

RAkenteellisen
KEhittämisen
Tukena
Tietohallinto

Korkeakoulujen ja
opetusministeriön
yhteinen
tietohallintohanke,
jota CSC koordinoi



Thomson Reutersin julkaisudatan suomalainen
aineisto

Yrjö Leino, CSC – Tieteen tietotekniikan keskus



Thomson Reutersin datan käsittely



- Mikä Thomson Reuters ja mikä data?
- Raakadatan rakenne ja osoitteiden määräytyminen; ongelmat
- Osoitteiden analysointi ja luokittelu
- Analyysin heikkoudet
- Jatkotyö

Thomson Reutersin data



- Kansainvälinen informaatioalan yritys.
- Kerää ja jalostaa tietoa eri aloilta (talous, terveydenhuolto, tiede, uutismaa, laki).
- Myy sekä raakadataa että jalostettua tietoa.
- Tieteellisten julkaisujen tietokannat koottu myös selattaviksi verkossa.
- Suomessa useiden yliopistojen ja esimerkiksi Suomen Akatemian käytössä.
- Opetusministeriö: tutkimuksen laadunarvioinnin tueksi Thomson Reutersin keräämään dataan pohjautuva bibliometrinen analyysi. Arviointi tapahtuu **tieteenaloittain** ja **organisaatiokohtaisesti**.

Thomson Reutersin data



1140 Fysiikka

Yliopistot	Vuosi	2003	2004	2005	2006	2007	2008
Åbo Akademi		34	31	15	17	20	21
Helsingin yliopisto		301	324	338	340	371	273
Joensuun yliopisto		41	71	33	65	47	60
Jyväskylän yliopisto		166	157	166	164	234	189
Kuopion yliopisto		13	24	20	20	32	25
Lappeenrannan teknillinen yliopisto		21	12	26	28	21	24
Oulun yliopisto		74	82	88	100	97	92
Sibelius-akatemia		0	0	0	1	0	0
Teknillinen korkeakoulu		374	399	331	388	373	360
Tampereen teknillinen yliopisto		69	80	68	92	81	95
Turun yliopisto		49	57	57	70	79	67
Vaasan yliopisto		3	5	1	1	0	2



Thomson Reuters raakadata

- Kerätty alkuperäisistä julkaisuista puoliautomaattisesti.
- Kattaa yli 15 000 julkaisusarjaa, jotka luokiteltu n. 250 tieteenalaan.
- Jakautuu useampiin luokkiin:
 - tieteelliset jurnaalit
 - konferenssijulkaisut
 - täydennystiedostot
 - korjaustiedostot

Opetusministeriön koekäyttöön saatu (Thomson Reutersin luvalla) Aalto-yliopiston hankkima raakadata vuosilta 2003-2009.

Thomson Reutersin raakadata



- Tieteelliset jurnaalit: 4 tiedostoa/vuosi, jokaisessa
n. 15 000 – 22 000 lehden numeroa,
n. 300 000 - 400 000 artikkelia ja
n. 70 – 115 miljoonaa riviä
- Artikkeleista talletettu mm.
 - Otsikko
 - Kirjoittajat
 - Osoitteet
 - Organisaatiot (**määräytyy osoitteesta !**)
 - Tiivistelmä
 - Viittaukset

Thomson Reutersin raakadatta



Raakadatatiedoston rakenne

Otsikkotiedot, tiedosto

Otsikkotiedot, lehti 1

Artikkelin 1 tiedot

Artikkelin 2 tiedot

Artikkelin 3 tiedot

Otsikkotiedot, lehti 2

Artikkelin1 tiedot

Artikkelin 2 tiedot

Artikkelin 3 tiedot

Thomson Reutersin raakadata



Raakadatatiedoston rakenne rivitasolla: kaksikirjaiminen kentän tunnus, sen jälkeen varsinainen data

TI Timing and location of phenomena during auroral breakup: A case study

AU Yahnin, AG

AU Pulkkinen, TI

NF Finnish Meteorol Inst, FIN-00101 Helsinki, Finland ← Osoitekenttä

NC Finnish Meteorol Inst ← Organisaatio suoraan osoitekentästä!

