iRODS functionality within the Grassroots Infrastructure

Simon Tyrrell, Xingdong Bian and Robert P. Davey
Earlham Institute, Norwich Research Park, Norwich, NR4 7UZ, UK
http://www.earlham.ac.uk/

Earlham Institute

Background

Grassroots is part of the Wheat Information System (WheatIS) to build a system that responds to the needs of the international wheat community

- Promotion of an open-access model for data exchange
- Reliance on a distributed system
- Facilitate sharing data and tools
- Promotion of the visibility of each participating platform to contribute to their sustainability.

Taken from the Wheat Information System website http://wheatis.org

Challenges

- Geographic disparity
  - Researchers spread out across the world
- Code reusability
  - Each set of developers re-inventing the wheel
- Data interoperability
  - Different custom formats for storing data
- Service interoperability
  - Connecting similar services
  - Allow data to be shared between services when possible

Goals

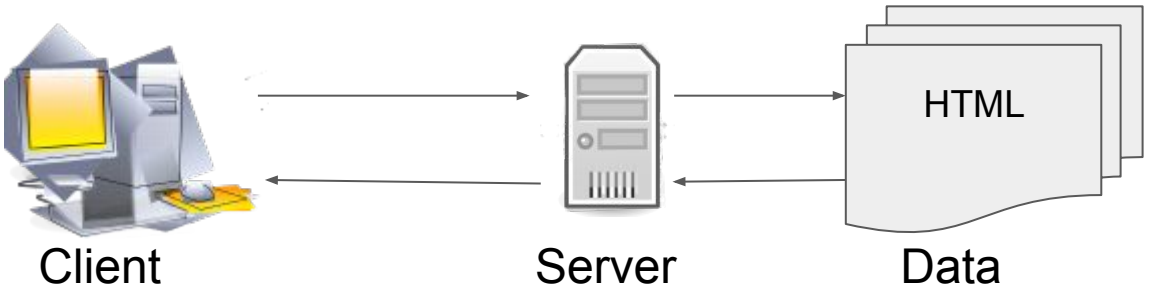
An infrastructure for a distributed set of servers to transparently share:

- Data
  - Well-described
  - Reuse
  - Federation
- Services
  - Can be added to analysis tools and pipelines
  - Integration

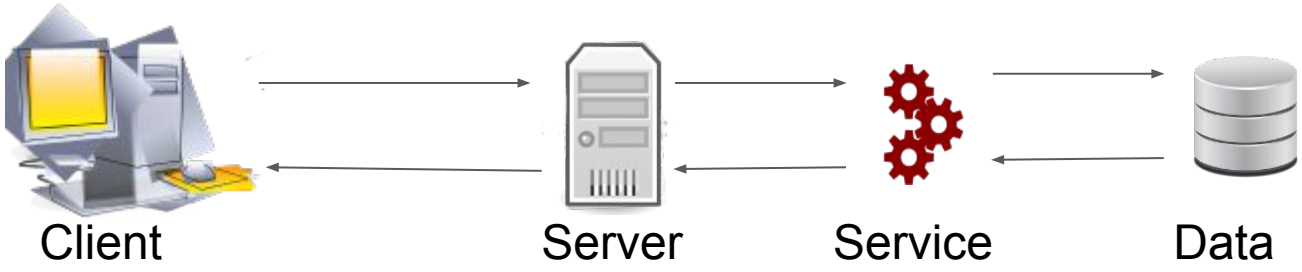
And make it as user-friendly as possible!

Typical Web Server-Client Interaction

- User requests data from a web server such as static html pages...



- … or dynamically-generated content such as BLAST or text search results via a Web Service.



Grassroots Server Infrastructure - Apache

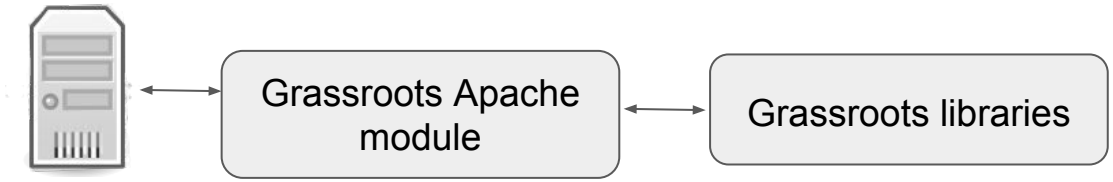
Apache httpd
Web Server

Apache httpd is the most commonly used Web Server

- Open source
- Very configurable
- Robust
- Widely supported
- Easily extensible by adding functionality as modules such as *e.g.*
  - SSL for secure connections
  - Authorisation and Authentication
  - CGI scripts

Server Infrastructure - Grassroots

Apache httpd
Web Server



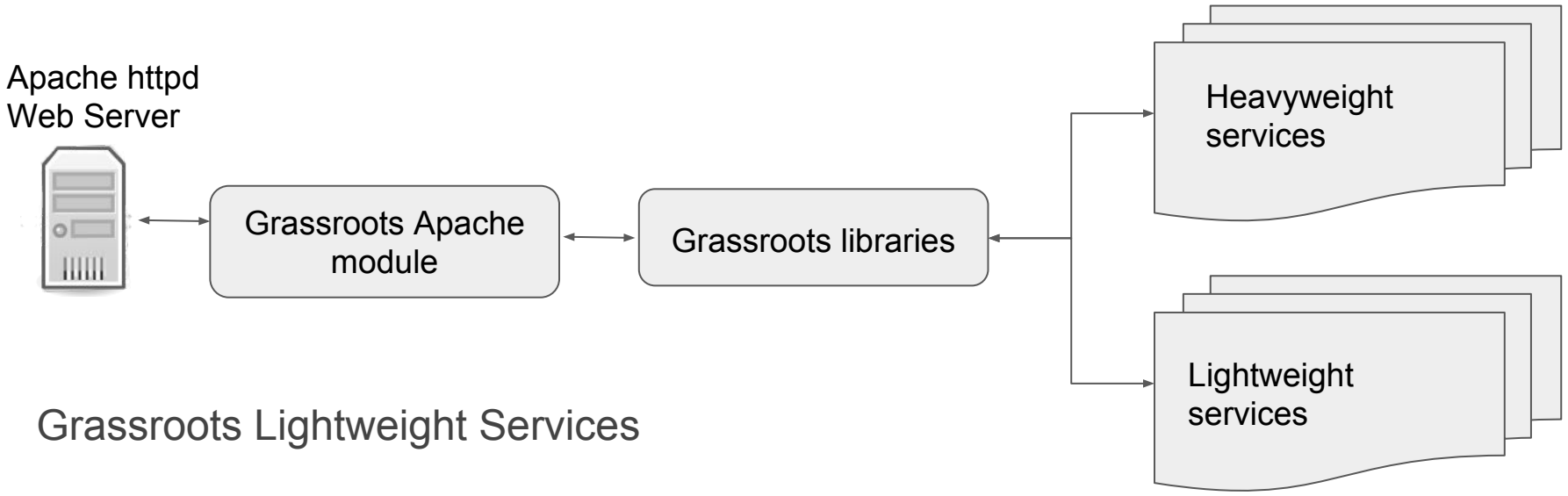
- Grassroots Apache module acts as a bridge between Apache and the Grassroots Infrastructure.
- A set of cross-platform libraries that can be used with the Apache web server via a Grassroots module including
  - Networking code to access code and services across the web
  - Server and Service management tools
  - Standardising access to/from our web services and their parameters
  - Read and write data from different resources *e.g.*
    - iRODS
    - Local files
    - Dropbox
    - Google drive
- Can run bespoke Grassroots Services to access and process data

Server Infrastructure - Heavyweight Services

Grassroots Heavyweight Services

- Programmer-level tools that conform to the Grassroots Services API, which is a strict set of standards to access underlying tools and data
  - BLAST
  - iRODS Search
  - Field Pathogenomics
  - SamTools

Server Infrastructure - Lightweight Services



Grassroots Lightweight Services

- Structured text files
- Scripts that use Grassroots libraries to access information from other web services *e.g.*
  - Call web searches and aggregate results

