršymo, duomenų atsissiuntimo galimybių užtikrinimas).



Duomenų įkėlimas (1)

- Duomenų pateikimo į Archyvą procesas:
 - Užpildoma **Duomenų aprašo forma**, prieinama per portalo vartotojo sąsają.
 - Užpildoma **Duomenų pateikimo sutarties forma**, prieinama per portalo vartotojo sąsają.
 - Per portalo vartotojo sąsają įkeliami duomenų **failai ir jų aprašai (metaduomenys)**.
 - Per portalo vartotojo sąsają įkeliami kiti su tyrimu susiję dokumentai (**Tyrimo metodikos aprašas, Tyrimo dalyvių sutikimo dokumentai, Duomenų sąrašas**).
 - Įkelti duomenys perduodami į Sistemą tolimesniam sutvarkymui (apdorojimui).



Duomenų įkėlimas (2)

Pagalba depozitoriams (tyrėjams), mažinanti laiko sąnaudas, būtų:

- Duomenų įkėlimo į sistemą proceso aprašas.
- Metaduomenų formavimo metodika
- Duomenų aprašo formos pildymo taisyklės
- Kokybiškai veikiančios informacinės sistemos tezaurai ir klasifikatoriai,
- Rekomendacijos duomenų parengimui (pvz., optiniam atpažinimui, fotografavimui) naudojamos techninės ir programinės įrangos, būtinų bylų formatų ir pan.).

Įvedus duomenis į sistemą, atliekamas jų **pilnumo patikrinimas**:

- Patikrinama, ar užpildytos visos tyrimo dokumentavimui reikalingos formos;
- Patikrinama, ar formose užpildyti visi reikalingi laukai;
- Pagal duomenų sąrašą patikrinama, ar įkelti visi duomenų failai;
- Patikrinama, ar duomenų failams teisingai įvesti metaduomenys.



Autorių teisių apsauga

- Archyvas privalo užtikrinti autorių teisių apsaugą:
 - depozitorius turėtų patvirtinti (pvz. dialogo būdu), kad turi intelektinės nuosavybės teises į įvedamus duomenis ir kad dalijasi jomis su sistema ir jos vartotojais aiškiai apibrėžtu būdu.
 - **problema: seni, atrinkti duomenys.** Reikės papildomų teisininkų išaiškinimų.
- Formuojant archyvą, turi būti numatyta gauti tyrimo dalyvių (ar kitų susijusių asmenų) sutikimą
 - tyrimo dalyvių sutikimo formos,
 - tipinės sutartys tarp tyrimo dalyvių ir archyvo,
 - **problema** - šiandieninė praktika tokia, kad daugeliu atvejų kokybinių tyrimų metu **sutartys ar sutikimo formos su duomenų teikėjais nėra sudaromos**



Duomenų anonimizavimas

- Tikslinga įvesti asmeninių duomenų deidentifikacijos lygmenis:
 - visiška anonimizacija (visi asmens duomenys deidentifikuojami);
 - dalinė anonimizacija (pvz., deidentifikuojamas informatorius, o užrašytojo, transkripcijos autorių ar kt. asmenų duomenys paliekami);
 - deidentifikacija netaikoma (tais atvejais, jei informantas pageidauja, kad jo duomenys / autorystė būtų vieši prieinami / skelbiami ir pan.).
- Anonimizuojant žmonių vardai keičiami pseudonimais, dažniausiai jau transkribavimo metu.
- Taip pat anonimizavimas reikalauja pakeisti erdvinius žmonių minimus referentus - geografinių, ypač lokalių vietovių, organizacijų pavadinimus.
- Bet konkrečios erdvinės ir socialinės lokalizacijos praradimas neretai daro duomenis mažai informatyviais. Tokiais atvejais tikslinga vietovardžius palikti, bet labiau apriboti duomenų prienamumą ir platinimą.
- Anonimizavimo reikmėms į sistemos sudėtį turėtų būti įtraukiamas kompiuterinės lingvistikos priemonėmis parengtas įvardintų esybių (*named entity*) atpažinimo modulis lietuvių kalbai.



