Jarmo Saarti

Kirjastojuridiikan ajankohtaispäivä 4.11.2014

Kirjastot ja datamining, tutkijan ja kirjaston näkökulmat

ITÄ-SUOMEN YLIOPISTO

Sisältö

- Data-yhteiskunnasta tietoyhteiskunnan kautta tietämysyhteiskunnaksi
- Datanlouhinta tutkimuksen työkaluna
- Kirjastojen oikeudet ja velvollisuudet

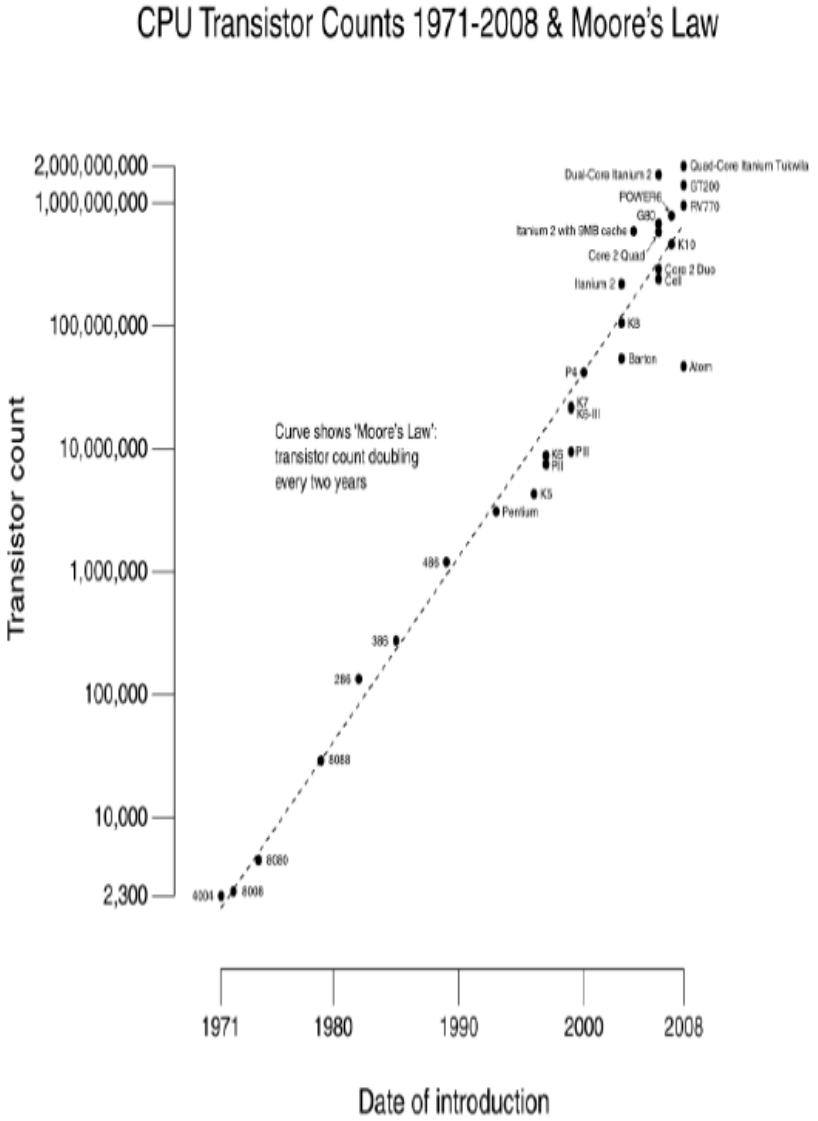

DATA
is the new oil

How Much Data is there?

2013

1.8 zetabytes?

And 80% is unstructured.



