Opin Gögn.is

Hagnýtar leiðbeiningar fyrir opinbera aðila

Gagnapakki

Í einum gagnapakka geta verið ein eða fleiri gagnaskrár sem allar tengjast innbyrðis. Til dæmis gæti gagnapakki fyrir bókaútgáfu frá árunum 2000-2014, innihaldið 15 gagnaskrár, eina fyrir hvert ár sem inniheldur lista af bókum sem gefnar voru út á því ári. Gagnapakkinn gæti alveg eins innihaldið eina stóra gagnaskrá fyrir öll árin (þó það gæti gert uppfærslur erfiðari). Annað dæmi gæti verið gagnapakki yfir kaffitegundir. Sá gagnapakki gæti innihaldið eina skrá sem innihéldi lista af kaffibaunategundum, meðmælum um bestu brennsluaðferðina, upprunaland og fleira. Gagnapakkinn gæti svo innihaldið aðra skrá með lýsingum á brennsluaðferðunum og svo nokkrar skrár með kortagögnum yfir kaffibaunaakra í hverju landi.

Gagnapakkinn á að gefa sem heillegasta mynd af “sértæku viðfangsefni”, til að forðast rangtúlkanir og gera notkun á gögnunum eins auðvelda og mögulegt er. Til þess er leikurinn gerður, að auðvelda endurnýtingu á gögnunum á sem réttastan hátt, eins mikið og mögulegt er.

Nöfn og aðrar upplýsingar gagnapakka og skráa

Annað sem þarf að hafa í huga til að auðvelda endurnýtingu á gögnunum er að gefa gagnapakkanum og tilheyrandi skrám skiljanleg, lýsandi nöfn og forðast sérfræðiorð. Nafnið er það fyrsta sem mögulegir notendur gagnanna sjá og þá kemur það oftast fyrir í lista sem viðkomandi notandi rennir hratt yfir til að finna gagnapakkann sem leitað er að. Þá er mikilvægt að viðkomandi þurfi ekki að stoppa og spá í hvað nafnið þýðir heldur skilji það undir eins.

Sem dæmi væri *Bjúgaldinaútflutningsskrá* of flókið þar sem það krefst þess að sá sem les nafnið þurfi að stoppa, þótt það sé ekki nema stutta stund, til þess að skilja við hvað er átt. Þótt bjúgaldin sé mögulega rétt orð er það ekki eitthvað sem almenningur notar daglega. *Útflutningur banana* er mun betra nafn sem fólk þekkir almennt til. Það skiptir ekki máli þótt opinbera heiti gagnaskráarinnar sé *bjúgaldinaútflutningsskrá* ef almenningur skilur ekki við hvað er átt í fljótu bragði. Sérfræðingar þurfa að horfa framhjá orðanotkun sem þeir eru vanir og nota frekar einfaldari og almennari orð.

Það sem ber að hafa í huga þegar nöfn eru valin er:

- Nöfnin þurfa að vera skýr og lýsandi fyrir almenning.
- Nöfnin mega ekki vera of löng.
- Aðalorðin eiga að standa fremst, sbr. *Tekjur ríkissjóðs: mánaðarlegt yfirlit*.

Flóknari orð, sérfræðiheiti og nánari upplýsingar eru svo settar í lýsingu gagnapakkans, t.d. *Útflutningi banana er haldið til haga í bjúgaldinaútflutningsskrá*. Það er einnig mikilvægt að gera sér grein fyrir því að í leitarviðmóti Opin gögn.is kemur ekki fram hvaðan gagnapakkinn kemur. Því er mikilvægt að heiti viðkomandi stofnunar eða sveitarfélags komi fram í lýsingunni (eftir því sem við á).

Leitin á síðunni leitar eftir atriðum sem koma fram í nafni og lýsingu gagnapakka og skráa sem getur valdið vandræðum. Tölvur eru ekki nægilega góðar í íslenskri beygingu eða samheitum. Til dæmis, ef við höldum okkur innan matarkyns gagnapakkadæma þá gæti verið til gagnapakki sem héti *Avókadóuppskriftir*. Einhverjir gætu notað íslenska heitið *lárpera* í staðinn fyrir *avókadó* en leit að orðinu *lárpera* myndi ekki skila neinum niðurstöðum. Lausnin á þessu vandamáli felst meðal annars í að nota efnisorð, sem eru einnig kölluð *tög*.

Með efnisorðum er hægt að bæta við stökum orðum eða mjög stuttum orðasamsetningum sem auka leitarhæfi gagnapakkans.