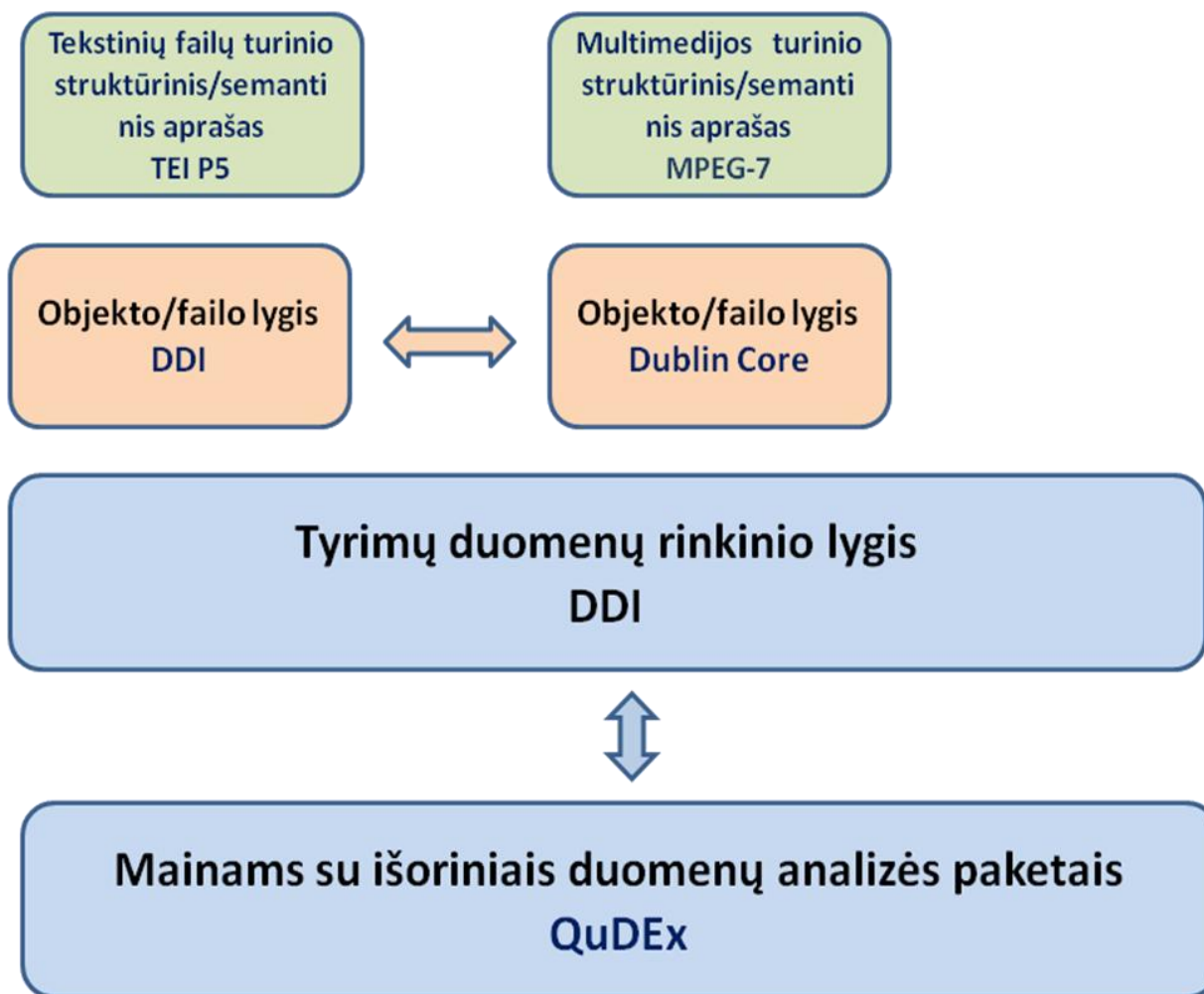
Kokybinių duomenų žymėjimo/anotavimo būdai

