Hafslund SESAM

2012-09-06

Lars Marius Garshol, <larsga@bouvet.no>
http://twitter.com/larsga

bouvet

Lars Marius Garshol

- Consultant in Bouvet since 2007
  - focus on information architecture and semantics
- Worked with semantic technologies since 1999
  - mostly with Topic Maps
  - co-founder of Ontopia, later CTO
  - editor of several Topic Maps ISO standards 2001-
  - co-chair of TMRA conference 2006-2011
  - developed several key Topic Maps technologies
  - consultant in a number of Topic Maps projects
- Published a book on XML on Prentice-Hall
- Implemented Unicode support in the Opera web browser

My role on the project

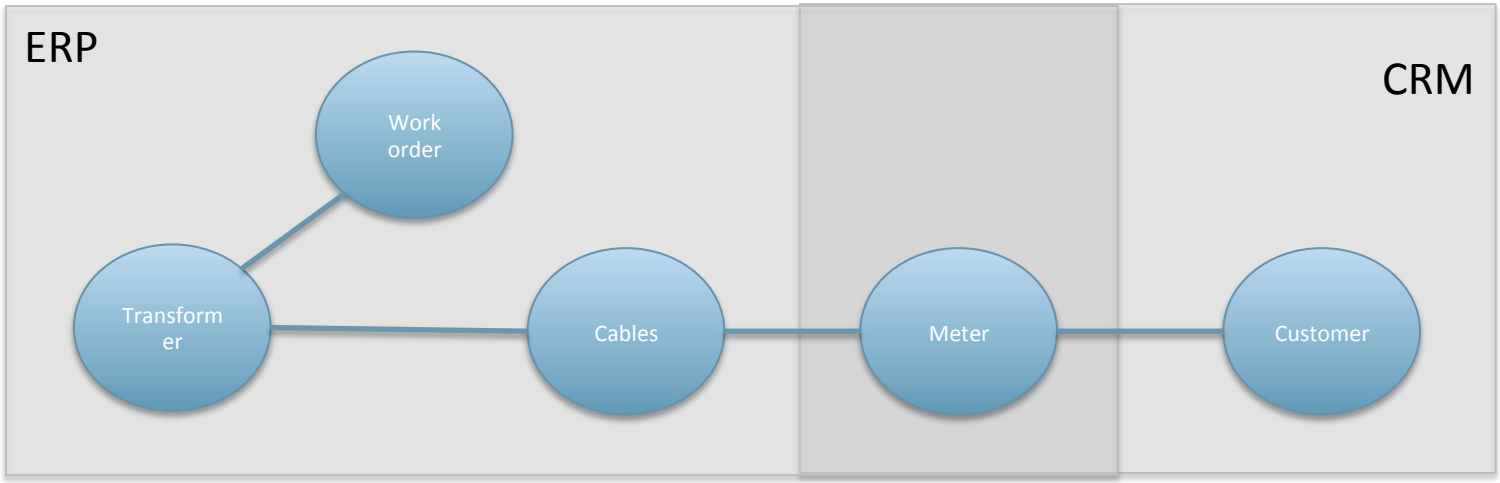
- The overall architecture is the brainchild of Axel Borge
- SDshare came from an idea by Graham Moore
- I only contributed parts of the design
  - and some parts of the implementation
- Don’t actually know the whole system

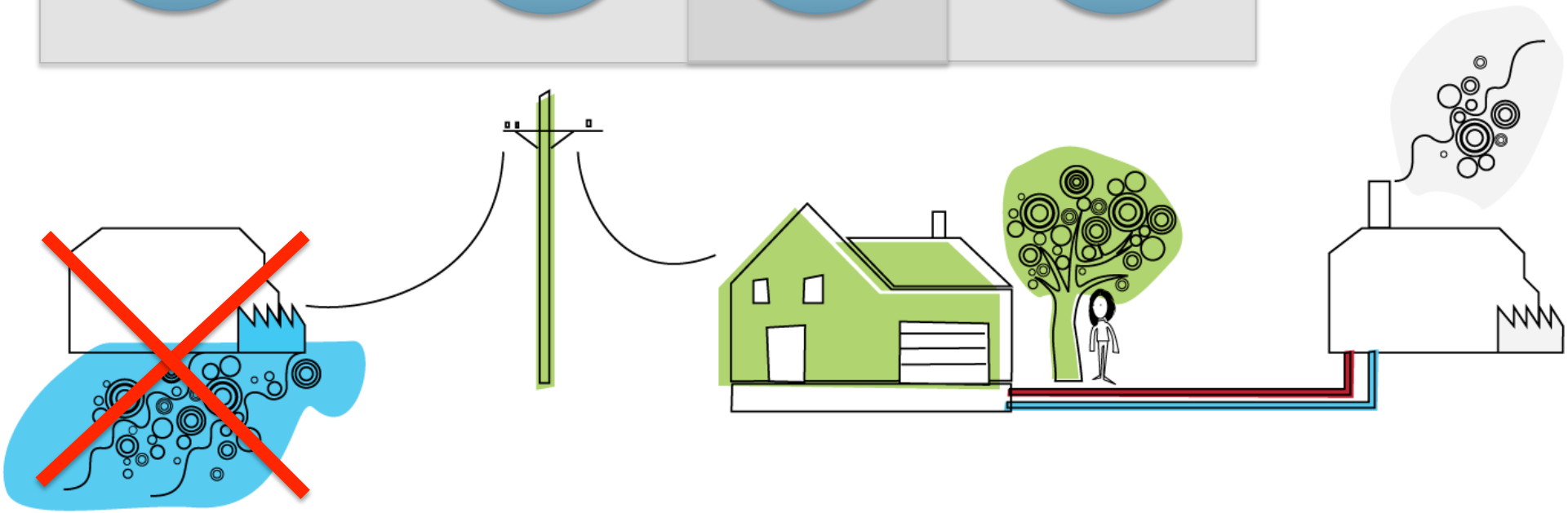
Hafslund SESAM

Hafslund ASA

- Norwegian energy company
  - founded 1898
  - 53% owned by the city of Oslo
  - responsible for energy grid around Oslo
  - 1.4 million customers
- A conglomerate of companies
  - Nett (electricity grid)
  - Fjernvarme (remote heating)
  - Produksjon (power generation)
  - Venture
  - …

What if...?





bouvet

How hard can it be?

- Design a single data model for the enterprise
- Appoint a master for each type of information
  - get rid of duplicate systems, convert old data
- Synchronize data into systems which need copies

Information utopia

- Reaching agreement is slow
  - slow is expensive
- Migrating to single masters is slow
  - new systems get added faster than you can replace the old
- This is a long and hard slog
  - but it's not necessary for search purposes

Hafslund SESAM

- An archive system, really
- Generally, archive systems are glorified trash cans
  - putting it in the archive effectively means hiding it
- Because archives are not important, are they?
- Except, when you need that contract from 1937 about the right to build a power line across…

Frykter at skapene er stjålet

Nav har mistet tre arkivskap med sensitive personopplysninger

Dokumenter og opplysninger fra 130 saker som enten er under etterforskning for trygdesvindel eller er henlagt, er borte.

Arild Færaas
Publisert: 18.jan. 2012 (10:04) **Oppdatert:** 18.jan. 2012 (11:45)

- Nav blåste sensitivinformasjon på nett
- Nav roter bort papirer

Tre arkivskap med mapper om 130 personer er borte. Det inneholder ifølge Nav opplysninger som ligger til grunn for behandling og utbetaling fra Nav, notater, korrespondanse med brukeren, i tillegg til vurderinger og vedtak. De forsvant i forbindelse med flytting på sensommeren i fjor.

