HAKURATKAISUN ANATOMIA - KURKISTUS PELLIN ALLE

Jukka Huhta
Nikke Myöhänen
Ville Tenhunen

5.11.2014

HELSINGIN YLIOPISTO

AGENDA

MITÄ?
MIKSI?
ARKKITEHTUURI
KAHLAUS
INDEKSIT
INTEGRAATIOT
KÄYTTÖLIITYMÄT
RAUDAT
KÄYTTÖ NYT JA JATKOSSA
SEURAAVAKSI

HELSINGIN YLIOPISTO

MITÄ?

- Entinen www-hakukone Googlen Search Appliance vuodesta 2008
- Niiden lisenssit umpeutuivat loppusyksystä 2014
- Uudeksi hakukoneeksi avoimen lähdekoodin Apache Solr

HELSINGIN YLIOPISTO

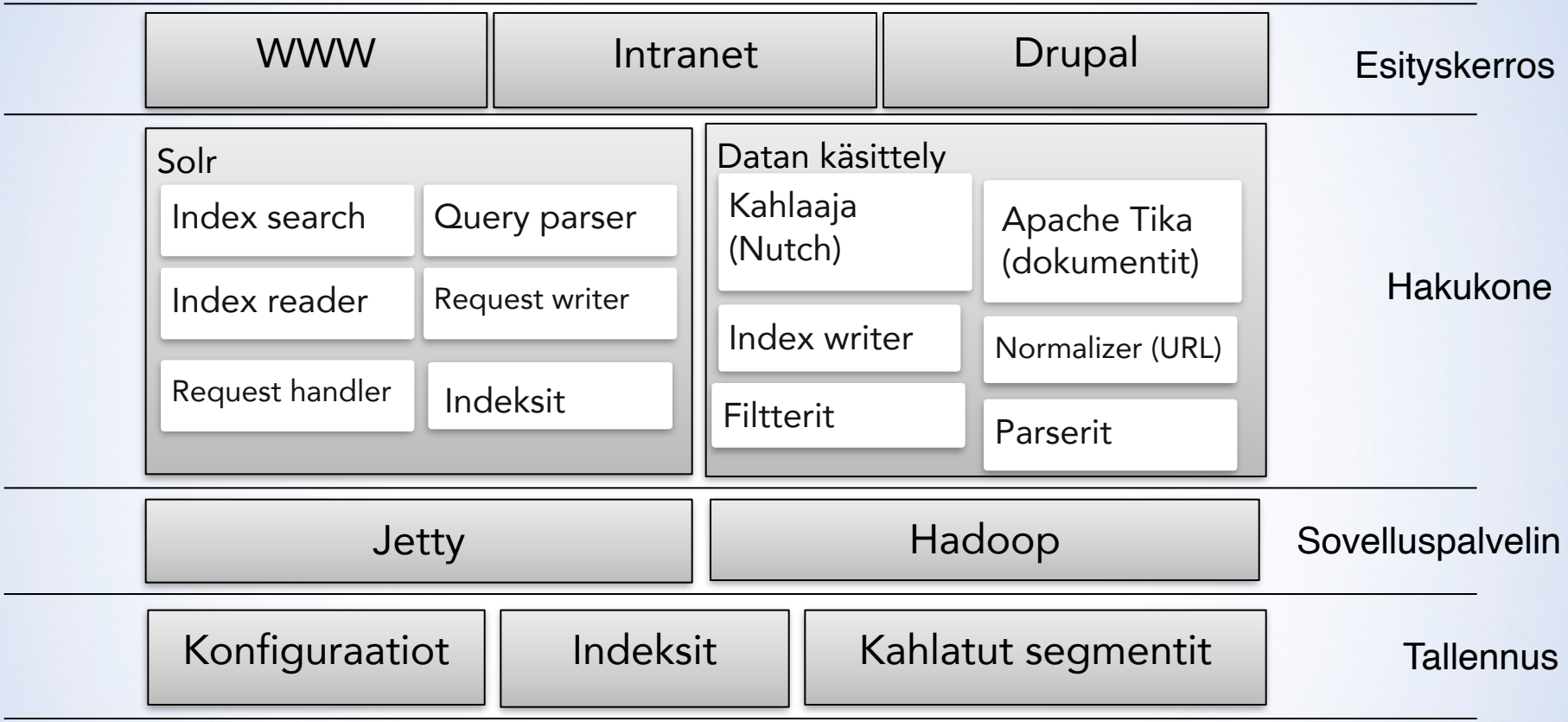
Apache Solr

MIKSI?