Metaduomenų standartai parenkami, sprendžiant šias problemas:

- Duomenų, atitinkančių įvairius tyrimų metodus, aprašymas;
- Skirtingų duomenų formatų aprašų formavimas (tekstiniai, statinio vaizdo, vaizdo įrašo, garso įrašo, daugialypės terpės duomenų);
- Skirtingų tyrimo duomenų lygių (duomenų rinkinio, duomenų failo bibliografijos, duomenų failo kokybinio turinio) aprašymas.
- Duomenų aprašymas mainų su kitomis sistemomis tikslais.



Rekomenduojama metaduomenų schema





Duomenų aprašo standartai (1)

- Atsižvelgiant į HSM kokybinių duomenų archyvų įvairiose šalyse analizę bei į jau egzistuojančią praktiką Lietuvoje (LiDA), kokybinių duomenų archyvui rekomenduojama naudoti DDI, kaip pagrindinį duomenų aprašo standartą.
- Juo būtų aprašoma šių lygių duomenys apie tyrimą:
 - Pats tyrimo aprašo metaduomenų dokumentas.
 - Tyrimo specifika (nurodoma, formuojant tyrimų aprašą).
 - Duomenų sąrašo failas, aprašantis visus tyrimą sudarančius failus.
 - Duomenų aprašas, naudojamas kintamų apibūdinimui mišraus kokybinio+kiekybinio tyrimo atvejui, taip pat duomenų sąrašo įrašų struktūros apibūdinimui.



Duomenų aprašo standartai (2)

- Šalia pagrindinio DDI standarto, papildomai rekomenduojama Sistemoje naudoti šiuos duomenų aprašo standartus:
 - Dublin Core (DC) – archyvo duomenų pateikimui į kitas išorines sistemas pagal OAI (*Open Archive Initiative*) apibrėžtą OAI-PMH automatinį duomenų surinkimo būdą. OAI-PMH palaikymas leidžia automatizuoti kreipinius į archyvą, apjungti skirtingų archyvų duomenis.
 - TEI P5 – kokybinio tyrimo tekstinių dokumentų turinio aprašymui (pvz., atskiriant skirtingų kalbėtojų informaciją interviu pokalbio transkripcijos įrašė).
 - MPEG-7 – aprašant kokybinį multimedijos failų turinį.
 - QuDex – duomenų importui/exportui iš/į CAQDAS programinių paketų (mišrių tyrimų atveju).



TEI P5 standarto panaudojimas

TEI schema kokybinių duomenų archyve gali būti naudojama tokių elementų aprašymui:

- dialogų (interviu) transkripcijos turinio,
- vardų ir vietų žymėjimui,
- teminių elementų (pvz., pagal tezaurą arba kitą klasifikatorių) žymėjimui,
- papildomų tyrėjų pastabų (interpretacijų) žymėjimui,
- anonimizuotų vietų žymėjimui,
- trūkstamų duomenų žymėjimui,
- ir t.t.



