A data storage for generic and heterogenous scientific data

Ketil Malde, Tomasz Furmanek, and Esmael Hassen

Data Submission

Is simple and easy

Uses domain specific file formats
Auto-generates most metadata
Free-text descriptions/
(almost) no mandatory fields

Provenance and links

Unique, persistent
IDs (citation)
Link datasets

Search/services

Self-contained services

Technology-agnostic (i.e. use
any relational database system)
Extract and index only relevant
data from data store
Independent and modular

Data Access

File-based storage

Access through HTTP,
FTP, rsync, bittorrent....

Easy replication

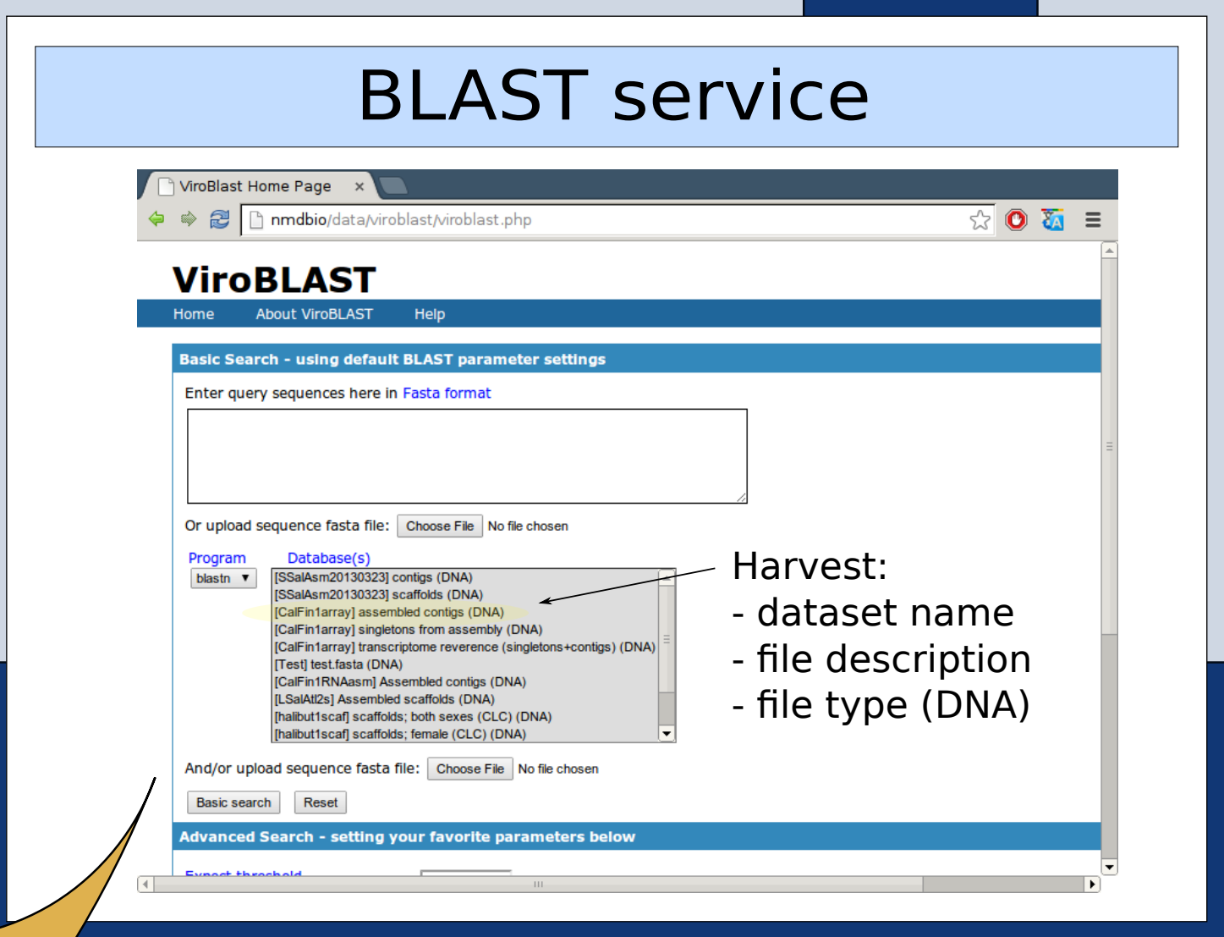
Domain-specific formats

Easy for domain
experts
No data conversion

Provenance

Identify origins of data

BLAST service



ViroBlast Home Page
nmdbio/data/viroblast/viroblast.php

ViroBLAST