+ Skaalautuvuus suhteessa kustannuksiin
+ Mahdollisuudet vastata muuttuviin vaatimuksiin
+ Yhteentoimivuus
+ Muokattavuus, säädettävyys ja avoin lähdekoodi
+ Komponenttien saatavuus, monipuolisuus ja vaihtoehdot
+ Dokumentaatio ja yhteisötuki

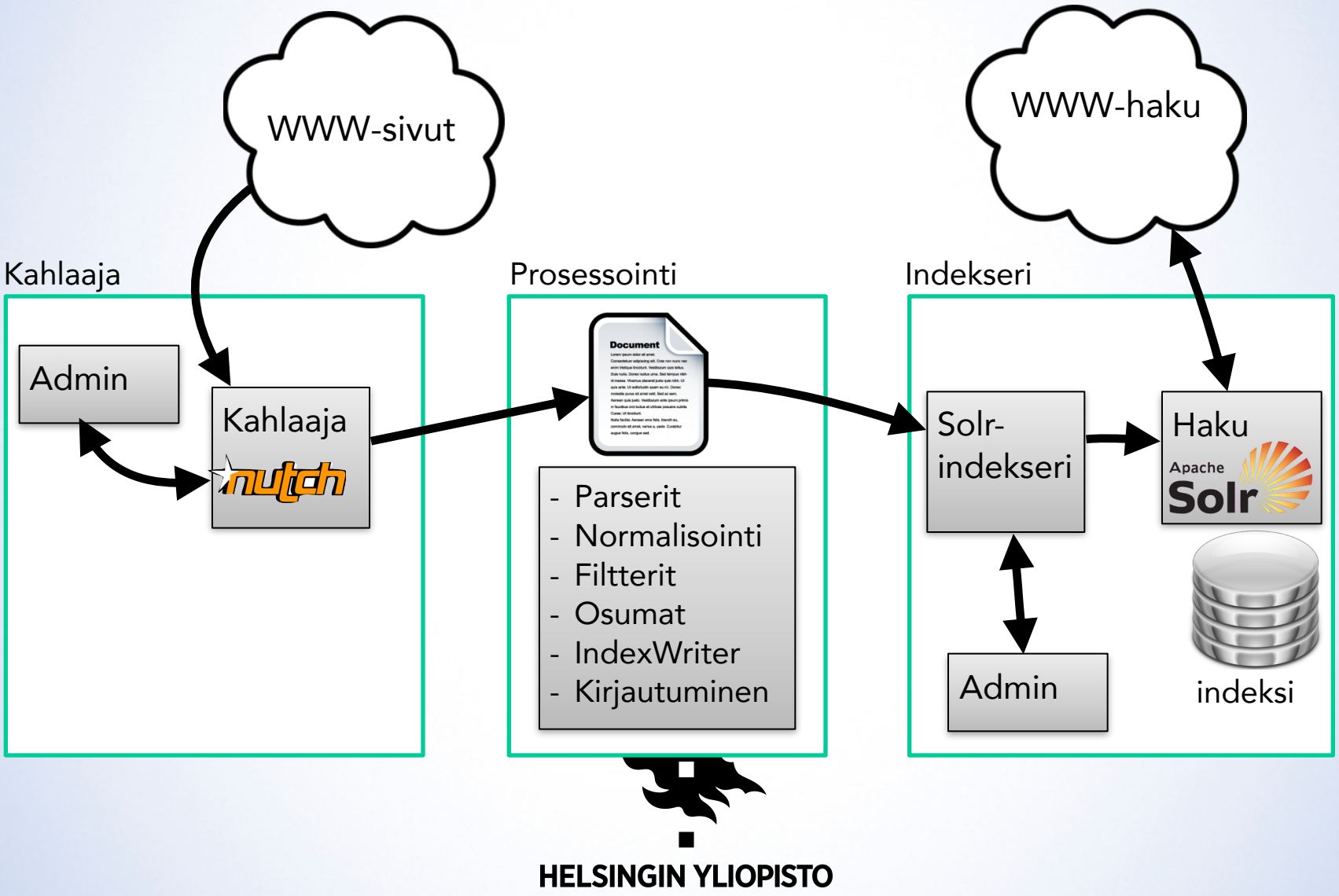
HELSINGIN YLIOPISTO