TEI P5 standarto panaudojimo pavyzdys interviu transkripcijos aprašymui

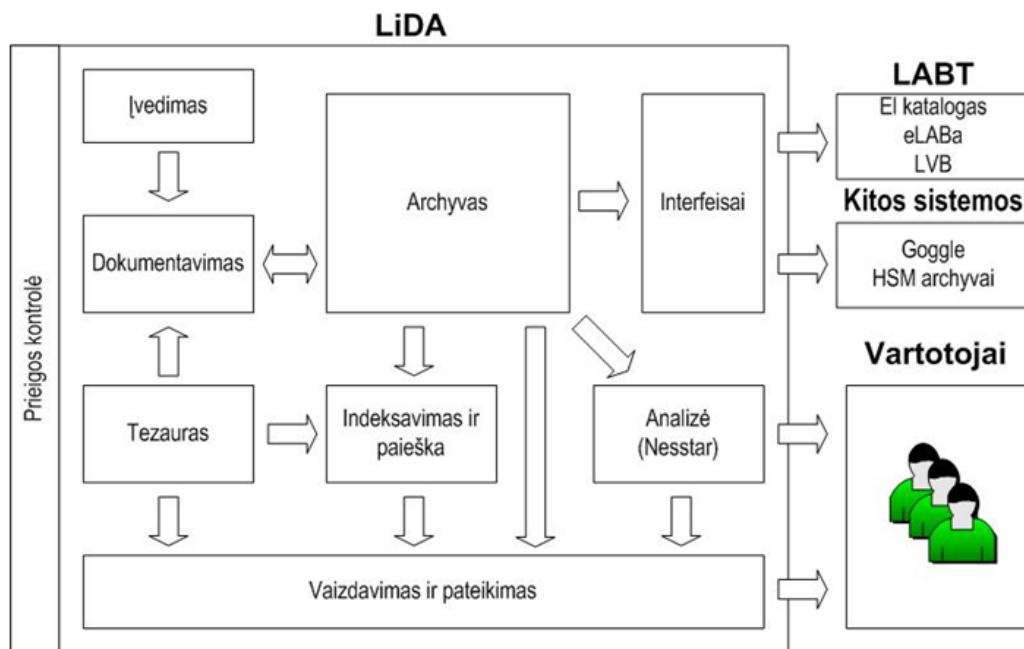
```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader> <!--TEI antraštės elementai --> </teiHeader>
  <text>
    <body>
      <div>
        <!--skyrelis visam interviu arba jo daliai-->
        <u who="#interviewer" xml:id="u1"><!--Interviuotojas klausia --></u>
        <u who="#subject" xml:id="u2"><!--Dalyvis atsako --></u>
        <u who="#interviewer" xml:id="u3"> <!--Interviuotojas klausia --></u>
        <u who="#subject" xml:id="u4"> <!--Dalyvis atsako --></u>
        .....
      </div>
    </body>
  </text>
</TEI>
```



Sklaida

- Svarbiausi sklaidos (duomenų pateikimo) būdai:
 - Tyrimo aprašo ir duomenų pavyzdžių peržiūra portale (neregistruotam vartotojui);
 - Išsamių tyrimo duomenų peržiūra portale (registruotam vartotojui);
 - Tyrimo duomenų atsisiuntimas iš portalo (*HTTP download*);
 - Tyrimo duomenų atsisiuntimas fiziniėje laikmenoje (CD-R, DVD-R).
- Reikalingą tyrimą vartotojas gali rasti šiais būdais:
 - Pasirinkdamas portale skelbiamų naujienų sąrašę;
 - Naršydamas portale pateikiamų temų sąrašę;
 - Atlikdamas paiešką pagal portale numatytą paieškos formą.
- Sistemoje racionalu numatyti kelis registruotų vartotojų lygius pagal informacijos prieinamumo lygį.
- Prieš gaudamas išsamius tyrimo duomenis, vartotojas turi sutikti su tyrimo duomenų peržiūros taisyklėmis.

Architektūra – LiDA pagrindu



- Kokybiniam duomenų archyvui reikėtų modifikuoti šiuos komponentus:
 - Įvedimo,
 - Dokumentavimo,
 - Indeksavimo ir paieškos,
 - Vaizdavimo ir pateikimo.



