

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/>

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

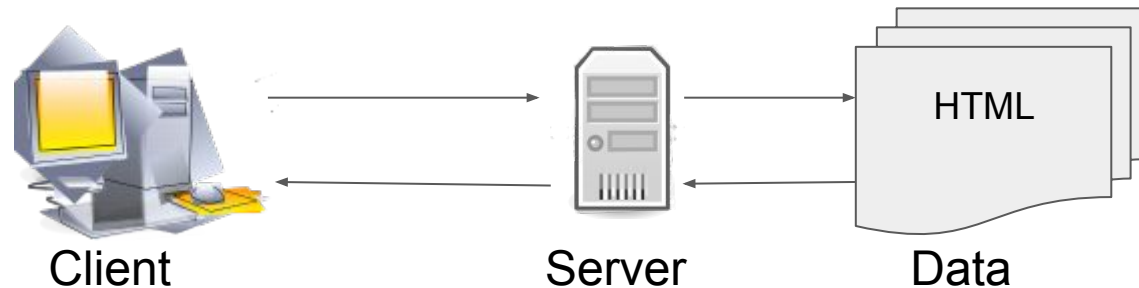
An infrastructure for a distributed set of servers to transparently share:

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

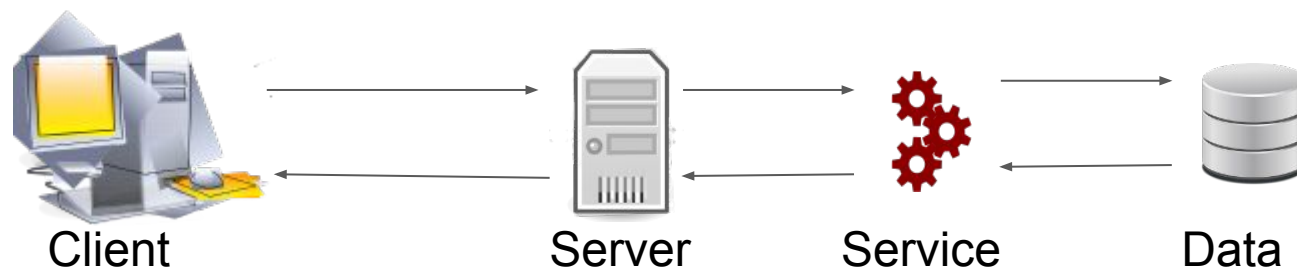
And make it as user-friendly as possible!