Home | About ViroBLAST | Help

Basic Search - using default BLAST parameter settings

Enter query sequences here in Fasta format

Or upload sequence fasta file: Choose File No file chosen

Program: blastn
Database(s):
[SSalAsm20130323] contigs (DNA)
[SSalAsm20130323] scaffolds (DNA)
[CalFin1array] assembled contigs (DNA)
[CalFin1array] singletons from assembly (DNA)
[CalFin1array] transcriptome reverence (singletons+contigs) (DNA)
[Test] test.fasta (DNA)
[CalFin1RNAasm] Assembled contigs (DNA)
[LSalAtl2s] Assembled scaffolds (DNA)
[halibut1scaf] scaffolds; both sexes (CLC) (DNA)
[halibut1scaf] scaffolds; female (CLC) (DNA)

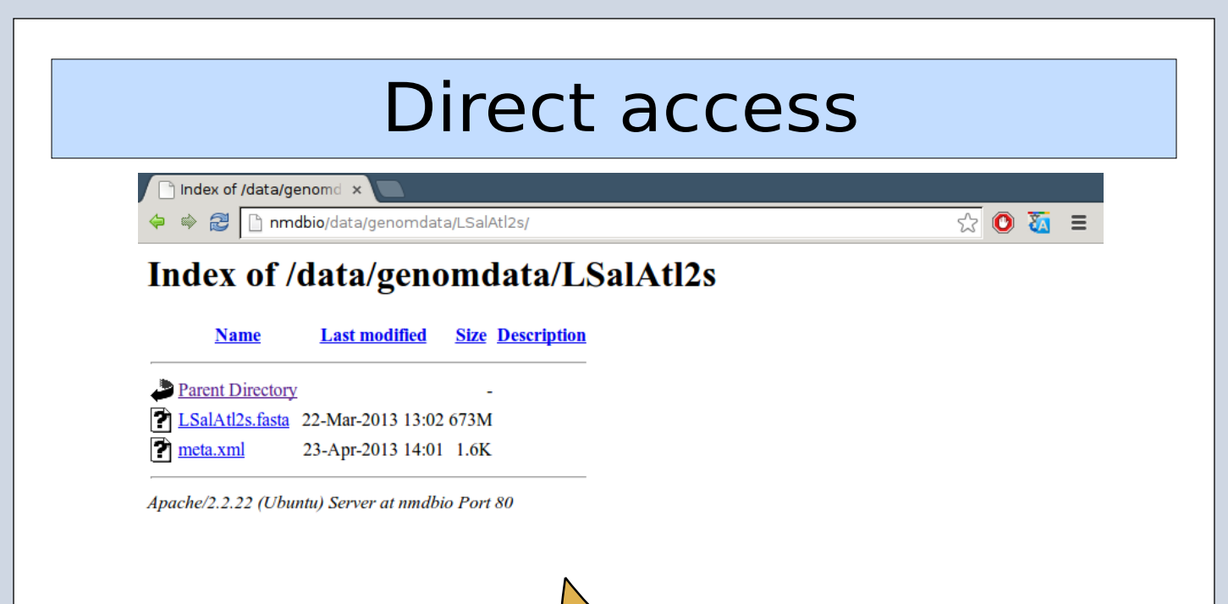
And/or upload sequence fasta file: Choose File No file chosen

Basic search | Reset

Advanced Search - setting your favorite parameters below

Harvest:
- dataset name
- file description
- file type (DNA)

