Fennican RDF-konversio ja teosten eristäminen

Osma Suominen 1.6.2016

Viime kerralla päätettiin yrittää omatoimisesti eristää teoksia Fennicasta

FRBR Work-Set Algorithm

- OCLC:n julkaisema v. 2009 (versio 2.0)
- Olennaisesti PDF, jossa kuvataan tarkasti, miten joukosta bib-tietueita saadaan lähinnä nimekkeitä vertaamalla irti teokset

“The research work-set algorithm generates a key for each bibliographic record. These FRBR keys can then be used to bring work-sets together. The current algorithm ignores format so that the generated work-sets are sometimes at a higher level than a FRBR work.

A work-set is a group of bibliographic records having the same FRBR key, generated according to the algorithm in this paper.

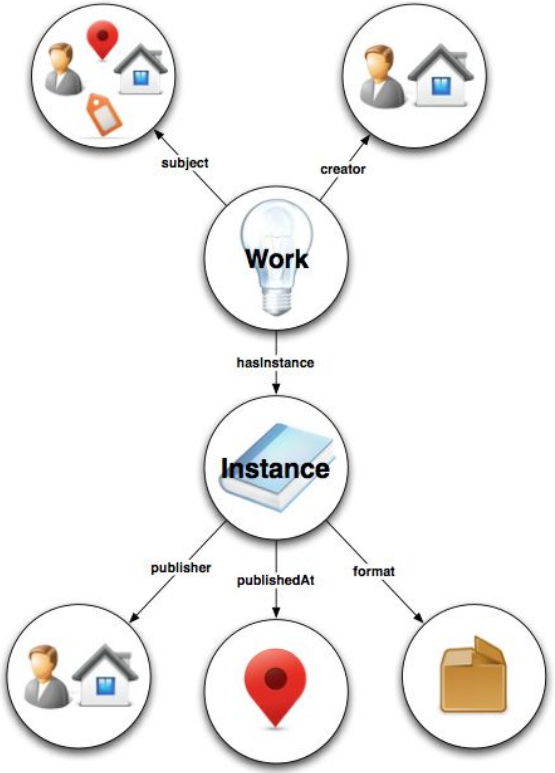
Authors and titles that match variant headings in the mapping files are changed to their preferred form. This means that building the mapping files is a prerequisite for building FRBR keys.”