På forsiden akkurat nå

Siste nytt

20:34 - Uakseptabel ungdomsledighet i Hellas
20:30 Clinton tror på snarlig løsning i Egypt-strid
20:19 Derfor eksisterer fenomenet skuddår
19:41 Tre elever døde etter skyting i Ohio
19:40 Tysk seriemorder dømt til livsvarig fengsel

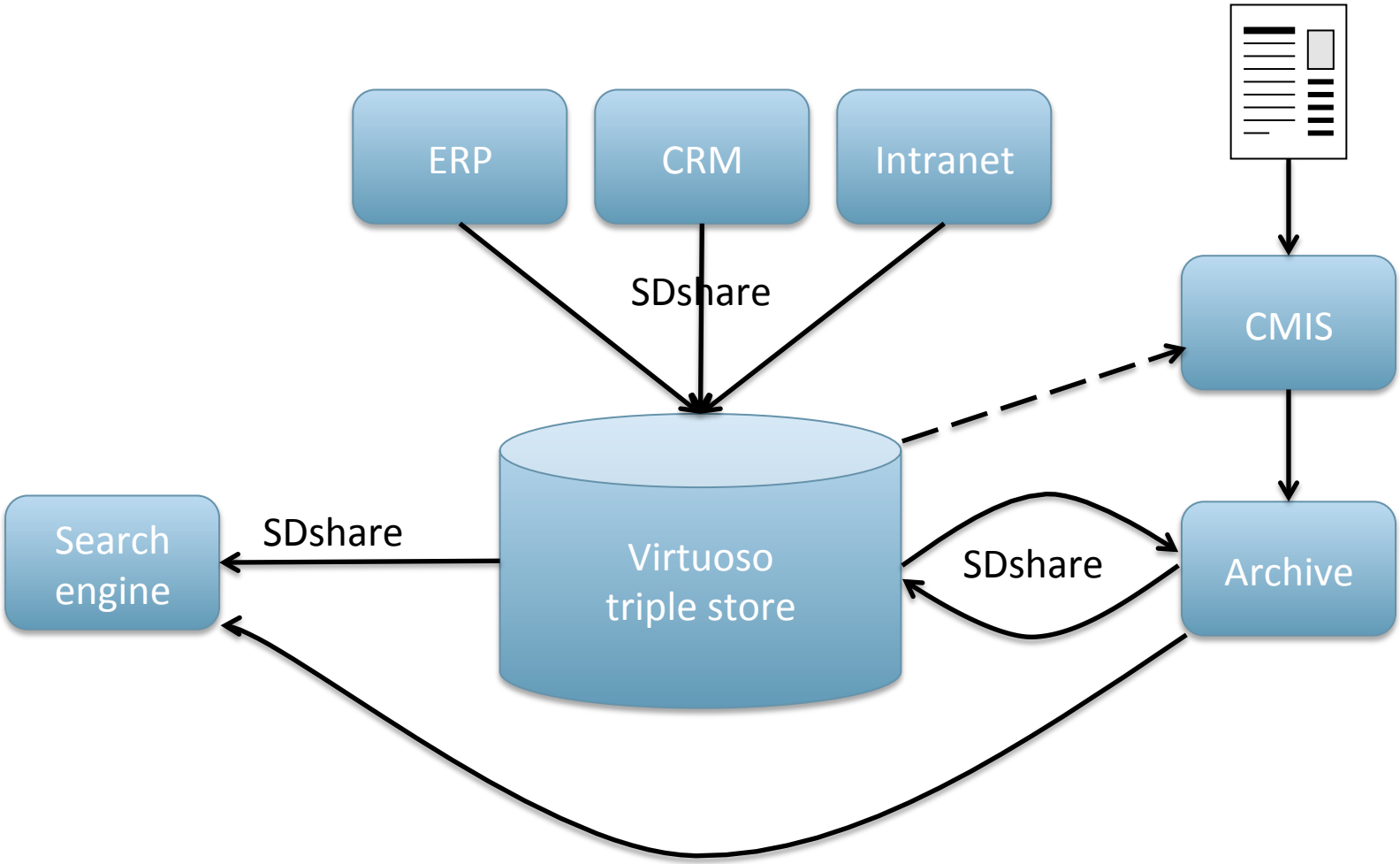
Problems with archives

- Poor metadata, because nobody bothers to enter it properly
  - yet, much of the metadata exists in the user context
- Not used by anybody
  - strange, separate system with poor interface
  - (and the metadata is poor, too)
- Contains only documents
  - not connected to anything else

Our goals

- Collect metadata automatically, from context
- Connect to context from enterprise systems
- Enrich with background knowledge
- Present it in an attractive, intuitive way
- Long term:
  - become a major part of the intranet
  - become *the* internal search solution

High-level architecture

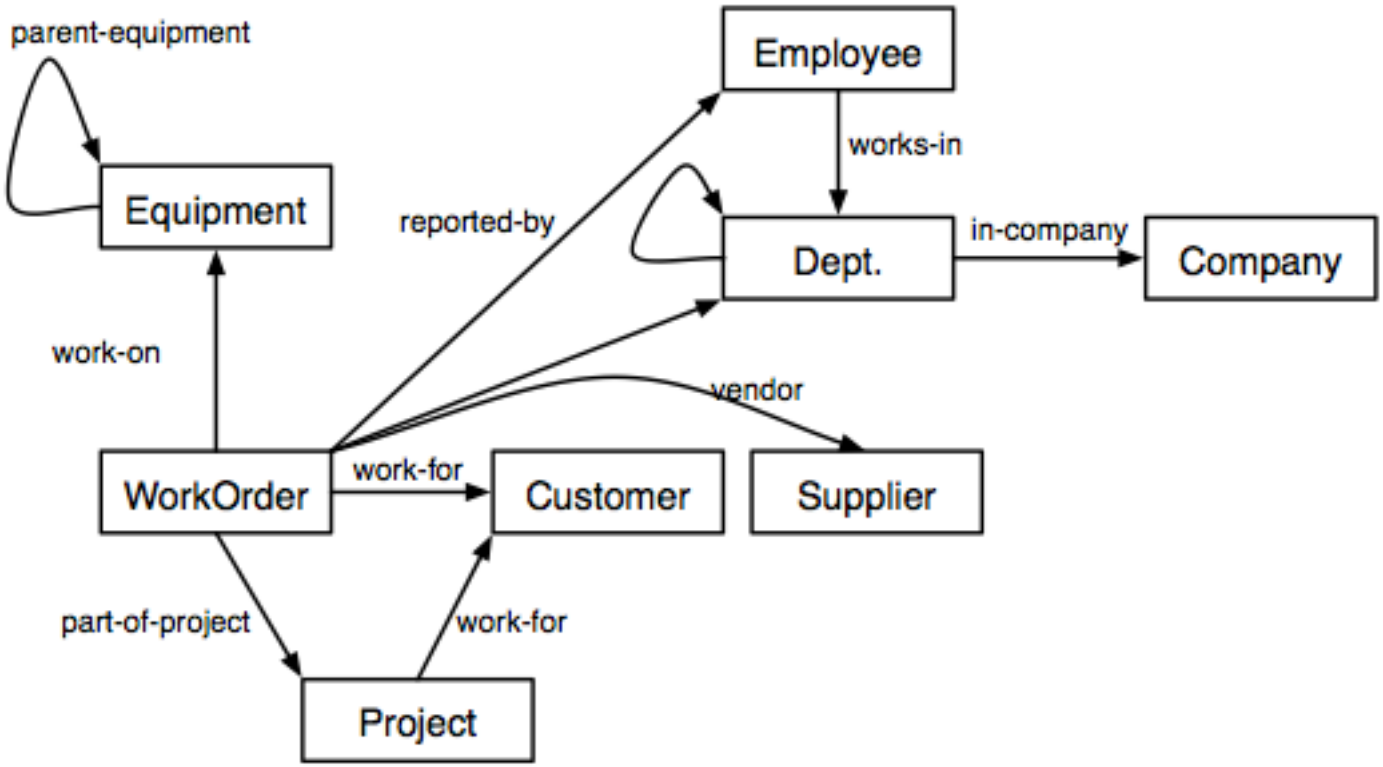


bouvet

Main principle of data extraction

- No canonical model!
- Instead, data reflects model of source system
- One ontology per source system
  - subtyped from core ontology where possible
- Vastly simplifies data extraction
  - for search purposes it loses us nothing
  - and translation is easier once the data is in the triple store