Direct access



Index of /data/genomd
nmdbio/data/genomdata/LSalAtl2s/

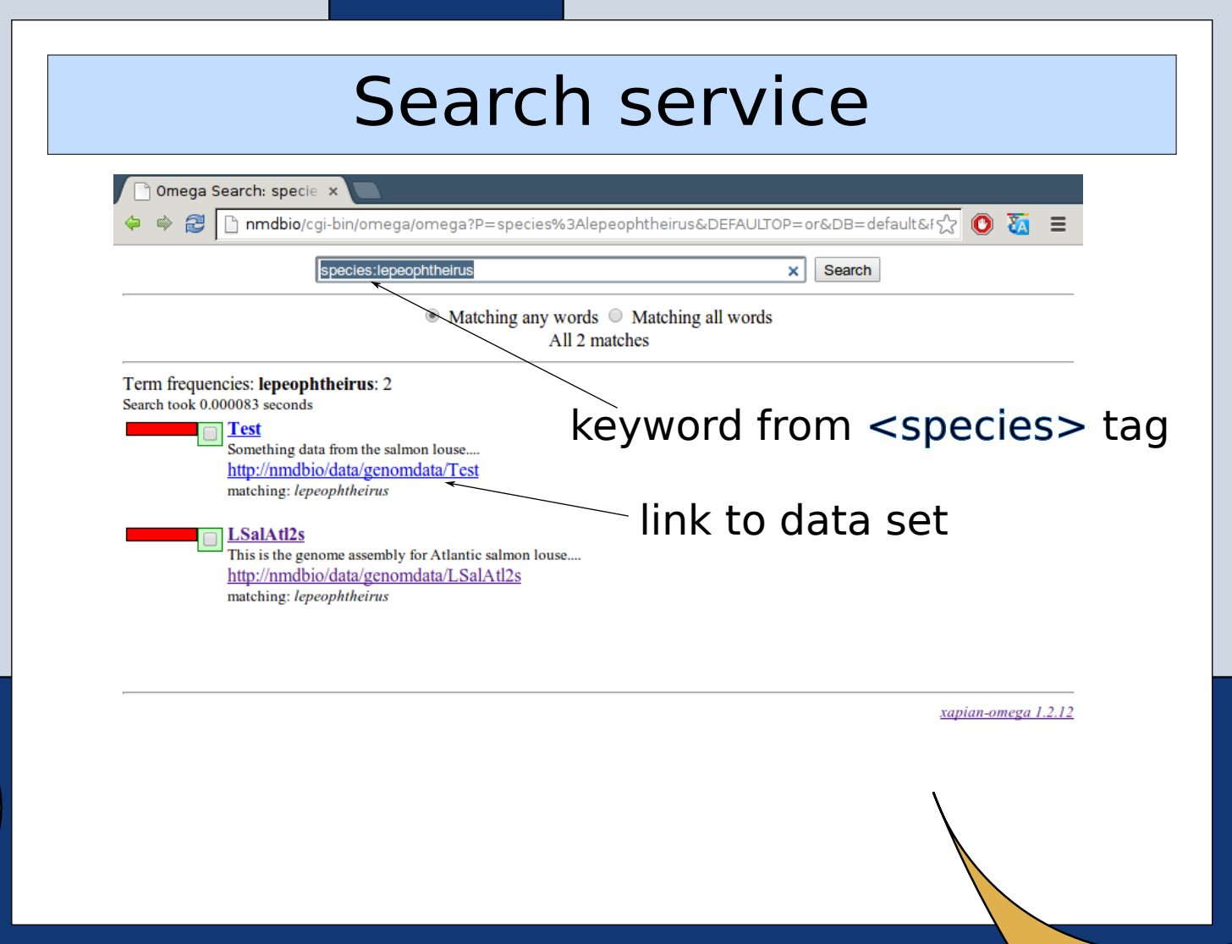
Index of /data/genomdata/LSalAtl2s

```
Name              Last modified      Size  Description
Parent Directory                      -
LSalAtl2s.fasta   22-Mar-2013 13:02  673M
meta.xml          23-Apr-2013 14:01  1.6K
```

Apache/2.2.22 (Ubuntu) Server at nmdbio Port 80

HTTP access
(just apache)