Julkaisujen määrän kasvu

- 1990-luvulta alkaen digitaalisten tieteellisten ja muiden julkaisujen määrä on kasvanut räjähdysmäisesti:
  - Maailmassa on julkaistu noin 130 miljoonaa kirjaa (Nosowitz 2010)
  - Noin 2 miljoonaa tieteellistä artikkelia vuosittain
  - Verkossa 114 miljoonaa englanninkielistä julkaisua. Julkaisut ovat artikkeleita, konferenssijulkaisuja, väitöskirjoja, opinnäytteitä, kirjoja, raportteja ja työpapereita. Google Scholarilla voi löytää näistä 100 miljoonaa (n. 87%) (Giles 2014)
- Tämän lisäksi muut digitaaliset julkaisemisen tavat ja formaatit ovat kasvaneet
- Määrät ovat ylittäneet normaalin inhimillisen käsityskyvyn – tarvitaan uusia työkaluja

FIVE KEY TRENDS WHICH WILL CHANGE OUR INFORMATION ENVIRONMENT

TREND 1:

NEW TECHNOLOGIES WILL BOTH EXPAND AND LIMIT WHO HAS ACCESS TO INFORMATION

An ever-expanding digital universe will bring a higher value to information literacy skills such as basic reading and competence with digital tools. People who lack these skills will face barriers to inclusion in a growing range of areas. The nature of new online business models will heavily influence who can successfully own, profit from, share or access information in the future.

TREND 2:

ONLINE EDUCATION WILL DEMOCRATISE AND DISRUPT GLOBAL LEARNING

The rapid global expansion in online education resources will make learning opportunities more abundant, cheaper and more accessible. There will be increased value on lifelong learning and more recognition of non-formal and informal learning.

TREND 3:

THE BOUNDARIES OF PRIVACY AND DATA PROTECTION WILL BE REDEFINED

Expanding data sets held by governments and companies will support the advanced profiling of individuals, while sophisticated methods of monitoring and filtering communications data will make tracking those individuals cheaper and easier. Serious consequences for individual privacy and trust in the online world could be experienced.

TREND 4:

HYPER-CONNECTED SOCIETIES WILL LISTEN TO AND EMPOWER NEW VOICES AND GROUPS

More opportunities for collective action are realised in hyper-connected societies – enabling the rise of new voices and promoting the growth of single-issue movements at the expense of traditional political parties. Open government initiatives and access to public sector data will lead to more transparency and citizen-focused public services.

TREND 5:

THE GLOBAL INFORMATION ECONOMY WILL BE TRANSFORMED BY NEW TECHNOLOGIES

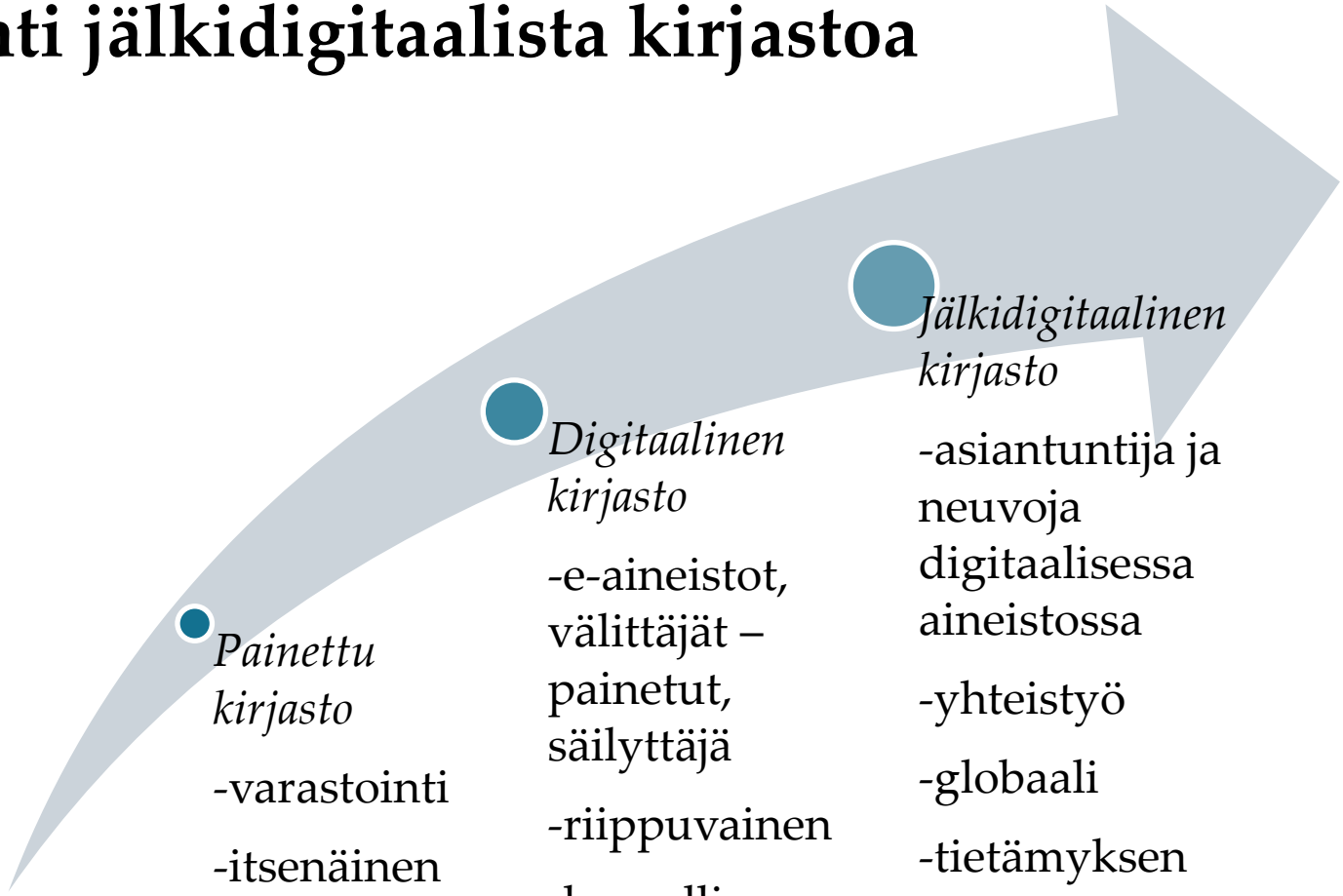
Proliferation of hyper-connected mobile devices, networked sensors in appliances and infrastructure, 3D printing and language-translation technologies will transform the global information economy. Existing business models across many industries will experience creative disruption spurred by innovative devices that help people remain economically active later in life from any location.

TRENDI 1:

UUDET TEKNOLOGIAT SEKÄ LAAJENTAVAT ETTÄ RAJOITTAVAT YKSILÖIDEN TIEDONSAANTIA

Digitaalisen maailman jatkuva laajeneminen lisää informaatiolukutaitojen kuten peruslukutaidon sekä digitaalisten välineiden käyttötaitojen arvoa. Ihmiset, joilla näitä taitoja ei ole, ovat vaarassa syrjäytyä yhä useammilla elämän alueilla. Uusien verkkoliiketoimintamallien luonne vaikuttaa voimakkaasti siihen, keillä tulevaisuudessa on pääsy tietoon, ketkä voivat omistaa, hyödyntää ja jakaa informaatiota.

Kohti jälkidigitaalista kirjastoa



Datan louhinta

- Data mining = datan louhintaa
- Tämän perusteella voidaan tieteessä luoda uutta osaamista ja uutta tieteellistä tutkimustulosta
- Tietoa ei voi louhia, tieto on tulkittua dataa ja vaatii – ainakin esitelmöijän mielestä – tietoisen subjektin tulkitsijaksi

ITÄ-SUOMEN YLIOPISTO

Tiedonlouhinta

Tiedonlouhinta (engl. *data mining*) tarkoittaa joukkoa menetelmiä, joilla pyritään oleellisen tiedon löytämiseen suurista tietomassoista.