NY Helsinki

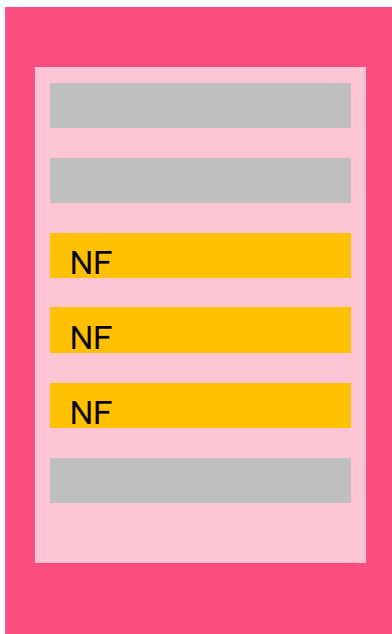
NU Finland

NZ FIN-00101

Thomson Reutersin raakadata



Artikkelissa useita osoitekenttiä, jokaisessa mahdollisesti useita osoitteita:



NF Univ 1, Dept, Univ 2, Dept

NF Univ 1, Dept, Company A

NF Company A

Thomson Reutersin raakadata



Mahdollisia ongelmia:

1) osoitteen ensimmäinen osa ei olekaan varsinainen organisaatio

NF **Matemat Tilastotieteen Laitos**, PL 68, FI-00014 Helsinki, Finland

NC **Matemat Tilastotieteen Laitos**

2) osoitteen siirtäminen TR:n tietoihin epäonnistunut

NF **Temple Univ**, Dept Comp Sci, FIN-33014 Tampere, Finland

NC **Temple Univ**

3) Annettu pelkkä katuosoite

NF **Mailante 109**, FI-08800 Lohja, Finland

NC **Mailantie 109**

Thomson Reutersin raakadata



Suomessa tuotettuja artikkeleita aineistossa yli 60 000 kpl.
Yhteensä yli 50 000 erilaista osoitetta, joista **kymmenesosa**
virheellisiä tai puutteellisia.



