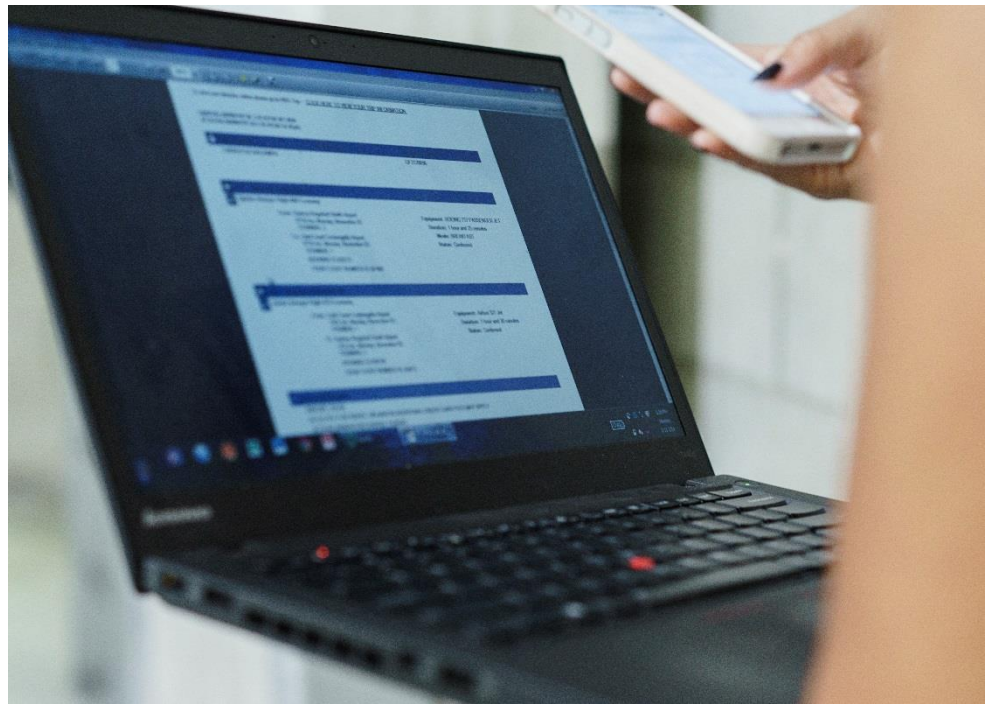


Informacinės visuomenės plėtros komitetas

Lietuvos atvirų duomenų formavimo metodologinių ir teisinio reglamentavimo priemonių įgyvendinimui bei tam reikalingų valstybės institucijų darbuotojų kompetencijų ugdymo paslaugos

Duomenų teikimo pakartotiniam naudojimui taikomos atvirųjų elektroninių formatų bei standartų rekomendacijos (R8)

2018 m. rugpjūčio 31 d.
Vilnius



Duomenų teikimo pakartotiniam naudojimui taikomos atvirųjų elektroninių formatų bei standartų rekomendacijos (R8)

Dokumento pakeitimai

Lentelė Nr. 1. Dokumento pakeitimų valdymas

Data	Aprašymas	Redaguotos dalys
2018-08-31	Pateiktas dokumento turinys derinimui	-
2018-09-10	Tikslinimas pagal gautas pastabas	

Duomenų teikimo pakartotiniam naudojimui taikomos atvirųjų elektroninių formatų bei standartų rekomendacijos (R8)

Informacinės visuomenės plėtros komitetui
Gedimino pr. 7,
LT-01103 Vilnius

2018 m. rugpjūčio 31 d.

Projekto „Lietuvos atvirų duomenų formavimo metodologinių ir teisinio reglamentavimo priemonių įgyvendinimui bei tam reikalingų valstybės institucijų darbuotojų kompetencijų ugdymo paslaugos“ Duomenų teikimo pakartotiniam naudojimui taikomos atvirųjų elektroninių formatų bei standartų rekomendacijos (R8)

Gerbiamieji,

Šią Ataskaitą parengė UAB „PricewaterhouseCoopers“, (toliau – Paslaugų teikėjas arba PwC), remiantis 2017 m. gruodžio 11 d. sudaryta pirkimo sutartimi Nr. 6F-52 (toliau – Sutartis), sudarytą su Informacinės visuomenės plėtros komitetu prie Susisiekimo ministerijos (toliau – Perkančioji organizacija, IVPK).