Hver gagnapakki getur haft fleiri en eitt efnisorð/-orðasamsetningu sem eru þá vanalega aðskilin með kommu í innslætti, t.d. *landmælingar, uppskipting lands*. Í dæminu um Avókadóuppskriftir væri hægt að setja inn efnisorðið *lárperur* og þar með gætu notendur sem eru vanir orðinu *lárpera* einnig fundið gagnapakkann.

Ákveðin venja hefur myndast um hvernig skuli byggja upp efnisorð:

- Efnisorð/tög skulu alltaf vera skrifuð í nefnifalli, fleiritölu og án greinis nema almennur ritháttur segir til um annað sbr. fyrir jörð ætti að nota *jarðir* nema átt við sé um jörðina okkar (heiti reikistjörnunnar) því þá ætti að nota efnisorðið *jörðin* þar sem það er almennur ritháttur fyrir reikistjörnuna.
- Mælt er með að efnisorð/tög séu skrifuð með lágstaf og gildir það þá líka um sérnöfn sbr. *háskóli íslands*

Annar kostur þess að nota efnisorð er að með þeim tengjast gagnapakkar sem auðveldar notendum að finna fleiri gagnapakka innan sama áhugasviðs. Þar af leiðandi ætti einnig að endurtaka orð úr titli eða lýsingu. Gagnapakkadæmið okkar um *Avókadóuppskriftir* gæti þar með verið með efnisorðin: *lárperur, avókadó, uppskriftir, matur*. Út frá því gæti notandi smellt á efnisorðið *uppskriftir* til þess að sjá alla gagnapakka sem einnig innihalda efnisorðið *uppskriftir* og út frá því fundið *ástaraldinuppskriftir*. Þetta auðveldar notendum að flakka um síðuna og finna áhugaverða gagnapakka. Í gagnaumsjónarkerfi eins og Opin Gögn.is, sem með tímanum mun innihalda fjölda gagnapakka, er mikilvægt að auðvelda notendum að finna gögnin strax frá upphafi með því að hugsa aðkomuna að gögnunum út frá sjónarhorni notenda/almennings frekar en sjónarhorni sérfræðings.

Gögn (tilföng)

Hver gagnapakki getur innihaldið eina eða fleiri gagnaskrá en hann getur auk þess innihaldið aðrar hjálplegar upplýsingar eins og til dæmis leiðbeiningar eða vörpunartöflur. Það er ástæðan fyrir því að Opin gögn.is notar orðið *tilföng*. Með tilföngum er átt við

gagnaskrár, leiðbeiningar og aðrar skrár sem tilheyra gagnapakkanum.

Skráarsnið

Gagnaskrár geta verið vistaðar í mismunandi skráarsniðum. Það mikilvægasta af öllu er að koma gögnunum í umferð en tefja ekki birtingu gagnanna vegna þess að ekki er búið að færa þau yfir á hentugt snið, eins og til dæmis CSV. Það er betra að setja þau út á því formi sem maður hefur þau og vera svo viðbúin því að taka á móti og verða við ábendingum frá notendum um hentugri snið eða uppsetningar.

Tim Berners-Lee, höfundur veraldarvefsins, setti saman fimm stjörnu kerfi opinna gagna sem opinberir aðilar geta horft til á vegferð sinni til tengdra opinna gagna (*e. Linked Open Data*). Þetta fimm stjörnu kerfi (gagnapakkar geta fengið 0-5 stjörnur í einkunn) hentar einnig mjög vel við opnun gagna almennt. Opinberir aðilar ættu að gera sér grein fyrir því að hugmyndin er ekki að allir gefi út fimm stjörnu gagnapakka strax í upphafi heldur má byrja með eina stjörnu og vinna sig svo upp í fimm stjörnur með tímanum.

Hér er er stutt lýsing á fimm stjörnu kerfi Tim Berners-Lee:

- ★ Gögnin eru á vefnum undir opnu leyfi (lykilatriðið er að gefa út undir opnu leyfi)
- ★★ Gögnin eru gefin út á véllæsilegu sniði (frekar en innskönnuð mynd af töflu eða PDF skjal sem er ekki véllæsilegt)
- ★★★ Gögnin eru gefin út á opnu skráarsniði (þ.e. snið án hugverkaréttinda, töflur eru til dæmis gefnar út sem CSV skrár frekar en Excel skrár)
- ★★★★ Notar opna staðla til að auðkenna færslur (önnur gagnasöfn geta bent á einstaka færslur í gögnunum)
- ★★★★★ Gögnin tengd inn í önnur gögn (gögnin benda á einstaka færslur í öðrum gagnasöfnum)

Sem sagt, fyrir opinbera aðila er gott að byrja á því að reyna að útbúa einnar stjörnu gagnapakka með það að markmiði að vinna

sig upp í þrjár til fimm stjörnur. Flestir notendur yrðu ánægðir með þriggja stjörnu gagnapakka því þá eru gögnin sjálf orðin nytsamleg. Með fjórum og fimm stjörnum verður gagnapakkinn skiljanlegri fyrir tölvur en í dag eru slík tengd gögn enn aðallega viðfangsefni rannsókna um hvernig megi byggja upp gögn á sama hátt og veraldarvefinn og jafnvel betur.