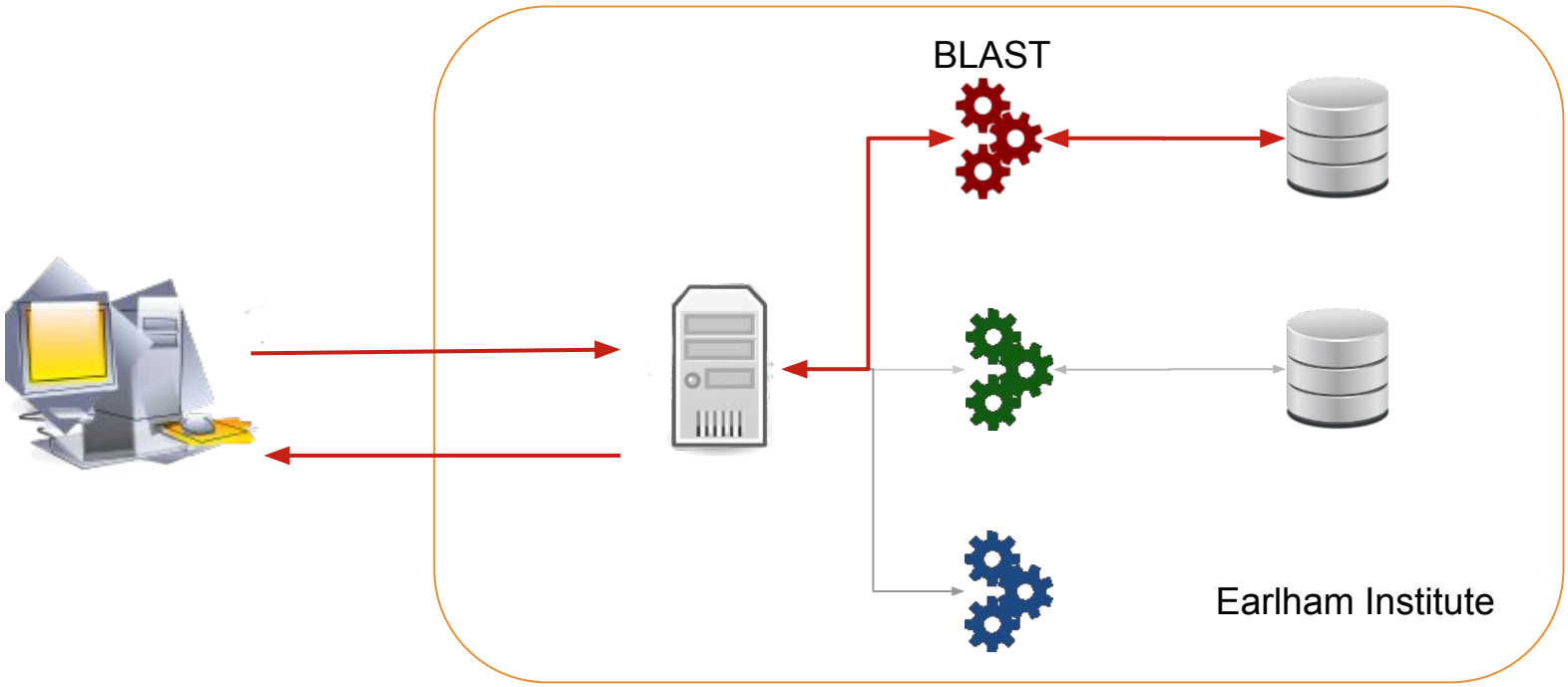
Grassroots architecture

- **Platform and programming language independent**
  - **Use any architecture that can produce and consume Grassroots information**
  - **Clear and easy JSON schema**

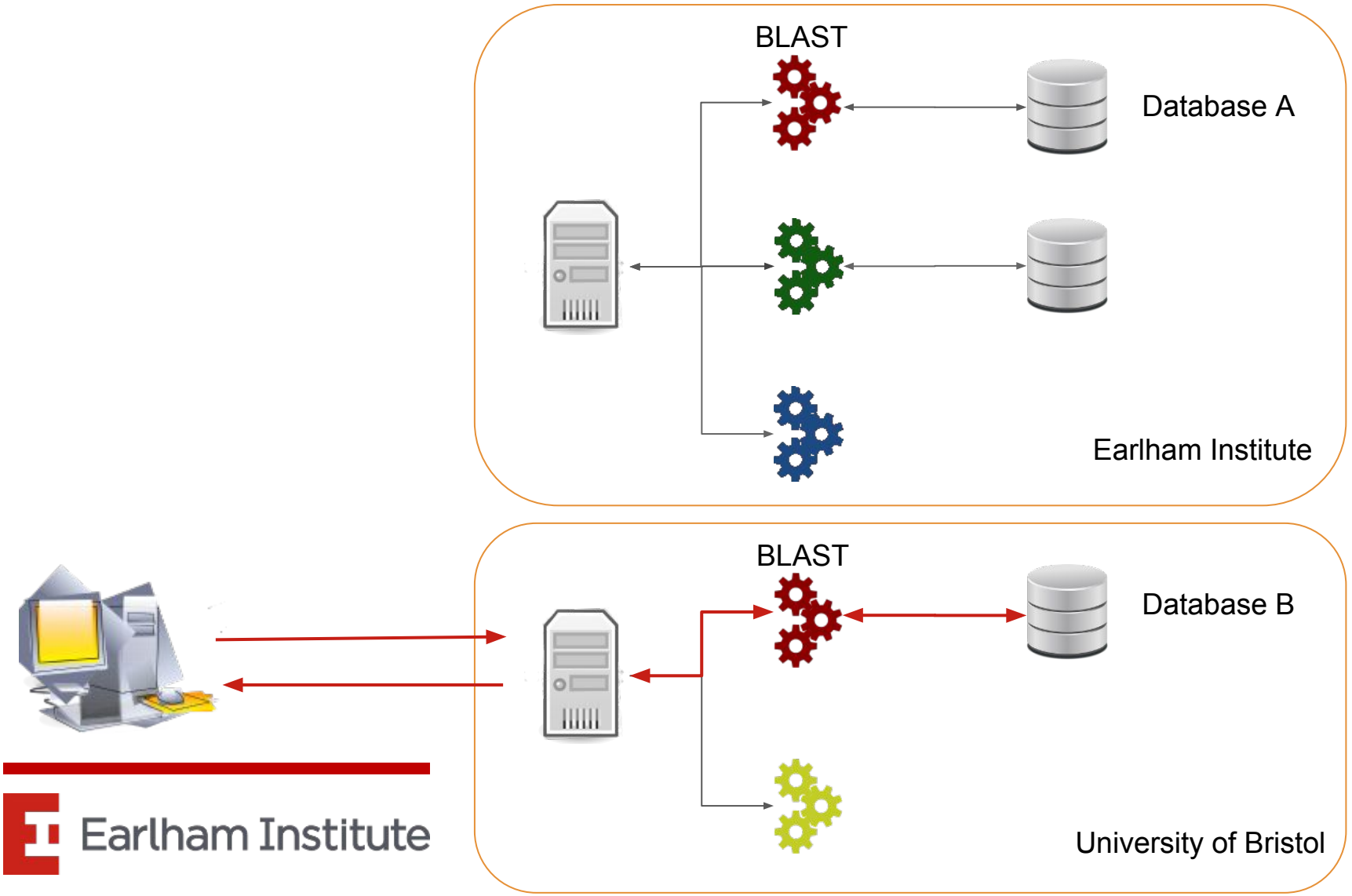
Grassroots architecture

- Platform and programming language independent
  - Use any architecture that can produce and consume Grassroots information
  - Clear and easy JSON schema
- **Distributed information exchange**
  - **Built upon interconnected web servers and services**
  - **Requires production and consumption of standardised information**
  - **Communicate through standardised REST API**

