Suured tekstiandmed humanitaarteadustes – kellele ja kuidas?

Liina Lindström (Tartu Ülikool)
liina.lindstrom@ut.ee
Ettekanne RR teadusnõukogus
6.9.2017

Sihtgrupid

I rühm:

- humanitaar- ja sotsiaalvaldkonnas tegutsevad teadlased (keel, kirjandus, ajalugu, kunstiajalugu, muusikaajalugu, ajakirjandus, jne.)
- samades valdkondades tegutsevad üliõpilased
- asjaarmastajad

II rühm: andmekaeve ja -teaduse spetsialistid

keeletehnoloogia, arvutilingvistika, informaatika, andmeteaduse valdkonnas tegelevad teadlased, õppejõud, üliõpilased

- Piirid I ja II rühma vahel ei ole aredad (digihumanitaaria)

Sihtgrupid II

I rühm: humanitaar- ja sotsiaalvaldkonna teadlased, üliõpilased, asjaarmastajad

- reeglina vähesed tehnilised oskused, ent suur vajadus andmete kiireks kättesaamiseks
- vajab eeltöödeldud materjale ja hästi läbimõeldud töökeskkonda; sõltuvad raamatukogu tootest

II rühm: andmeteaduse spetsialistid

- potentsiaalne edasiarendajate rühm, kes teeb asju tehnilistest huvidest lähtuvalt
- kasutab andmeid keeltehnoloogia/andmekaeve tehniliseks arendamiseks
- vajab andmeid võimalikult lihtsal, ent siiski korrastatud kujul
- Oluline on kiire ja lihtne ligipääs andmetele

Tekstiandmetest eeltöötlemine

- kõigist tekstidest peab olema saadaval lihtne txt-formaat: pdf-id, pildid vms on mõttetud (või ainult lisainfo): see on vajalik mõlemale sihtgrupile
- Tasub panustada tekstituvastuse (OCR) kvaliteeti (vanemad tekstid)
- Html-ist või muust koodisodist puhastamine (uuemad tekstid)
- Ühtlustamine, süstematiseeritud kujule viimine
- metainfo

Metaandmed

- Vähemalt sama oluline kui tekstiandmed ise
- peab olema kättesaadav igal töötluse astmel → läbimõeldud esitus
- Metaandmed võidakse kasutada ka iseseisvalt (tekstiandmetest eraldi)

Tekstiandmete tüübid I

- Vanad ajalehed jm perioodika
- Autoriõigustega katmata tekstid
- Autoriõigustega kaetud tekstid
- Veebitekstid
- Veel?

- Tekstitüüpidega tasub ringi käia erinevalt

Tekstiandmete tüübid II: Vanad ajalehed jm perioodika

- Tõenäoliselt kõige laiem kasutajaskond humanitaar- ja sotsiaalvaldkonna teadlaste hulgas
- tasub panustada korraliku veebipõhise keskkonna ülesehitamisele
- Tasub panustada kvaliteetsele tekstituvastusele
- Tasub panustada keeletehnoloogiliste vahendite rakendamisele

Tekstiandmete tüübid III: Vanad ajalehed jm perioodika

- OCRi tuunimine
- Päringuvõimaluste avardamine ja kohandamine tuvastusvigadega (nt Levensteini distantside rakendamine)
- Nimede tuvastus
- Lemmatiseerimine
- Metaandmete mitmekülgne kasutamine
- Vektorsemantika rakendamine päringute tegemisel
- Visualiseerimine (ajatelg, kaardid, võrgustikud, muu)
- Tulemuste allalaadimisvõimalused (txt, csv, muu)
- Kvantitatiivsete andmete kasutamise võimalus

Tekstiandmete tüübid IV: autoriõiguste alt vabad tekstid (raamatud)

- Heterogeenne materjal erinevatest ajastutest
- Heterogeenne kasutajaskond
- Võib läheneda samamoodi kui vanadele ajalehtedele
- Tõenäoliselt suurem osa kasutajaid eelistab kasutada terviktekstidena

Tekstiandmete tüübid V: autoriõigustega kaetud tekstid

- Tõenäoline sihtgrupp: keele- ja kirjandusteadlased; keeletehnoloogia ja keeleressursside arendajad
- Piiratud ligipääs luua autoriseerimise kaudu
- Teine viis piirata on piirata väljundi mahtu: nt otsitavale reale/lausele võib saada konteksti päritud mahus; suuremaid üksusi ei väljastata

Kus andmeid hoida?

- Kõik andmetüübid peavad olema kättesaadavad veebi kaudu (mida pole veebis, seda pole olemas)
- Erinevate autoriõigusega seotud küsimuste puhul tasub lahendusi otsida pigem
  - kasutajatuvastuse ja kasutajalepingute läbi mõeldud süsteemi kaudu
  - väljastatava tekstihulga piirangute kaudu
- Koha peal kasutatavad töömasinad ei ole lahendus!
  - Piiratud ajaressurss
  - Piiratud vahendid

Mõtteid, soovitusi

- Digari edasiarendamine tekstipõhjaliselt otsitavaks (kõigist tekstiosadest)
- Tulemused allalaetavaks (koos metainfoga), nt csv-formaadis
- Kõigist tekstidest lihtsad *plain text* versioonid kättesaadavaks
- Koostöö keeletehnoloogia ja korpuste arendamisega tegelevate inimestega
- Ärge unustage, et sihtgrupid ja nende vajadused on erinevad
- Kõik peab olema kättesaadav veebi kaudu