Það er langbest að setja sér markmið um að klifra upp stjörnustigann og byrja strax á því að næla sér í eina stjörnu. Með því að taka á móti athugasemdum gefst notendum færi á að hjálpa til við að fjölga stjörnum gagnapakkans.

Tengill og upphal

Opinber aðili sem ætlar að birta gögn á Opin gögn.is getur bætt við tilföngum á tvo vegu:

- Með því að hala upp viðkomandi skrá
- Með því að gefa upp tengil á skrána á öðrum netþjóni

Í flestum tilfellum er hentugast að nota tengla frekar en upphal. Þá er hlutverk Opin gögn.is aðallega að taka saman á einn stað yfirlit um þau opnu gögn sem mismunandi stofnanir gefa út. Þjóðskrá Íslands, sem heldur utan um rekstur Opin gögn.is, veitir góðfúslega umsjónarmönnum gagna möguleika á að hala skránni upp á þeirra netþjóna en það ætti aðeins að nota í þeim tilfellum þar sem gögnin geta ekki verið geymd á öðrum stað.

Með tengli er bæði þægilegra að uppfæra gagnaskrárnar án þess að þurfa að fara í gegnum Opin gögn.is (svo framarlega sem sama vefslóðin sé notuð, þ.e. að vefslóð gagnanna sé varanleg). Auk þess er það betra fyrir Þjóðskrá að netþjónar hennar fyllist ekki af stórum skrám. Með þessu mætti í raun segja að gögnunum sé streymt af netþjónum opinbera aðilans en þau gerð finnanleg/aðgengileg í gegnum Opin gögn.is. Tengillinn virkar líka vel fyrir þá opinberu aðila sem opna á forritaskil (*e. API*) að gögnunum. Í þeim tilfellum þar sem ætlunin er að hala upp skránni er best að ráðfæra sig við Þjóðskrá og biðja um leyfi, að minnsta kosti svo hægt sé að undirbúa netþjóna undir hýsingu fleiri

gagnaskráa.

Á Opin gögn.is snýst þetta um að velja *Hala upp* eða *Tengill* þegar nýju tilfangi er bætt við gagnapakka.

Uppfærslur á tilföngum

Ef gagnaskrá breytist og eldri útgáfu var halað upp á netþjóna Opin gögn.is í stað þess að vera sett inn sem tengill eða ef tengillinn á gögnin breytist, þarf að uppfæra tilfangið. Það ferli er alveg eins og ferlið þegar nýju tilfangi er bætt við. Notendum gefst tækifæri til að velja á milli *Hala upp* eða bæta við/breyta *Tengli*. Aðeins kerfisstjórnendur Þjóðskrár eða stjórnandi gagnapakka (hjá viðkomandi stofnun) geta uppfært tilföngin.

Í flestum tilfellum ætti að forðast slíkar uppfærslur. Það er betra að bæta við skrám frekar en að uppfæra skrár því notendur taka ekki endilega eftir því að skrár hafi verið uppfærðar. Ef svo færi að notandi fylgist vel með og sækir nýjustu útgáfu gagnanna þegar einhverjum línum hefur verið bætt við þá þýðir það að notandinn hefur líklegast nú þegar sótt meirihluta gagnanna. Það er í raun bara eyðsla á tíma, bandbreidd og vélarafli að sækja mörgum sinnum sömu gögnin og þurfa svo að sía þau frá til að koma í veg fyrir að eldri gögn séu óvart notuð mörgum sinnum í úrvinnslu.

Þess vegna er gott að hugsa um hvernig væri best að brjóta upp tilföngin þannig að þegar uppfæra á gögnin sé gagnapakkinn uppfærður frekar en eitt tilfang. Þetta er best að meta út frá uppfærslutíðninni. Til dæmis, ef ný gögn birtast í hverjum mánuði væri gott að brjóta tilföngin upp eftir mánuðum þannig að *í næsta mánuði* sé nýju tilfangi bætt við frekar en að það gamla sé uppfært og að notandi þurfi að ná aftur í alla gömlu mánuðina og sía þá burt. Þetta er líka auðveldara fyrir umsjónarmann gagnanna.

Þumalputtareglan er að uppfærslur á tilföngum ætti aðeins að vera notuð til að leiðrétta gögn en ekki til þess að bæta við gögnum.