Darbo apimtis

Duomenų teikimo pakartotiniam naudojimui taikomos atvirųjų elektroninių formatų bei standartų rekomendacijos (R8) (toliau – Ataskaita) apimtis buvo apibrėžta pirkimo „Lietuvos atvirų duomenų formavimo metodologinių ir teisinio reglamentavimo priemonių įgyvendinimui bei tam reikalingų valstybės institucijų darbuotojų kompetencijų ugdymo paslaugos“ techninėje specifikacijoje (toliau – Techninė specifikacija).

Darbo metodika

Ataskaita parengta remiantis Techninėje specifikacijoje nurodytais dokumentais bei reikalavimais, PwC pateiktu techniniu pasiūlymu ir susitikimų metu gauta medžiaga. PwC nesiėmė priemonių šaltinių patikimumui nustatyti ir papildomai netikrino pateiktos informacijos. Todėl PwC nesuteikia tiesioginių ar numanomų garantijų kitiems asmenims (išskyrus IVPK pagal Sutartį) dėl Ataskaitos tikslumo ar išsamumo.

Darbus pagal Sutartį atlikome iki 2018 m. rugpjūčio 31 d. Ataskaitoje neaprašomi vėlesnių įvykių ar aplinkybių padariniai, taip pat vėliau paaiškėjusi informacija. Negalime nustatyti, kokią įtaką šie darbai arba tyrimai būtų turėję Ataskaitos parengimui.

Atkreipiame Jūsų dėmesį į Ataskaitoje išdėstytas svarbias mūsų pateikiamas prielaidas. Ataskaitos teikėjas neatsako kitiems asmenims (išskyrus IVPK pagal Sutartį) už Ataskaitos rengimą. Be to, Ataskaitos teikėjas neprisiima sutartinės, deliktinės ir kitokios atsakomybės (neatsižvelgiant į ieškinio formą) taikytinos teisės leistina apimtimi ir neprisiima atsakomybės už kitų asmenų patirtus padarinius (išskyrus IVPK Sutarties pagrindu) ar kitus pagal šią Programą priimtus sprendimus ar atsisakymą juos priimti.

Dokumento platinimas ir atsakomybė

Asmuo, susipažinęs ir perskaitęs šią Ataskaitą, sutinka ir įsipareigoja laikytis šių sąlygų:

- Ataskaitos skaitytojas supranta, kad Ataskaita pagrįsta ekspertiniu vertinimu;
- Ataskaitos skaitytojas supranta, kad Ataskaitos teikėjo darbai yra atlikti pagal mūsų kliento – IVPK – nurodymus ir skirti tik mūsų klientui;
- Ataskaitos skaitytojas pripažįsta, kad Ataskaita parengta mūsų kliento – IVPK – nurodymu, todėl į ją gali būti įtraukta ne viskas, kas svarbu skaitytojui;
- Ataskaitos skaitytojas sutinka, kad PwC, jos partneriai, įgaliotojai, darbuotojai ir atstovai neprisiima sutartinės ar deliktinės atsakomybės (įskaitant aplaidumą ir teisės aktuose nustatytą įsipareigojimų

Duomenų teikimo pakartotiniam naudojimui taikomos atvirųjų elektroninių formatų bei standartų rekomendacijos (R8)

pažeidimą ir kt.) ir neatsako už žalą, išlaidas ar nuostolius, patirtus skaitytojui vadovaujantis šiuo dokumentu ar su juo susipažinus.

Bendroji informacija

Jeigu kiltų klausimų dėl šios Ataskaitos turinio, prašome kreiptis į Justą Urboną telefonu +37065511056 arba el. paštu justas.urbonas@pwc.com.