Prognozuojamos modifikacijos

- Įvedimo komponento korekcijos būtų susijusios su kokybiniam tyrimui būdingų duomenų failų ir jų meta-duomenų failų, bei papildomų tarnybinių dokumentų įkėlimu.
- Dokumentavimo komponentui reikėtų įtraukti papildomus modulius:
 - Anonimizavimo;
 - Įvardintų esybių atpažinimo ir žymėjimo tekste;
 - Žymėjimo formatų konvertavimo (DDI-DC, DC-DDI, DDI-TEI antraštė, DC-MPEG-7);
 - Automatinio atitikimo žymėjimo formatams įvertinimo.
- Indeksavimo ir paieškos komponento korekcijos:
 - papildomas indeksavimas pagal duomenų turinio žymėjimą (interviu struktūrą, įvardintas esybes ir t.t.),
 - konteksto nurodytam raktažodžiui (*KWIC*) paieškos įtraukimas.
- Papildomos vaizdavimo ir pateikimo komponentės funkcijos:
 - Tyrimo duomenų žemėlapiu pavaizdavimas (grafinis tyrimo katalogo vaizdas).
 - Atskirų tyrimo duomenų failų peržiūra (pažymint struktūrinius ir turinio metaduomenis, leidžiant atlikti pagal juos papildomą paiešką/navigaciją).
 - Žymėtų objektų (įvardintų esybių) pasirodymo dažnumų vaizdavimas žodžių debesiu, objektų ryšių vaizdavimas medžio struktūra.



Apibendrinimas

- Kuriant kokybinių duomenų archyvą, svarbiausios spręstinos problemos yra:
 - tinkamų metaduomenų standartų parinkimas
 - duomenų kokybės užtikrinimas
 - duomenų anonimizavimas ir autorinių teisių apsauga
 - funkcionali paieškos sistema
- Šias problemas įvairiais būdais sprendžia egzistuojantys kitų šalių kokybinių duomenų archyvai
 - Gerų, išbaigtų sprendimų nėra, daugeliu atveju eksperimentuojama
 - Galima formuluoti kai kurias rekomendacijas, kuria patirtimi būtų galima pasinaudoti
- Duomenų žymėjimui, įvertinus kokybinio archyvo poreikius bei kitų archyvų patirtį, siūloma naudoti DDI, TEI, MPEG-7, QuDEX metaduomenų standartų rinkinį (atskiri standartai atskiriems anotavimo lygiams)



Ačiū už dėmesį!

d.vitkute@if.vdu.lt

a.utka@hmf.vdu.lt



UNIVERSITAS
VYTAUTI MAGNI
MCMXXII

Back-up



Qualidata pavyzdys

- **Qualidata** yra JK socialinių mokslų kokybinių duomenų archyvo paslauga
- Integruoja egzistuojančius akademinius ir viešus repozitoriumus, kurie gali priimti kokybinius duomenis archyvavimui, pasiūlo paieškos ir naršymo paslaugas.
- Duomenys talpinimui parenkami pagal nustatytą tvarką (*Collections Development Policy*), kuri atsižvelgia į santykinę duomenų svarbą, formatą, medžiagos naudingumą ir kokybę, pernaudojamumo galimybes, intelektualinės nuosavybės teises ir konfidencialumą.
- Priima visų formatų duomenis – skaitmeniniu formatu, popieriniu formatu (spausdintinius ir rankraštinius), garso ir vaizdo įrašų medžiagą, fotografijas ir t.t.
- Talpinant duomenis, stengiamasi juos papildyti paaiškinančia informacija - talpinama papildoma informacija, dokumentacija, interviu su depozitoriais medžiaga.



