UDK 004.656PREMIS

Skaitmeninių išteklių išsaugojimo metaduomenų standartai: atlikti darbai ir siekiai

Sally H. McCALLUM
Kongreso biblioteka, Vašingtonas, JAV, el. p. smcc@loc.gov

Svarbiausias sėkmingo skaitmeninių išteklių išsaugojimo komponentas yra metaduomenys, nes jie leidžia automatiškai išsaugoti tokius išteklius. Žmogiškieji ištekliai tampa nebereikalingi tvarkant daugelį skaitmeninių dokumentų, nes protingiau elektroninių išteklių išsaugojimo darbus atlikti kompiuteriu. Praėjusį dešimtmetį daryta nemažai bandymų sukurti skaitmeninę saugyklą. Jų metu buvo taikomi skirtingi metodai, sukurti ir vartoti skirtingi duomenų modeliai, pastūmėję pažangos link. Šiame straipsnyje akcentuojama naujausia iniciatyva, vadinama PREMIS, kurios pagrindą sudaro šiuolaikinės sąvokos ir naujausia patirtis. Ji vertinga tuo, kad atlieka nuodugnų tyrimą, atskleidžiantį, ar identifikuoti metaduomenys gali būti vartojami bendrame kontekste, ir sudaryti pagrindą detalesniems metaduomenims, padeda išsiaiškinti, kiek išteklių dar prireiks išsaugojimo darbams atlikti. Straipsnyje taip pat aptariama iniciatyva sudaryti papildomų techninių metaduomenų ir dokumentų formatų registrus.

Reikšminiai žodžiai: metaduomenų standartai; PREMIS.

Pagrindiniai metaduomenys, vartojami išsaugojimo metu: PREMIS

Ištakos

Projektas „Išsaugojimo metaduomenys: taikymo strategijos“ (*Preservation Metadata: Implementation Strategies – PREMIS*) buvo parengtas, remiantis praėjusį dešimtmetį įgyta patirtimi[1]. Bibliotekų bendruomenė, o ypač ICABS priklausančios institucijos ir jų partneriai, vykdo reikšmingą veiklą, kurdami informacijos išsaugojimo sistemas. Buvo kuriami formalių ir neformalių duomenų modeliai, identifikuojami duomenų elementai, skirti išsaugojimo funkcijai atlikti, nors dažnai buvo siekiama daugiau nei išsaugojimo, akcentuojant prieigos ir platinimo klausimus. Tarp tokių projektų verta paminėti NEDLIB (*Networked European Deposit Library*) projektą, kuriam vadovavo Olandijos ir Prancūzijos nacionalinės bibliotekos, CEDARS (*CURL Exemplars in Digital Libraries*) projektą, kurį parengė specialistai iš Jungtinės Karalystės, Australijos nacionalinės bibliotekos vykdytą PANDORA projektą bei dar daug įvairių organizacijų, pavyzdžiui, OCLC pasiūlytas iniciatyvas, Kongreso bibliotekos vykdytą Nacionalinės skaitmeninės bibliotekos projektą ir kt.

Įdomu, kad visuose projektuose tam tikrais etapais buvo taikomas standartinis OAIS modelis[2], kuris pirmiausia buvo išbandytas kosmoso duomenų sistemose, o vėliau paskelbtas ISO standartu (ISO 14721). OAIS modelis darė vienijantį poveikį tyrimams, atliktiems praėjusį dešimtmetį, jau vien tuo, kad pasiūlė bendrą kalbą, vartojamą diskusijose. Paprastai omenyje turimi informacijos archyvavimo, pateikimo ir platinimo paketai (atitinkamai AIP, SIP ir DIP), kaip svarbiausi abstraktieji skaitmeninių saugyklų kūrimo komponentai. Šiuos informacijos paketus sudaro keturios dalys, susijusios su apdorojamu informacijos objektu: turinio informacija, pakavimo informacija, aprašo informacija ir išsaugojimo informacija. 2002 m. OCLC ir RLG (Mokslinių bibliotekų grupė) remto projekto metu modeliai ir metaduomenys, apibrėžti anksčiau minėtų projektų metu, buvo puikiai sujungti į vieną struktūrą ir pritaikyti platesniame standartinio OAIS modelio kontekste[3]. Todėl svarbiausia PREMIS darbo grupės užduotis buvo naudojantis duomenų žodynu išskirti šias gijas ir paversti jas pritaikomų duomenų elementų rinkiniu.

