CLARIN-miðstöð á Stofnun Árna Magnússonar í íslenskum fræðum

1 Inngangur

Síðla árs 2018 fól mennta- og menningarmálaráðuneytið Stofnun Árna Magnússonar í íslenskum fræðum að hafa forystu um þátttöku Íslands í evrópska rannsóknarinnviðaverkefninu CLARIN ERIC (https://www.clarin.eu/) – CLARIN stendur fyrir „Common Language Resources and Technology Infrastructure“ og ERIC stendur fyrir „European Research Infrastructure Consortium“. CLARIN ERIC er undir hatti Evrópusambandsins og starfar eftir samþykktum sem hafa verið staðfestar af framkvæmdastjórn þess.

Flest ríki Evrópu taka þátt í þessu samstarfi, þ. á m. öll Norðurlönd, og ber mennta- eða vísindamálaráðuneyti hvers ríkis yfirleitt ábyrgð á þátttökunni. Ráðuneytið felur síðan einni stofnun að hafa forystu um þátttöku ríkisins í CLARIN ERIC og standa fyrir myndun landshóps (e. *national consortium*) helstu hagsmunaaðila, einkum háskóla og rannsóknarstofnana, en sums staðar einnig málnefnda, landsbókasafna og annarra safna. Ráðuneytið tilnefnir einnig landsfulltrúa (e. *national coordinator*) sem er yfirleitt starfsmaður forystustofnunarinnar og stýrir starfinu í viðkomandi landi.

Orð og tunga 22 (2020), 119–125, https://doi.org/10.33112/ordogtunga.22.9

2 Hvað er CLARIN ERIC?

CLARIN ERIC er stofnað til að halda utan um stafræna innviði – gögn og hugbúnað – til nota við rannsóknir í félags- og hugvísindum. Eftir að almenn tölvuvæðing hófst fyrir 40 árum eða svo hefur orðið til gífurlega mikið af stafrænum gögnum af ýmsu tagi – textasöfn, orðasöfn, og alls kyns skrár. Sumt af þessu hefur verið byggt upp frá grunni á undanförnum áratugum, en einnig hefur verið gert mikið átak í því að koma eldri gögnum á stafrænt form.

Stafræn gögn bjóða vitaskuld upp á margvíslega möguleika umfram pappírsgögn. Það er margfalt fljótlegra að leita í þeim og vinna ýmiss konar skrár og töflur upp úr þeim. Stafræn gögn eru líka margfalt sveigjanlegri en pappírsgögn – auðvelt að lagfæra villur í þeim, uppfæra þau, raða þeim á mismunandi hátt o.s.frv. Notendur eru ekki lengur háðir einu eintaki á tiltekinni stofnun eða safni – það er auðvelt að afrita gögnin og dreifa þeim, eða gera þau aðgengileg á netinu.

Þetta stórbætta aðgengi að gögnum leiðir vitanlega til þess að miklu fleiri fræðimenn geta nýtt þau en áður, og eflir þannig og styrkir margvíslegar rannsóknir. En þetta þýðir líka að fólk er oft að skoða og vinna með gagnasöfn sem það þekkir ekki fyrir. Söfnin eru mjög margbreytileg, framsetning þeirra misjöfn, leitarmöguleikar ólíkir, og svo mætti lengi telja. Það getur verið mjög flókið og tímafrekt fyrir ókunnuga að setja sig inn í þetta og átta sig á því hvernig hægt er að finna það sem leitað er að í gögnunum.

Meginmarkmið CLARIN ERIC er að nýta þá möguleika sem stafræn málleg gögn, málföng (e. *language resources*), bjóða upp á og bæta aðgengi að þessum gögnum og hugbúnaði sem gerður er til að vinna með þau. Þetta krefst margvíslegs undirbúnings sem mikilvægt er að sem víðtækust samvinna sé höfð um. Jafnframt er markmið CLARIN ERIC að notendur geti nýtt notandanafn og aðgangsorð við heimastofnun sína til að fá aðgang að þessum gögnum og búnaði (e. *single sign-on*).