Sovelluskohteet [muokkaa | muokkaa wikitekstiä]

Tiedonlouhintaa voidaan soveltaa hyvin laaja-alaisesti, sillä lähtökohdaksi tarvitaan ainoastaan dataa. Tyypillisesti tiedonlouhinnassa käytetty data on esimerkiksi mittauksia teollisuusprosessista, otteita asiakastietokannasta tai vaikkapa web-palvelimen loki-tiedostoja.

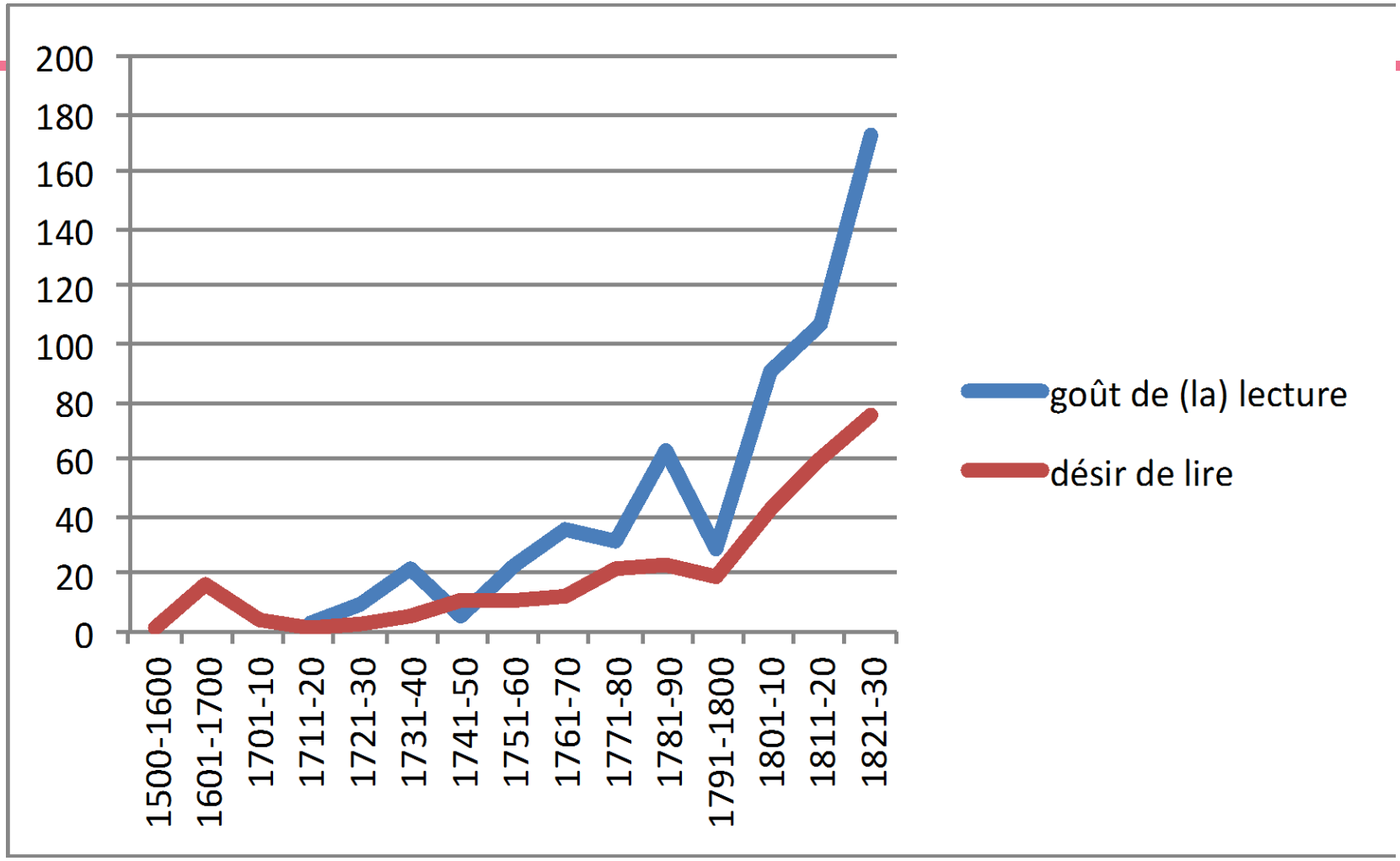
Menetelmät [muokkaa | muokkaa wikitekstiä]