Tikslai

PREMIS projektas – tai daugiametės darbo grupės pastangos, įgyvendinamos kartu su svarbius projektus vykdančiomis institucijomis iš viso pasaulio. Atstovai iš Australijos, Naujosios Zelandijos, JAV, Didžiosios Britanijos, Olandijos ir Vokietijos prisidėjo įvairiais būdais. Veikla, suplanuota vieneriems metams, buvo įvykdyta per dvejus, rezultatas – gautas elementų rinkinys gali būti panaudotas kaip projektų įgyvendinamo pagrindas.

Siekta kelėto tarpusavyje susijusių tikslų, kurių visi yra praktiniai ir skirti suteikti pagrindą idėjų įgyvendinimui. Pradiniai tikslai buvo tokie: nustatyti pagrindinį metaduomenų rinkinį ir sudaryti metaduomenų žodyną, ir šiuo metu abu tikslai sėkmingai pasiekti. Bandymai naudotis metaduomenų žodynu sudarys puikiausią pagrindą trečiajam tikslui pasiekti, t. y. kurti alternatyvias įgyvendinimo strategijas. Šiuo metu vykdoma veikla padės pasiekti galutinius tikslus, išmėginti duomenų žodyną ir sukurti pagrindiniais elementais grindžiamas bendradarbiavimo programas.

Tyrimas

Pirmiausia buvo atliktas skaitmeninių saugyklų projektų tyrimas, siekiant nustatyti šiuo metu taikomas praktikas ir skaitmeninių projektų kryptis. Tyrimo metu buvo gauti besiformuojančiai sričiai neblogi rezultatai – 48 atsakymai iš 13 šalių. Toliau pateikiamos pagrindinės išvados[4], kurios suteikė daugiau duomenų tuo pačiu metu ir vėliau vykdytiems duomenų žodyno kūrimo darbams:

– saugyklos projektavimui ir pradiniams darbams dažniausiai taikomas OAIS standartinis modelis;

– paprastai saugojimo sistemose kaupiama per didelis kiekis metaduomenų; metaduomenys XML arba sąryšinėse duomenų bazėse saugomi tam, kad būtų greitai randami ir lanksčiai pateikiami kartu su turinio objektu apibūdinimo ir išsaugojimo tikslams;

– METS (*Metadata Encoding and Transmission Standard*) plačiai taikomas koduoti skaitmeniniams objektams reikalingiems metaduomenims, tarp jų ir išsaugojimo metaduomenims; METS sistemoje techniniams atvaizdų metaduomenims koduoti taikomas MIX (*Metadata for Images in XML*);

– šiuo metu siekiama išsaugoti originalą ir keletą sunormintų ir (arba) perkeltų turinio objekto versijų su susijusiais metaduomenimis;

– įvairiarūšių strategijų taikymas net ir institucijos viduje neretas šioje eksperimentinėje ir besiformuojančioje srityje.

Be to, tyrimas parodė, kad su skirtingų rūšių objektais (bitų srautais, failais, loginiais objektais ir t. t.) susiję metaduomenys buvo atskiriami, o informacija, nurodanti santykius tarp objektų, dažnai užrašoma. Kadangi tiriamoji priemonė šioje naujoje srityje nėra galutinė, rezultatai buvo ir įdomūs, ir naudingi duomenų žodyno kūrimui.

Duomenų žodynas

Remdamasi ankstesniu pamatiniu projektu (ir netiesiogiai keliais didžiausiais paskutinio dešimtmečio projektais) bei informacija, gauta skaitmeninių saugyklų tyrimo metu, PREMIS darbo grupė sudarė pagrindinių duomenų elementų žodyną[5]. Pradiniais projekto etapais priimta keletas reikšmingų sprendimų, svarbių praktiškai įgyvendinant projektą.