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

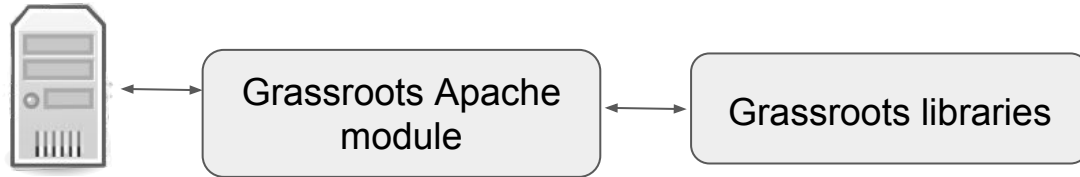


Apache httpd is the most commonly used
Web Server

- 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

Apache httpd
Web Server



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

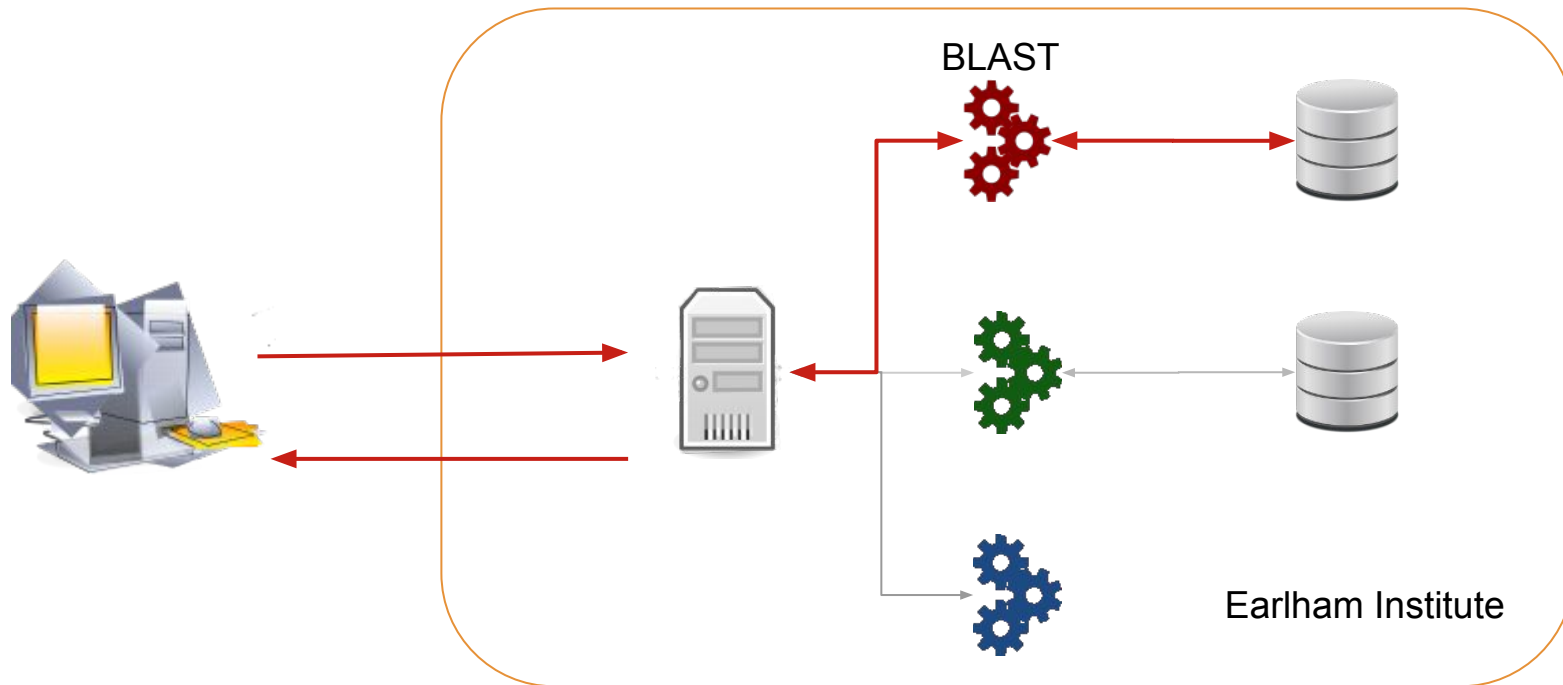
Grassroots architecture

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

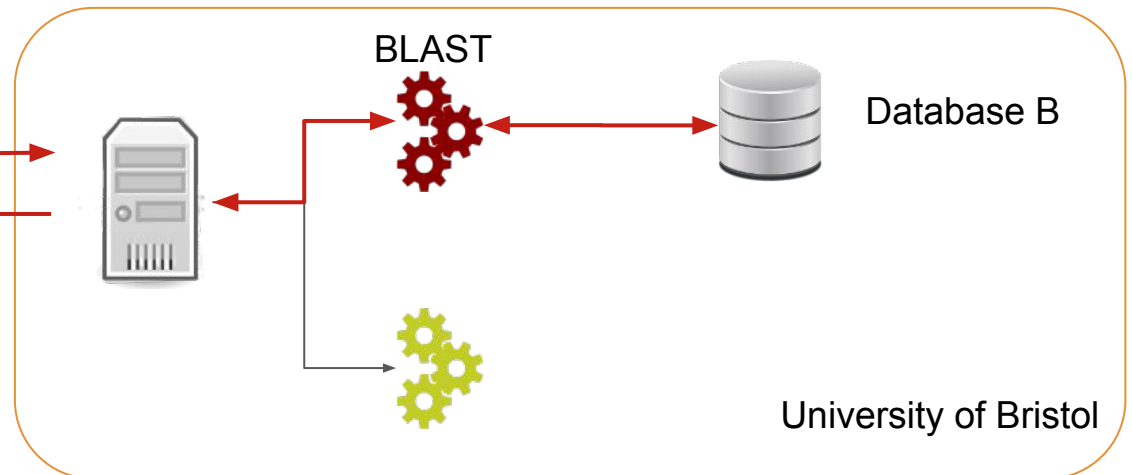
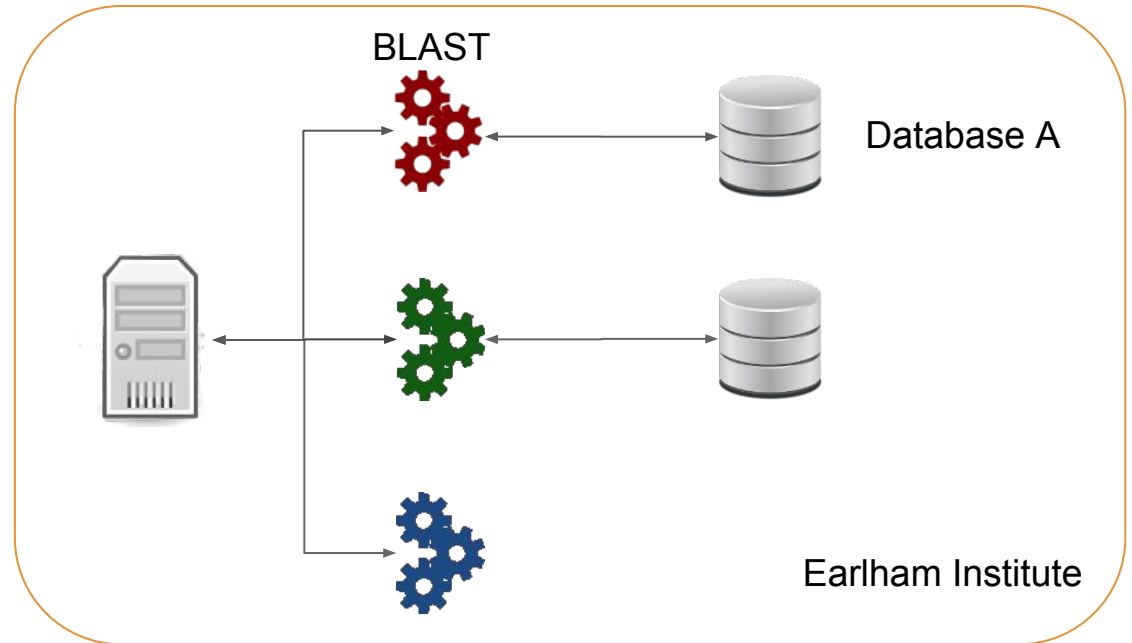
Grassroots architecture