Search service



Omega Search: specie
nmdbio/cgi-bin/omega/omega?P=species%3Alepeophtheirus&DEFAULTOP=or&DB=default&

species:lepeophtheirus | Search

○ Matching any words ○ Matching all words
All 2 matches

Term frequencies: **lepeophtheirus**: 2
Search took 0.000083 seconds

Test
Something data from the salmon louse....
http://nmdbio/data/genomdata/Test
matching: *lepeophtheirus*

LSalAtl2s
This is the genome assembly for Atlantic salmon louse....
http://nmdbio/data/genomdata/LSalAtl2s
matching: *lepeophtheirus*

xapian-omega 1.2.12

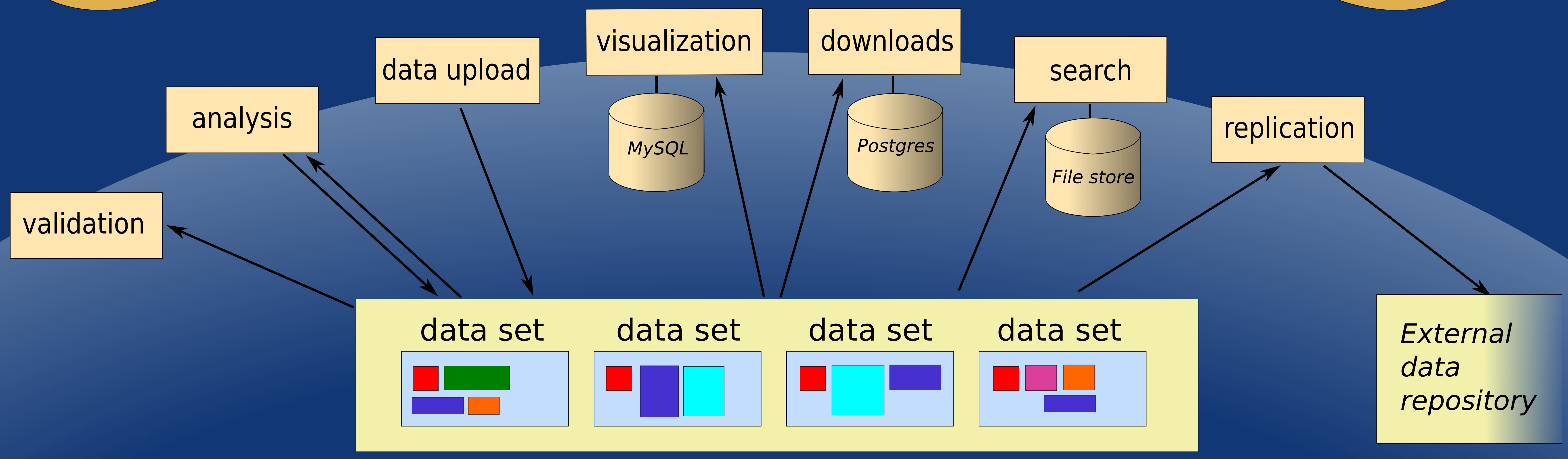
keyword from <species> tag

link to data set

Specialized search service - scans *relevant* data, ignores rest

Generic search service scans and indexes *metadata* only

specialized services



generic file storage

Metadata

XML format

automatic validation
tagging (TSN, geoloc, etc)
free text descriptions

Directory listings

Checksums for integrity
File type tagging

Data set relationships

Obsolescense and replacement
Dependencies
Aggregation and extraction

Extensibility

New data types

Technology advances means
frequent new data types
Adding a new data type is a
two minute operation

New technologies

Services are independent,
can use different technologies

Easy integration of off-the-shelf
products

Separation of concerns

Separate roles with separate
skill sets

System Admin
data copying
replication, extraction,
maintenance of
services

Data Manager
metadata analysis,
service design

Domain expert
analyze data,
provide relevant metadata,
submit results

Source code and documentation
`http://malde.org/~ketil/datastore`

Contact:
`ketil.malde@imr.no`

Norwegian Marine Data Centre

HAVFORSKNINGSINSTITUTTET
INSTITUTE OF MARINE RESEARCH