Run a Service



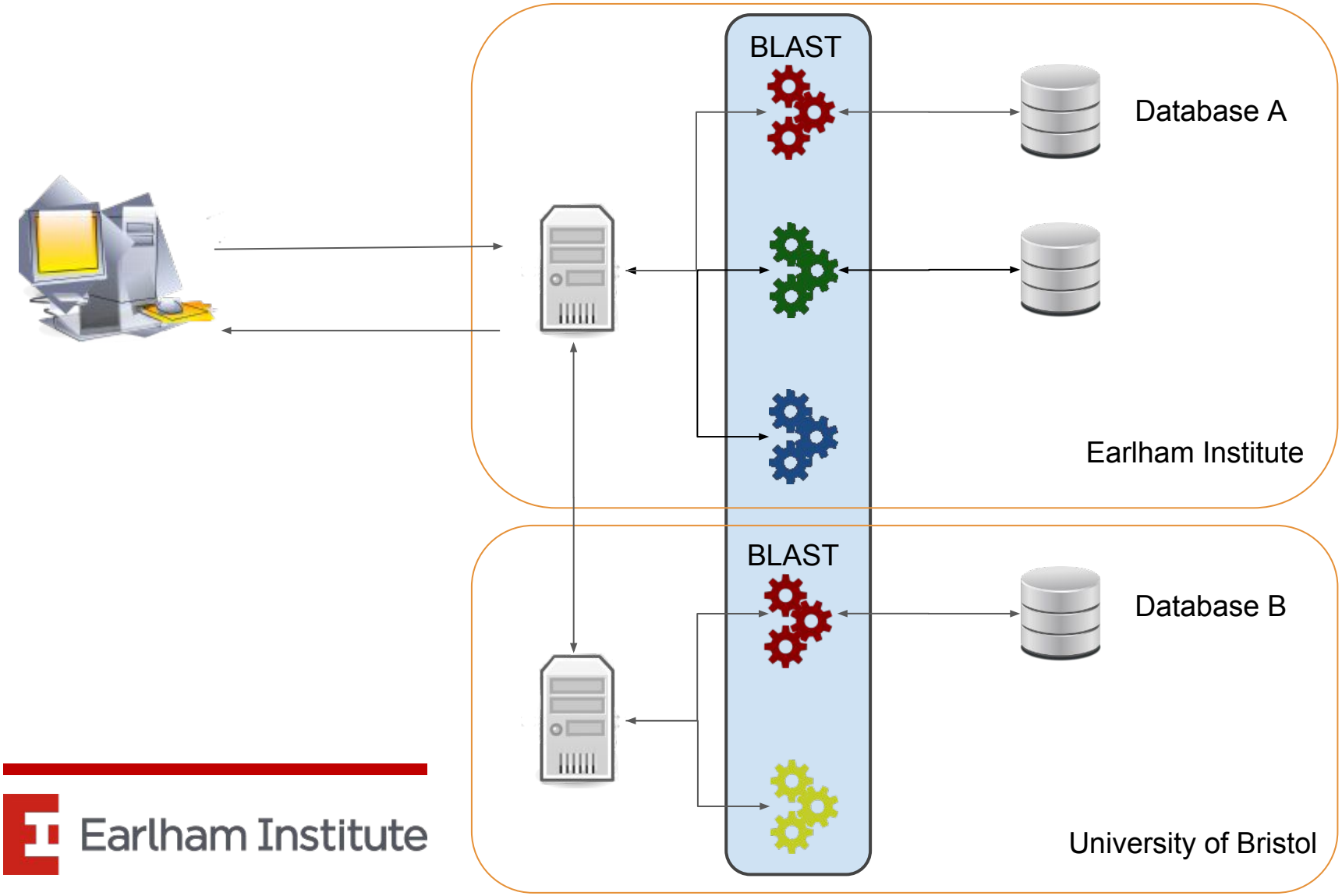
Run the same Service on another Server



Issues

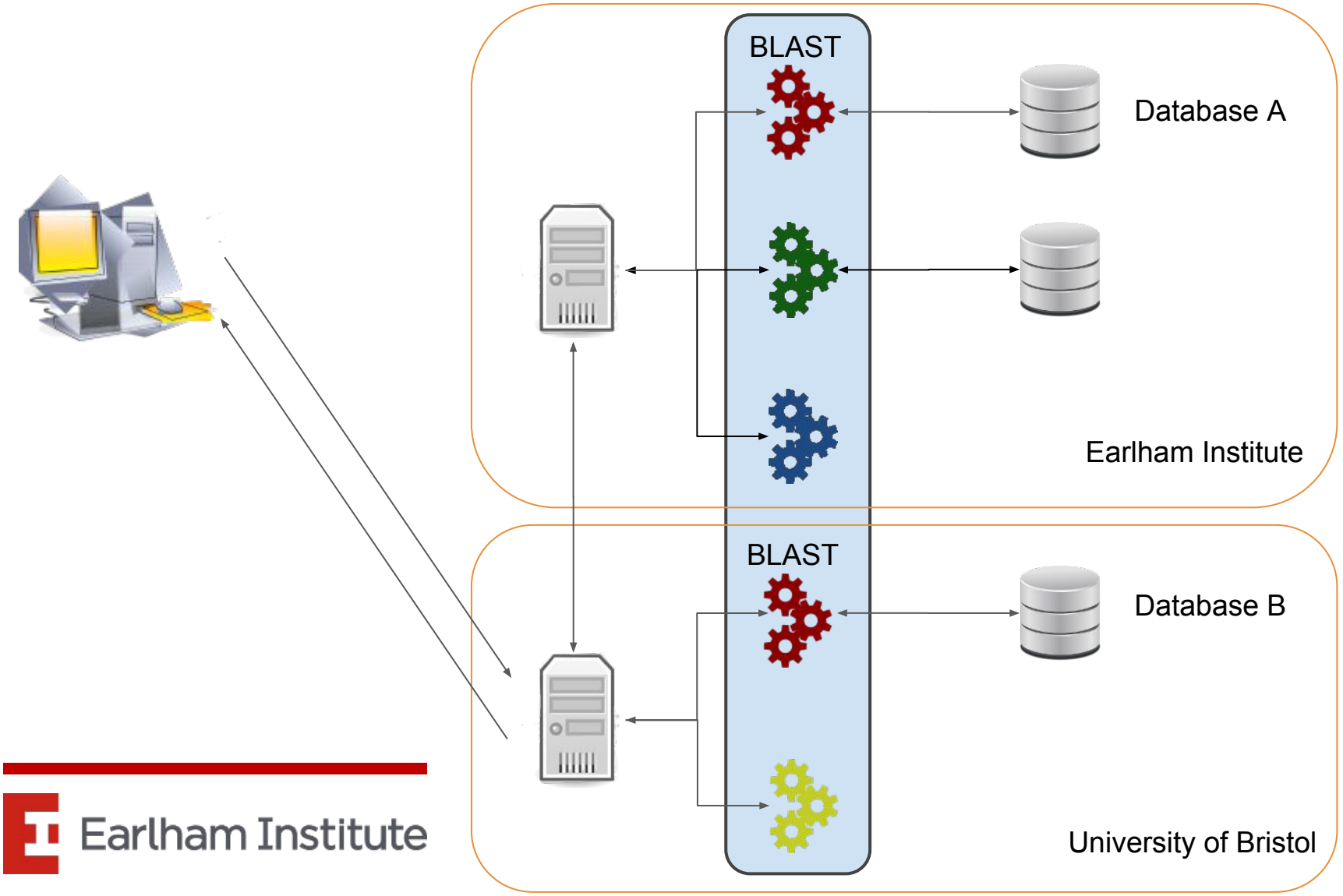
- Manually having to access each Service individually
- Collation of results
- Human error
  - Not running each service with the same parameters
  - Mistakes when putting the results together
- Time consuming