Darbo grupė pagrindinius duomenų elementus apibūdina kaip „aspektus, kuriuos dauguma veikiančių saugyklų turi žinoti, norėdamos tęsti skaitmeninį išsaugojimą“[6]. Grupė specialiai nenagrinėjo gerai žinomų išsaugojimo aspektų, pvz., detaliųjų techninių metaduomenų, skirtų skirtingoms medijoms. PREMIS darbo grupė toliau gvildeno tik tuos techninius metaduomenis, kurie yra dažniausiai taikomi failų formatams.

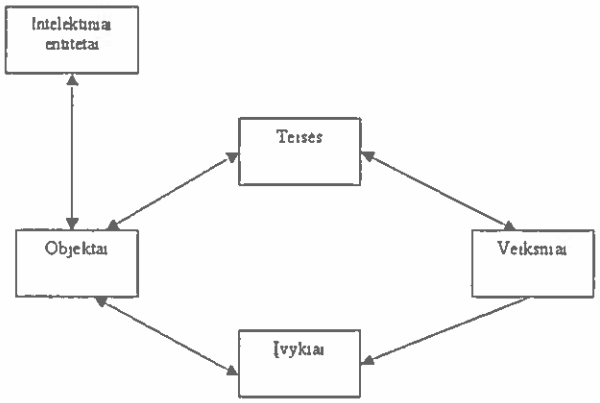
Darbo grupė taip pat nusprendė, kad nurodytieji metaduomenys turi būti automatiškai pateikiami ir vartojami, kiek tai įmanoma. Todėl pirmenybė buvo teikiama autoritetinių sąrašų reikšmėms, o ne tekstiniams aprašams. Pasirinkimas taip pat siejosi su darbo grupės siekiu savarankiškai įdiegti duomenų žodyną. Tyrimas parodė, kad saugyklos jau šiuo metu yra kuriamos, o planuojamų sukurti saugyklų sistemų terpės gali pasižymėti specifiniais bruožais. PREMIS pagrindiniai elementai, kurie bus prieinami saugyklai, nebūtinai turi būti joje saugomi. Elementus būtų galima laikyti pagalbinėse sistemose, jie galėtų atsispindėti saugyklos taisyklėse arba saugomi vietinėje duomenų bazėje ar vidiniu formatu. Svarbu, kad pagrindiniai duomenys būtų pasiekiami, kai juos reikės konvertuoti į atitinkamus standartus, jei jais reikės pasikeisti. Duomenys taip pat turi būti prieinami bet kokia programine įranga, kurią naudoti gali pasirinkti saugykla, tikėdamasi prieiti prie PREMIS pagrindinių duomenų. Sistemų nereikia dar kartą įdiegti arba specialiai pertvarkyti tam, kad būtų išsaugoti PREMIS pagrindiniai duomenys viename iš standartinių formatų. Todėl darbo grupė „metaduomenų elementus“ duomenų žodyne pakeitė „semantiniais vienetais“.

Duomenų modelis

Nesiekiant smulkiai apibūdinti visą duomenų modelį, svarbu išskirti kelis jo bruožus. (Modelis yra smulkiai aprašytas PREMIS ataskaitoje – žr. 5 nuorodą.)

Modelis yra *nesudėtingas*. Jame pateikiamos tik penkios entitetų rūšys: *objektai, įvykiai, veiksniai, teisės* ir pats *intelektinis entitetas*. Į duomenų žodyną įtrauktos informacijos esmingumas buvo kruopščiai ištirtas. Todėl, pavyzdžiui, aprašomieji metaduomenys, apibūdinantys in-

telektinį entitetą (knygą, žemėlapį, tinklalapį ir t. t.) sudaromi pagal vieną iš jau veikiančių standartų, pvz., MARC, MODS (*Metadata Object Description Standard*) ir DC (*Dublin Core*). Detalieji duomenys apie veiksnius taip pat sudaromi pagal MARC, MADS (*Metadata Authority Description Standard*), *vCard* ir kitus standartus. Teisių duomenis sudaro duomenys, susiję su leidimais vykdyti išsaugojimo veiklą, nes teisės, siejamos su prieiga prie *objekto* ir jo platinimu, nėra esminės išsaugojimui. Detalūs techniniai metaduomenys, medijos ir techninės įrangos dokumentacija nėra įtraukti, tačiau formato ekspertai turi juos apibrėžti.