Määritelmänä tiedonlouhinta ei rajaa käytettäviä menetelmiä. Useimmiten käytettäviä algoritmeja ovat mm. erilaiset klusteroinnit, korrelaatiot, neuroverkot, itseorganisoituvat kartat, jne. Yleisesti ottaen tiedonlouhinnan menestyksellisessä hyödyntämisessä kaikkein oleellisinta on datan ja sen eri suureiden kokonaisvaltainen ymmärtäminen. Myös pelkkä innovatiivinen lähestymistapa esimerkiksi datan visualisoinnissa voi auttaa näkemään tietovaraston hyötyjä täysin uudesta perspektiivistä.

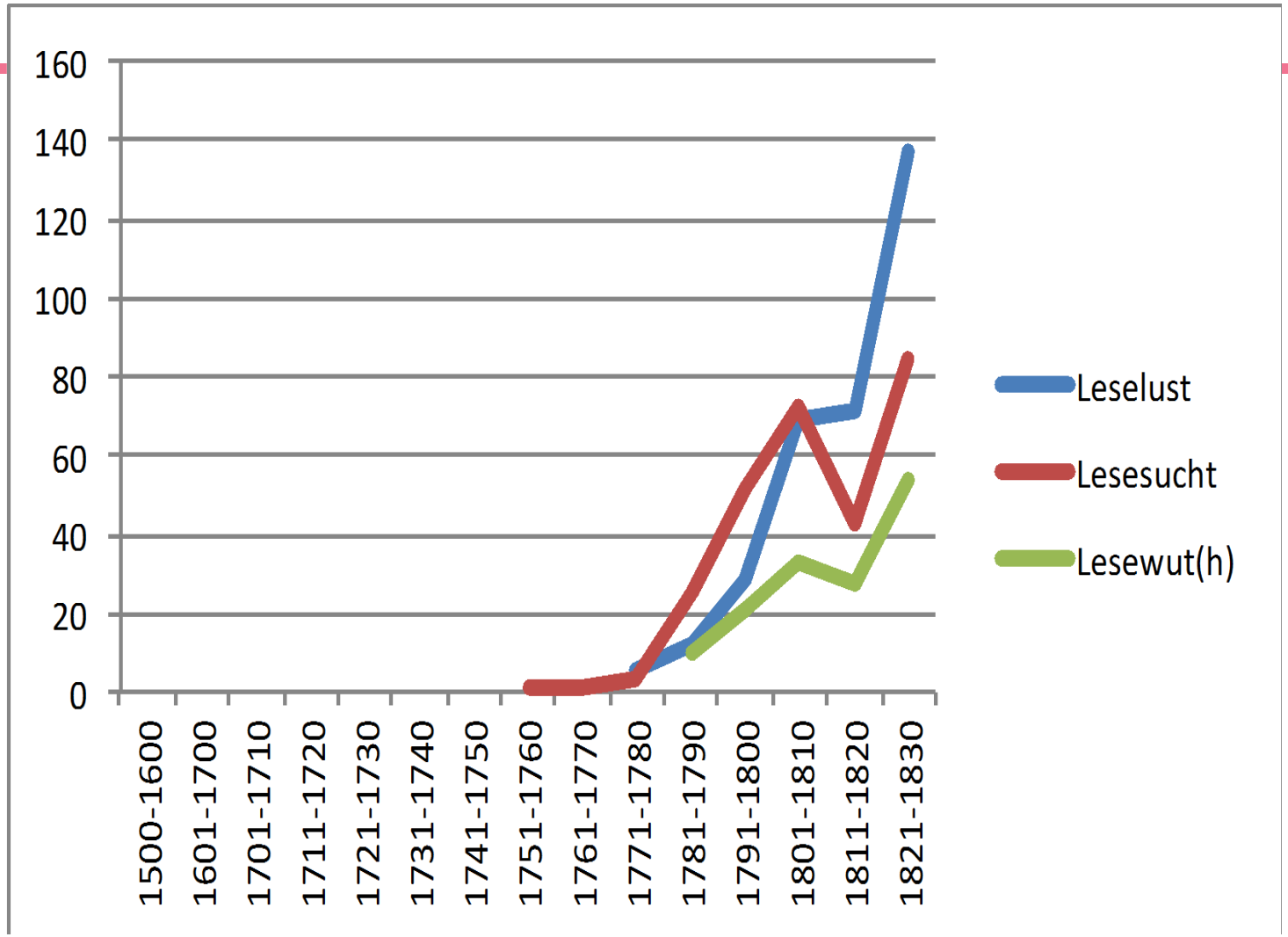
Tutkijan datan louhinta, osa 1

- Perustiedonhaku on datan louhintaa, kun tallennettu tekstimassadata on tarpeeksi iso
- Tutkimuskysymyksiin vastaaminen vs. uusien tutkimuskysymysten etsiminen
- Käsitteiden historia konkreettisena esimerkkinä

ITÄ-SUOMEN YLIOPISTO



Lukemiseen, erityisesti lukuhaluun liittyvien fraasien käytön kasvu alkaa Euroopan voimakkailta kieli- ja kirjallisuusalueilta. Suurten kielialueiden kehitystä kuvaavat Google Books -aineiston perusteella laatimani kuviot keskeisimpien lukuhalua tarkoittavien fraasien esiintymisestä (absoluuttisia lukuja) (tiedot artikkelista Mäkinen 2013a):

