From documents to datasets

Leif Johansson
TF-storage NDN2014
Moving files...

- Back in 2008 we started to think about moving files
- Lots of stuff already existed
  - Box
  - Dropbox
  - Filesender
- We (thought that we) needed to make something new...
Enter Lobber

- A “federation-enabled” torrent tracker
- Share massive files
- Decentralized storage (storage nodes)
- Storage nodes running deluge/transmission
There were some problems...

- Upload from web is ... a challenge
- Java-applet implementation of torrent ... not perfect
- Which BT client should we integrate with?
  - ctorrent
  - rtorrent
  - transmission
  - deluge
Then our customers came to our aid

- Re-focused our efforts on commodity services
- SUNET synchronization service tender launched in 2011
- Several bids including Box
- Box won (on price)
- We launched the SUNET Box service in 2012
- By 2013 NDN had duplicated the tender and now all Nordic countries share the same framework w. Box
The Box setup

- Single framework contract covering
  - Price
  - Integration
  - Data protection
  - Liability
  - etc

- Each country does a separate call-of-contract

- All countries share the same technical infrastructure
**Technical integration**

- Single IdP proxy (for all the Nordics)
- Access control on per-domain basis
  - Eg uio.no can include all students, while chalmers.se only allows staff
- schacHomeOrganization optionally overrides Shibboleth scope
- On-boarding done by NDN NOC team
- Not very useful for very large datasets
  - Box is for documents, not datasets
Limitations

- At first only a single email per user was supported (now fixed)
- Only a single IdP per customer (fixed using IdP proxy)
- Windows installer hard to package for site-wide distribution (getting better)
Some numbers...

- TODO
The Kinderegg problem

- Very Large Files, low cost or simple: Pick any 2
- Box is low cost and simple
- Lobber was low cost (you guess the rest)
Datasets, not documents

- KB.se wanted help with a small problem...
  - distribute large datasets to an unknown set of consumers
  - ... “and we really like torrents”
Enter SUNET Datasets

- An experiment
- A rewrite of lobber (aka lobo2)
- A public API (w. OAuth2 and all the trimmings)
- No Java
- A federation-enabled tracker
- All open source
  - https://github.com/SUNET/lobo2
  - https://github.com/SUNET/lobo2a
Future of this stuff @ SUNET

- Definitely a filesender instance
  - maybe w lobo2 integration
  - maybe w btsync integration
- Probably a lot more Box users
- Maybe a lobo2 instance
Conclusions

- No tool is good for everything
  - We have Box and we still probably want filesender & lobo2
- Good tools *may* get used
  - The payoff has to warrant the investment
  - The remaining 20% may be too hard to get to
- Bad tools will *never* get used
  - Quality is king
  - Java as a client tool is dead
Q & A