Pagarbiai

Audrius Leipus
UAB „PricewaterhouseCoopers“
Konsultacijų skyriaus vyr. projektų vadovas

Turinys

1	Rekomenduojamų atvirų duomenų metaduomenų publikavimo formatai	6
2	Rekomenduojami atvirų duomenų formatai.....	6
3	Pagrindiniai atvirų duomenų formatai	7
3.1	Geografinių duomenų formatai	10
3.2	Multimedijos duomenų formatai.....	11

1 Rekomenduojamų atvirų duomenų metaduomenų publikavimo formatai

Atvirų duomenų rinkinių metaduomenys į atvirų duomenų portalą turi būti perduodami taikant DCAT standartą ir XML, JSON arba kitus su Atvirų duomenų portale palaikomus formatus.

Rekomenduojama naudoti šiuos atvirų duomenų rinkinių metaduomenų publikavimo formatus.

Formatas	Aprašymas
RDF/ XML	RDF/ XML yra pagrindinis susietųjų duomenų formatas, naudojamas pateikti duomenis tinkle. Šis formatas skirtas išreikšti RDF grafo duomenis XML tipo dokumentu/failu.
RDFa	RDFa yra HTML5 plėtinys, kuris gali būti naudojamas į HTML panašaus tipo dokumentuose, pvz., XML, SVG, XHTML5 ir kituose. Šis formatas publikuoti struktūrizuotus duomenis portaluose ir juos efektyviau nuskaityti paieškos varikliams.
JSON¹	JSON yra duomenų keitimosi formatas, kurį lengva skaityti, analizuoti ir sugeneruoti. JSON yra teksto formatas, paremtas JavaScript programavimo kalba, kuris yra optimizuotas duomenų mainams. JSON yra kuriamas naudojant struktūras: 1. pavadinimų/reikšmių porų ir 2. reikšmių sąrašo.
XML	XML yra plačiai naudojamas duomenų mainų formatas. XML yra tekstinių duomenų formatas, kuris palaiko Unicode koduotę. XML formatas turi apibrėžtą struktūrą ir yra orientuotas į dokumentus, tačiau šis formatas dažnai naudojamas ir žiniatinklio paslaugose.
HDF5²	HDF5 yra duomenų modelis, biblioteka ir failų formatas, skirtas saugoti ir valdyti duomenims. Jis palaiko neribotą skaičių duomenų tipų ir yra sukurtas lanksčiai ir efektyviai įvesčiai/išvesčiai ir dideliems duomenų kiekiams.

2 Rekomenduojami atvirų duomenų formatai

Atvirus duomenis rekomenduojama skelbti atvirais formatais. Institucijos turėtų orientuotis į tuos formatus, kurie leistų pasiekti kuo aukštesnį atvirų duomenų brandos lygį. Institucijos turėtų atverti duomenis ne žemesniu nei 3 brandos lygį atitinkančiais formatais.

Pagrindiniai duomenų atvėrimui naudojami formatai pateikiami Aprašo 1 priede, tačiau pažymėtina, kad jie neapibrėžia visos formatų aibės. Todėl duomenis atveriančios organizacijos turėtų įvertinti naudotojų poreikį ir duomenis atverti tokiu formatu, kuris yra paklausiausias tarp atvirų duomenų naudotojų.

¹ Daugiau apie JSON formato: <https://www.w3.org/TR/json-ld/>

² Šaltinis: <https://support.hdfgroup.org/HDF5/>

Atvirų duomenų rinkinys gali būti publikuojamas keliais formatais iš karto.

Rekomenduojama atvirų duomenų rinkinio formatą parinkti taip, kad jo parengimo kaštai būtų proporcingi naudotojų poreikiui, o pats formatas atviras ir tinkamas naudotojų pakartotiniam panaudojimui, t. y. jeigu atvirų duomenų naudotojų poreikis specifiniam duomenų rinkiniui yra didelis, organizacijos turėtų skirti papildomų resursų rinkinio atvėrimui keliais formatais (tokių, kuriuos naudoja atvirų duomenų naudotojai). Atviri formatai leidžia programinės įrangos kūrėjams sukurti įvairius įrankius ir pakartotinai panaudoti, modifikuoti ar keisti atvertų duomenų rinkinių duomenimis.

Atnaujinant ar papildant duomenų rinkinius, rekomenduojama išlaikyti pasirinktą (-us) duomenų rinkinio formatą (-us).