- 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



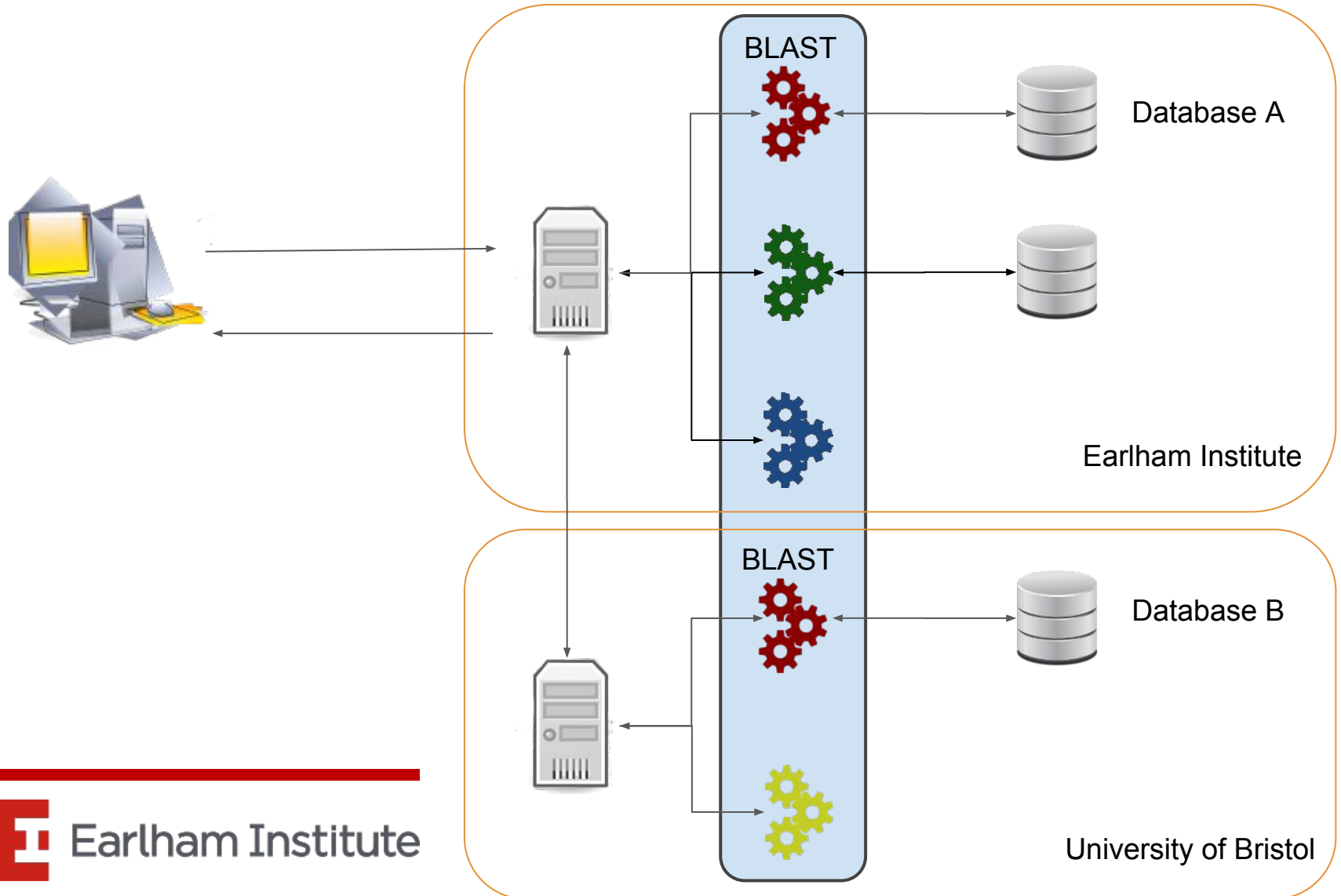
Run the same Service on another Server