Simplified core ontology

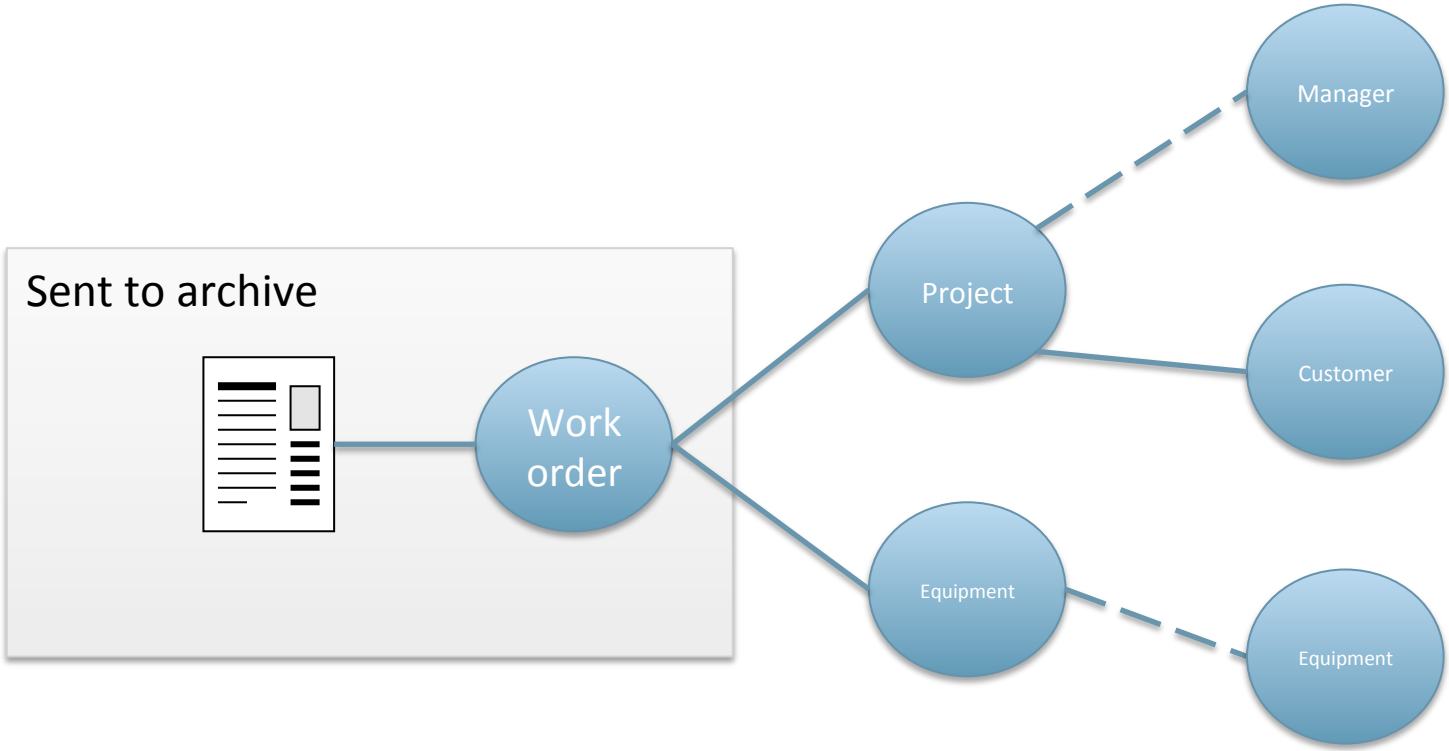


bouvet

When archiving

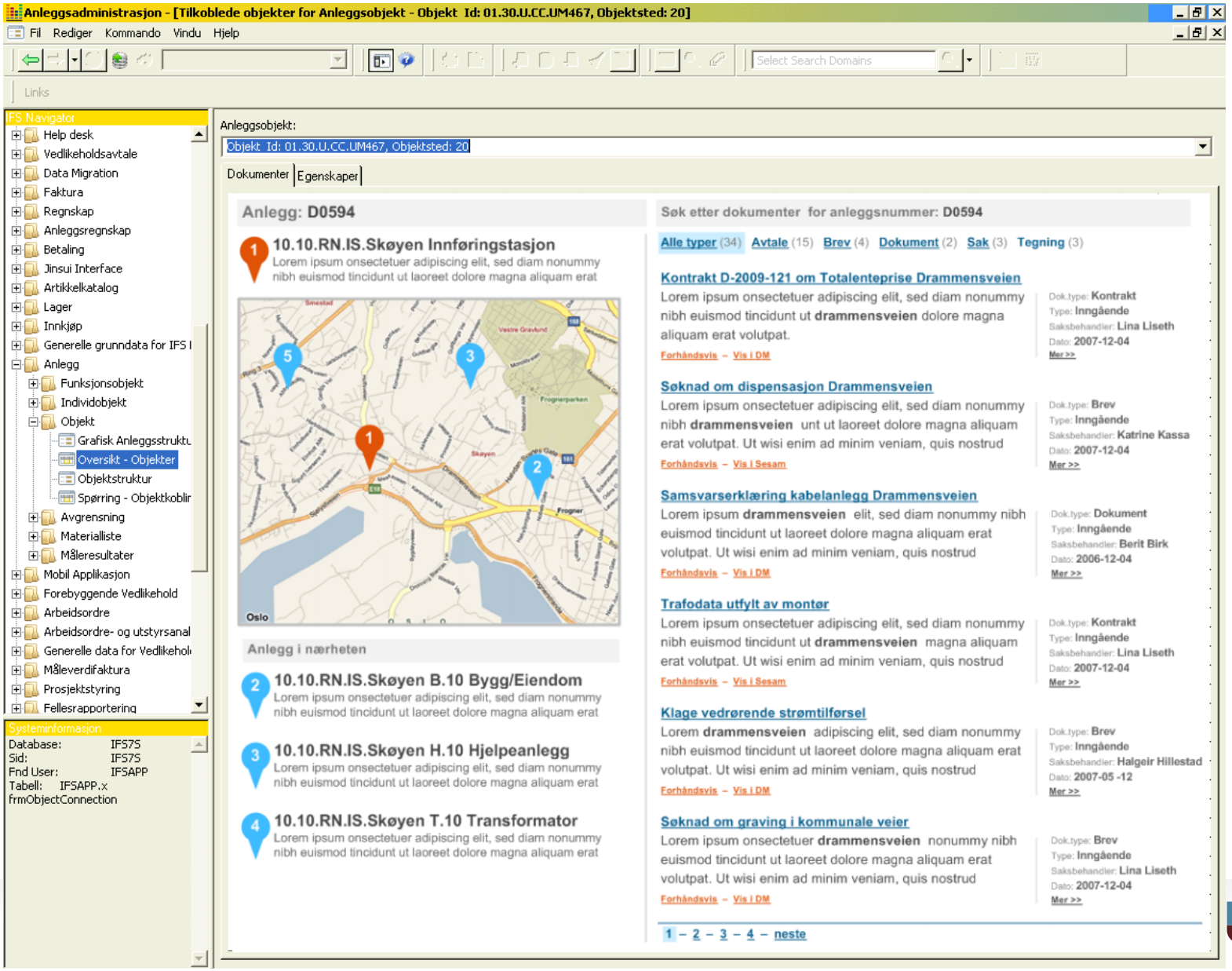
- The user works on the document in some system
  - ERP, CRM, whatever
- This system knows the context
  - what user, project, equipment, etc is involved
- This information is passed to the CMIS server
  - it uses already gathered information from the triple store to attach more metadata