Distributed Services



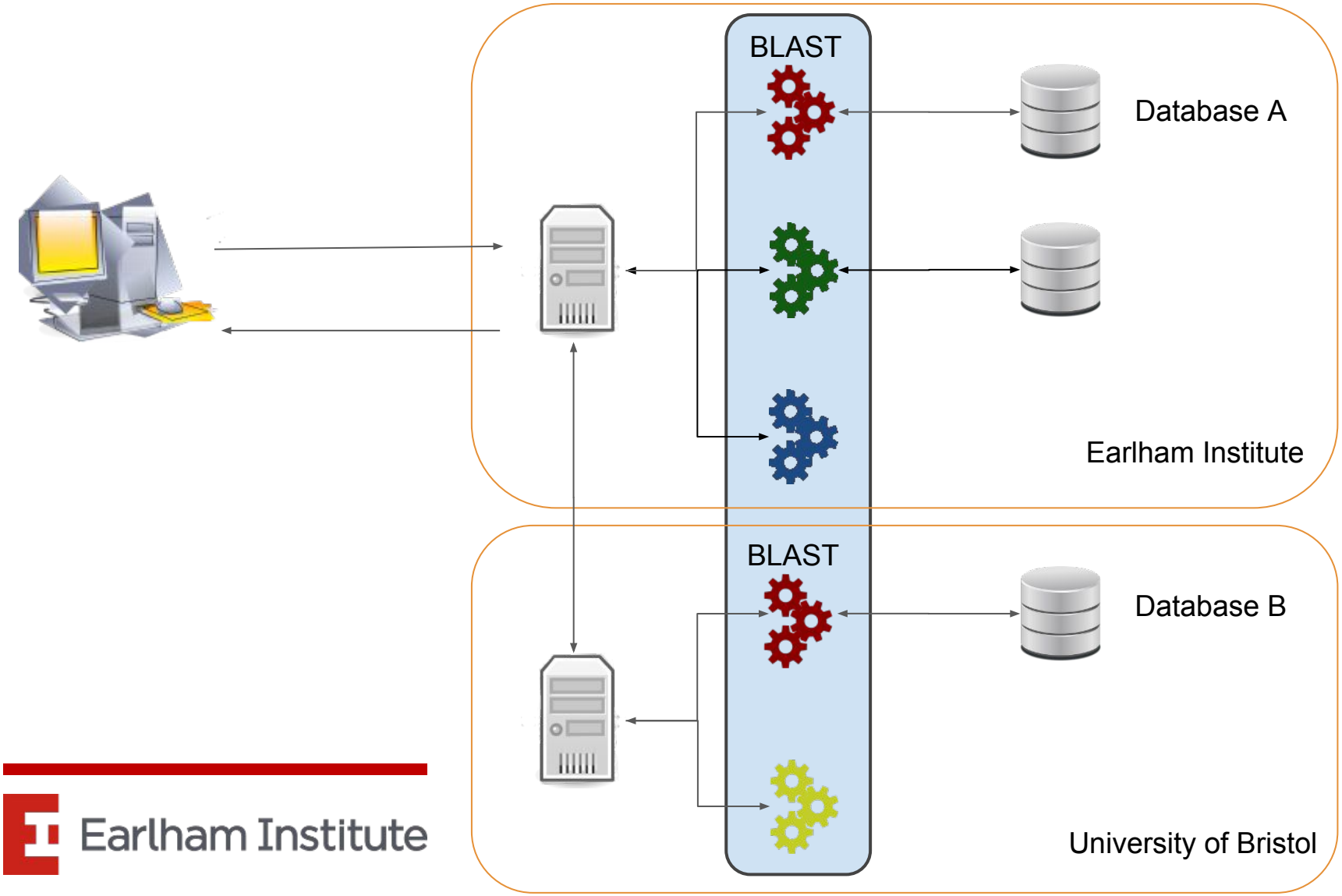
Earlham Institute

Different Server, Same List of Services



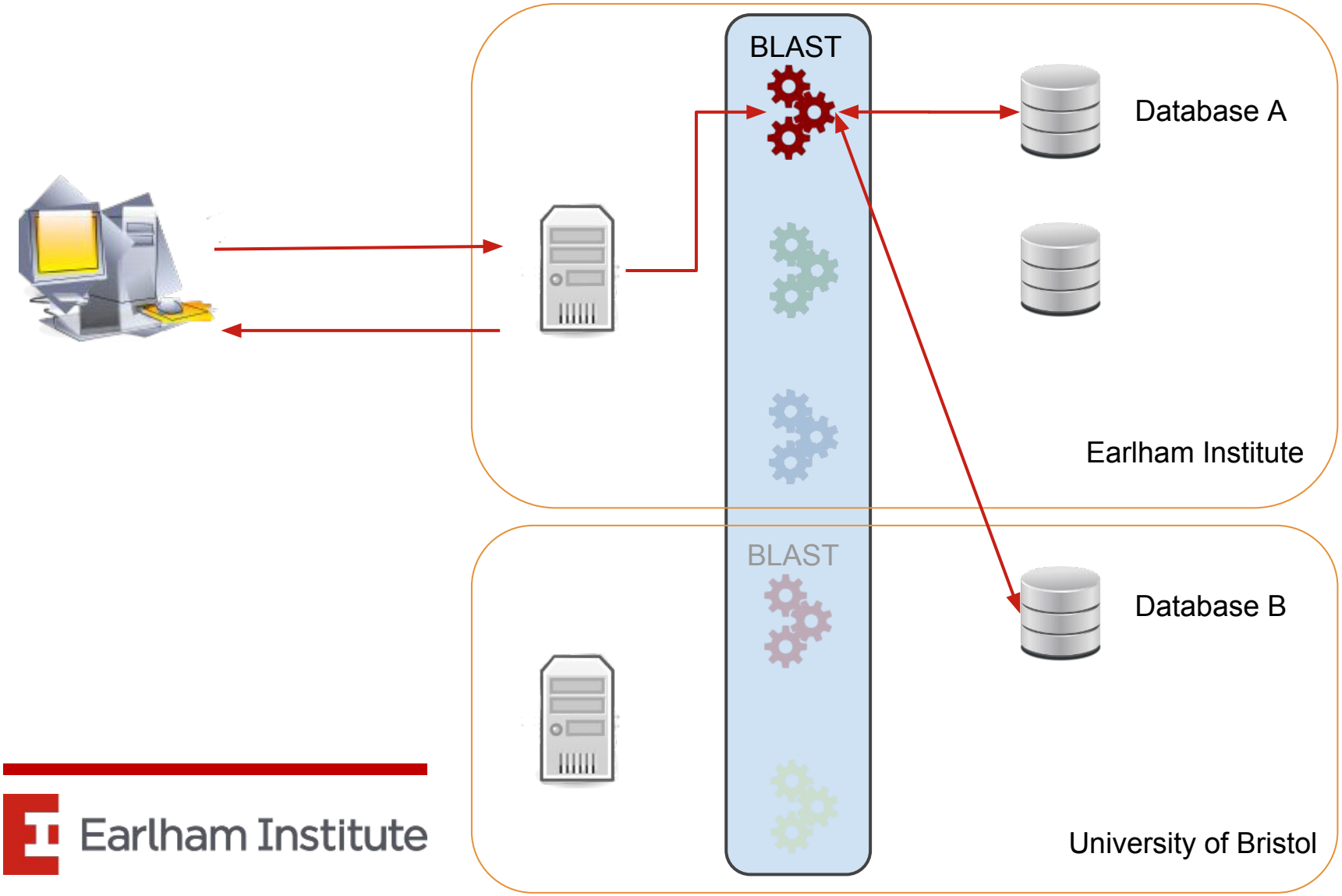
Earlham Institute

Duplicated Services...



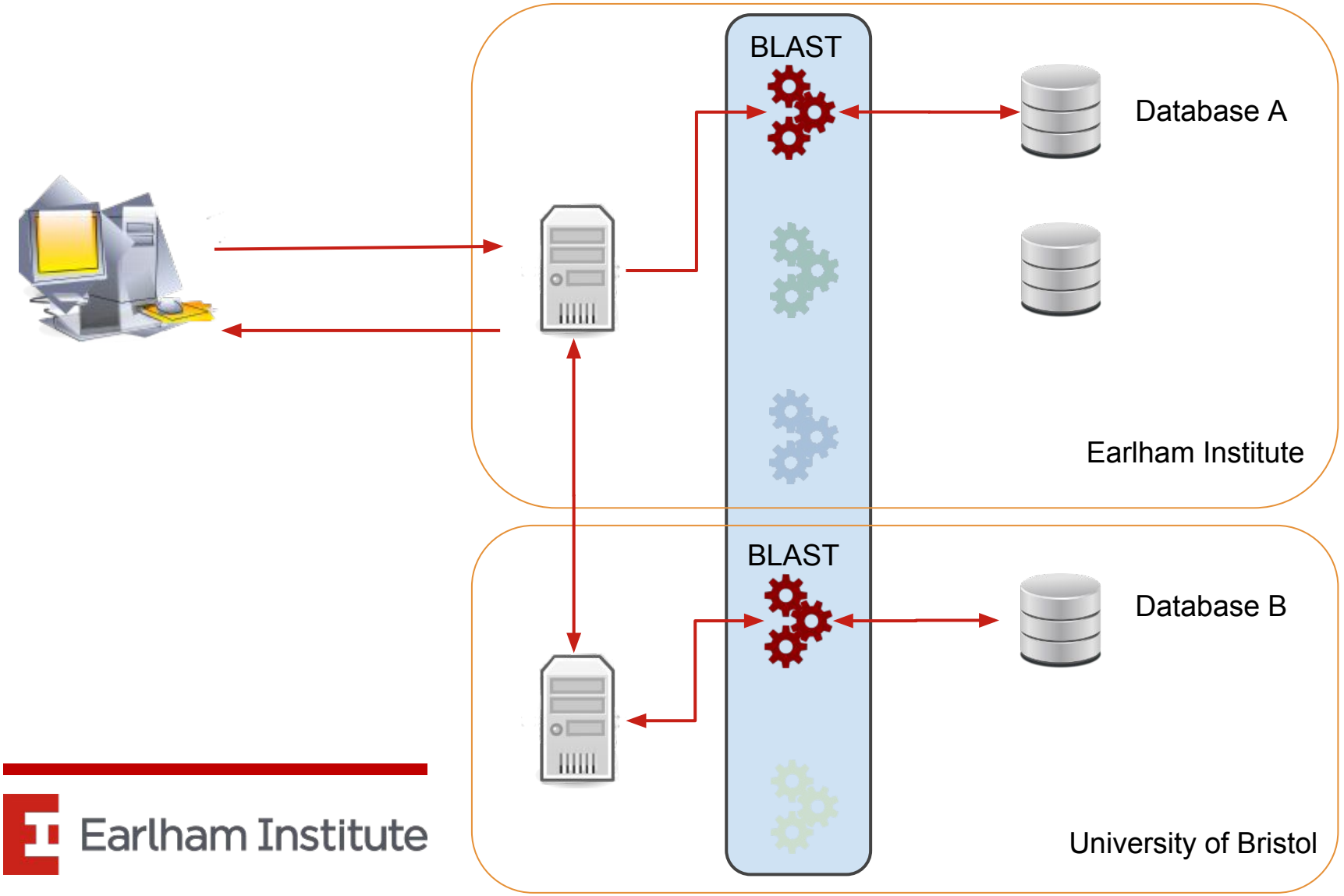
Earlham Institute

… Get Amalgamated



Earlham Institute

Consolidate Services - Under the hood



Earlham Institute

Issues running further Services

- Manually having to extract relevant values from each set of results
- Human error
  - Not running each service with the same parameters
  - Mistakes when putting the results together
- Time consuming

Running Further Services

```
Database: databases/Triticum_aestivum_CS42_TGACv1_scaffold.annotation.gff3.cds.fa

> TRIAE_CS42_6DS_TGACv1_542925_AA1732620.1   gene=TRIAE_CS42_6DS_TGACv1_542925_AA1732620
Length=1674

 Score = 159 bits (198),  Expect = 2e-38
 Identities = 100/101 (99%), Gaps = 0/101 (0%)
 Strand=Plus/Minus

Query  1     CTGTAGATGTGCACCTTGATGGTATCCTCGGCGATGAGCTCGAAGACGCAAACNTCGAAC  60
             |||||||||||||||||||||||||||||||||||||||||||||||||||| ||||||
Sbjct  1610  CTGTAGATGTGCACCTTGATGGTATCCTCGGCGATGAGCTCGAAGACGCAAACATCGAAC  1551

Query  61    TTCTCCAGATTGTTGCCGATCGAGAACTGGCTCCAGCCTCT  101
             |||||||||||||||||||||||||||||||||||||||||
Sbjct  1550  TTCTCCAGATTGTTGCCGATCGAGAACTGGCTCCAGCCTCT  1510

Lambda     K         H
   0.634   0.408     0.912

Gapped
Lambda     K         H
   0.550   0.210     0.460
```