Qualidata pavyzdys - metaduomenų standartai

- Naudoja DDI, kaip pagrindinį kokybinių tyrimų duomenų archyvo metaduomenų standartą.
 - Leidžia tinkamai aprašyti patį tyrimą ir su juo susijusius dokumentus (failus), tačiau nesudaro galimybių aprašyti dokumento turinio (pvz., interviu struktūros).
- Dokumentų turinio aprašymui taip pat naudoja ir TEI standartą.
- DDI ir TEI standartų integravimo darbai buvo atlikti ESRC SQUAD projekto rėmuose

Qualidata pavyzdys – anonimiškumo užtikrinimas

- Anonimizavimo lygis nustatomas individualiai kiekvienam duomenų rinkiniui ir priklauso nuo tyrimo pobūdžio.
- Reikalaujama, kad anonimizavimas būtų atliktas prieš pateikiant duomenis archyvui. Atskirais atvejais, duomenis anonimizuoja Qualidata darbuotojai.
- Panaikinamos ir pakeičiamos pseudonimais svarbiausios identifikacinės detalės (realūs vardai, vietų, kompanijų vardų, gatvių pavadinimai ir t.t.).
- Naudojamos automatinės paieškos ir keitimo technologijos, po kurių seka rankinė peržiūra (*proofreading*).
- Tarnybinėje informacijoje išsaugomi ryšiai tarp pseudonimų ir realių vardų.
- Atskirais atvejais, gali būti gautas respondento leidimas panaudoti duomenis be anonimiškumo užtikrinimo.
- Tais atvejais, kai neįmanoma užtikrinti anonimiškumo, leidžiama ribota prieiga, paprastai reikalaujant atskiro depozitoriaus leidimo.



Qualidata pavyzdys – sklaida

- Vartotojai naudojami internetine paieškos ir naršymo sistema, leidžiančia taip pat ir atsisiųsti duomenis.
- Prieš įsigydamas duomenis, vartotojas turi galimybę naršymo būdu peržiūrėti pavyzdžius (*web based samplers*), taip pat perskaityti susijusią metodologinę informaciją.
- Sistemoje yra daug pavyzdžių apie tai kaip naudoti sukauptus duomenis – teminiai web puslapiai, DUK informacija, straipsniai, panaudojimo atvejų analizė ir t.t.



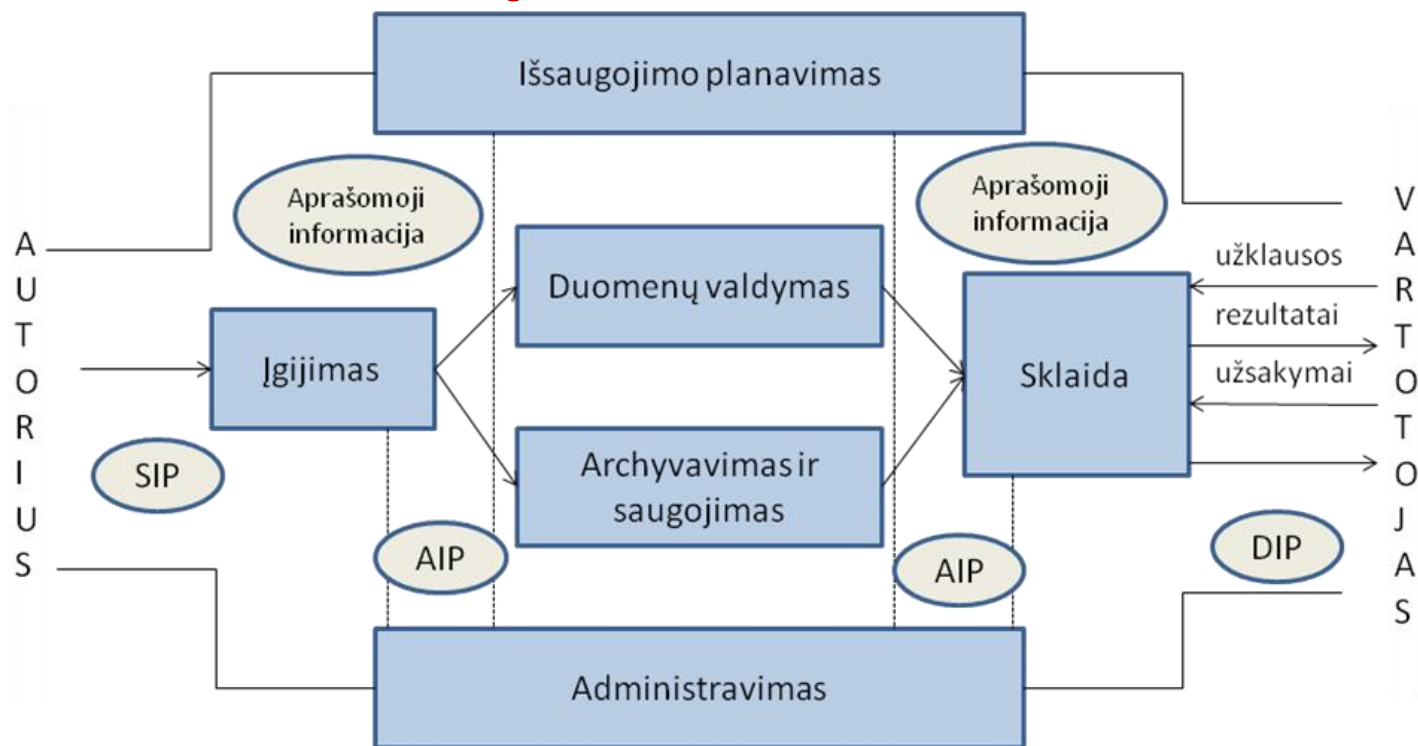
AQuA (ADA Qualitative) pavyzdys (1)