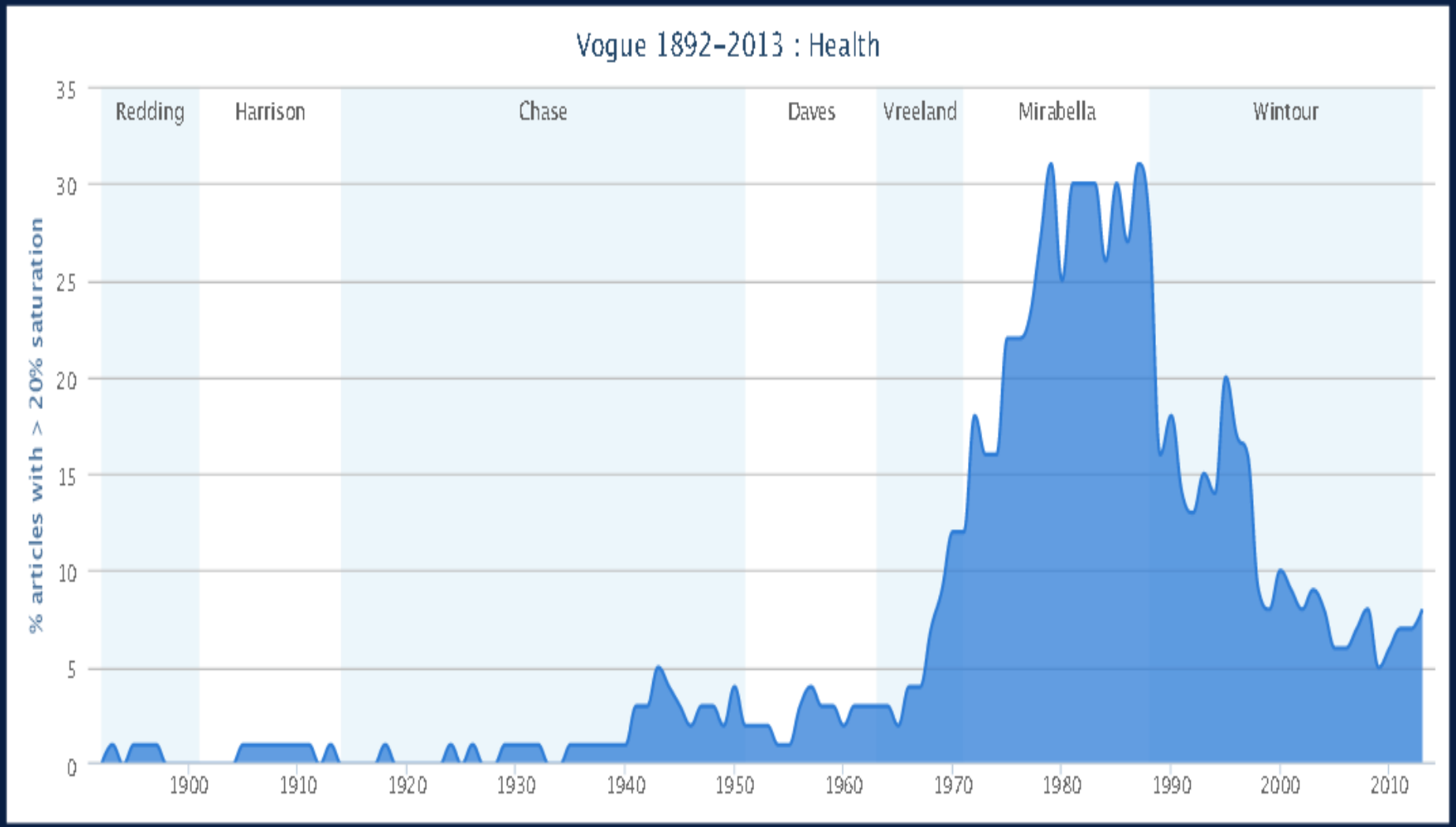


Lukemiseen, erityisesti lukuhaluun liittyvien fraasien käytön kasvu alkaa Euroopan voimakkailta kieli- ja kirjallisuusalueilta. Suurten kielialueiden kehitystä kuvaavat Google Books -aineiston perusteella laatimani kuviot keskeisimpien lukuhalua tarkoittavien fraasien esiintymisestä (absoluuttisia lukuja) (tiedot artikkelista Mäkinen 2013a):

Tutkijan datan louhinta, osa 2

- Itsejärjestäytyvät tekniikat
- Avaavat jotakin, mitä ihminen ei itse pysty hahmottamaan suurista datamassoista
- Esimerkkinä Peter Leonardin Vogue lehden analyysi naisten terveyteen liittyen
- http://www.ifla.org/files/assets/academic-and-research-libraries/publications/2014-08_ifla-reduced.pdf

ITÄ-SUOMEN YLIOPISTO



http://www.ifla.org/files/assets/academic-and-research-libraries/publications/2014-08_ifla-reduced.pdf

Kirjastojen ja kirjastoaineistojen datan louhinta

- Visualisointitekniikat:
- Tiedonhaun tägi/asiasanapilvet
- Tilasto/kokoelmatietojen visuaalinen hahmottaminen tiedonhakua varten
- Esimerkkinä Microsoftin Academic Search