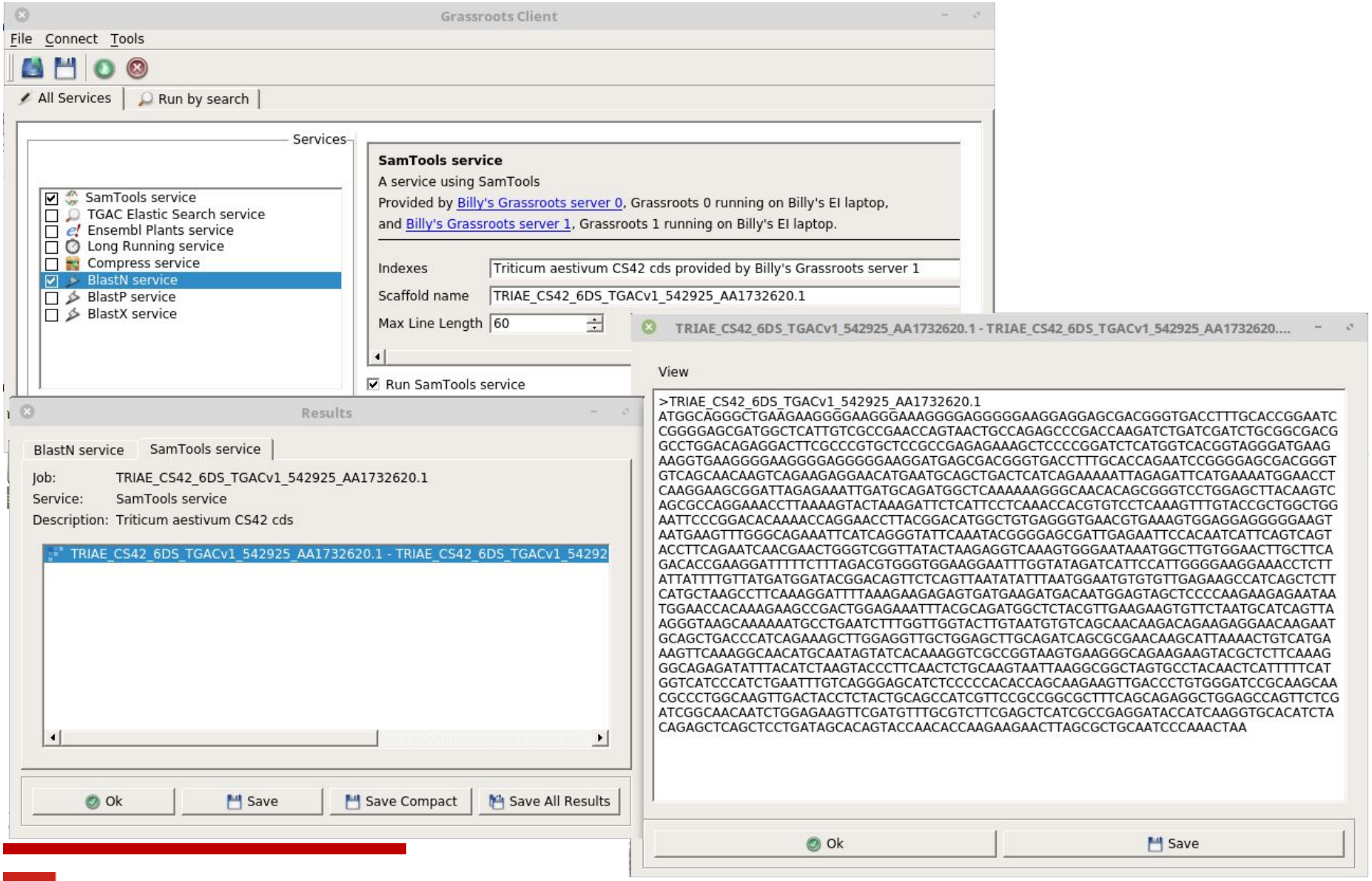
Earlham Institute

... Parse Service Output...

```
Database: databases/Triticum_aestivum_CS42_TGACv1_scaffold.annotation.gff3.cds.fa

> TRIAE_CS42_6DS_TGACv1_542925_AA1732620.1
Length=1674

 Score = 159 bits (198),  Expect = 2e-38
 Identities = 100/101 (99%), Gaps = 0/101 (0%)
 Strand=Plus/Minus

Query  1     CTGTAGATGTGCACCTTGATGGTATCCTCGGCGATGAGCTCGAAGACGCAAACNTCGAAC  60
             |||||||||||||||||||||||||||||||||||||||||||||||||||| ||||||
Sbjct  1610  CTGTAGATGTGCACCTTGATGGTATCCTCGGCGATGAGCTCGAAGACGCAAACATCGAAC  1551

Query  61    TTCTCCAGATTGTTGCCGATCGAGAACTGGCTCCAGCCTCT  101
             |||||||||||||||||||||||||||||||||||||||||
Sbjct  1550  TTCTCCAGATTGTTGCCGATCGAGAACTGGCTCCAGCCTCT  1510

Lambda     K         H
   0.634   0.408     0.912

Gapped
Lambda     K         H
   0.550   0.210     0.460
```

Earlham Institute

… To Run Another Service Automatically



Earlham Institute

Grassroots architecture

- Platform and programming language independent
  - Use any architecture that can produce and consume Grassroots information
  - Clear and easy JSON schema
- Distributed information exchange
  - Built upon interconnected web servers and services
  - Requires production and consumption of standardised information
  - Communicate through standardised REST API
- Run computational tasks through local/HPC services
- **Semantic metadata support**
  - **Ontologies / controlled vocabularies**
  - **Data description consistency**

Example Ontology data


```
"@context" : "http://schema.org",
"Date collected" : {
        "@type" : "Date",
        "date" : "2013-05-16"
},
"Name/Collector" : {
        "@type" : "Person",
        "name" : “Lemmy"
},
"Company" : {
        "@type" : "Organization",
        "name" : "FooBar Inc"
},
"location" : {
        "location" : {
                "@type" : "GeoCoordinates",
                "latitude" : 53.0668342,
                "longitude" : -0.5540889
        },
                "north_east_bound" : {
                        "@type" : "GeoCoordinates",
                        "latitude" : 53.0703866,
                        "longitude" : -0.5396723
                },
                "south_west_bound" : {
                        "@type" : "GeoCoordinates",
                        "latitude" : 53.0551367,
                        "longitude" : -0.5623362
                }
},
"Address" : {
        "@type" : "PostalAddress",
        "postalCode" : "LN5 0QG",
        "addressLocality" : "Welbourn",
        "addressRegion" : "Lincolnshire",
        "addressCountry" : "GB"
}
```

Example Ontology data

```
"@context" : "http://schema.org",
"Date collected" : {
        "@type" : "Date",
        "date" : "2013-05-16"
},
"Name/Collector" : {
        "@type" : "Person",
        "name" : “Lemmy"
},
"Company" : {
        "@type" : "Organization",
        "name" : "FooBar Inc"
},
"location" : {
        "location" : {
                "@type" : "GeoCoordinates",
                "latitude" : 53.0668342,
                "longitude" : -0.5540889
        },
                "north_east_bound" : {
                        "@type" : "GeoCoordinates",
                        "latitude" : 53.0703866,
                        "longitude" : -0.5396723
                },
                "south_west_bound" : {
                        "@type" : "GeoCoordinates",
                        "latitude" : 53.05513670000001,
                        "longitude" : -0.5623362
                }
},
"Address" : {
        "@type" : "PostalAddress",
        "postalCode" : "LN5 0QG",
        "addressLocality" : "Welbourn",
        "addressRegion" : "Lincolnshire",
        "addressCountry" : "GB"
}
```

Other Ontologies

- Schema.org
  - A collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet[1]
- Sequence Ontology
  - A collaborative ontology project for the definition of sequence features used in biological sequence annotation[2]
- EDAM Ontology
  - A simple ontology of well established, familiar concepts that are prevalent within bioinformatics[3]
- FALDO
  - Lightweight interval-based genomic feature descriptors[4]

1. Taken from http://schema.org/
2. Taken from http://www.sequenceontology.org/
3. Taken from http://edamontology.org/page
4. Taken from https://github.com/JervenBolleman/FALDO

Grassroots architecture

- Platform and programming language independent
  - Use any architecture that can produce and consume Grassroots messages
  - Clear and easy JSON schema
- Distributed information exchange
  - Built upon interconnected web servers and services
  - Requires production and consumption of standardised information
  - Communicate through standardised REST API
- Run computational tasks through local/HPC services
- Semantic metadata support
  - Ontologies / controlled vocabularies
  - Data description consistency
- **Extensible**
  - **Adding and integrating services**
    - **Programming a service conforming to JSON schema API**
    - **Or adding JSON description of a service using a generic API**

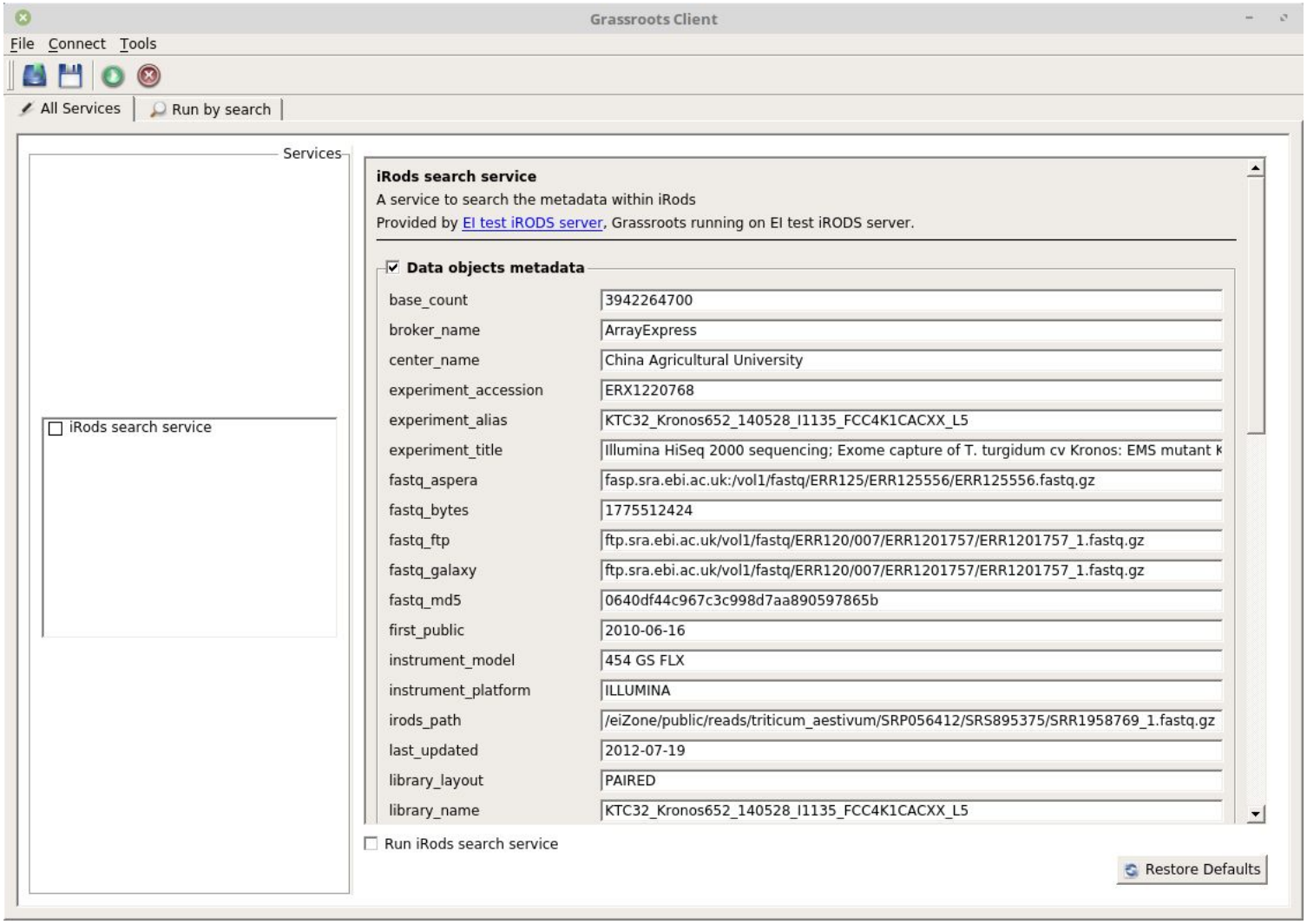
Services

- Installation is simply copying files into a given Grassroots folder
- Service can be configured by editing a text file
- No need to restart Apache to pick up any Service changes or additions
- Keyword-aware
    - Services know their data and how to interpret a general search term, similar to Google's search box.