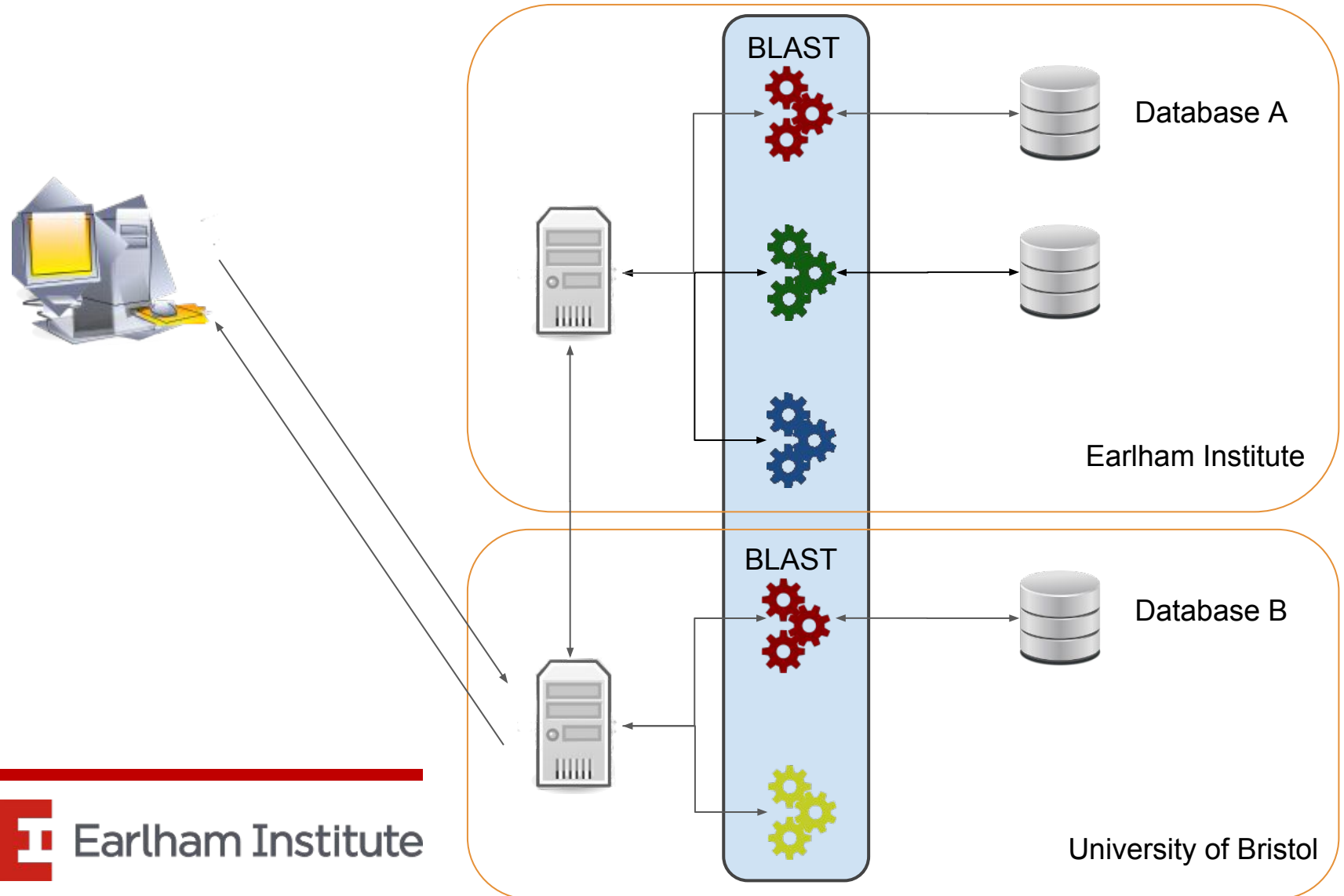
Issues

- 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