Institucijos turėtų atverti duomenis ir kitais formatais, kurie geriausiai atspindi atveriamų duomenų turinį. Todėl išskiriamos dvi papildomos formatų grupės - geografinių duomenų formatai ir multimedijos. Papildomos formatų grupės yra aprašomos Aprašo 2 priede ir 3 priede.

3 Pagrindiniai atvirų duomenų formatai

Formatas	Atvirumo įvertinimas	Formato ribojimai	Failo panaudojimas ³
DOC ar DOCX	1 brandos lygis	Ribotas automatinis apdorojimas ir naudojimas kitose nei Microsoft sistemose.	Klasikinių dokumentų tokiais formatais, kaip DOCX, ODF, OOXML arba PDF gali pakakti kai kurioms duomenų rūšims atverti. Jais skelbti duomenis gali būti pigu, kadangi dažnai tai yra pradinis duomenų formatas. Atveriant duomenis rekomenduojama nesinaudoti teksto saugojimo formatais, jei galima naudoti kitus formatus. Šie formatai nepalaiko vienodos duomenų struktūros, o tai reiškia, kad gali būti sudėtinga automatizuoti duomenų apdorojimą. Jeigu atviri duomenų rinkiniai yra atveriami teksto saugojimo formatais, tuomet turėtų būti naudojami dokumentų šablonai, leidžiantys suprasti, kokia yra duomenų struktūra, kad vėliau būtų įmanoma struktūrizuoti dokumentuose išsaugotus duomenis.
PDF	1 brandos lygis	Naudinga, kai reikia išsaugoti formatavimą keičiantis dokumentais, bet ribotas automatinis apdorojimas, ženklų atpažinimas ir naudojimas kituose projektuose.	
TXT	1 brandos lygis	Paprastas tekstinis formatas, perskaitomas daugumoje operacinių sistemų. Negalimas formatavimas.	

³ Šaltinis: <http://opendatahandbook.org/guide/lt/appendices/file-formats/>

Duomenų teikimo pakartotiniam naudojimui taikomos atvirųjų elektroninių formatų bei standartų rekomendacijos (R8)

Formatas	Atvirumo įvertinimas	Formato ribojimai	Failo panaudojimas ³
RTF	1 brandos lygis	Paprastas tekstinis formatas, perskaitomas daugumoje operacinių sistemų, išlaikomas tam tikras formatavimas.	<p>Tekstiniai failai (.txt) yra lengvai kompiuterių perskaitomi. Šių formatų failai neturi metaduomenų apie dokumentą, dėl šios priežasties failo kūrėjai turėtų failą parengti taip, kad nuskaičius informaciją būtų galima interpretuoti ją tokią, kokia ji yra, be papildomo žinojimo apie dokumento struktūrą (metaduomenis).</p> <p>Gali būti problemų, naudojant grynojo teksto failus skirtingose operacinėse sistemose. MS Windows, Mac OS X ir Unix skirtingai interpretuoja failo pabaigą.</p> <p>Atveriant duomenis rekomenduojama nesinaudoti grynojo teksto formatais, jei galima naudoti kitus formatus.</p>
HTML	1 brandos lygis	Dažniausiai naudojamas duomenų internetiniuose puslapiuose publikavimo formatas. Naudojant šį formatą, duomenys neturi privalomos struktūros.	<p>Didelė dalis duomenų įvairiose svetainėse skelbiami HTML formatu. Šio formato duomenų gali pakakti, jei duomenys stabilūs ir nedidelės apimties. Šiuo formatu saugomus duomenis sudėtingiau atsisiųsti ir su jais dirbti, tačiau pigu ir paprasta nukreipti. Šis formatas gali būti vienu pradinių duomenų atvėrimo būdų.</p> <p>Atveriant duomenis rekomenduojama suformuoti ir papildomą failą duomenų atvėrimui.</p>
XLS ar XLSX	2 brandos lygis	Ribotas automatinis apdorojimas ir naudojimas kitose ne Microsoft sistemose.	<p>Daugelis valstybės institucijų saugo informaciją XLS arba XLSX formatų failuose. Tokie duomenys dažnai gali būti iš karto panaudojami, jeigu teisingai aprašytos stulpelių reikšmės.</p>
ODT	2 brandos lygis	Ribotas automatinis apdorojimas	<p>Kai kuriais atvejais skaičiuoklių dokumentuose gali atsirasti papildomos informacijos arba formulių, kurias sudėtinga interpretuoti. Todėl rekomenduojama skaičiavimus dokumentuoti papildomai.</p> <p>Atveriant duomenis, rekomenduojama atlikti papildomą duomenų konvertavimą ir failus pateikti mašininiam skaitymui patogesniu CSV formatu.</p>
CSV	3 brandos lygis	Universalus formatas, skirtas struktūrizuotiems duomenims (pvz., skaičiuoklės).	<p>CSV yra kompaktiškas formatas, tinkamas dideliems duomenų rinkiniams, turintiems vienodą struktūrą. Šio formato duomenis dažnai sudėtinga interpretuoti, todėl labai svarbu, kad kablelių atskirto turinio failuose būtų tiksliai dokumentuotos laukų reikšmės. Būtina griežtai laikytis failo struktūros, nes praleista viena reikšmė gali neleisti nuskaityti viso likusio failo informacijos.</p>