- ADA (Australian Data Archive) Qualitative, anksčiau vadintas AQuA (Australian Qualitative Archive) – tai Australijos kokybinių duomenų archyvas
- Naudoja DDI 2 metaduomenų standartą, bandydama derinti su QuDEX schema duomenų mainams su CAQDAS paketais
- Naudojamos kompiuterinės lingvistikos technologijos paieškos galimybių išplėtimui - sąvokų gavyba (*concept mining*), naudojant specialų Leximancer įrankį:
 - klasifikavimo metaduomenys yra automatiškai priskiriami, analizuojant tekstą ir naudojant Leximancer tezaurą
 - leidžia atsiriboti nuo subjektyvumo problemų rankiniu būdu klasifikuojant medžiagą

AQuA (ADA Qualitative) pavyzdys (2)

- Aiškiai aprašyta prieigos teisių, intelektualios nuosavybės teisių ir duomenų apsaugos politika
- Depozitorius gali pasirinkti teises sau ir kitiems:
 - nėra ribojimų duomenų naudojimui, depozitorius nepageidauja gauti informacijos apie duomenų vartojimą,
 - depozitorius pageidauja informacijos apie jo duomenų naudojimą, kad galėtų bendrauti su kitais panašių interesų vartotojais,
 - vartotojas turi gauti raštišką depozitoriaus leidimą prieš skelbdamas su duomenimis susijusias interpretacijas,
 - depozitorius nori gauti informaciją apie kiekvieną prašymą pasinaudoti duomenimis, kad galėtų duoti leidimą,
 - duomenims gali būti nustatytas tam tikras embargo periodas, t.y. iki depozitoriaus nurodytos datos neleidžiama jokia prieiga prie duomenų
- Tyrėjai gali leisti kitiems tyrėjams naudoti medžiagą su tam tikromis sąlygomis, pvz. apribojant prieigą tam tikroms institucijoms, arba reikalaujant papildomo patvirtinimo, nurodžius tyrimų paskirtį
- Depozitorius turi įrodyti, kad turi intelektualios nuosavybės teises ir leidžia archyvui naudoti duomenis (sutartis).

Architektūra – pagal OAIS funkcinių modelių



3 informacinių paketų tipai, atitinkantys skirtingus informacijos gyvavimo sistemoje etapus :

- įgijimo informacinis paketas SIP (*Supply Information Package*),
- archyvo informacinis paketas AIP (*Archive Information Package*),
- sklaidos informacinis paketas DIP (*Distribution Information Package*).