Pagrindinis PREMIS duomenų modelis

Pagrindinė modelio koncepcija – semantiniai vienetai objektams apibūdinti, kurie gali būti apibrėžti trimis lygmenimis, taip suteikiant galimybę laisvai pasirinkti, kokią informaciją pagal medžiagą atitinkantį lygmenį įtraukti, bei lanksčiai naudotis saugykla. *Bitų srautas* yra pirmasis lygmuo, įeinantis į kitą, *failo* (arba *failų srauto*) lygmenį. Aukščiausiąjį, pateikimo lygmenį sudaro failų rinkinys, būtinas galutiniam intelektinio objekto *pateikimui*.

Įvykio entitetas, aprašantis su objektu susijusius veiksmus, yra svarbi modelio dalis. Didelė veiksmų įvairovė daro įtaką skaitmeninės medžiagos išsaugojimo procesui, taip pat objekto pakeitimui, tinkamumo ir integralumo patikrinimui, užklausų platinimui ir ataskaitoms. Dažnai įvykiai taip pat siejasi tarpusavio santykiais, nes derivacinio įvykio metu sudaromas naujas objektas, o objektų tarpusavio santykis paprastai svarbus įrašyti išsaugojimo tikslams. Duomenų žodyne pateikiami keletą santykio rūšių apibūdinantys semantiniai vienetai, susiję su įrašo derivatu ir struktūrine santykio informacija, priklausomybės ir kitais santykiais.

Reikšmingą duomenų modelio aspektą darbo grupė pavadino 1:1 principu. Iš esamų objektų sukurti nauji objektai (kopijos, versijos, transformacijos ir t. t.) traktuojami kaip nauji objektai ir siejami su „senuoju" objektu įvykio ir santykio informacija. Vienas iš tyrimo rezultatų parodė, kad saugyklose dažnai saugoma daug objekto kopijų, todėl išsaugojimo tikslams svarbu, kad duomenys apie tokį objektą būtų. Todėl santykio informacija sudaro sąsają ir nekliudo užrašyti visą išsaugojimo informaciją apie derivaciją. Nors saugykloje gali susidaryti duomenų medžiai, taip išvengiant informacijos pertekliaus, keisdamasi duomenimis saugykla turi turėti galimybes perleisti atskirą objektą su visais išsaugojimo metaduomenimis.

Kitas etapas – išbandymas

PREMIS yra kruopščiai suplanuoto tarptautinio bendradarbiavimo padarinys. Jo metu buvo parengtas metaduomenų žodynas, galintis sudaryti sąlygas standartiniams išsaugojimo informacijos mainams su skaitmenine medžiaga iš elektroninių archyvų. Dėl to saugyklose nebūtina diegti specialios infrastruktūros, tačiau pateikiamos gairės, nusakančios, kaip sudaryti pagrindinius išsaugojimo metaduomenis. PREMIS projekte dalyvavo specialistai iš viso pasaulio, o jį rėmė OCLC ir RLG, tuo tarpu Kongreso biblioteka buvo atsakinga už kito etapo oficialios svetainės administravimą[7]. Visus projekto dokumentus ir naujienas galima rasti šioje svetainėje.