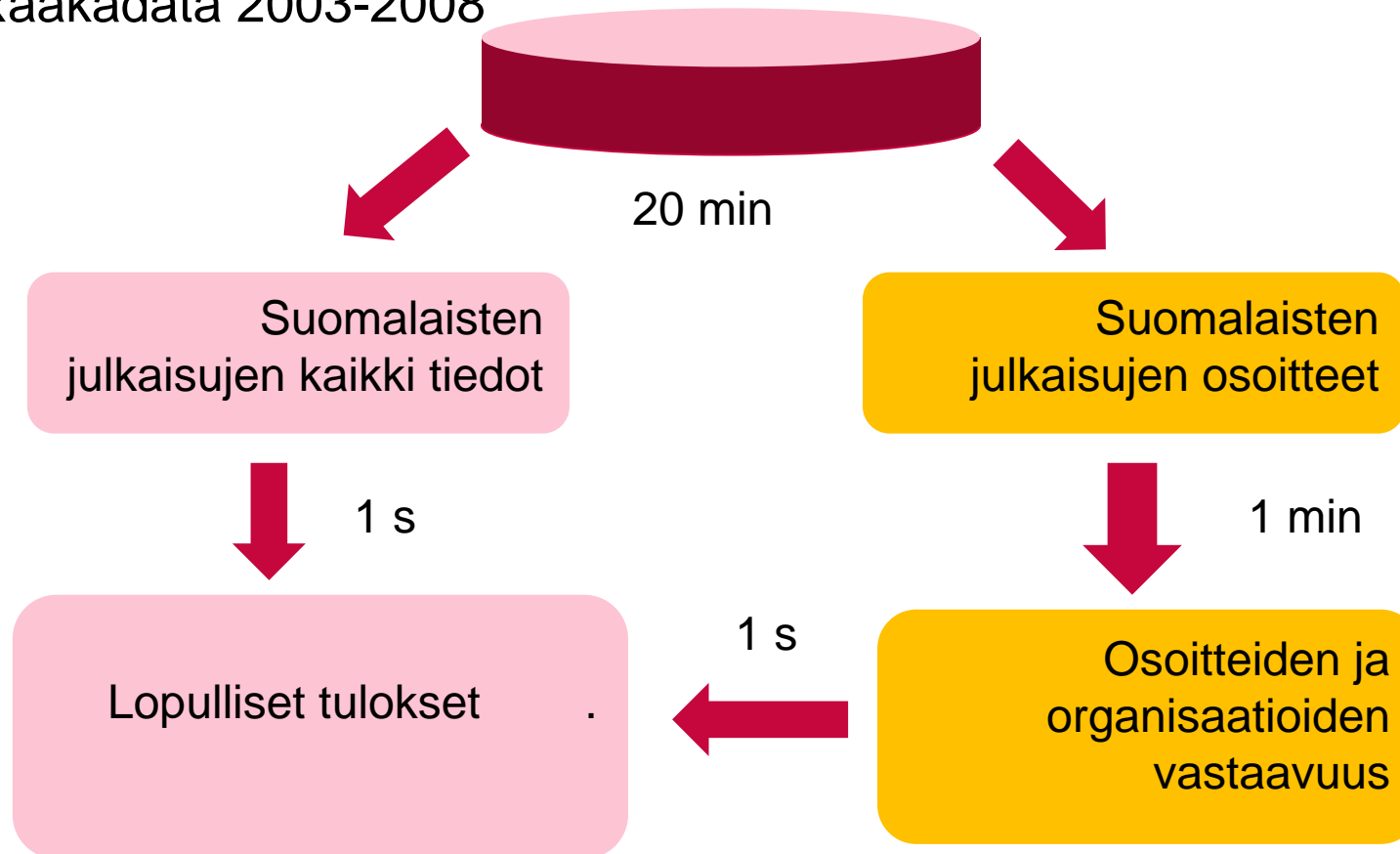
Virheelliset osoitteet analysoitava ja yhdistettävä
oikeisiin organisaatioihin ennen kuin tuloksia
lasketaan.

Thomson Reutersin datan käsittely



RAKETTI

Raakadata 2003-2008



Thomson Reutersin datan käsittely



- Organisaatioiden ja osoitteiden välinen yhteys pyritty selvittämään päättelysäännöillä.
 - Tavoitteena löytää pienehkö joukko sääntöjä, joilla saadaan iso joukko osoitteita luokiteltua oikein.
 - Sääntöjä täydennettävä vuosittain.
-
- Selvät organisaationimet: Univ Kuopio, Finnish Def Force
 - Vakiintuneet lyhenteet: TKK, VTT, THL
 - Erilliset laitokset: Tvarminne Zool Stat, Tuorla Observ
 - Postinumerot (ehdollisesti): FI-02015, FI-00014

Thomson Reutersin datan käsittely



Mahdollisia ongelmia:

– Kirjoitusvirheet:

Lyvaskyla, Jyvskyl, Jyvaskyl, Jyvaskala, Jyvaeskylae,
Jynaskyla, Jyaskyla, Jvaskyla, Juvascyla, Javaskyla,
Hyvaskayla, Jyuvaskula, Jyavaskyla,....