Í hverju þátttökulandi eru settar upp CLARIN-miðstöðvar, ein eða fleiri. Þessar miðstöðvar eru af mismunandi tegundum. Einfaldasta tegundin eru svokallaðar C-miðstöðvar (e. *CLARIN C-Centre*) sem varðveita lýsigögn (e. *metadata*), en aðaltegundin er svokallaðar B-miðstöðvar (e. *CLARIN B-Centre*) sem varðveita gögn ásamt lýsigögnum og veita ákveðna þjónustu, s.s. upplýsingar um gögn og tæknilega ráðgjöf. Þriðja tegundin er svo K-miðstöðvar (e. *CLARIN K-Centre*) sem eru upplýsingaveitur um tiltekin málleg efni, t.d. einstakt tungumál.

3 Gögn í CLARIN ERIC

Áður en gögn eru skráð í miðlægan gagnagrunn CLARIN ERIC þarf í fyrsta lagi að útbúa lýsigögn þar sem innihaldi gagnanna er lýst og upplýsingar gefnar um ýmis atriði sem þau varða – höfunda, tungumál, gagnasnið, notkunarskilmála o.s.frv. Þessi lýsigögn þurfa að vera á samræmdu sniði til að auðvelda notkun þeirra og leit í þeim. CLARIN ERIC hefur útbúið sniðmát fyrir lýsigögn til að leiðbeina notendum um hvaða upplýsingar þurfi að fylgja gögnunum.

Í öðru lagi þarf að ákveða notkunarskilmála gagnanna – hvort þau eru öllum opin og aðgengileg án takmarkana, eða hvort einhverjar takmarkanir eru á aðgengi og notkun, og þá hverjar. Það er t.d. algengt að óheimilt sé að nýta gögn í hagnaðarskyni eða breyta þeim á einhvern hátt. Til eru ýmsir staðlaðir leyfisskilmálar sem hægt er að velja á milli, t.d. svonefnd *Creative Commons*-leyfi (https://creativecommons.org/), en einnig er hægt að gera gögn aðgengileg með sérsniðnum leyfum.

Í þriðja lagi getur þurft að breyta gagnasniðinu. Ýmis samræmd snið hafa verið sett fram fyrir mismunandi tegundir mállegra gagna – textasöfn, orðasöfn, handrit, uppskriftir hljóðskráa o.s.frv. Þar má ekki síst nefna margvísleg snið frá *Text Encoding Initiative* (https://tei-c.org/). Æskilegt er að gögn séu á einhverju slíku þekktu sniði eftir því sem kostur er, en lágmarkskrafa er að sniði gagnanna sé nákvæmlega lýst þannig að auðvelt sé fyrir notendur að átta sig á því.

Í fjórða lagi þarf að gera gögnin aðgengileg, ásamt lýsigögnum. Það er hægt að gera á ýmsan hátt. Að sumum gögnum er eingöngu leitaraðgangur gegnum ákveðið leitarviðmót. Notendur geta þá leitað að orðum og orðasamböndum en það er misjafnt eftir gagnasniði og leitarviðmóti hversu nákvæm leitin getur verið, og eftir hvaða atriðum er hægt að leita. Í öðrum tilvikum er hægt að sækja gögnin í heild, stundum með ákveðnum skilyrðum sem kveðið er á um í leyfisskilmálum sem þarf að samþykkja áður en gögnin eru sótt.

Miðlægt tölvukerfi CLARIN ERIC skannar reglulega allar lýsigagnaskrár sem vistaðar eru á öllum CLARIN-miðstöðvum. Upplýsingar úr þessum skrám fara inn í miðlægan gagnagrunn, sýndarsafn málfanga (e. *Virtual Language Observatory*, https://vlo.clarin.eu/) og þar er hægt í einni leit að leita í lýsigögnum meira en milljón málfanga um alla Evrópu. Öllum gögnum sem eru lögð inn til einhverrar CLARIN-miðstöðvar er gefið varanlegt auðkenni (e. *Persistent Identifier*, *PID*).

Það tryggir að ávallt sé hægt að finna gögnin enda þótt vistun þeirra og hefðbundin vefslóð (URL) kunni að breytast.

4 Ísland og CLARIN ERIC

Upphaf CLARIN má rekja aftur til 2008 þegar undirbúningsfasi þess hófst. Ísland var ekki með frá byrjun en komst inn í samstarfshóp undirbúningsfasans árið 2010, en án fjárhagslegs stuðnings. Þegar eðli CLARIN breyttist árið 2012 og CLARIN ERIC varð til varð Ísland ekki stofnaðili. Íslandi var þó boðin þátttaka í sérstöku norrænu CLARIN-neti, Nordic CLARIN Network, sem kostað var af NordForsk á árunum 2014–2017. Íslenskir fræðimenn tóku þátt í ýmsum fundum og vinnustofum sem netið skipulagði.