Grassroots - iRODS metadata search service

- Expose iRODS data as a Grassroots service
- Expose all user-accessible metadata keys
- Get all possible values for each key

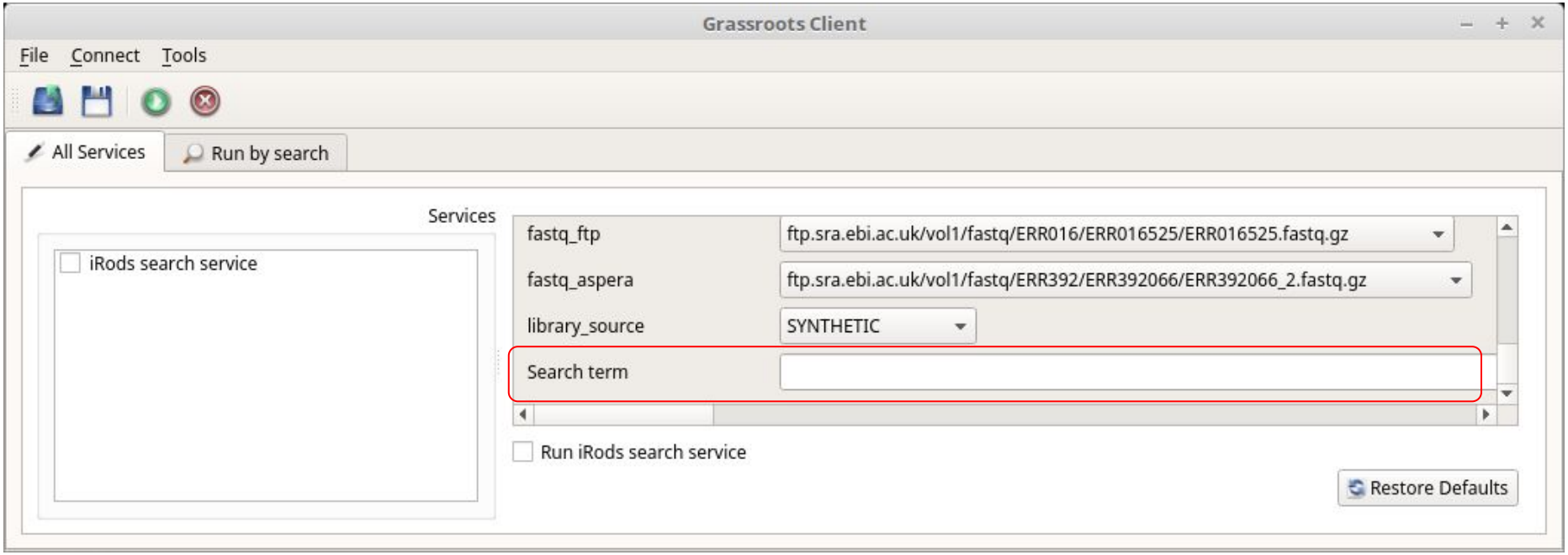
iRODS - Metadata search service



Earlham Institute

iRODS - Metadata search service

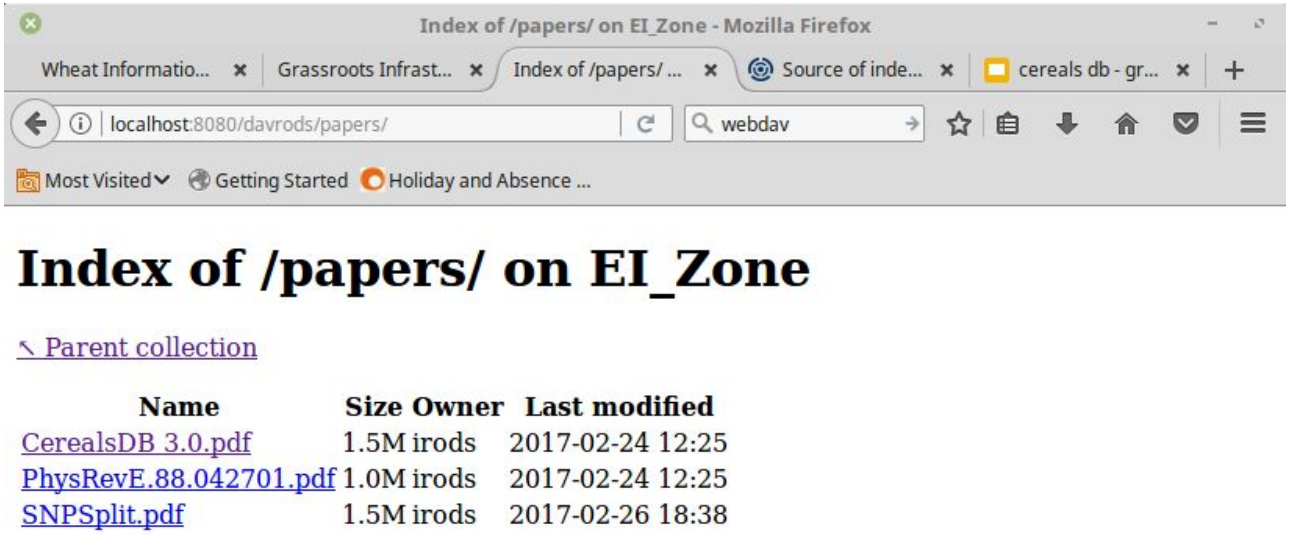
- Global keyword search across all metadata keys and values



iRODS - Davrods

Open source project available at https://github.com/UtrechtUniversity/davrods by Ton and Chris Smeele

- Apache module to access iRODS server
- Web Distributed Authoring and Versioning (WebDAV) interface to iRODS repositories



iRODS - Davrods customisation

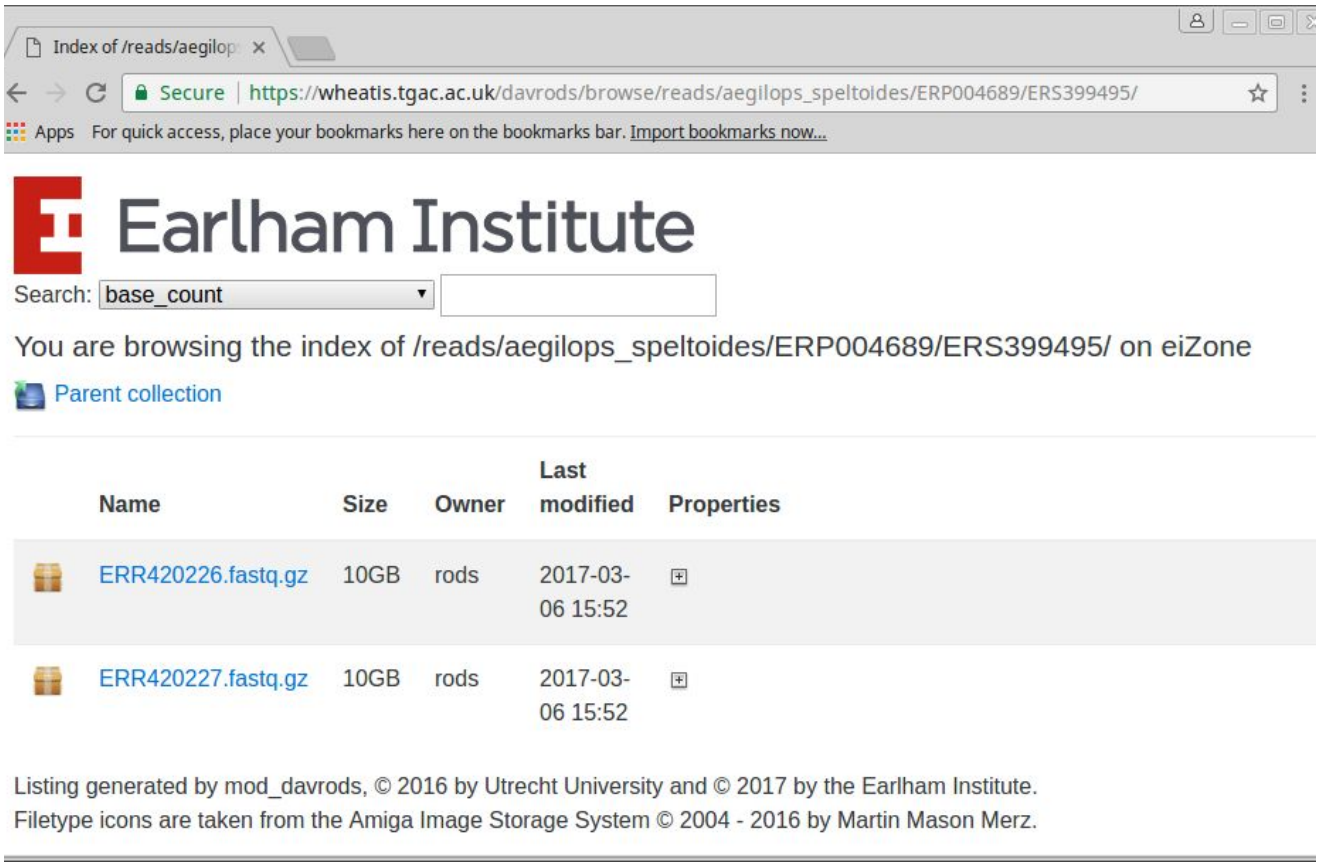
Open source project available at https://github.com/billyfish/davrods