ARKKITEHTUURI



HELSINGIN YLIOPISTO

TOIMINTA

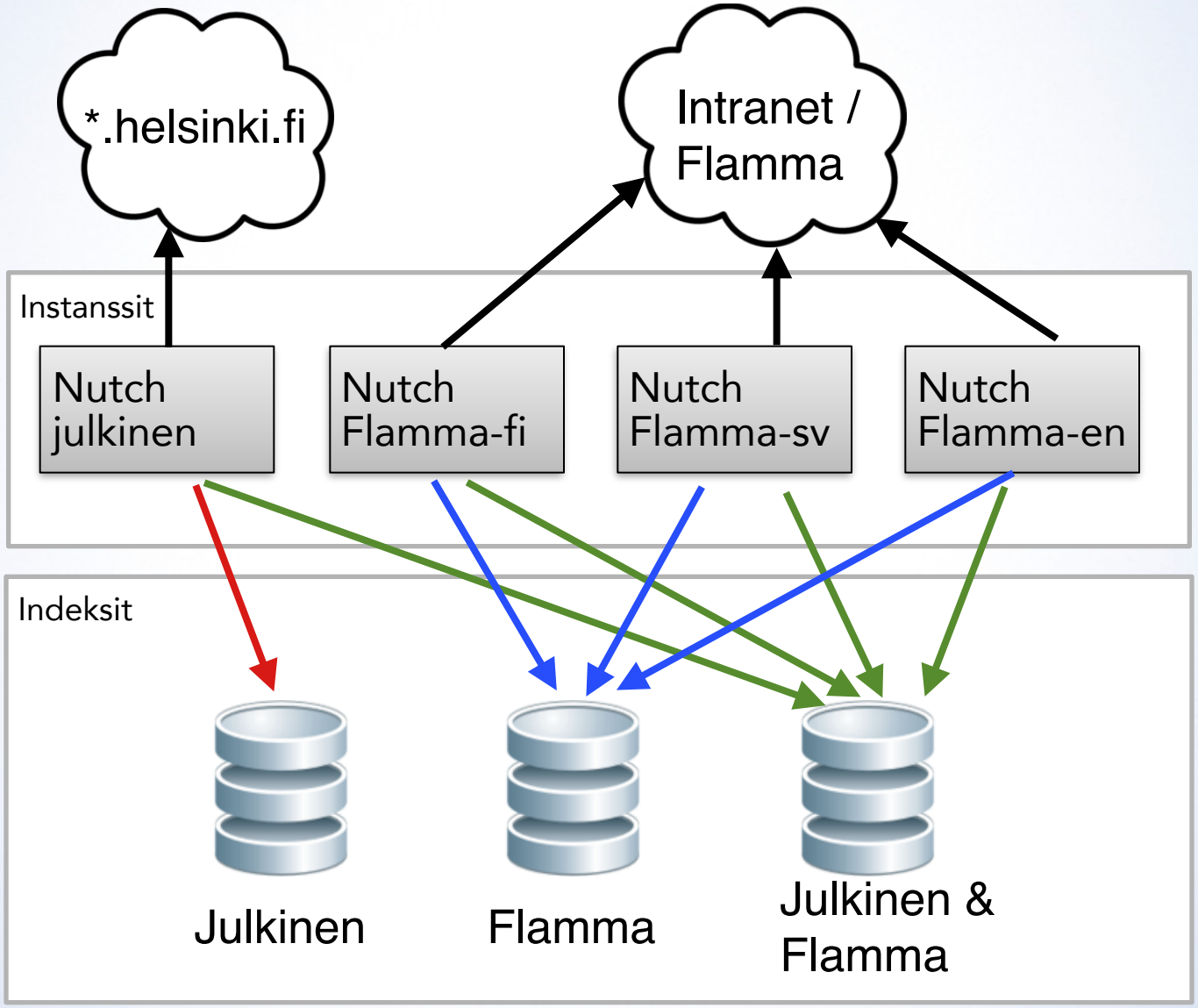

WWW-sivut
WWW-haku
Kahlaaja
Prosessointi
Indekseri
Admin
Kahlaaja
nutch
Document
- Parserit
- Normalisointi
- Filtterit
- Osumat
- IndexWriter
- Kirjautuminen
Solr-
indekseri
Haku
Apache
Solr
Admin
indeksi