Duomenų teikimo pakartotiniam naudojimui taikomos atvirųjų elektroninių formatų bei standartų rekomendacijos (R8)

Formatas	Atvirumo įvertinimas	Formato ribojimai	Failo panaudojimas ³
XML	3 brandos lygis	<i>Extensible Markup Language</i> (XML) yra aprašinėjimo kalba, kuri apibrėžia duomenų kodavimo taisyklių rinkinį, naudojant žymes, atributus ir turinį. Šio formato failai yra sudėtingi ir turi apibrėžtą struktūrą.	XML yra plačiai naudojamas formatas dalijantis duomenimis, nes jis suteikia galimybę išlaikyti duomenų struktūrą ir įrašyti dokumentaciją kartu su duomenimis neatliekant pačių duomenų nuskaitymo.
XHTML	3 brandos lygis	XHTML (<i>eXtensible HyperText Markup Language</i> – „išplečiama hiperteksto žymėjimo kalba“) yra žymėjimo kalba, kuri turi panašias į HTML žymėjimo taisykles, tik jos sintaksė yra griežtesnė. XHTML remiasi XML, todėl XHTML dokumentas turi būti struktūriškai tvarkingas (angl. <i>well-formed</i>).	XHTML naudojamas duomenų publikavimui tiesiogiai portaluose papildomai pateikiant ir duomenų struktūrą. XHTML dokumentas, atitinkantis kurią nors iš oficialių XHTML specifikacijų, vadinamas korektišku dokumentu (angl. <i>valid document</i>). Nors XHTML turėtų būti vienodai atvaizduojamas visose naršyklėse, taip nėra. Vis dėlto, kiekvieną XHTML dokumentą rekomenduojama patikrinti tam skirtu tikrinimo įrankiu (angl. <i>validator</i>). Dauguma XHTML failų yra suderinami su HTML, t. y. juos gali apdoroti naršyklė, skirta HTML dokumentams. Tokie dokumentai turi būti perduoti su MIME tipu <code>text/html</code> . Naršyklė tokį dokumentą apdoros kaip „žymų sriubą“ (angl. <i>tag soup</i>). XHTML dokumentą perduodant su MIME tipu daugumai senesnių naršyklių iškils problemų su XHTML dokumento apdorojimu.
JSON	3-5 brandos lygis	JavaScript Object Notation, yra atviras standartinis formatas, kuris naudoja žmonėms suprantamą tekstą reikšminių verčių duomenų porų perdavimui.	JSON yra paprastas failų formatas, kurį labai lengva skaityti bet kuriai programavimo kalbai. Jo paprastumas reiškia, kad kompiuteriams jį apdoroti lengviau nei kitus formatus, pavyzdžiui, XML. JSON struktūra yra apibrėžta, tačiau ji neturi tokių griežtų struktūros reikalavimų kaip XML.
RDF XML, RDF TURTLE	3-5 brandos lygis	RDF failo formatas yra skirtas aprašyti duomenų šaltinius ir nuorodas į kitus šaltinius struktūrizuotais formatais.	W3C rekomenduoja RDF formatą naudoti duomenų saugojimui taip, kad būtų galima apjungti duomenis gautus iš skirtingų šaltinių. RDF duomenys gali būti išsaugoti XML arba JSON formatu, bei kitais formatais. RDF skatina internetinių nuorodų naudojimą kaip unikalų raktą, kuris įgalina įvairių atvirų duomenų iniciatyvų sujungimą internete. RDF vis dar nėra labai stipriai išpopuliarėjęs, tačiau tai dažniausiai naudojamas formatas tarp Atviros Vyriausybės iniciatyvų. Atveriant duomenis rekomenduojama siekti RDF formato panaudojimo.