- Themeable listings similar to mod_autoindex
- Metadata displayed
- Can be configured to make data public without the need to log in
- Public demo available at https://wheatis.tgac.ac.uk/davrods/browse/reads/

iRODS - Davrods themeable listings

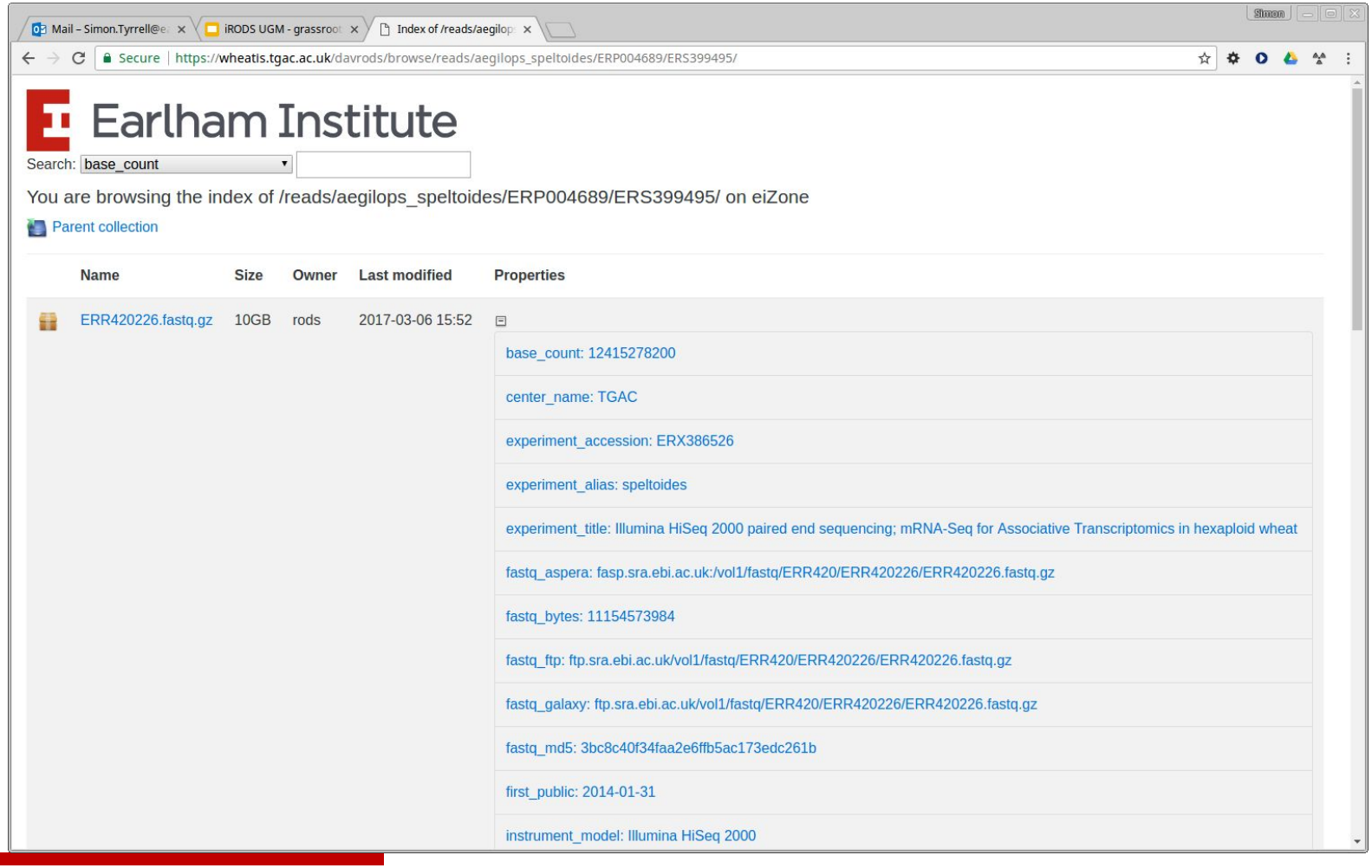
Added themeable listings

- Html <head> additions
- Header and footer around the listings
- Filetype icons



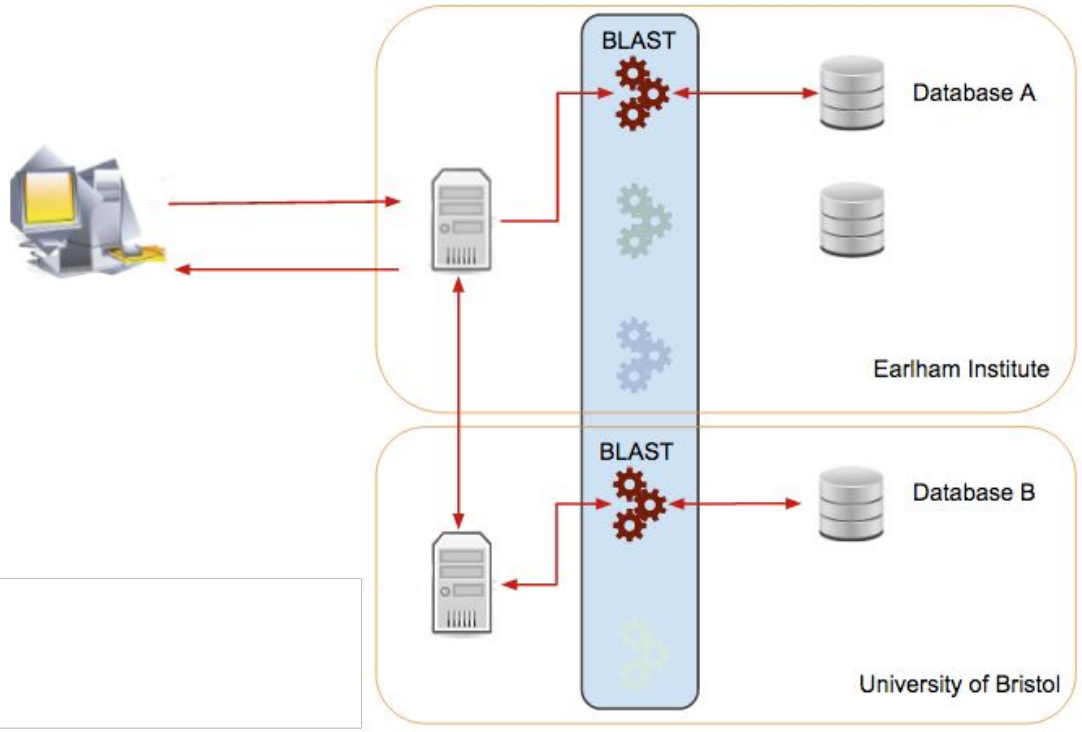
iRODS - Davrods metadata enhancements

- Expandable metadata key-value pairs as clickable links
- Search API and form



BLAST web service

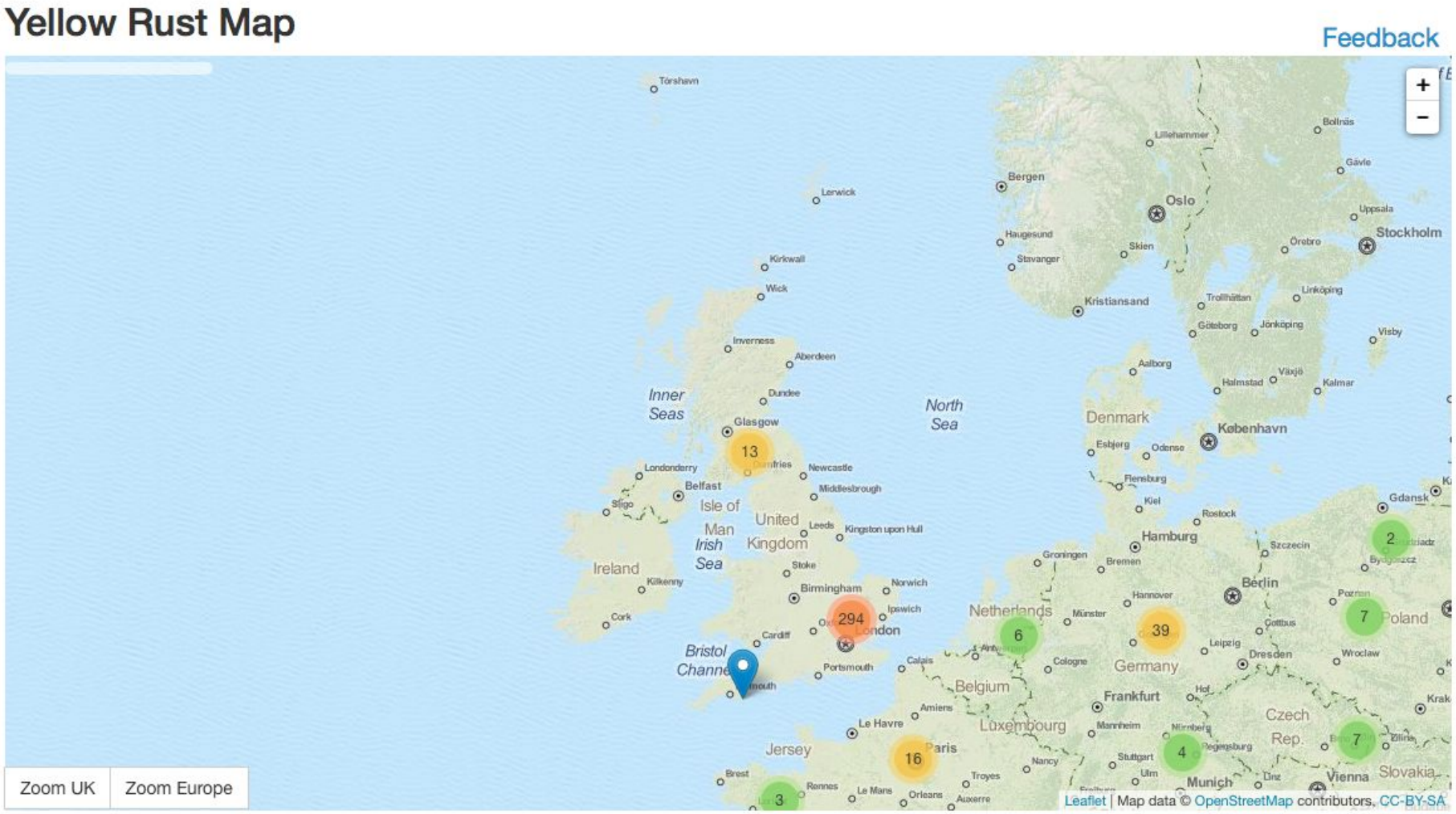
- Basic Local Alignment Search Tool (BLAST) finds regions of similarity between biological sequences.
- Utilises EI's high performance computing cluster to run jobs
- Enables high-throughput BLAST searches
- Across multiple databases
- Across multiple sites
  - University of Bristol
- Since 12 Nov 2015 launch
  - 8000+ page visits
  - More than 12000 jobs
- Dynamic web front-end
- Semantic marked up output