KAHLAUS

- WWW-ilmentymien kahlauksesta vastaa Nutch (versio 1.9)
- Neljä instanssia
    - Helsinki.fi
    - Flamma-fi, Flamma-sv ja Flamma-en
- Kahlaaja seuraa annettua URLia etsien uusia linkkejä sivuilta. Hakua rajataan FQDN domain-osion perusteella
- Kahlaukset esim. 30 min välein, 2h välein ja 1 krt / päivä (syvyys vaihtelee). Syväkahlaus 1 krt / päivä.
- Hakutulokset talletetaan palvelimen levylle erillisiin segmentteihin, joista ne kirjoitetaan indekseihin.

HELSINGIN YLIOPISTO

KAHLAAJAT JA INDEKSIT

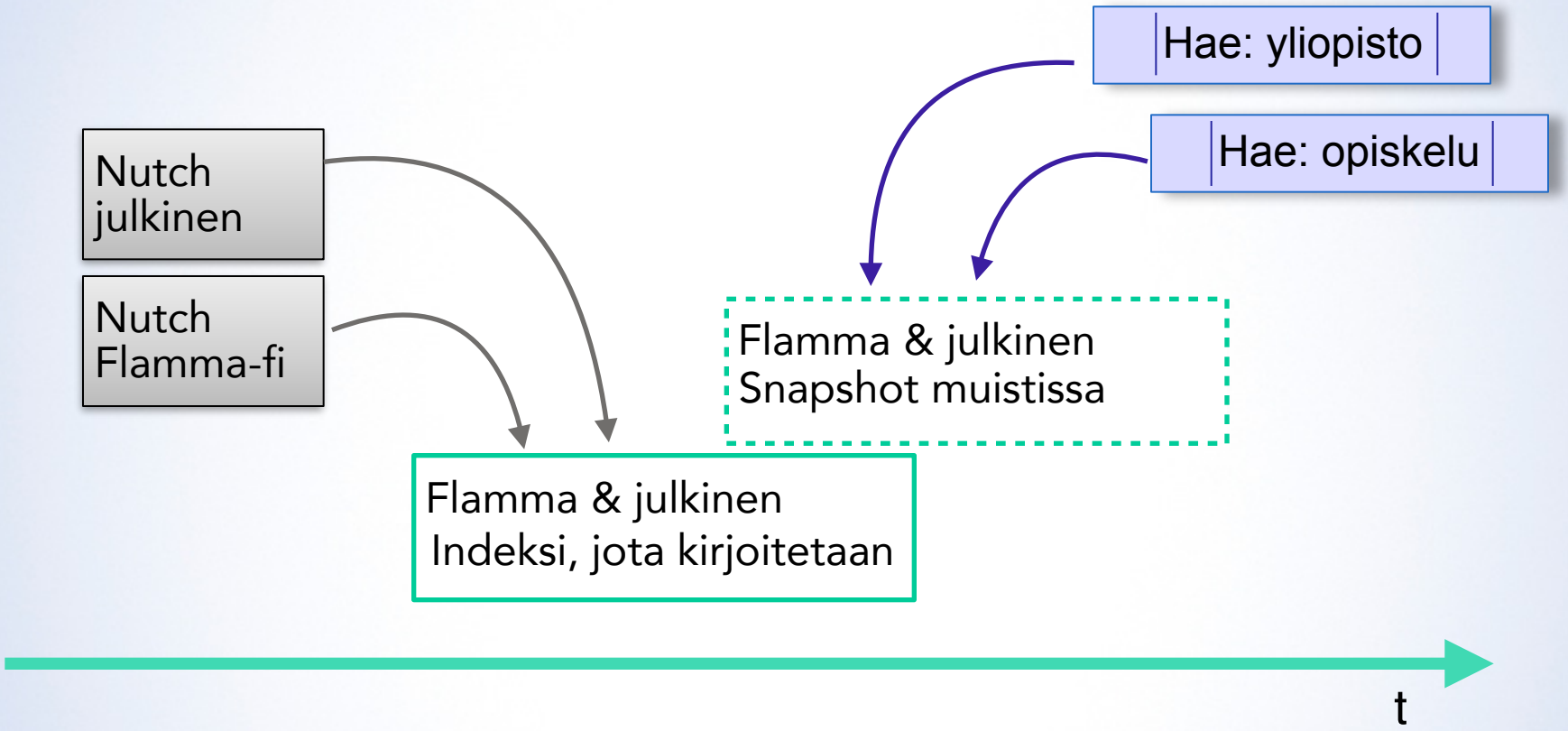


INDEKSI

- Indeksejä 3 kpl (kahlaajan indeksiä päivittävät instanssit)
    - Flamma (flamma-fi, flamma-sv, flamma-en)
    - Helsinki.fi (helsinki.fi)
    - Kaikki [Flamma + Helsinki.fi] (flamma-fi, flamma-sv, flamma-en ja helsinki.fi)
- Indeksin "Kaikki" voisi myös yhdistää Flamma- ja helsinki.fi -indekseistä, mutta toimenpide on liian hidas. Nopeampi tapa on indeksoida kahlatut segmentit suoraan kahteen erilliseen indeksiin.
- Indeksiä voi kirjoittaa yksi index-writer kerrallaan. Samaan aikaan indeksistä voidaan kuitenkin hakea hakutuloksia.

HELSINGIN YLIOPISTO

INDEKSOINTI JA HAUT



HELSINGIN YLIOPISTO

INTEGRAATIOT

- Aluksi vain HTTP(S) rajapintaa, ei varsinaisia integraatioita
    - Käyttöliittymät standalone
    - Flamma vaatii kahlaajalta kirjautumisen sisältöihin pääsemiseksi.