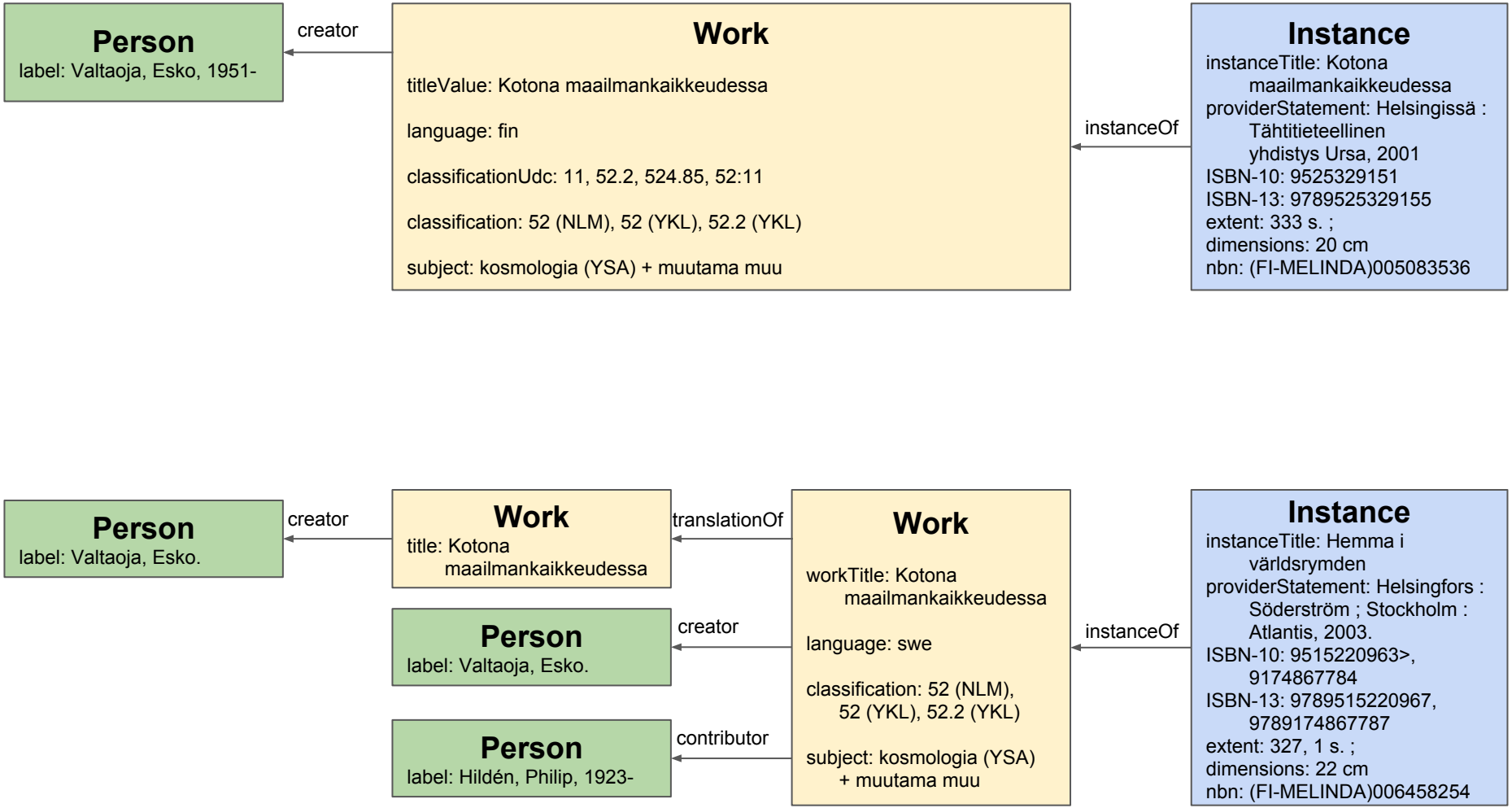
LOC BIBFRAME 1.0-muunnin marc2bibframe

- Pilkkoo bib-tietueen BIBFRAME-mallin mukaisesti: Work / Instance

- Work’eja tulee usein useita per bib-tietue
  - FRBR teos ja ekspressio erikseen Workeja
  - sarjakin on Work, esim. “WSOY pokkari”

- Work’eille lasketaan avainmerkkijono, joka vastaa suunnilleen Work-Set algoritmin tuottamaa avainta
  - ei kuitenkaan kaikille Work’eille, vaan ilmeisesti vain ekspressiotason Work’eille tai sellaisen puuttuessa teostaso-Workille
  - **ei tällaisena kelpaa meille, tehdään oma**




Person
label: Valtaoja, Esko, 1951-
creator
Work
titleValue: Kotona maailmankaikkeudessa
language: fin
classificationUdc: 11, 52.2, 524.85, 52:11
classification: 52 (NLM), 52 (YKL), 52.2 (YKL)
subject: kosmologia (YSA) + muutama muu
instanceOf
Instance
instanceTitle: Kotona maailmankaikkeudessa
providerStatement: Helsingissä : Tähtitieteellinen yhdistys Ursa, 2001
ISBN-10: 9525329151
ISBN-13: 9789525329155
extent: 333 s. ;
dimensions: 20 cm
nbn: (FI-MELINDA)005083536
Person
label: Valtaoja, Esko.
creator
Work
title: Kotona maailmankaikkeudessa
translationOf
Work
workTitle: Kotona maailmankaikkeudessa
language: swe
classification: 52 (NLM), 52 (YKL), 52.2 (YKL)
subject: kosmologia (YSA) + muutama muu
instanceOf
Instance
instanceTitle: Hemma i världsrymden
providerStatement: Helsingfors : Söderström ; Stockholm : Atlantis, 2003.
ISBN-10: 9515220963>, 9174867784
ISBN-13: 9789515220967, 9789174867787
extent: 327, 1 s. ;
dimensions: 22 cm
nbn: (FI-MELINDA)006458254
Person
label: Valtaoja, Esko.
creator
Person
label: Hildén, Philip, 1923-
contributor