Auto-tagging



bouvet

Showing context in the ERP system

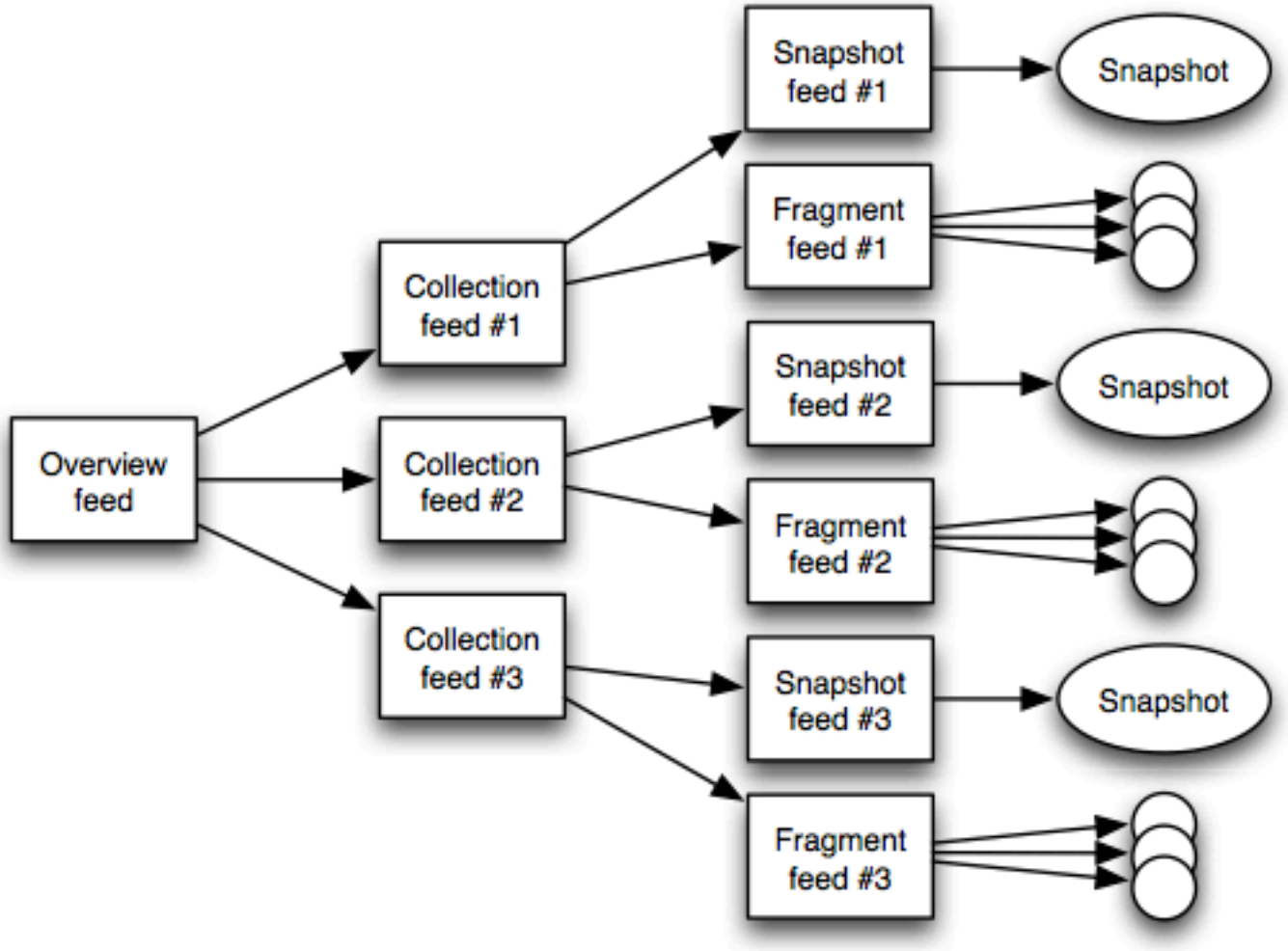


bouvet

The data integration

- All data transport done by SDshare
- A simple Atom-based specification for synchronizing RDF data
  - http://www.sdshare.org
- Provides two main features
  - snapshot of the data
  - fragments for each updated resource

SDshare service structure



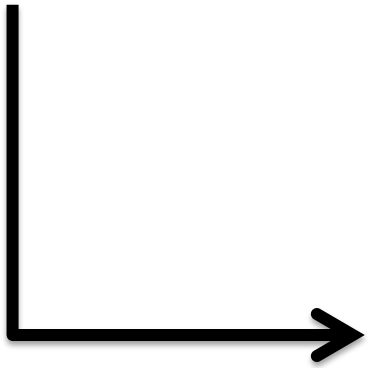
bouvet

Typical usage of SDshare

- Client downloads snapshot
  - client now has complete data set
- Client polls fragment feed
  - each time asking for new fragments since last check
  - client keeps track of time of last check
  - fragments are applied to data, keeping them in sync

Implementing the fragment feed

```
select objid, objtype, change_time
from history_log
where change_time > :since:
order by  change_time asc
```

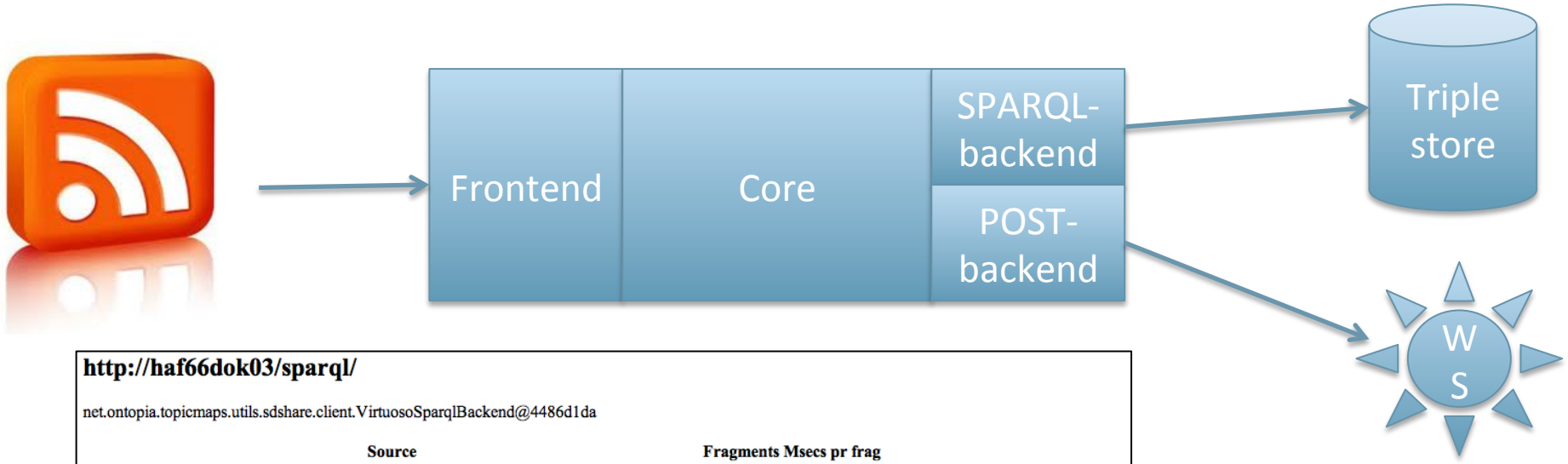


```
<atom>
 <title>Fragments for ...</title>
 ...

 <entry>
  <title>Change to 34121</title>
  <link rel=fragment href="..."/>
  <sdshare:resource>http://...</sdshare:resource>
  <updated>2012-09-06T08:22:23</updated>
 </entry>
 <entry>
  <title>Change to 94857</title>
  <link rel=fragment href="..."/>
  <sdshare:resource>http://...</sdshare:resource>
  <updated>2012-09-06T08:22:24</updated>
 </entry>
...
```