- Solr-indeksiin voitaisiin syöttää tietoa useiden järjestelmien kuten Drupalin kautta (onnistuisi myös esim. Djangon tai WordPressin)
- Suunnitelmia: Matterhorn, Drupal, Koulutushaku

HELSINGIN YLIOPISTO

SOVELLUKSET

- Solr versio 4.8.1 (indeksi)
  - Jetty 6.1.26
- Nutch versio 1.9 (kahlaaja)
  - Hadoop (hadoop-core 1.2.0)
- Käyttöliittymät
  - PHP (ml. solr-php-client -kirjasto)
- Joitain omia skriptejä (bash, Perl) hallintaan

HELSINGIN YLIOPISTO

RAUDAT

- 2 kpl virtuaalipalvelimia
  - CPU: 2 x 2560 Mhz
  - Muisti: 16 Gt
  - Systeemilevy: 10 Gt
  - Datalevy: 900 Gt
  - Käyttöjärjestelmä: Red Hat Enterprise Linux Server release 6.5 (Santiago)
  - Kahdennus: Fail over

HELSINGIN YLIOPISTO

KÄYTTÖ NYT JA JATKOSSA

Käyttö alkuvaiheessa

- Ensin WWW-haku julkisille sivuille ja intranetiin
- Sivustokohtaiset haut

Mahdollisia kehitysideoita

- Sivustojen omaa skeemaa ja indeksiä vaativat haut
- Aineistojen hallinta
- Yhteistyö kirjastojen kanssa

HELSINGIN YLIOPISTO

SEURAAVAKSI

- Versiopäivitykset
- Arkkitehtuurimuutokset
  - Klusterointi ja hajautus kuormantasauksen taakse?
  - Tietokantaratkaisut?
  - Rautapalvelimet?
  - Nutchin ja Solr:n eriyttäminen koneilla?
- Jatkokehitys WWW-konsepteihin
- Uudet käyttötavat ja kohteet
- Suorituskyvyn ja palvelutason jatkuva kehittäminen

HELSINGIN YLIOPISTO

Kiitos!

Lisätiedot projektista: ville.tenhunen@helsinki.fi

HELSINGIN YLIOPISTO