Teosavainten luonti ~ FRBR Work-Set algorithm

Tietueelle luodaan yksi/useampi avain ensimmäisellä mahdollisella tavalla:

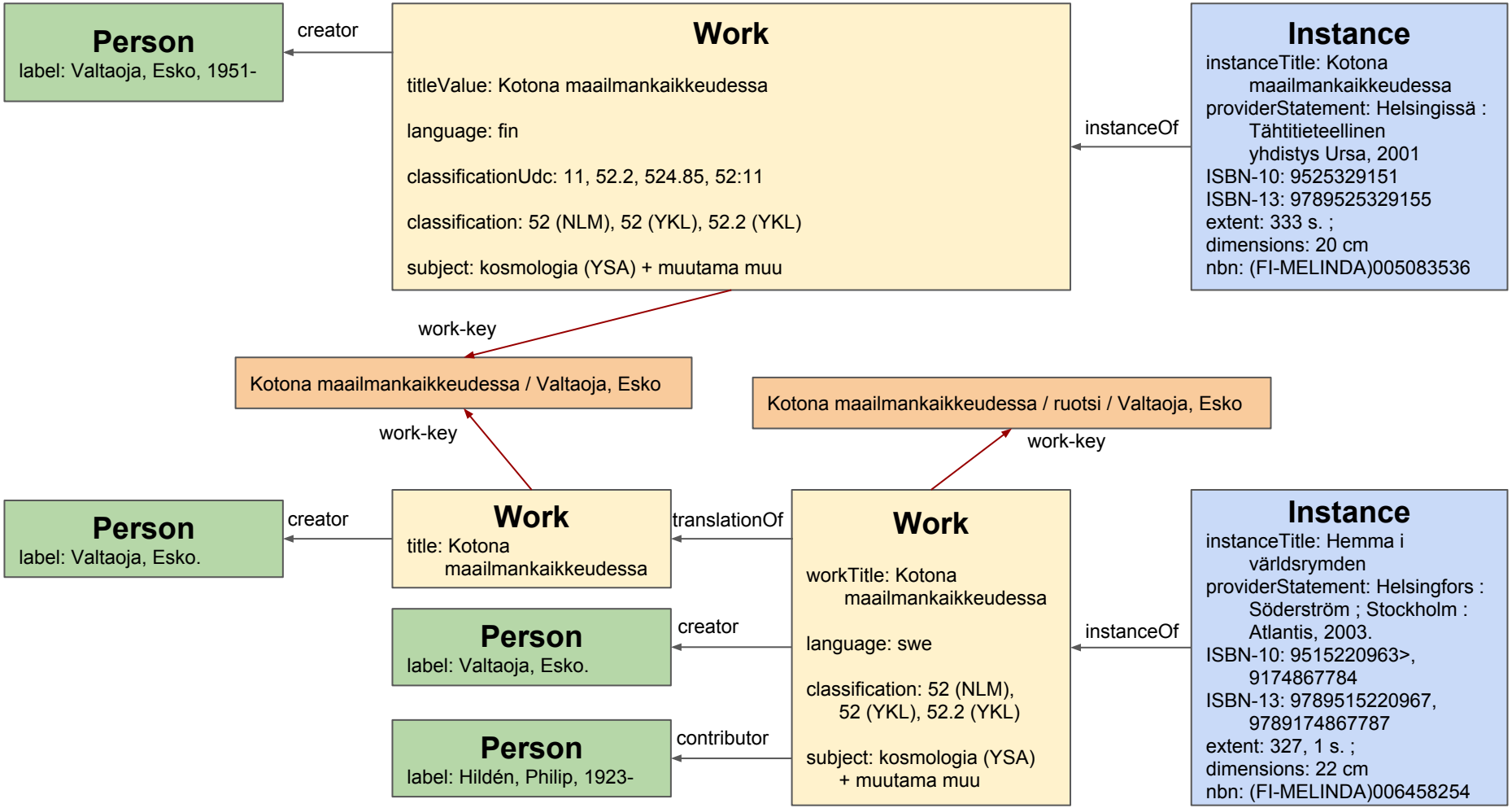
1. Jos on otsikko (24x) ja tekijä (100/110/111), käytetään näiden yhdistelmää
   a. esim. “Kotona maailmankaikkeudessa / Valtaoja, Esko”
   b. jos kysymys käännöksestä, lisätään otsikkoon myös kohdekieli