bouvet

The SDshare client



http://haf66dok03/sparql/

net.ontopia.topicmaps.utils.sdshare.client.VirtuosoSparqlBackend@4486d1da

Source	Fragments	Msecs pr frag
http://esb/document/umic/duke/rest/SDShareService/duke/collections?collectionId=duplicates	0	
http://haf66dok04:9090/sdshare/collections/IFSData	301	444.1661
http://haf66dok04:9090/sdshare/collections/SiebelCustomer	230000	418.77567
http://haf66dok04:9090/sdshare/collections/SiebelAddress	1320000	103.94148
http://haf66dok04:9090/sdshare/collections/SiebelAsset	1990000	87.86205
http://haf66dok04:9090/sdshare/collections/SiebelServicepoint	1590000	116.458565
http://haf66dok04:9090/sdshare/collections/SiebelMeter	763805	119.147026
http://haf66dok01:8090/collection.aspx?collectionId=case	0	
http://haf66dok01:8090/collection.aspx?collectionId=groups	5656	182.04791
http://haf66dok01:8090/collection.aspx?collectionId=users	4725	181.53143
http://haf66dok01:8090/collection.aspx?collectionId=contact	4	132.5
http://haf66dok01:8090/collection.aspx?collectionId=codetables	6498	65.040474

jdbc:oracle:thin:@172.19.4.174:1521:WLSPRD

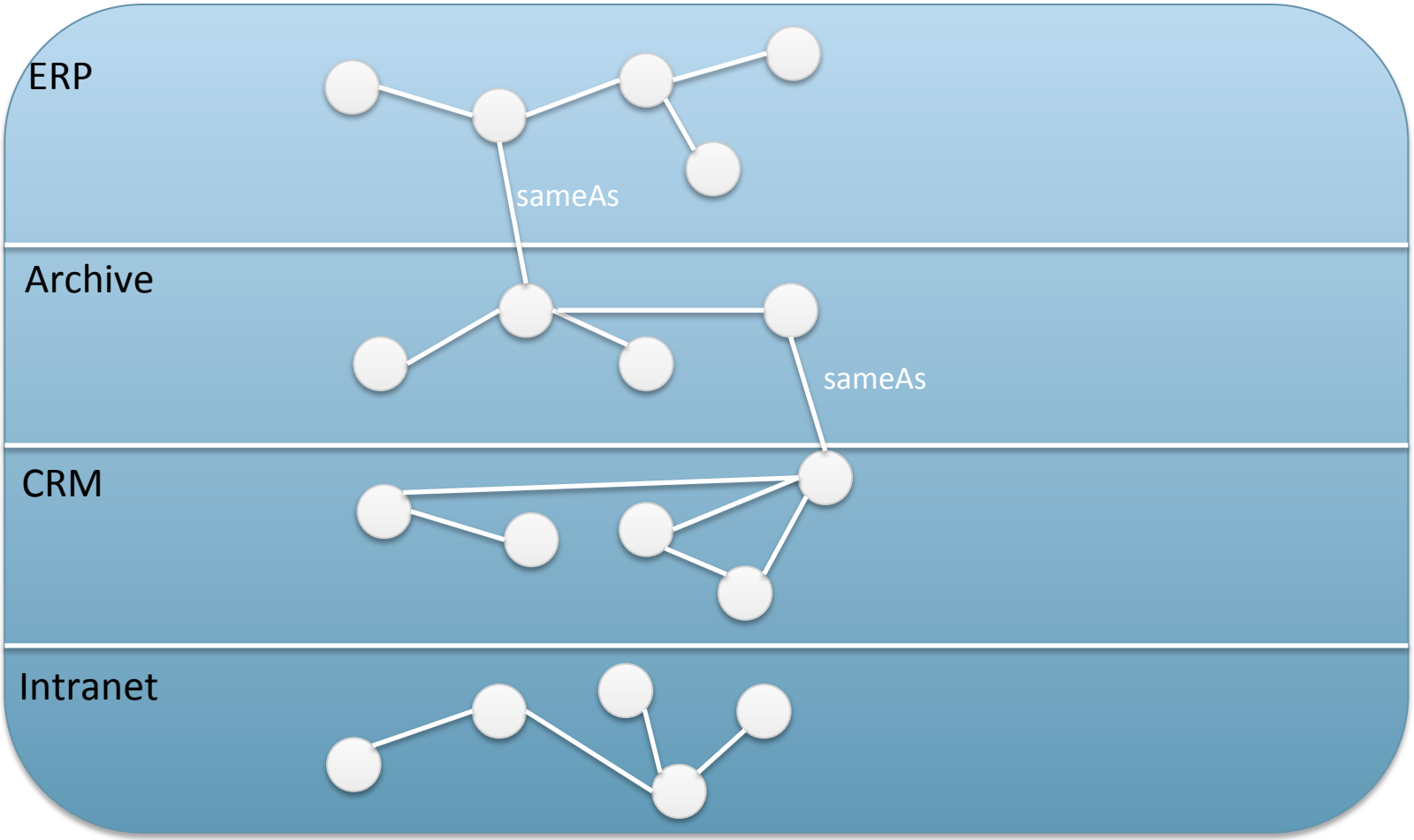
net.ontopia.topicmaps.utils.sdshare.client.JDBCQueueBackend@63b5a40a

Source	Fragments	Msecs pr frag
http://haf66dok04:8080/umic-sdshare/standardumicfeed/collections?collectionId=http%3A%2F%2Fpsi.hafslund.no%2Fsesam%2Ffeeds%2Fifs%2Fdata	120	46.158333

http://code.google.com/p/sdshare-client/

bouvet

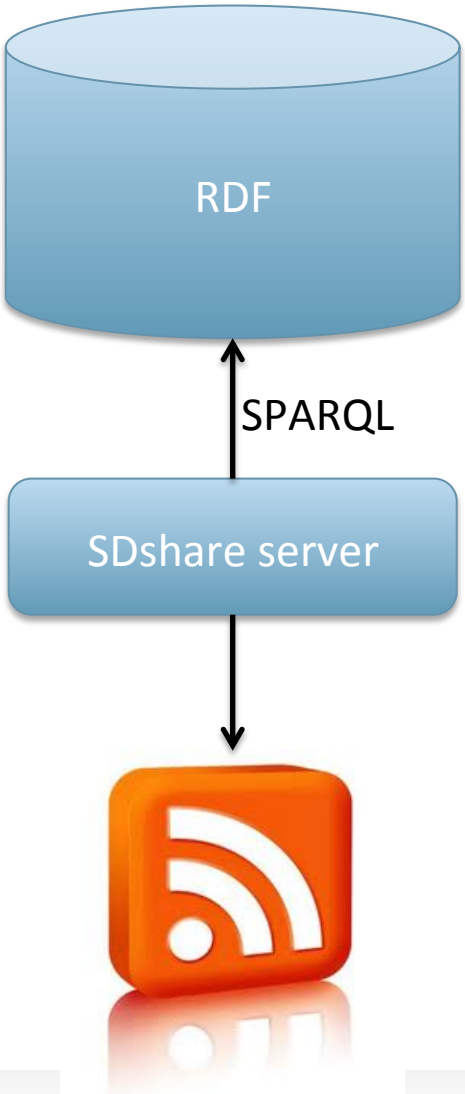
Data structure in triple store



Triple store

bouvet

Getting data out of the triple store



- Set up SPARQL queries to extract the data
- Server does the rest
- Queries can be configured to produce
  - any subset of data
  - data in any shape

Contacts into the archive

- We want some resources in the triple store to be written into the archive as “contacts”
  - need to select which resources to include
  - must also transform from source data model
- How to achieve without hard-wiring anything?

