@yoyehudi

@intermineorg

InterMine

Open source success thanks to open data

UNIVERSITY OF CAMBRIDGE

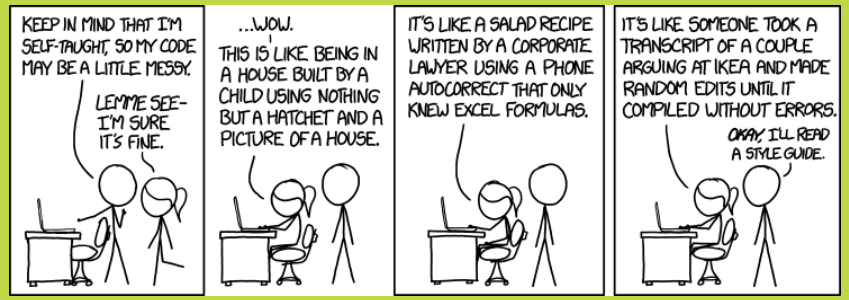
InterMine

Open Data

We all love **data**, but...

Messy Data

(Technically this comic is about code, but it works for messy data too...)



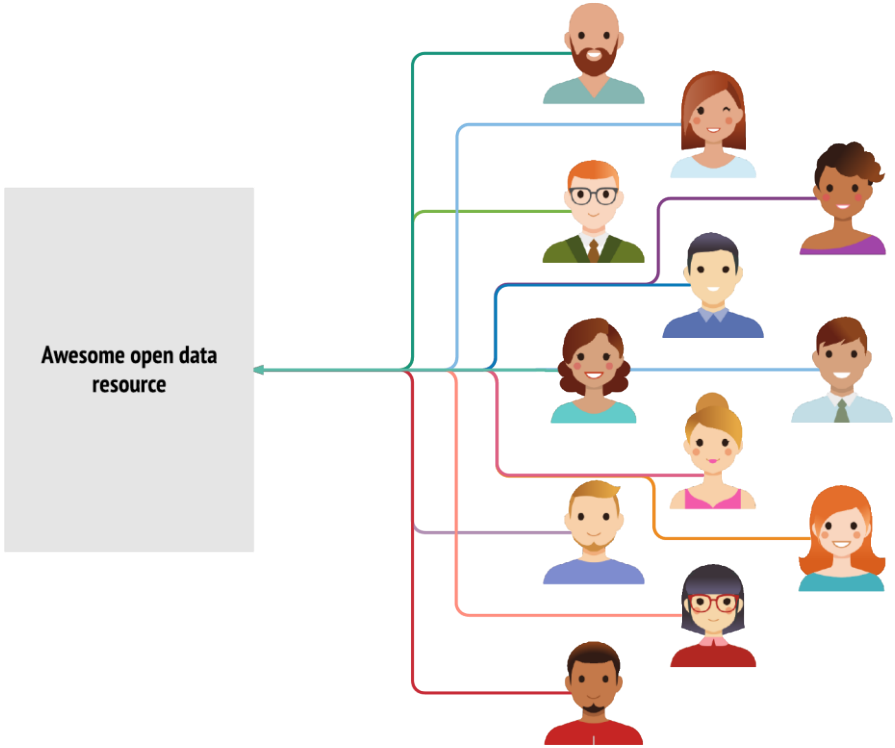
Examples

- Mysterious 0-width whitespace
- Inconsistent number of columns in the same file
- Abstruse file names
- No clear primary identifiers, so it's nearly impossible to integrate
- Special characters that make your script fail inexplicably

UNIVERSITY OF CAMBRIDGE

InterMine

We shouldn't have to do the same task again & again



Every single person who downloads a data set may end up tidying / integrating the same data sources. That's a lot of unnecessary duplication.