Field Pathogenomics Service

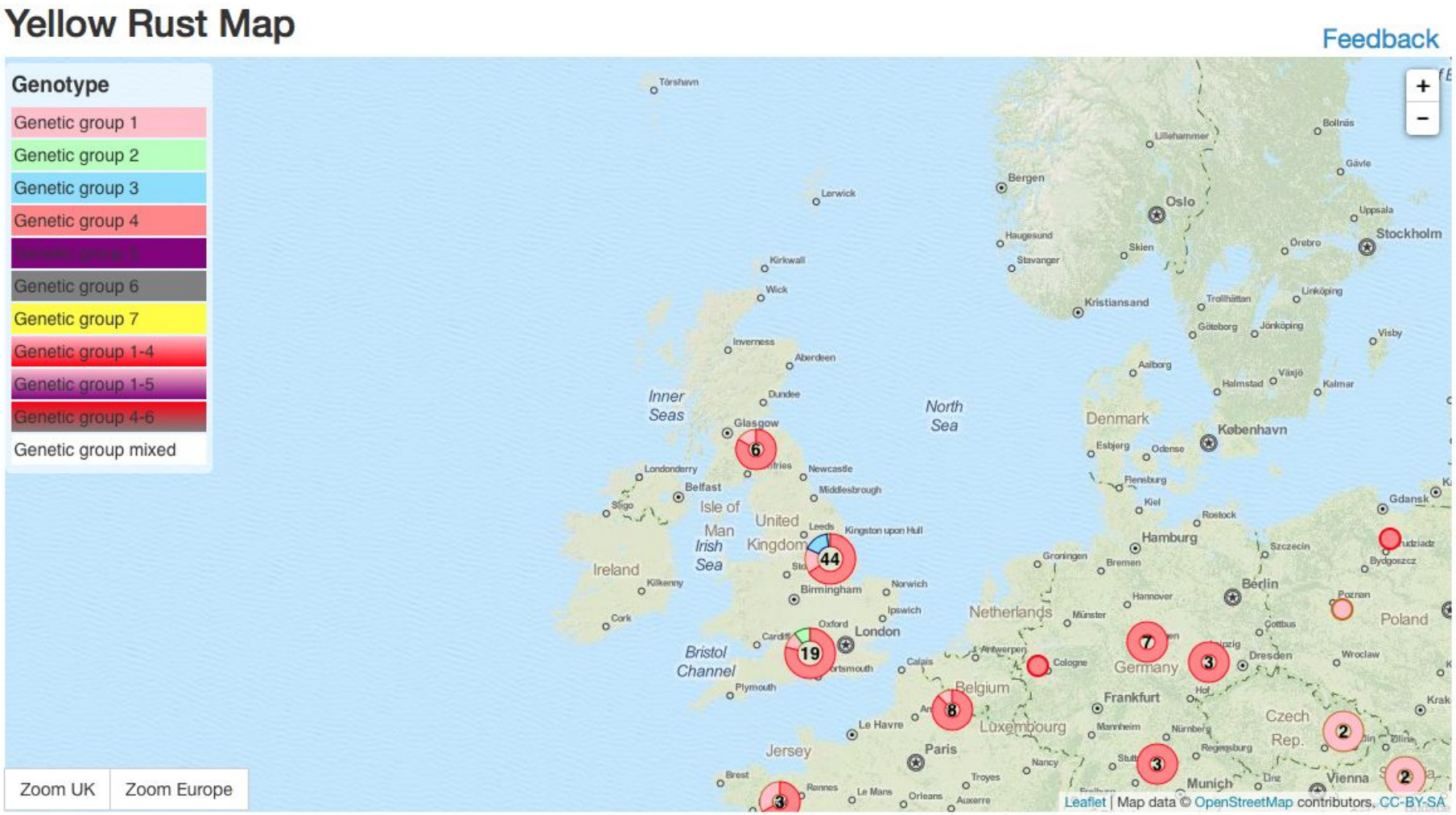
- Collaboration with Diane Saunders
- An intuitive and informative breeder's tool
- Tracking yellow rust pathogen spread over time and location
- Genotype data
  - Sequencing of infected varieties gives pathogen and host information
- Phenotype data
  - Scoring matrices of resistance

Field Pathogenomics - Samples view



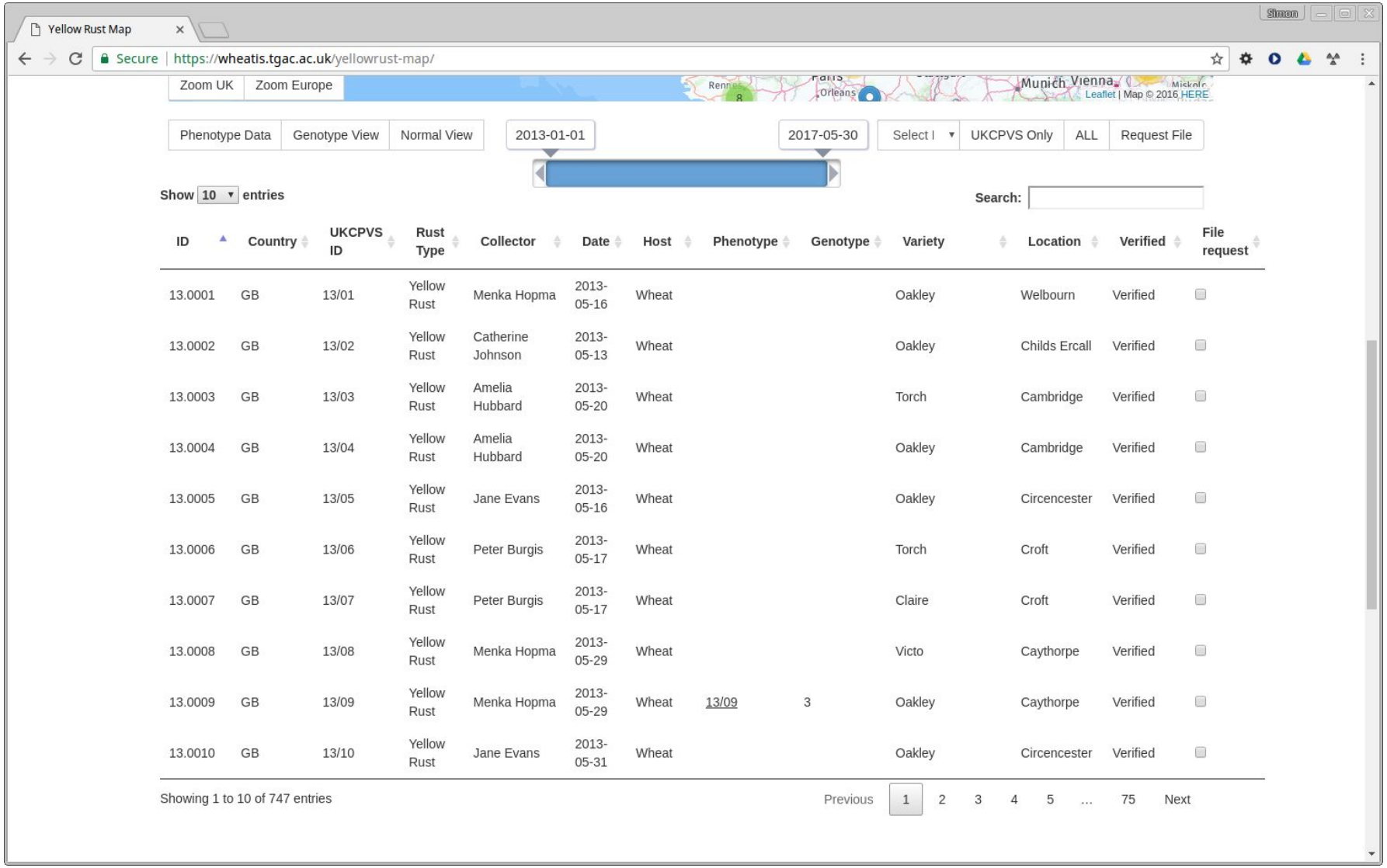
Earlham Institute

Field Pathogenomics - Genotype view



Earlham Institute

Field Pathogenomics - table



Earlham Institute

Future work

- More iRODS integration
  - Storage of service results on iRODS shares
    - Automatic metadata detailing job parameters to facilitate reproducibility
  - Federation
    - Bristol University
    - INRA
- More Services
  - Marker design tool
  - SNPs position indexes

Acknowledgements

- University of Bristol
  - Paul Wilkinson
  - Mark Winfield
  - Keith Edwards
  - Gary Barker
  - CerealsDB Team
- INRA-URGI (France)
  - Michael Alaux
  - Raphael Flores
  - Hadi Quesneville
- John Innes Centre
  - Ricardo Ramirez Gonzalez
- Earlham Institute
  - Ksenia Krasileva
  - Diane Saunders
  - Matt Clark
  - Erik van den Bergh
  - Toni Etuk
  - Felix Shaw
  - Jon Wright
  - Paul Bailey
  - Bernardo Clavijo
  - Luis Yanes
  - Rob Davey

Availability

- Source files available at https://github.com/TGAC?q=grassroots
- Portal at https://wheatis.tgac.ac.uk/