Dabar galima planuoti, kaip pasiekti galutinius projekto tikslus, išbandyti duomenų žodyną ir pradėti bendradarbiauti, koncentruojantis ties metaduomenimis. Neseniai buvo sukurta XML schema duomenų žodyne nurodytiems semantiniams vienetams identifikuoti[8]. Žodyną būtina išbandyti ir pritaikyti naujiems projektams bei duomenų mainams. Tačiau tikimasi, kad jau įkurtos saugyklos arba planuojami specifinių struktūrų projektai taip pat bus įtraukti į bandymus, analizuojant jų numanomus bei tikslius metaduomenis ir lyginant juos su semantiniais duomenų žodyno vienetais. Tuo tarpu duomenų žodyno ir XML schemos forma paliekama tokia, kokia yra, tačiau bet kada gali būti peržiūrėta, pritaikant bandymų metu įgytą patirtį.

Kiti tikslai

Kaip jau minėta, egzistuoja ir kiti išsaugojimo metaduomenų skaitmeninėse medijose aspektai, kurių neaptarė PREMIS darbo grupė, pavyzdžiui, išplėstiniai teisių metaduomenys ir detalieji techniniai metaduomenys, taip pat skaitmeninio formato informacija.

Teisių metaduomenys

PREMIS apytiksliai apibūdino teisių metaduomenis, todėl būtų galima ginčytis, ar dalis prieigos ir platinimo informacijos yra svarbi išsaugojimo tikslams. Tačiau dauguma iniciatyvų atkreipia dėmesį į teisių išraiškos kalbą ir idėjines standartų problemas, susijusias su prieiga ir

platinimu. Tarp kelių didžiausių tiriamųjų darbų galima paminėti Europos Sąjungos vykdomą „Indecs“ projektą, leidėjų grupių veiklą ONIX projekte ir Skaitmeninių bibliotekų federacijos (*Digital Library Federation – DLF*) Elektroninių teisių valdymo iniciatyvą (*Electronic Rights Management Initiative – ERMI*).

Techniniai metaduomenys

PREMIS tyrimo metu nustatyta, kad daugelis saugyklų taiko METS sistemą jose saugomų skaitmeninių objektų metaduomenims susieti, be to, jose saugomi įvairių rūšių techniniai metaduomenys įvairiais kiekiais, priklausomai nuo to, kiek jų saugykla gali automatiškai surinkti. Didelė pažanga buvo padaryta atvaizdų išteklių metaduomenų srityje, kurioje taikomi standartai. NISO (*National Information Standards Organisation*) parengė standartinį duomenų žodyną ir leido jį išbandyti 2002 metais[9]. Tačiau šiuo metu jau dažnai taikoma išplėstinė METS schema MIX, pagrįsta NISO duomenų žodynu[10]. Tai, kad ši schema buvo taip greitai pradėta naudoti, rodo, kad saugyklos labai domisi standartais ir detalios techninės informacijos gairėmis. Norėdama turėti detalius techninius metaduomenis, bibliotekų bendruomenė turi bendradarbiauti arba bent jau įdėmiai sekti naujai atsirandančius pramoninius standartus, nes šio lygmens metaduomenis būtina išvesti iš objektų net labiau nei PREMIS lygmens informaciją. METS svetainėje pateikiama keletas vietinių techninių metaduomenų schemų, skirtų įvairių rūšių medžiagai. Jas galima panaudoti kuriant atvaizdų duomenims tinkamus standartus[11].

Formatų registrai

Antrasis pakankamai vertingas išsaugojimo metaduomenų aspektas yra lengva priciga prie elektroninių duomenų formatų. Tokią informaciją dažnai galima rasti už įvairius duomenų formatus atsakingų įmonių svetainėse, jei tokios egzistuoja, tačiau šis informacijos gavimo būdas nėra efektyvus. Žvelgiant iš išsaugojimo perspektyvos, žinios apie duomenų formatus praverčia patvirtinant skaitmeninius objektus arba tikrinant jų integralumą, jos padeda įvertinti riziką, susijusią su įvairiais skaitmeniniais formatais, bei nurodo tinkamas skaitmeninių objektų perkėlimo trajektorijas. Failų formatų suvokimas taip pat gali padėti nustatyti metaduomenis, kuriuos būtų galima išgauti iš skaitmeninio objekto, ir užpildyti PREMIS bei detaliųjų techninių metaduomenų duomenų bazes.