2. Jos on yhtenäistetty nimeke (130), käytetään tätä:
   a. esim. “Raamattu. Uusi testamentti”

3. Jos on sekä otsikko (24x) että lisäkirjauksina tekijöitä (70x/71x), käytetään näiden yhdistelmiä:
   a. esim. “Maastokartta : peruskartta 1:20000 / Maanmittauslaitos”

4. Jos mikään ylläoleva ei onnistu, tietue muodostaa yksinään teoksen

Teos muodostuu joukosta tietueita, joilla on päällekkäisiä teosavaimia


Person
label: Valtaoja, Esko, 1951-
creator
Work
titleValue: Kotona maailmankaikkeudessa
language: fin
classificationUdc: 11, 52.2, 524.85, 52:11
classification: 52 (NLM), 52 (YKL), 52.2 (YKL)
subject: kosmologia (YSA) + muutama muu
instanceOf
Instance
instanceTitle: Kotona maailmankaikkeudessa
providerStatement: Helsingissä : Tähtitieteellinen yhdistys Ursa, 2001
ISBN-10: 9525329151
ISBN-13: 9789525329155
extent: 333 s. ;
dimensions: 20 cm
nbn: (FI-MELINDA)005083536
work-key
Kotona maailmankaikkeudessa / Valtaoja, Esko
work-key
Kotona maailmankaikkeudessa / ruotsi / Valtaoja, Esko
work-key
Person
label: Valtaoja, Esko.
creator
Work
title: Kotona maailmankaikkeudessa
translationOf
Work
workTitle: Kotona maailmankaikkeudessa
language: swe
classification: 52 (NLM), 52 (YKL), 52.2 (YKL)
subject: kosmologia (YSA) + muutama muu
instanceOf
Instance
instanceTitle: Hemma i världsrymden
providerStatement: Helsingfors : Söderström ; Stockholm : Atlantis, 2003.
ISBN-10: 9515220963>, 9174867784
ISBN-13: 9789515220967, 9789174867787
extent: 327, 1 s. ;
dimensions: 22 cm
nbn: (FI-MELINDA)006458254
Person
label: Valtaoja, Esko.
creator
Person
label: Hildén, Philip, 1923-
contributor