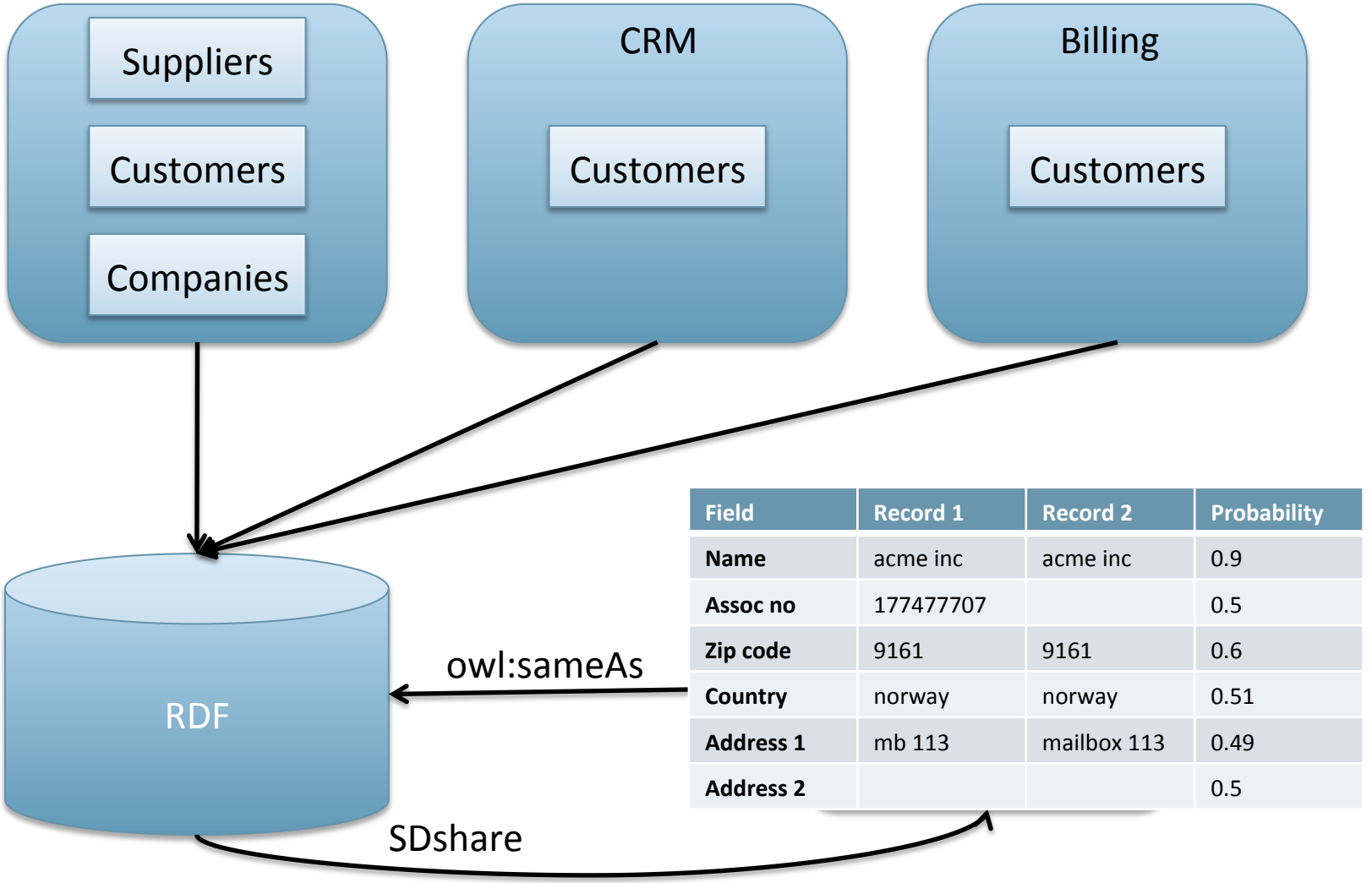
Contacts solution

- Create a generic archive object writer
  - type of RDF resource specifies type of object to create
  - name of RDF property (within namespace) specifies which property to set
- Set up RDF mapping from source data
  - type1 maps-to type2
  - prop1 maps-to prop2
  - only mapped types/properties included
- Use SPARQL to
  - create SDshare feed
  - do data translation with CONSTRUCT query

Access control

- Implemented by search engine
  - on login a SPARQL query lists user's access control group memberships
  - search engine uses this to filter search results
  - user only sees what they have access rights to
- In some cases, complex access rules are run to resolve ACLs before loading into triple store

Duplicate suppression



Field	Record 1	Record 2	Probability
Name	acme inc	acme inc	0.9
Assoc no	177477707		0.5
Zip code	9161	9161	0.6
Country	norway	norway	0.51
Address 1	mb 113	mailbox 113	0.49
Address 2			0.5