Įgyvendinami du gerai žinomi projektai, kurių metu buvo kolektyviai kuriami nuolat atnaujinami katalogai, tačiau nėra visiškai aišku, ar galima šiuos projektus paremti. Vienas iš projektų yra Jungtinės Karalystės nacionalinių archyvų kuriamas PRONOM[12]. Šis registras pradžioje kurtas kaip lokali priemonė, kurios Nacionaliniams archyvams reikėjo kovojant su programinės įrangos senėjimu dokumentų perdavimo procesui reguliuoti. 2004 m. registras buvo pateiktas internete, o 2005 m. buvo sukurta patobulinta nauja jo versija. Kadangi daug dėmesio skiriama viešai prieinamiems įrašams, registras ypač tiko į tekstą orientuotiems programinės įrangos formatams.

Kitas projektas, pasiekęs įgyvendinamumo etapą, yra Pasaulinis skaitmeninių formatų registras (*Global Digital Format Registry – GDFR*), kurio idėja kilo DLF remto susitikimo metu 2003 metais[13]. Kai tik Harvardo universiteto darbuotojai išplatino registro modelį, Pensilvanijos universitetas sukūrė prototipinę formato paslaugą ir pavadino ją Formatų registro demonstracine versija (*Format Registry Demonstration – FRED*)[14]. Naudodamiesi šia versija saugyklų kūrėjai gali eksperimentuoti ir aiškintis, kiek naudinga gali būti ši paslauga, kokias paslaugas reikėtų pasiūlyti, kaip ją reikėtų prižiūrėti ir t. t.

Ši sritis nėra patraukli, tačiau paaiškėjo, kad yra svarbi duomenų išsaugojimui visose medijose, o bendras registras būtų labai naudingas visai bendruomenei.

Išvados

Palaipsniui, remiantis ankstesniais konceptualiais modeliais ir patirtimi, gairės ir standartai metaduomenims padeda išsaugoti saugyklas. Saugyklų kūrėjams jau nebereikia pradėti darbą nuo nulio. Dabartinė darbotvarkė, siekiant raidos ateityje, yra PREMIS pagrindinių elementų išbandymas, detaliųjų techninių reikalavimų paisymas ir bendradarbiavimas, kuriant duomenų formatų registrą.

Iš anglų kalbos vertė S. Racevičiūtė

Straipsnis parengtas pagal pranešimą, skaitytą 2005 m. Osle (Norvegija) vykusioje 71-ojoje IFLA konferencijoje „Libraries – A Voyage of Discovery“.

---

[1] PREMIS oficiali svetainė: http://www.loc.gov/standards/premis

[2] *Reference Model for an Open Archival Information System (OAIS).* Washington, DC: Consultative Committee for Space Data Systems, 2002. (http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf).

[3] *A Metadata Framework to Support the Preservation of Digital Objects.* Dublin, Ohio: OCLC Online Computer Library Center, 2002. (http://www.oclc.org/research/projects/pmwg/pm_framework.pdf).

[4] *Implementing Preservation Repositories for Digital Materials: Current Practice and Emerging Trends in the Cultural Heritage*

Community. Dublin, Ohio: OCLC Online Computer Library Center, 2004. (http://www.oclc.org/research/projects/pmwg/surveyreport.pdf).

[5] *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group, May 2005*. (http://www.oclc.org/research/projects/pmwg/premis-final.pdf).

[6] Ibid., p. ix.

[7] PREMIS oficiali svetainė: www.loc.gov/premis/

[8] PREMIS schemas žr.: http://www.loc.gov/standards.premis/schemas.html

[9] *Data Dictionary B Technical Metadata for Digital Still Images, NISO Z39.87-2002/AIIM 20-2002*. (http://www.niso.org/standards/resources/z39_87_trial_use.pdf).

[10] MIX: http://www.loc.go/mix

[11] Žr.: http://www.loc.gov/mets

[12] Daugiau informacijos: http://www.nationalarchives.gov.uk/pronom/

[13] Daugiau informacijos: http://hul.harvard.edu/gdfr/

[14] Daugiau informacijos: http://tom.library.upenn.edu/fred/