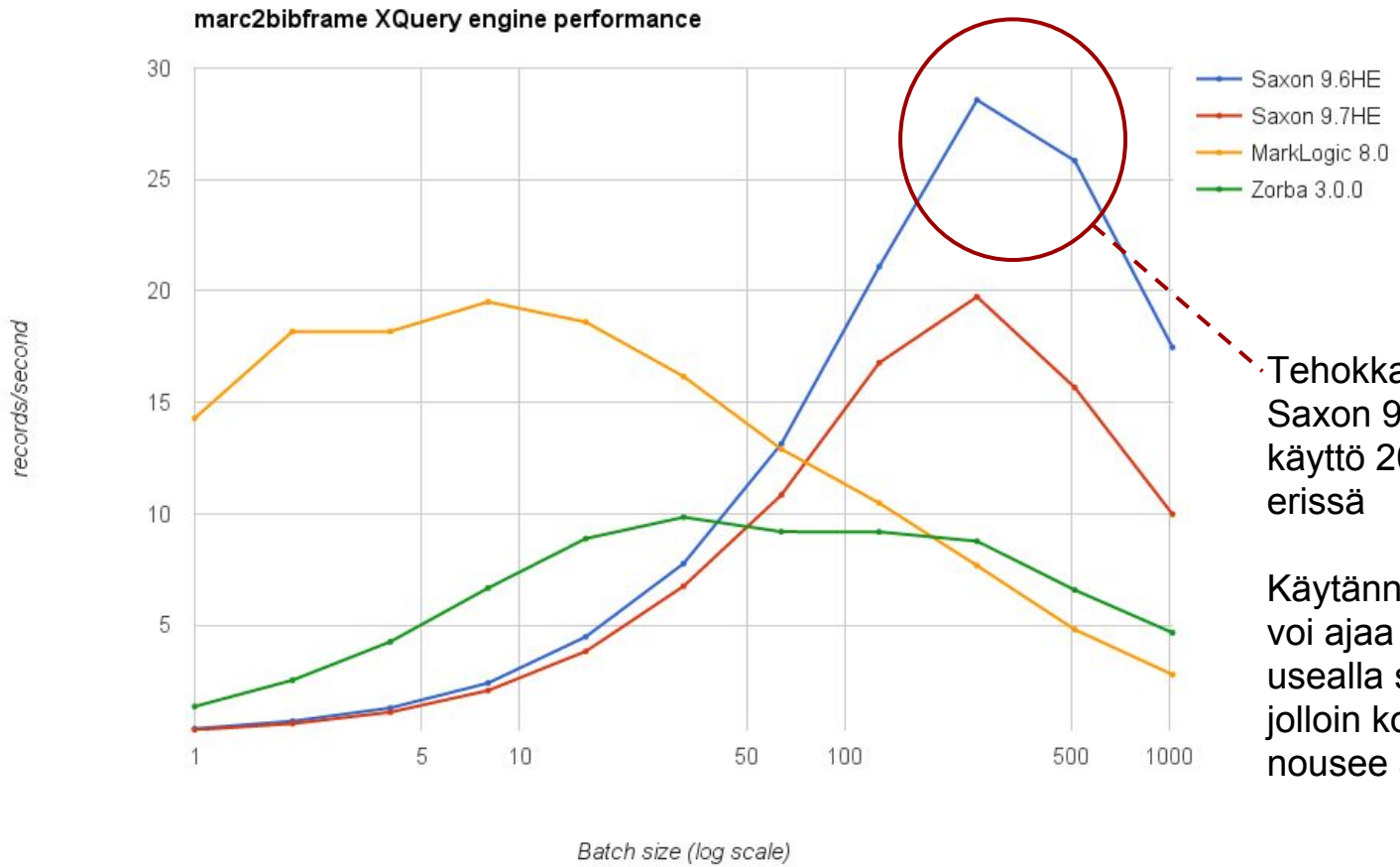
Top 20 teokset Fennicassa (alustava)

1. Maastokartta : peruskartta 1:20000 / Maanmittauslaitos (2769)
2. Peruskartta 1:25000 / Maanmittauslaitos (678)
3. Raamattu (459)
4. Maastokartta 1:50000 / Maanmittauslaitos (365)
5. Kalevala (362)
6. Meditationes sancti evangelii / Petraeus, Aeschillus Olai (199)
7. Sinuhe egyptiläinen / Waltari, Mika (173)
8. Raamattu. Uusi testamentti (167)
9. Seitsemän veljestä / Kivi, Aleksis (139)
10. Meditationes sanctarum epistolarum / Petraeus, Aeschillus Olai (131)
11. Chronicon episcoporum Finlandensium / latina / Porthan, Henrik Gabriel (113)
12. Meditationes sancti evangelii / latina / Petraeus, Aeschillus Olai (97)
13. Teknillisen alan opetussuunnitelmatoimikunnan mietintö / Teknillisen alan opetussuunnitelmatoimikunta (97)
14. Peittoalue Suomessa / Karttakeskus (83)
15. Tuntematon sotilas / Linna, Väinö (83)
16. Merikartta. 18, Helsingin edusta. (74)
17. Merikartta. 29, Degerby-Berghamn. (71)
18. Merikartta. 21, Hanko-Jussarö. (70)
19. Raamattu. Valikoima (70)
20. Cars collection. (70)