http://code.google.com/p/duke/

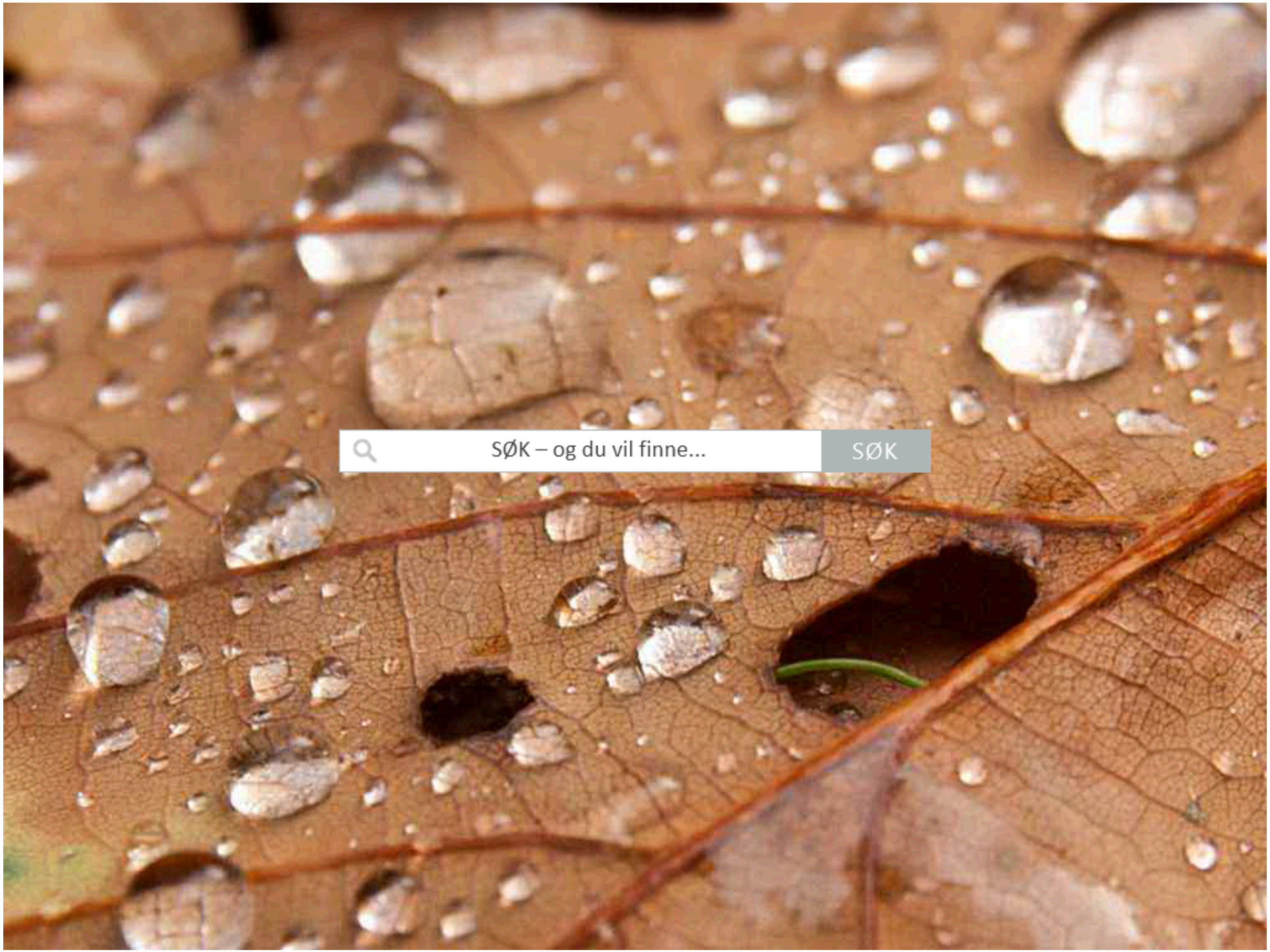
bouvet

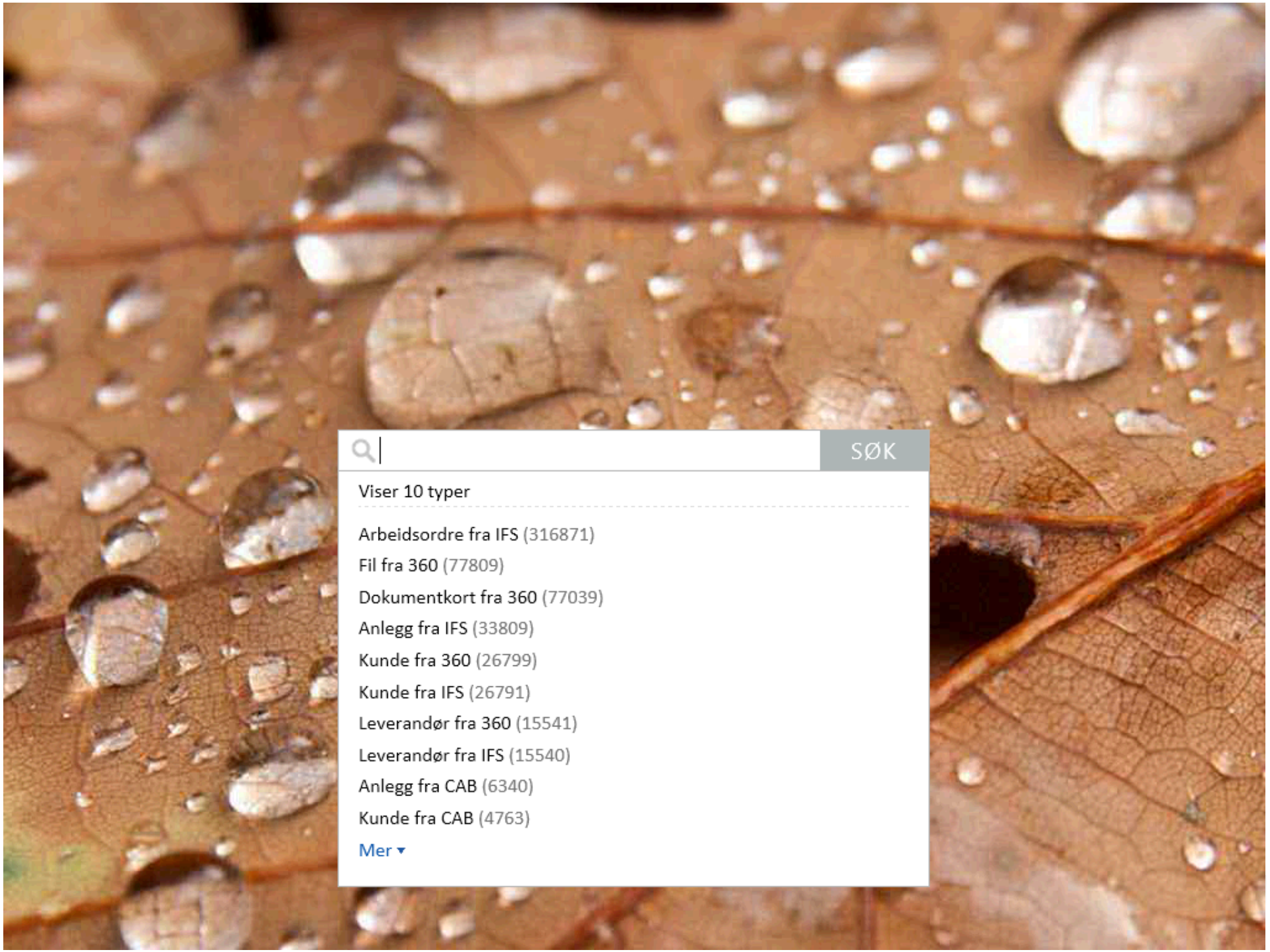
Properties of the system

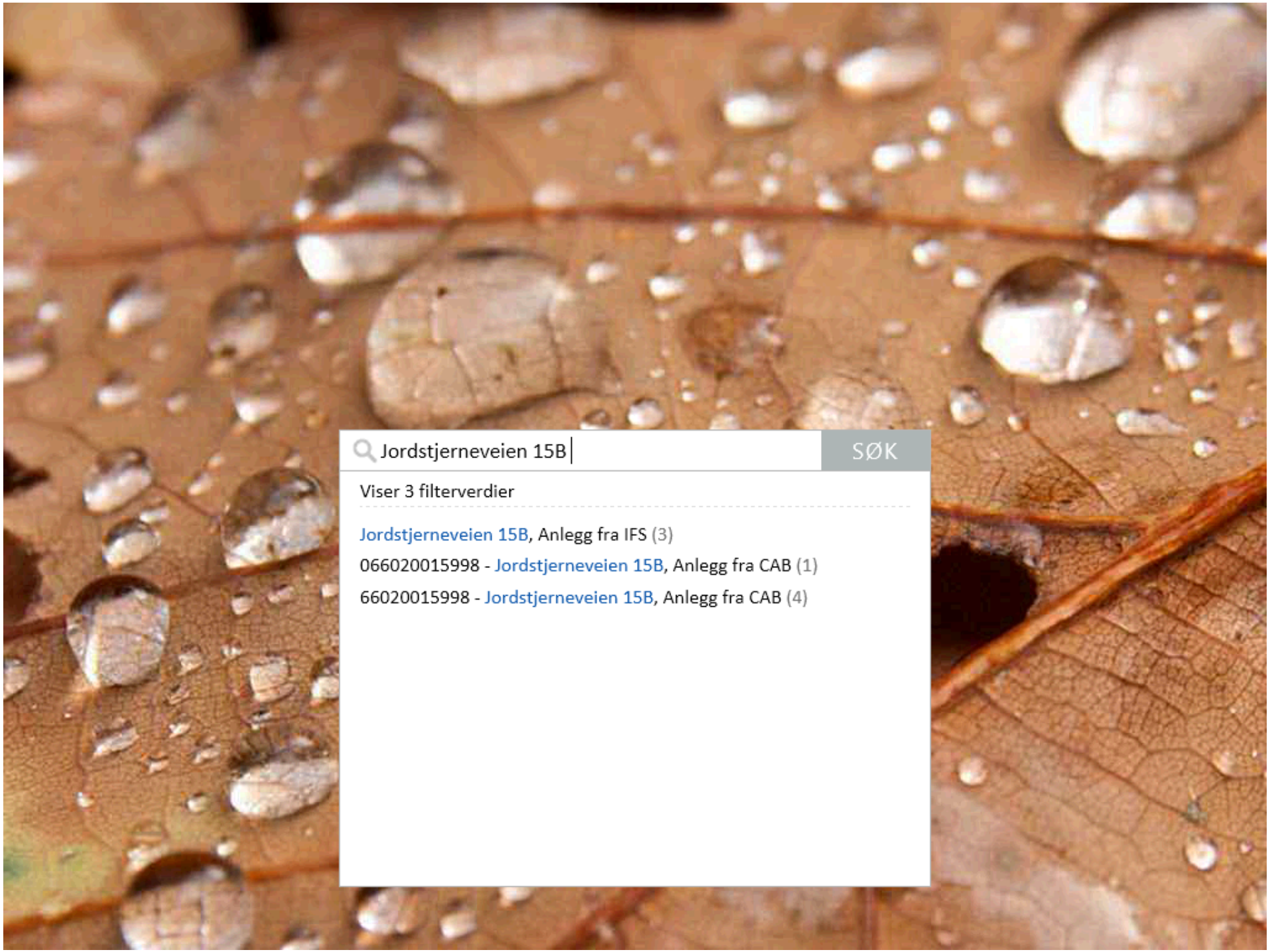
- Very little state
  - most components are stateless (or have little state)
- Idempotent
  - applying a fragment 1 or many times: same result
- Clear and reload
  - can delete everything and reload at any time
- Uniform integration approach
  - everything is done the same way
- Really simple integration
  - setting up a data source is generally very easy
- Adding integrations is easy
  - doesn't impact other integrations in any way