3.1 Geografinių duomenų formatai

Failo tipas	Aprašymas
Taškinės grafikos formatai	
GeoTIFF	Viešai prieinamas metaduomenų standartas, kuris leidžia pridėti georeferencinius duomenis TIFF faile.
GML	Formatas, skirtas atvaizduoti geografinius duomenis ir paremtas XML sintakse.
JPEG2000	Atviro kodo taškinės grafikos formatas. JPEG 2000 privalumas yra lankstumas. Kodavimo srautas, gautas suglaudinus vaizdą su JPEG 2000, gali būti keičiamas dekoduojant, pavyzdžiui, galima gauti atvaizdą mažesne skiriamąja geba. Nustatydamos kodavimo srautą sistemos gali reikšmingai padidinti našumą.
netCDF-CF	Dvejetainis atviro kodo duomenų saugojimo formatas, kuris leidžia tiesiogiai prisijungti prie žemėlapių per OPeNDAP protokolą. Šis formatas dažniausiai naudojamas klimatologijos, meteorologijos ir okeanografijos GIS sistemose.
Vektorinės grafikos formatai	
Shapefile	Dažnai naudojamas geoerdvinių vektorinių duomenų formatas, skirtas geografinėms informacinėms sistemoms (GIS).
Geography Markup Language (GML)	XML pagrindu veikiantis atvirasis standartas (OpenGIS) GIS duomenų mainams.
GeoJSON	Atviras standartas, skirtas atvaizduoti paprastus geografinius duomenis ir jų erdvinius parametrus.
Keyhole Markup Language (KML)	XML pagrindu veikiantis atvirasis standartas (OpenGIS) GIS duomenų mainams.
MapInfo format	TAB MapInfo vektorinių duomenų formatas naudojantis TAB, DAT, ID ir MAP failus
Spatialite	SQLite erdvinis plėtinys, suteikiantis vektorinės geoduomenų bazės funkcionalumą. Jis panašus į „PostGIS“, „Oracle Spatial“ ir „SQL Server“ su erdviniais plėtiniais.
Duomenų apsikeitimo formatai	
Geographic Data Files (GDF)	Geografinių duomenų apsikeitimo failo formatas.
GeoPackage (GPKG)	SQLite pagrindu sukurtas atvirasis formatas skirtas vektoriniams ir taškiniams duomenims.

3.2 Multimedijos duomenų formatai

Failo tipas	Aprašymas
Video formatai	
Gif	Formatas, palaikantis paprastą animaciją be garso.
MP4	Dažnai naudojamas skaitmeninis multimedijos formatas, skirtas video ir audio turiniui, bet taip pat gali būti naudojamas kitiems duomenims kaupti, tokiems kaip subtitrai ar vaizdai.
Audio formatai	
MP3	Dažniausiai naudojamas skaitmeninio audio turinio kodavimo formatas.
AAC	Audio turinio kodavimo standartas, skirtas failo suspaudimui su praradimais.
WMA	Audio turinio kodavimo standartas, skirtas failo suspaudimui su praradimais, sukurtas Microsoft.
Vaizdo formatai	
JPG	Fotografinių vaizdų išsaugojimo formatas (ir jų suspaudimo algoritmas).
TIF	Formatas, naudojamas aukštos kokybės rastriniam tipui saugoti grafika.
PNG	Bitų masyvo formatas, kuris gali būti suglaudintas „be nuostolių“.