991286 tietueesta muodostui 842159 teosta (vajaat 1,2 tietuetta per teos)

Koko Fennican BIBFRAME -konversio LOC:n marc2bibframe-konvertterilla

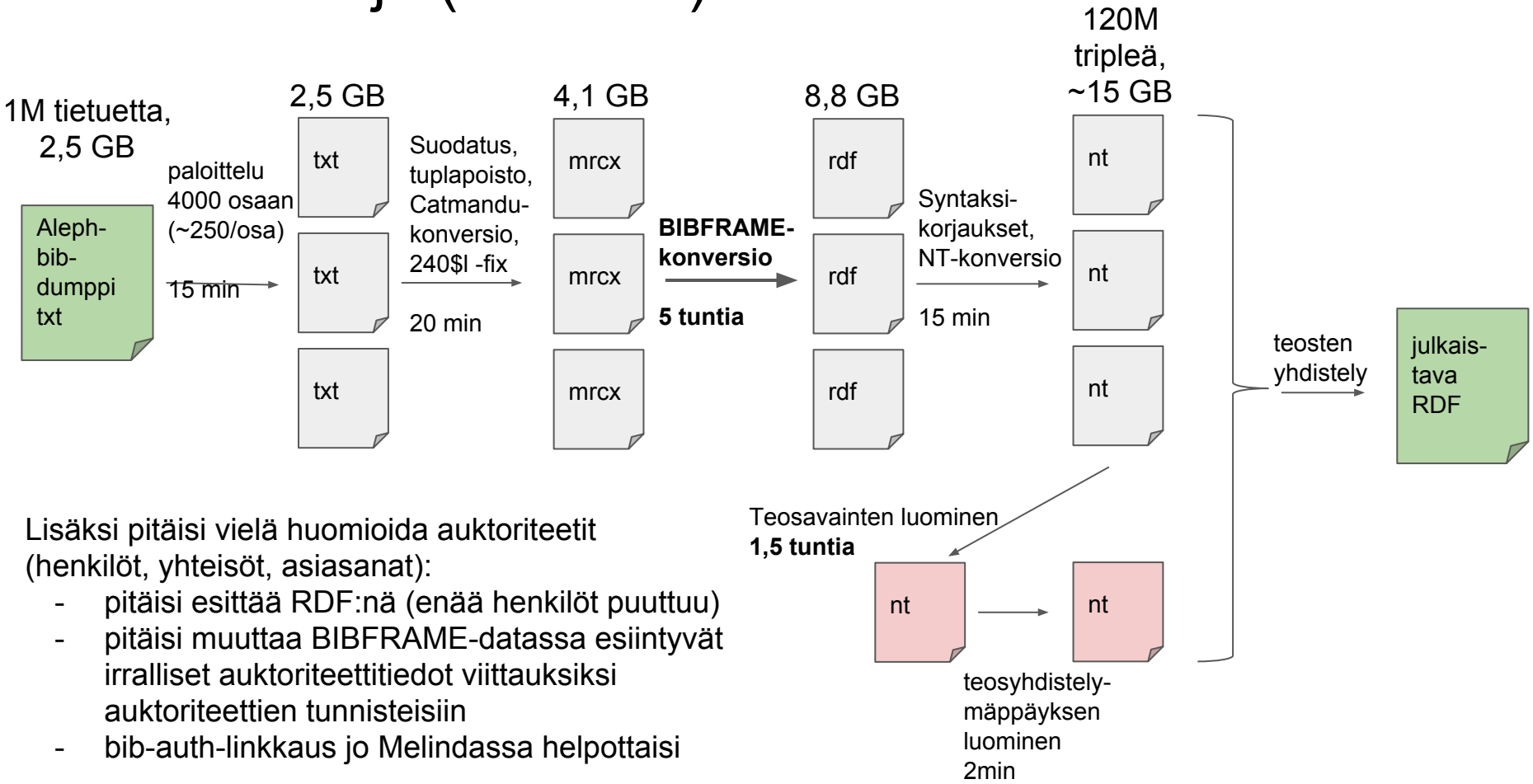
- Konvertteri vaatii MARCXML:ää ja tuottaa RDF/XML:ää
- Konvertteri “näkee” vain yhden tietueen kerrallaan, joten tulos on sama riippumatta siitä minkä kokoisissa erissä konversio tehdään
  - käytännössä konversio hidastuu ja muistinkäyttö nousee, jos tietueita on kerralla liikaa
- virheet lähtödatassa voivat rikkoa konversion
  - näiden selvittely on aikaavievää etsiväntyötä, virheilmoitukset ovat usein kehnoja
- marc2bibframe toteutettu XQuery-kyselyinä
  - voidaan ajaa eri XQuery-moottoreilla ja tuloksen pitäisi olla sama
  - standardinmukainen, mutta hitaanpuoleinen toteutustapa