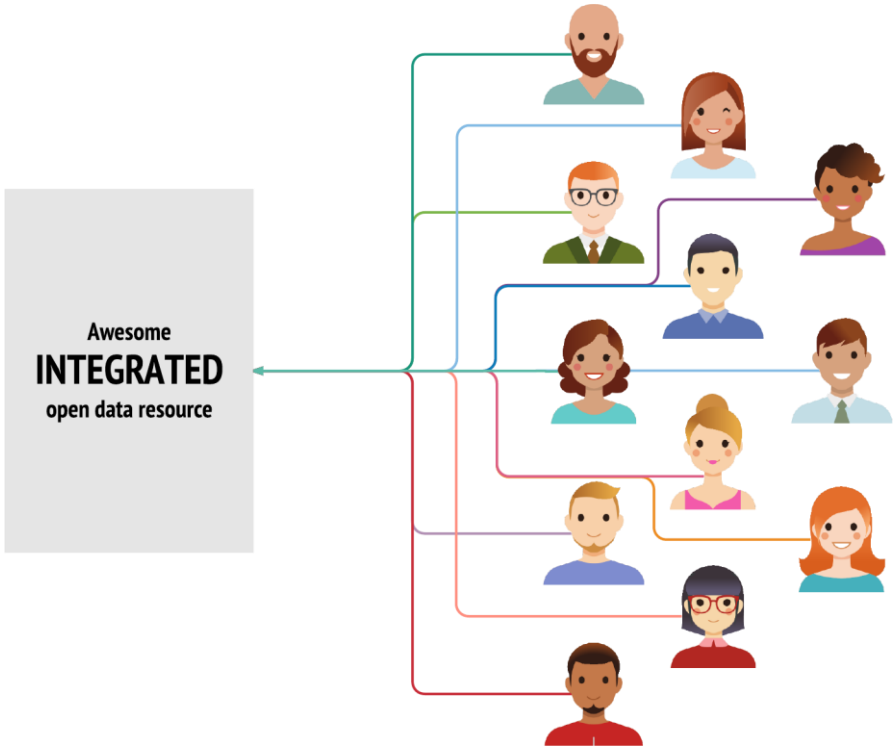
All these people are a tiny bit sad, because they all had to integrate & tidy the same data

Avatar images created by Freepik.com

Gos Micklem

Enter InterMine

We can do better!



Take the pain out of:

1. Integrating the data
2. People repeatedly tidying up the same data individually.

Happy people with pre-integrated easy-to query data

Avatar images created by Freepik.com

In 2002...

...we started integrating Fruit Fly data in FlyMine

Open Source

LGPL 2.1

- This licence is open source, so anyone can:
  - See our code
  - Modify our code
  - Re-use our code
- Unlike some open source licences, it has weak copyleft: proprietary applications can use InterMine.

UNIVERSITY OF CAMBRIDGE

InterMine

FlyMine -> InterMine

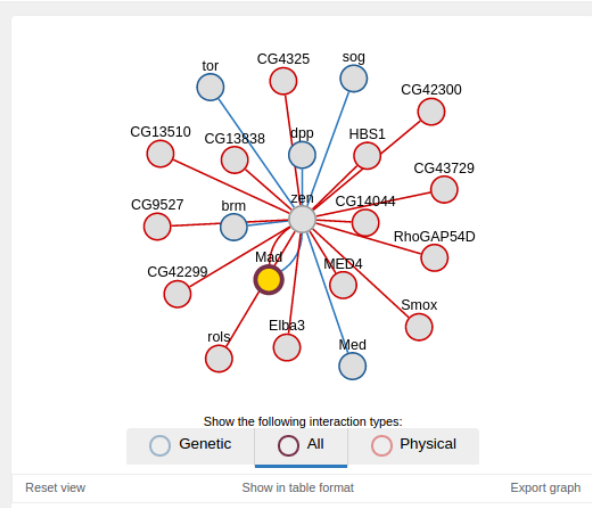
Next, we made it generic...
(no longer just for flies!)

UNIVERSITY OF CAMBRIDGE

InterMine

Over the years, InterMine added features

Data visualisations



Web Services
+ client libraries

python™

POWERED BY Perl

JavaScript

R

Ruby PROGRAMMING Language

Java

UNIVERSITY OF CAMBRIDGE

InterMine

.... and one InterMine became many!

There are around 30 public InterMine instances worldwide.