– Kirjavuus organisaation nimen kirjoitusasussa:

Helsinki Univ Cent Hosp,	Univ Helsinki Hosp,
Univ Helsinki Cent Hosp,	Helsinki Cent Univ Hosp,
Cent Univ Hosp Helsinki,	Univ Helsinki, Dept, Cent Hosp,
Univ Cent Hosp Helsinki	+ muunnelmat Ctr, Central

Thomson Reutersin datan käsittely



- **Organisaation nimen tai osoitteen muutokset:**

Sci Comp Ltd, Ctr Comp Sci, Finnish IT Ctr Sci,
Ctr High Performance Comp & Networking,
Tieteellinen laskenta Oy

- **Organisaatioiden yhdistely ja erottelu:**

THL = KTL + Stakes + Finoha + Alkoholitutkimussäätiö + ...

- **Yhteiset tutkimuslaitokset ja instituutit:** Biomedicum,
Helsinki Institute of Physics,...

Thomson Reutersin datan käsittely



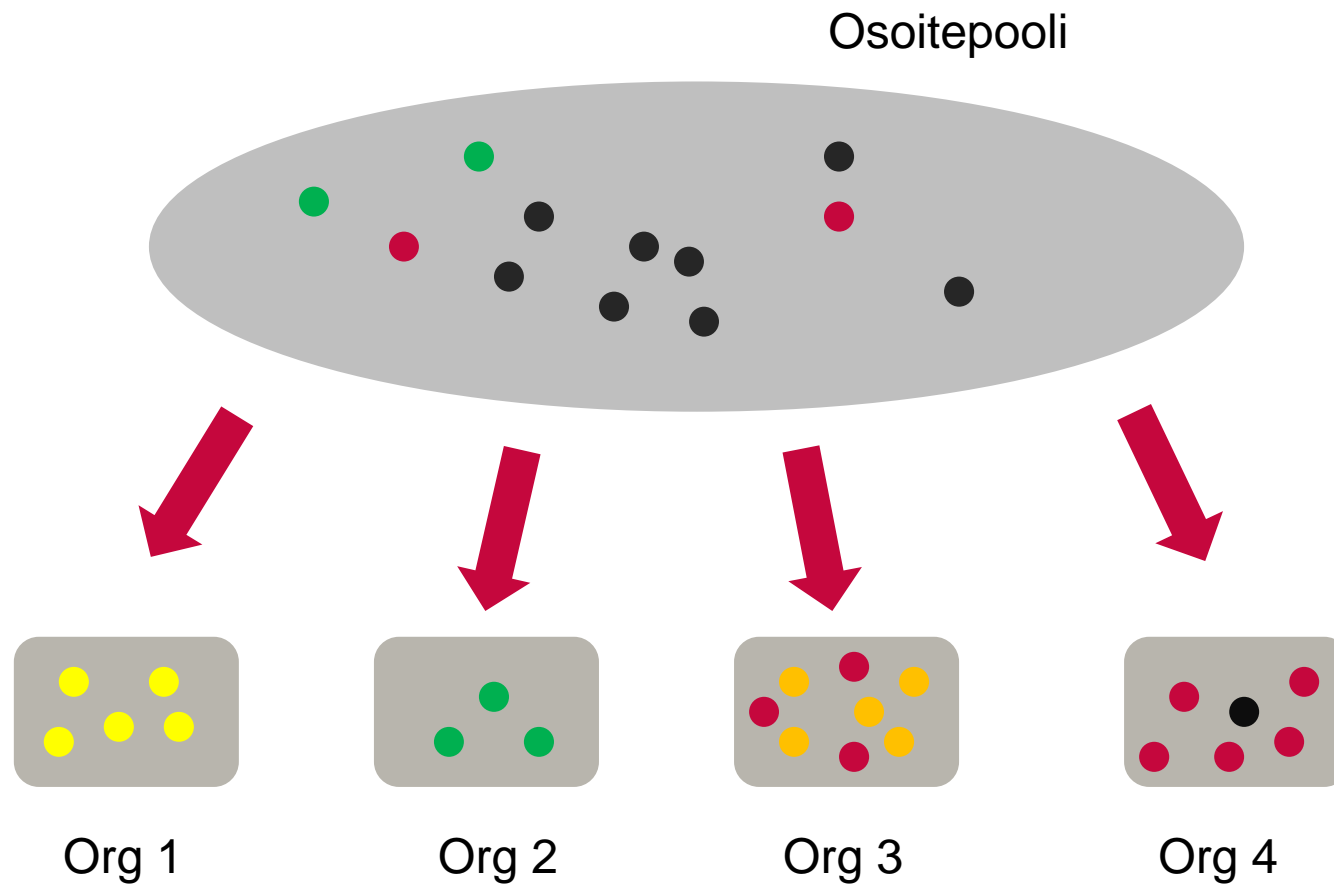
- **Nimet tai lyhenteet, joilla useita merkityksiä:**
 - Nokia, Outokumpu (paikkakunta / yritys)
 - TUCS (Turku Centre for Computer Science / Tammerfors universitets centralsjukhus)
- **Nimet, jotka ovat sisältävät toisten organisaatioiden nimiä:**