ITÄ-SUOMEN YLIOPISTO

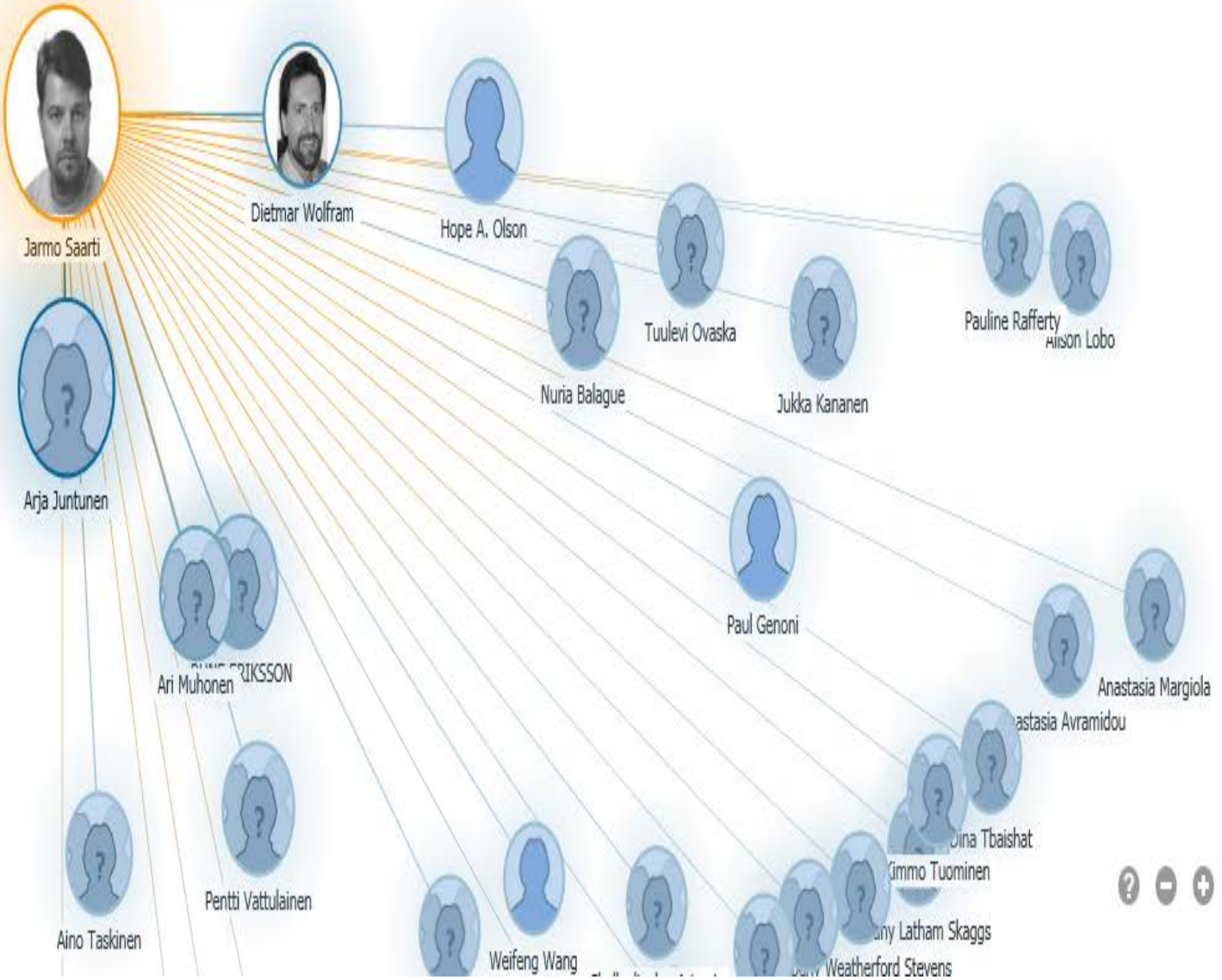
jarmo saarti

Result

Jarmo Saarti
University of Eas...

Jarmo Saarti

Co-author Graph | Co-author Path | Citation Graph



Kirjastot ja datamining - johtopäätöksiä

- Datamining on jo täällä
- Sitä voidaan hyödyntää monella tavalla tutkijan ja kirjaston arkea helpottamaan
- Kirjastoille ja akateemiselle tutkimukselle on taattava datan louhinnan oikeus tutkimus- ja opetustarkoitukseen
- Datan louhinta on rinnastettava nykyisiin kirjastoaineistojen lukemisen tapoihin

ITÄ-SUOMEN YLIOPISTO

Kiitokset,
kysymykset nyt tai
jarmo.saarti@uef.fi

ITÄ-SUOMEN
YLIOPISTO

www.uef.fi