

# Digital Humanities Hackathonissa syntyi uutta avointa koodia

Julkaistu 22.05.2015 klo 13:54 Kirjoittaja Jessica Parland-von Essen

Mitä hyötyä ihmistieteiden tutkijalle on aineistojen sähköisestä muodosta? Mitä iloa on datan visualisoinnista ja laskennallisista menetelmistä? Mitä käytännön apua tietojenkäsittelystä ihan oikeasti voi humanistille olla? Mitä sellaista voi koneella voi tehdä, joka tuo aitoa lisäarvoa esimerkiksi historian tutkijalle? Kaikkea kivaa datalla voi tietenkin tehdä, mutta hyödyttääkö se *oikeasti* humanistia?

Helsingin yliopistossa järjestettiin 11.-15.5.2015 työviikon mittainen digitaalisten ihmistieteiden tapahtuma, jossa opiskelijat ja tutkijat eri aloilta lähtivät kokeilemaan tutkimusta neljän eri aineiston tiimoilta. Aineistoina olivat englanninkielinen kirjekorpus, suomalaiset sanomalehdet, Blue Wings-lehdet sekä sadan vuoden takaisia teknisten alojen lehtiä. Tarkoituksena oli nimenomaan edetä tutkimuslähtöisesti ja kehittää teknisiä ratkaisuja varsinaisen kysymyksenasettelun ehdoilla.

Osallistujia oli lähes neljäkymmentä, mukana eri alojen humanisteja sekä tietojenkäsittelyalan opiskelijoita. Viikko oli mielenkiintoinen ja haastava. Työnteko ja luennot vuorottelivat ohjelmassa. Yhteisen kielen löytäminen ja mielekkäiden tutkimuskysymysten muotoilu täysin aineistolähtöisesti ei aina ollut helppoa, mutta yhteistyön kehittyessä päästiin moniin oivalluksiin. Joskus ennakkoon vaikealta tuntuneet asiat ratkesivat yllättävän helposti, toisaalta pienet yksityiskohdat saattavat joskus muodostua pulmallisiksi. Kokonaisuudessaan prosessi on yleensä kokeilun ja muuntelun vuorottelua, jossa edetään yhteistyössä iteroiden. Selvältä näyttää, että digitaaliset aineistot ja menetelmät voivat sekä nopeuttaa ja helpottaa humanistitutkijan työtä että auttaa uusien asioiden oivaltamisessa ja tutkimuskysymysten muotoilemisessa.



# SCOTLAND, ENGLAND, BRITAIN AND THE KINGDOM IN 18<sup>TH</sup>. CENTURY ENGLISH LETTERS

## THE CORPUS OF EARLY ENGLISH CORRESPONDENCE (CEEC)



Part of a letter written by Jane Austen. Austen is one of over 300 authors included in the corpus.

	Men	Women	Total
Informants	218	97	315
Letters	3,676	1,269	4,945
Words	1.61m	0.6m	2.22m

### CEEC in numbers by gender.

	Men	Women	Total
Royalty	10 (5%)	6 (6%)	16 (5%)
Nobility	18 (8%)	27 (29%)	45 (15%)
Gentry	55 (26%)	26 (28%)	81 (26%)
Clergy	46 (21%)	9 (10%)	55 (18%)
Professionals	41 (19%)	12 (13%)	53 (17%)

## RESEARCH QUESTIONS

- 1) How much is Scotland (and related concepts, especially England and Britain) discussed in the corpus before and after the Act of Union?
- 2) Who discusses Scotland?
- 3) What kinds of topics can be modelled based on (parts of) the data?

## ACT OF UNION, 1707

The Act turned the island into one political entity, reigned by Queen Anne (1665–1714).



The Royal Coat of Arms of the United Kingdom of Great Britain and Northern Ireland. The lion represents England, the unicorn Scotland.

## THE PRETENDERS OLD AND YOUNG



HELSINGIN  
HELSINGFORS U  
UNIVERSITY O  
HUMANISTINEN T  
HUMANISTISKA F  
FACUL

Depart

Depart

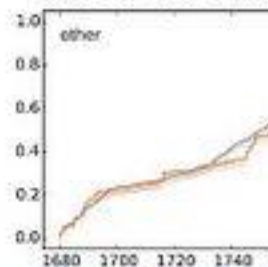
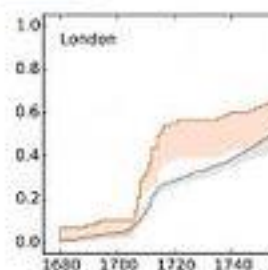
Depart

Depart

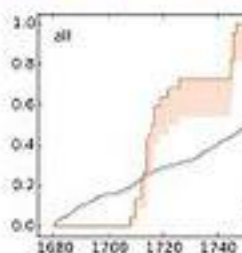
Depart

Depart

## SCOTLAND CEEC



## PRETENDER THE CORPU



## WORD CLOU THE WRITEN

Ihmistieteilijän tutkimusaineistoja ovat potentiaalisesti mitkä tahansa aineistot, ainoastaan tutkimuskysymys ratkaisee sen, miten dataa lähestytään. Keskustelua eri alojen asiantuntijoiden välillä vaatiikin sen keksiminen, mitä datalla voi tehdä. Tässä Hackathonissa käytetyt aineistot olivat osin tutkijoiden aiemmin tuottamia, osin Kansalliskirjaston dataa. Erilaiset korpuukset, kuten Kielipankin aineistot, tarjoavat mahdollisuuksia yhteiskunnan, historian ja kulttuurin tutkijoille käyttää laskennallisia menetelmiä. Kirjastojen, museoiden ja arkistojen sähköiset luettelot ja muut aineistot ovat suuria osin hyvin rakenteistettuja tietovarantoja, joista voisi louhia monenlaista tietoa. Silti suuri osa työstä on usein aineiston valmistelua. Tutkijoiden itsensä luomat aineistot tarvitsevatkin säilytystä, eikä tiedonhallintaa tulisi unohtaa, mikäli halutaan aineistolle jatkokäyttöä ja tutkimukselle toistettavuutta. Esitetyt kysymykset vaikuttavat datan ja menetelmien muotoon ja kehitykseen, joten riittävä kuvailu ja kontekstin dokumentointi on tärkeää.

Digital Humanities Hackathonin aikana karttui GitHubiin mielenkiintoista koodia, josta varmasti on hyötyä vastaisuudessa. Sekä aineiston että menetelmien uudelleenkäyttö on usein tehokasta, mutta vaikeutena saattaa olla niiden löytäminen.

Kokonaisuudessaan Mikko Tolosen, Eetu Mäkelän ja Timo Honkelan suunnittelema ja toteuttama tapahtuma oli innostava ja varmaan vastaavia käytännön tekemisestä lähteviä eri kokoisia hankkeita tarvitaan ja tullaan näkemään jatkossakin.

#### **Aiheesta lisää:**

- [Summary of Digital Humanities Hackathon #DHH15 Helsinki Digital Humanities -blogi.](#)
- [Kielipankki](#)
- [Yhteiskuntatieteellinen tietoarkisto](#)