Helsinki University of Technology

Thomson Reutersin datan käsittely



Kahden tyyppin virheitä



Thomson Reutersin datan käsittely



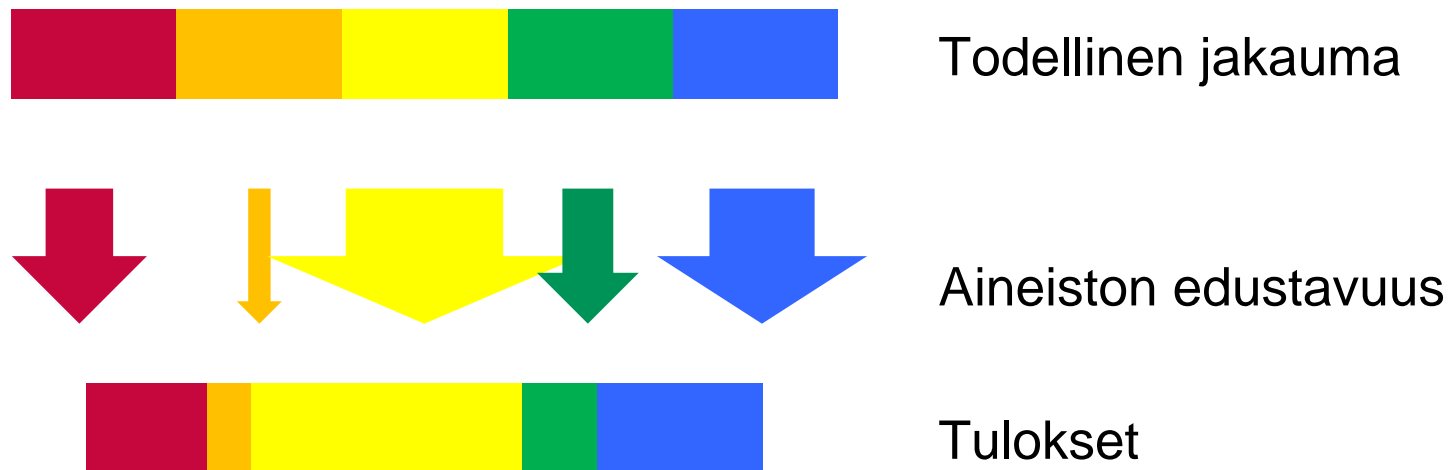
Tilanne luokittelun suhteen

- Kaikkiaan aineistossa 61512 suomalaista artikkelia, joissa n. 51 000 erilaista suomalaista osoitetta.
- Uusia suomalaisia osoitteita 7000 – 8000 vuosittain.
- Päätteily sääntöjä käytössä n. 1300
- Ilman luokiteltua organisaatiota jää alle 500 artikkelia
- Katuosoitteet, yhdistykset, virastot, pienyritykset, ulkomaalaiset osoitteet (väärä maakoodi), puutteelliset osoitteet, ristiriitaiset osoitteet,...