FlyMine - *Drosophila* genomics
modMine - fly and worm modENCODE data
MouseMine - at MGI
RatMine - at RGD
WormMine - at WormBase
YeastMine - at SGD
ZebrafishMine - at ZFIN
INDIGOmine - microbes
toxoMine - Toxoplasma gondii
ThaleMine - Araport Project with data for Arabidopsis thaliana

TargetMine - drug target discovery
MitoMiner - proteomic data for mitochondria
HumanMine - human
FlyTF.org - *Drosophila* transcription factors
PhytoMine - plants
MedicMine - *Medicago truncatula*
BovineMine - *Bos Taurus*
HymenopteraMine - Bees, Ants & Wasps
SoyMine - Soybase soy bean data

BeanMine - LegFed chado bean data
LegumeMine - String bean, Soy, and Peanut
PeanutMine - Peanut chado/GFF data
Shaare - Gene candidate prioritisation
PlanMine - Planarian flatworms
Wheat3BMine - Wheat chromosome 3B
GrapeMine - Grapevine
RepetDB - repetitive DNA elements
XenMine - Xenopus
TetraMine - *Tetrahymena thermophila*

See them all at `http://registry.intermine.org/`





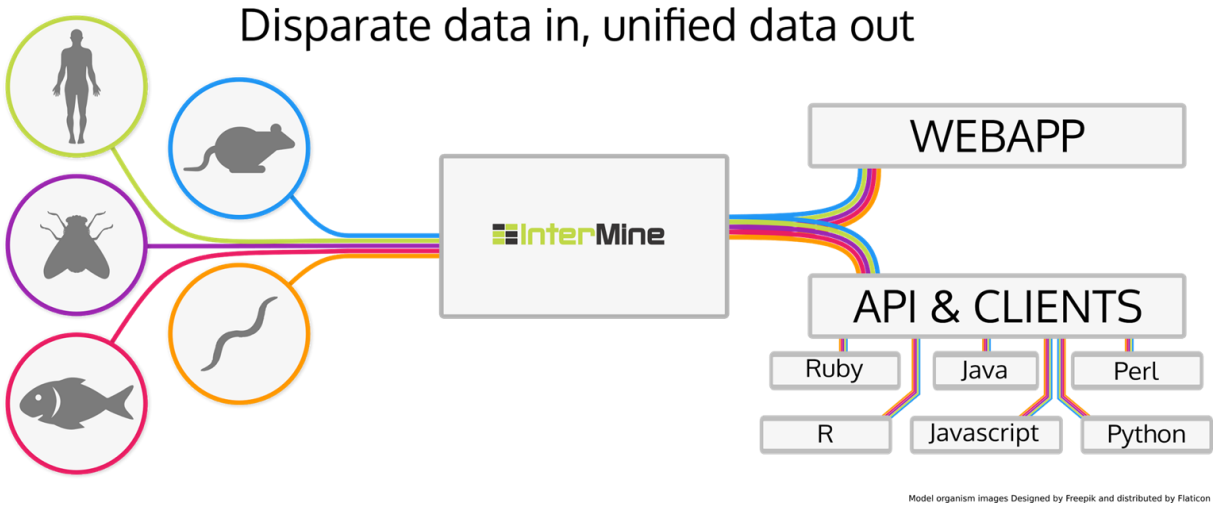
What is InterMine nowadays?

Open Source
Data Warehouse

InterMine

github.com/intermine

Star 83 Fork 282



UNIVERSITY OF CAMBRIDGE

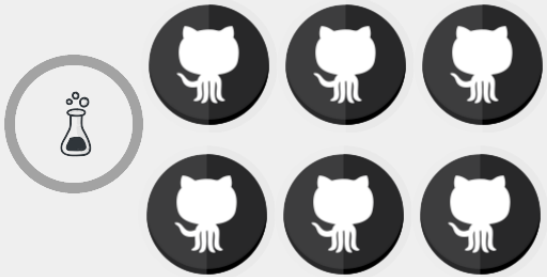
InterMine

Who Are We?

InterMine

UNIVERSITY OF CAMBRIDGE

Gos Micklem's group



Six devs, one biologist

- Started 2002, funded to 2021
- InterMine is an open source data warehouse and analysis system

Without **open data**,
InterMine probably
wouldn’t exist.

Thank you!

InterMine

intermine

info@intermine.org

intermine.org

@intermineorg

@yoyehudi