Tehokkain yhdistelmä on Saxon 9.6HE-moottorin käyttö 200-500 tietueen erissä

Käytännössä muunnosta voi ajaa rinnakkain usealla suorittimella, jolloin kokonaisteho nousee 3-4-kertaiseksi

Muunnosketju (alustava)



Lisäksi pitäisi vielä huomioida auktoriteetit (henkilöt, yhteisöt, asiasanat):

- pitäisi esittää RDF:nä (enää henkilöt puuttuu)
- pitäisi muuttaa BIBFRAME-datassa esiintyvät irralliset auktoriteettitiedot viittauksiksi auktoriteettien tunnisteisiin
- bib-auth-linkkaus jo Melindassa helpottaisi

Ongelmia konversiossa

- jotkin kentät toistuvat, vaikka ei saisi (LDR, 001, 005, 008, 100/110/111, 245)
  - saattaa olla kysymys siitä että Melindassa on yhdistetty tietueita joita ei pitäisi
- roolikoodeissa käytetty pilkkuja 700$e-kentässä esim. “aut,”
- puuttuvia ja rikkinäisiä kielikoodeja, esim. “u”
- monenlaisia yksittäisiä tai enintään 2-3 kertaa esiintyviä virheitä
- myös marc2bibframessa bugeja, tuottaa joskus rikkinäistä RDF:ää

- raportoitu tietueongelmista Tutkiin ja niitä on korjattu Melindaan
- ongelmia edelleen esim. väärin yhdistyneet tietueet / toistuva 245
- koko ajan löytyy lisää ongelmatapauksia…
- noin 99,8% tietueista tähän mennessä konvertoitu onnistuneesti, selvittämättömiä ongelmia enää arviolta kymmenen tietueen kanssa

Seuraavat askeleet

- konversioprosessin siirto virtuaalikoneelle (linkeddata-kk)
- auktoriteettilinkityksen parantaminen, BIBFRAME-auktoriteettiviittausten korvaaminen järkevämmillä linkeillä esim. SKOS-käsitteisiin
- RDF-tietokantojen arviointi (Jena TDB hidastelee jo 120M triplellä)
- pohdinta siitä, julkaistaanko tässä muodossa (BIBFRAME 1.0 + omat säädöt) vai halutaanko jotain muuta, esim. RDA-tietomalli tai BIBFRAME 2.0
- olisi kiva, jos LOC julkaisisi BIBFRAME 2.0 muunnostyökalun
  - marc2bibframen kehitys hyytynyt loka-marraskuussa

Kiitos