Data volumes

Graph	Statements
IFS data	5,417,260
Public 360 data	3,725,963
GeoNIS data	44,242
Tieto CAB data	138,521,810
Hummingbird 1	32,619,140
Hummingbird 2	165,671,179
Hummingbird 3	192,930,188
Hummingbird 4	48,623,178
Address data	2,415,315
Siebel data	36,117,786
Duke links	4,858
Total	**626,090,919**


SØK – og du vil finne...
SØK


SØK
Viser 10 typer
Arbeidsordre fra IFS (316871)
Fil fra 360 (77809)
Dokumentkort fra 360 (77039)
Anlegg fra IFS (33809)
Kunde fra 360 (26799)
Kunde fra IFS (26791)
Leverandør fra 360 (15541)
Leverandør fra IFS (15540)
Anlegg fra CAB (6340)
Kunde fra CAB (4763)
Mer


Jordstjerneveien 15B
SØK
Viser 3 filterverdier
Jordstjerneveien 15B, Anlegg fra IFS (3)
066020015998 - Jordstjerneveien 15B, Anlegg fra CAB (1)
66020015998 - Jordstjerneveien 15B, Anlegg fra CAB (4)

Hafslund

« Tilbake til søkeresultatet

Jordstjerneveien 15B
Anlegg fra IFS
MCH 66020015998
inngår i Mortensrudhøyden (Jordstjerneveien)
Lengdegrad 59.8457151

Åpne i IFS | Mer info

Avansert ▾

Søk videre eller start nytt søk

SØK

Axel Borge

Resultat 1 - 2 av 2

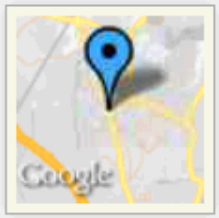
066020015998 - Jordstjerneveien 15B
Anlegg fra CAB
Adresse Jordstjerneveien 15B
Postnummer 1283
Poststed OSLO

66020015998 - Jordstjerneveien 15B
Anlegg fra CAB
Adresse Jordstjerneveien 15B
Postnummer 1283
Poststed OSLO

◂ 1 ▸

Hafslund

« Tilbake til søkeresultatet



Mortensrudhøyden (Jordstjerneveien)
Anlegg fra IFS
MCH KLUM410W107
inngår i Kundesentraler Klemetsrud

Åpne i IFS

Avansert ▾

Søk videre eller start nytt søk

SØK

Axel Borge

Resultat 1 - 10 av 150

Sekundærledninger Mortensrudhøyden
Arbeidsordre fra IFS
utført på Mortensrudhøyden (Jordstjerneveien)
innrapportert av Thormod Kvarme

pumpe stoppet
Arbeidsordre fra IFS
utført på Mortensrudhøyden (Jordstjerneveien)
innrapportert av Olaf Nilsen

Kundesentral for Mortensrudhøyden boligfelt
Arbeidsordre fra IFS
utført på Mortensrudhøyden (Jordstjerneveien)
innrapportert av Thormod Kvarme

Kundesentraler sekundærsiden og villavekslere
Arbeidsordre fra IFS
utført på Mortensrudhøyden (Jordstjerneveien)
innrapportert av Thormod Kvarme

Jordstjerneveien 52
Anlegg fra IFS
MCH KLUM410W107H
inngår i Mortensrudhøyden (Jordstjerneveien)

Jordstjerneveien 29C og 29D
Anlegg fra IFS
MCH KLUM410W107B
inngår i Mortensrudhøyden (Jordstjerneveien)

Jordstjerneveien 113
Anlegg fra IFS
MCH 66020113999
inngår i Mortensrudhøyden (Jordstjerneveien)

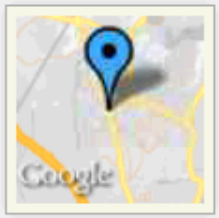
Jordstjerneveien 101
Anlegg fra IFS
MCH 66020101999
inngår i Mortensrudhøyden (Jordstjerneveien)

Jordstjerneveien 90

Jordstjerneveien 84

Hafslund

« Tilbake til søkeresultatet



Mortensrudhøyden (Jordstjerneveien)
Anlegg fra IFS
MCH KLUM410W107
inngår i Kundesentraler Klemetsrud

Åpne i IFS

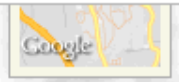
Avansert ▾

SØK

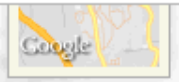
Axel Borge

Viser 4 typer

Anlegg fra IFS (102)
Arbeidsordre fra IFS (27)
Fil fra 360 (11)
Dokumentkort fra 360 (10)



inngår i Mortensrudhøyden (Jordstjerneveien)



inngår i Mortensrudhøyden (Jordstjerneveien)

Jordstjerneveien 113
Anlegg fra IFS
MCH 66020113999
inngår i Mortensrudhøyden (Jordstjerneveien)

Jordstjerneveien 101
Anlegg fra IFS
MCH 66020101999
inngår i Mortensrudhøyden (Jordstjerneveien)

Jordstjerneveien 90

Jordstjerneveien 84

Conclusion

How did it work out?

- RDF is great for information integration
- SDshare approach makes things even easier
- CMIS was not a success
  - Apache server immature, a real pain
- The archive product was a pain, too
  - lots of problems of various kinds
- Deduplication worked well
  - we see many uses for it in other contexts
- Getting access to data is slooooow
  - both at database level, and getting data into systems

My current project

- Integrate
  - Identity management system (IDM)
  - EPiServer CMS
  - Sharepoint
- starting August 13, ending November 1
- Right now we have
  - IDM
  - EPiServer CMS
  - Regjeringen.no
  - Sharepoint (lacking data)
  - ActiveDirectory (waiting for IT to open port)

Questions?

Have written a paper on the project, available on request. Looking for somewhere to publish it. Tips welcome.