Thomson Reutersin data

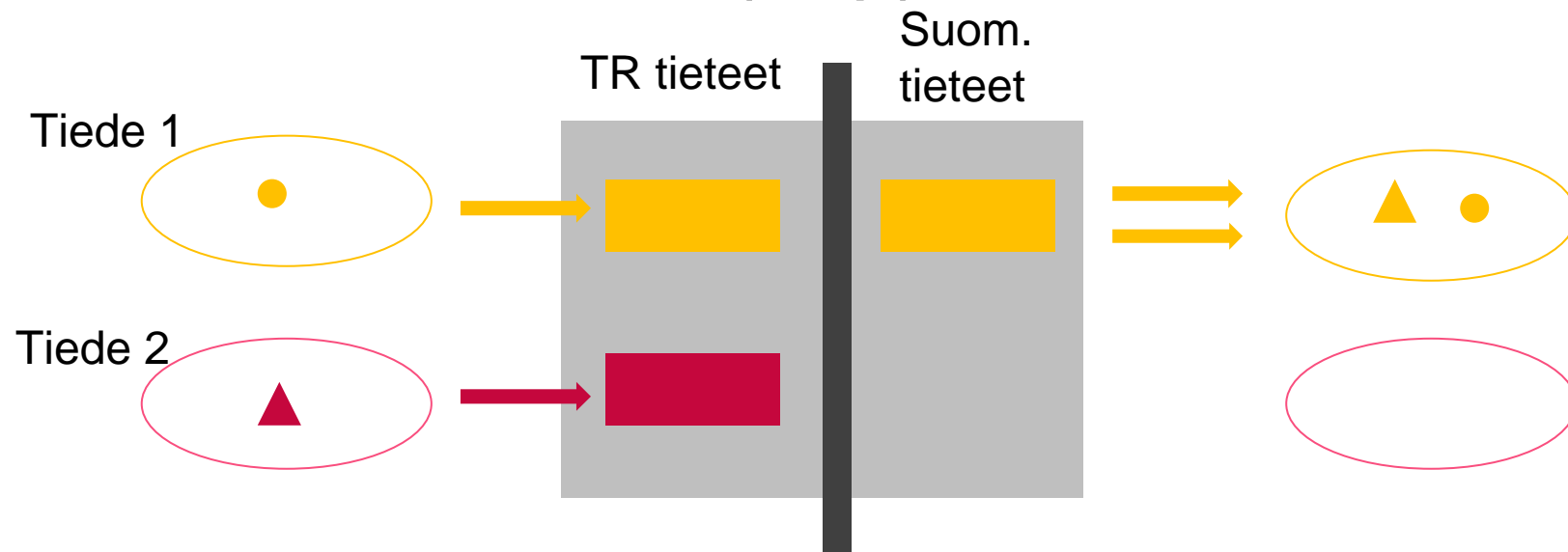
Tulosten luotettavuuteen vaikuttavia seikkoja:

1. Eri tieteenalojen erilainen edustavuus raakadatassa



Thomson Reutersin data

2. Thomson Reutersin tieteenalat (250 kpl) on määritelty julkaisusarjoittain. Vain yksi ala/lehti otettu huomioon, ja se on luokiteltu aina kokonaisuudessaan yhden suomalaisen tieteenalan (66 kpl) alle.



Jatko



Seuraavaksi tehtävä bibliometrinen analyysi. Pohdittavia kysymyksiä (kyllä, näitä on varmasti pohdittu aiemminkin ja myös muualla kuin Suomessa):

- Mitä on laadukas tutkimus? Onko laadussa eri dimensioita?
- Ovatko laadun dimensiot mitattavissa bibliometrisillä mittareilla, ja jos, niin millä?
- Tieteenalakohtaiset erot julkaisutavoissa ja viittauskäytännöissä?
- Oikeat aikaskaalat?