Í verkáætlun um íslenska máltækni sem gefin var út sumarið 2017 er sérstakur kafli um CLARIN. Þar er útskýrt hvernig aðild myndi gagnast Íslandi, með aðgangi að margvíslegum búnaði og gögnum, svo og að sérþekkingu á ýmsum sviðum. Innan máltækniáætlunarinnar á að þróa margs kyns gögn og búnað og það er mjög mikilvægt að gerð, lýsing og varðveisla þessara málfanga fylgi viðurkenndum stöðlum. Í áætluninni var því lagt til að Ísland gerðist aðili að CLARIN ERIC til að auðvelda vinnslu og varðveislu málfanganna.

Mennta- og menningarmálaráðuneytið féllst á þessa tillögu og ákvað að fjármagna þátttöku Íslands í CLARIN ERIC til fimm ára. Það kom þó í ljós að nauðsynlegt væri að breyta lögum til að Ísland gæti orðið fullgildur aðili og því var ákveðið að sækja um áheyrnaraðild (e. *observership*). Umsóknin var samþykkt á allsherjarþingi CLARIN ERIC í nóvember 2018 og áheyrnaraðild Íslands tók gildi 1. nóvember það ár.

Ráðuneytið fól Stofnun Árna Magnússonar í íslenskum fræðum að vera fulltrúi Íslands gagnvart CLARIN ERIC og leiðandi aðili (e. *leading partner*) í íslenskum CLARIN-landshópi, eins og áður segir. Eiríkur Rögnvaldsson, prófessor emerítus, var tilnefndur landsfulltrúi CLARIN á Íslandi. Flestar stofnanir sem málið varðar taka þátt í landshópi CLARIN-IS. Auk Stofnunar Árna Magnússonar í íslenskum fræðum eru það Háskóli Íslands, Háskólinn í Reykjavík, Landsbókasafn Íslands – Háskólabókasafn, Þjóðskjalasafn Íslands, Íslensk málnefnd, Ríkisútvarpið, og Almannarómur.

CLARIN-miðstöðin á Árnastofnun, CLARIN-IS (https://clarin.is/), tók til starfa í ársbyrjun 2019. Þar starfa Eiríkur Rögnvaldsson landsfulltrúi í 40% starfi og frá 1. apríl sama ár Samúel Þórisson tölvunar-

fræðingur í fullu starfi. Meginverkefni miðstöðvarinnar hafa verið tvö: Annars vegar þátttaka í samstarfi CLARIN ERIC, og hins vegar uppbygging varðveislusafns (e. *repository*) sem komst í gagnið síðla árs (https://repository.clarin.is/). CLARIN-miðstöðin hefur einnig verið skráð sem lýsigagnamiðstöð (e. *CLARIN C-Centre*).

Í júní 2019 voru ný lög um samtök evrópskra rannsóknarinnviða samþykkt á Alþingi. Í framhaldi af því ákvað mennta- og menningarmálaráðherra snemma árs 2020 að Ísland sækti um fulla aðild að CLARIN ERIC. Umsóknin var samþykkt í lok febrúar og Ísland er fullgildur aðili að CLARIN ERIC frá 1. febrúar 2020 en gengið var frá undirritun aðildarsamnings 10. mars. CLARIN-miðstöðin, sem hafði verið í húsnæði Árnastofnunar á Laugavegi 13, er nú flutt í Þingholtsstræti 29 þar sem máltæknihópur Árnastofnunar hefur aðsetur.

5 Starfsemin fram undan

Þótt megintilgangurinn með stofnun CLARIN ERIC hafi verið að styðja rannsóknir í hug- og félagsvísindum nýtast þau gögn sem komið hefur verið upp á ýmsum öðrum sviðum, ekki síst í máltækni sem er í örum vexti víðast hvar. Eins og áður segir er aðild Íslands að CLARIN ERIC fjármögnuð af máltækniáætlun stjórnvalda og í samningum Almannaróms f.h. ríkisins við SÍM – samstarf um íslenska máltækni, sem vinnur að framkvæmd máltækniáætlunarinnar, eru ákvæði um að allar afurðir máltækniverkefnisins, bæði gögn og hugbúnaður, verði lagðar inn í varðveislusafn íslensku CLARIN-miðstöðvarinnar.

Þetta er grundvallaratriði. Ein meginforsenda máltækniáætlunarinnar er að afurðir hennar verði öllum aðgengilegar og ókeypis, þannig að fyrirtæki og stofnanir sem vilja nýta þær við þróun máltæknibúnaðar geti gengið að þeim sér að kostnaðarlausu. Því er mjög mikilvægt að hægt sé að ganga að þeim á einum stað, ítarleg lýsing á þeim liggi fyrir, og þær séu á þekktu og vel skilgreindu sniði. Innlögn í CLARIN-miðstöðina tryggir þetta allt saman.

Fyrstu afurðum máltækniverkefnisins hefur þegar verið skilað og þær skráðar í safnið. Þar geta CLARIN-notendur hvar sem er fundið þær gegnum áðurnefnt sýndarsafn málfanga, og sótt þær þangað. Skráning gagna Árnastofnunar í varðveislusafnið er einnig hafin og verður unnið að henni á næstunni. Að því loknu verður farið að huga að skráningu gagna annarra þátttakenda í íslenska landshópnum í varðveislusafnið.

Fullgildum þátttakendum í CLARIN ERIC er skylt að koma upp a.m.k. einni tæknilegri þjónustumiðstöð (e. CLARIN *B-Centre*). CLARIN-IS vinnur að þessu en það er töluvert mál – slík miðstöð þarf að fullnægja ýmsum skilyrðum og fá sérstaka vottun. Auk þess er áhugi á því hjá íslensku CLARIN-miðstöðinni að koma upp þekkingarmiðstöð (e. *CLARIN K-Centre*) um íslenskt mál, þar sem hægt væri að sækja hvers kyns gögn og upplýsingar um íslensku. Undirbúningur þessa er á frumstigi, en slík miðstöð yrði hugsanlega rekin í samvinnu við aðra aðila, t.d. Íslenska málnefnd sem hefur lýst áhuga á því að koma upp upplýsingaveitu af þessu tagi.

Einnig liggur fyrir að kynna CLARIN fyrir hugsanlegum notendum, einkum fræðimönnum í ýmsum greinum hug- og félagsvísinda. Það er ljóst að innan CLARIN ERIC eru margvísleg gögn, innlend og erlend, sem geta gagnast málfræðingum, bókmenntafræðingum, sagnfræðingum, heimspekingum, félagsfræðingum, mannfræðingum, stjórnmálafræðingum, þjóðfræðingum, og mörgum öðrum. Fáir vita hins vegar af þessum gögnum og þeim möguleikum sem í þeim felast, og það er hlutverk CLARIN-miðstöðvarinnar að kynna þetta.

Enn fremur er stefnt að öflugri þátttöku í ráðstefnum og viðburðum á vegum CLARIN ERIC. Sú þátttaka er þegar hafin – sjö Íslendingar sóttu ársráðstefnu CLARIN ERIC í Leipzig haustið 2019 og voru þar með einn fyrirlestur og þrjú veggspjöld. CLARIN ERIC kostar þátttöku fimm fulltrúa frá hverju aðildarlandi, auk þeirra sem eru með erindi eða veggspjöld. Einnig er einum doktorsnema frá hverju landi boðin þátttaka sér að kostnaðarlausu. Auk þessa stendur CLARIN ERIC fyrir vinnustofum af ýmsu tagi sem Íslendingar geta nú sótt – og eru þegar farnir að gera.

6 Lokaorð

Mikilvægi stafrænna gagna í hvers kyns rannsóknum í hug- og félagsvísindum hefur aukist hröðum skrefum á undanförnum árum. Þar nægir að nefna vefinn *Tímarit.is*, sem óhætt er að segja að hafi gerbreytt aðstöðu til rannsókna á íslenskri málfræði og sögu, svo að dæmi séu tekin. Risamálheildin (https://malheildir.arnastofnun.is/) hefur einnig nýst á margvíslegan hátt á þeim stutta tíma sem liðinn er síðan hún var opnuð. En fjölmörg önnur íslensk málleg gagnasöfn eru til þótt ekki séu þau jafnþekkt eða aðgengileg.

Það er hlutverk íslensku CLARIN-miðstöðvarinnar að gera íslensk málföng sem aðgengilegust og vekja athygli á notagildi þeirra. Að því verður unnið eftir megni á næstu árum.

Eiríkur Rögnvaldsson
landsfulltrúi CLARIN
Stofnun Árna Magnússonar í íslenskum fræðum