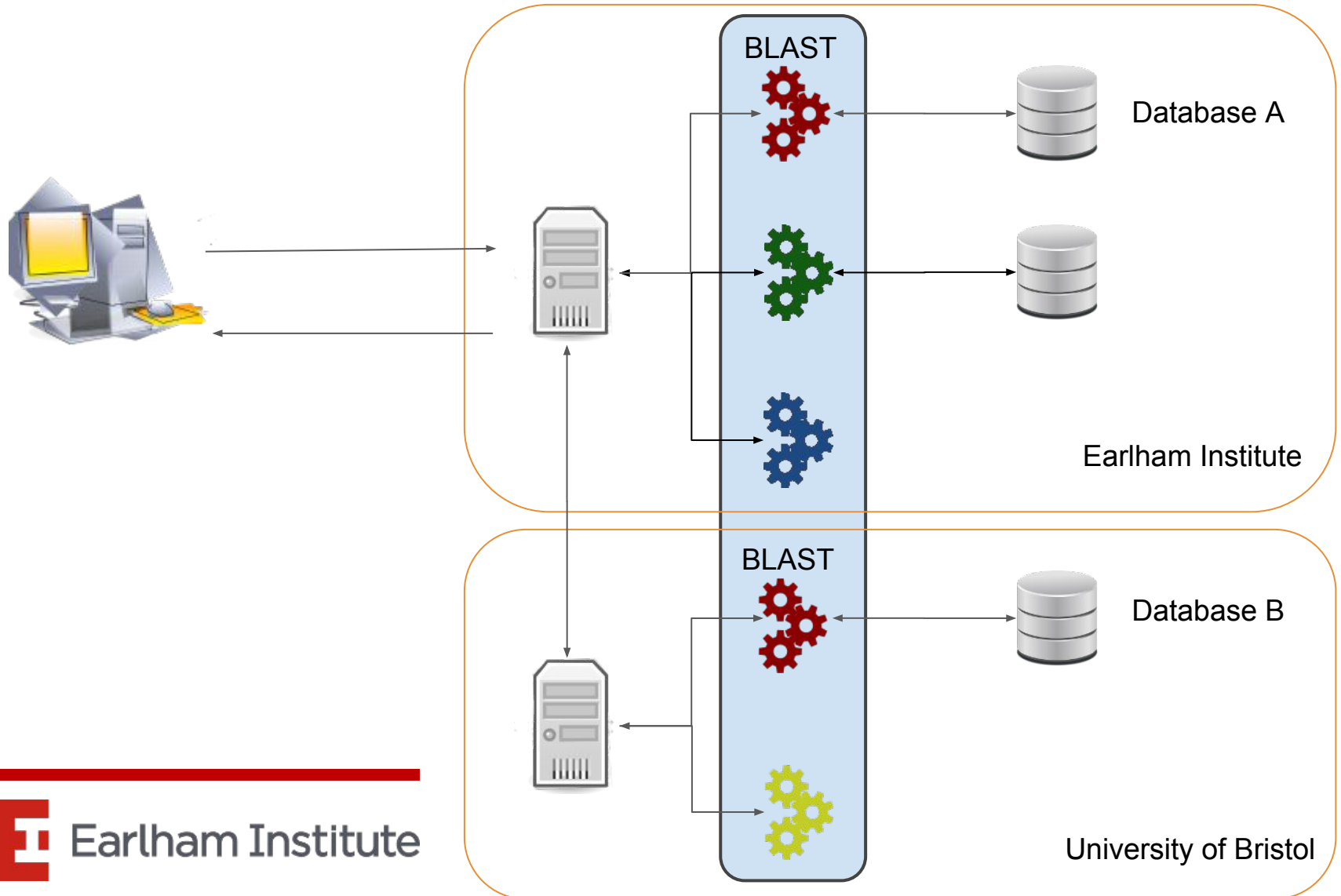
Distributed Services



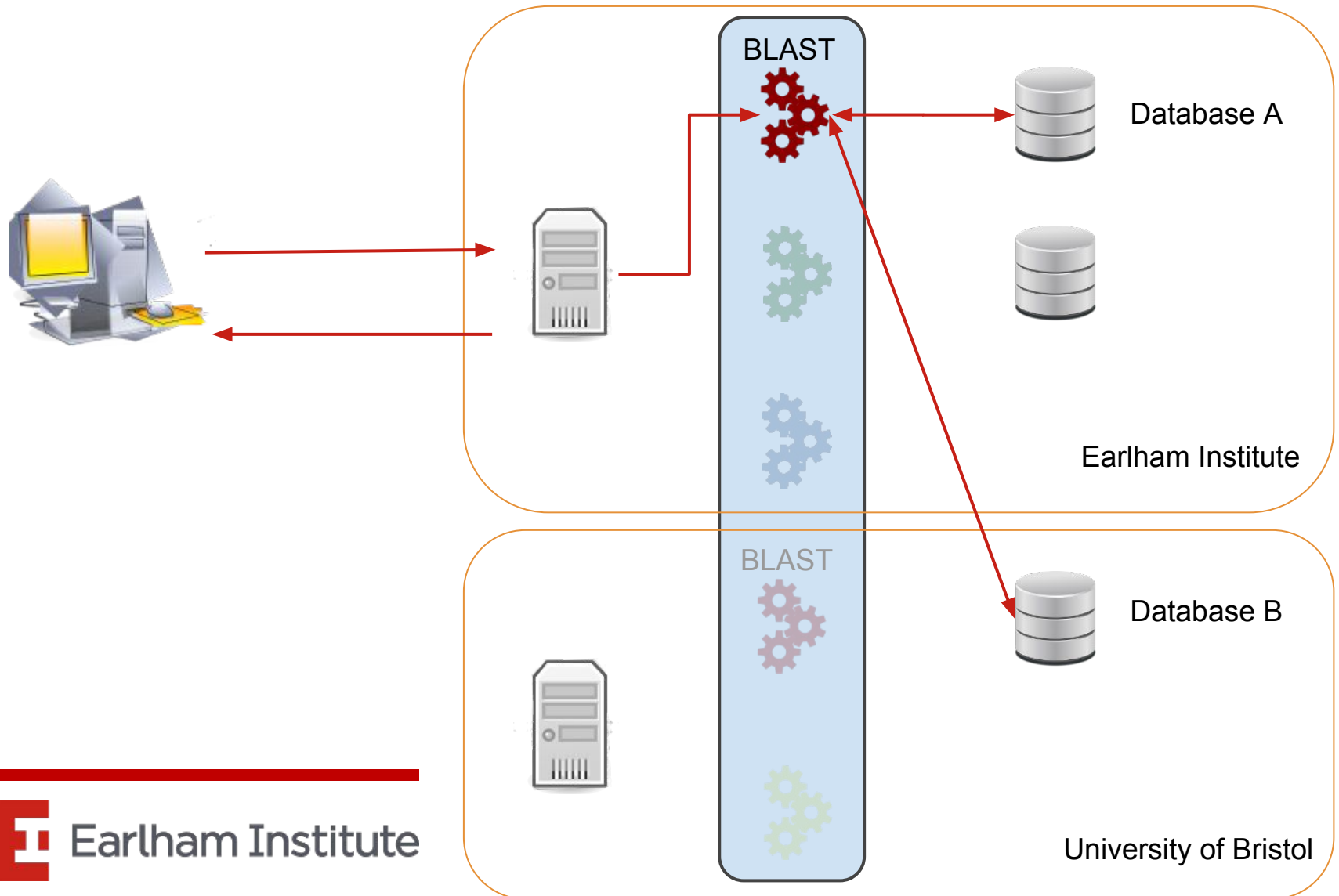
Different Server, Same List of Services



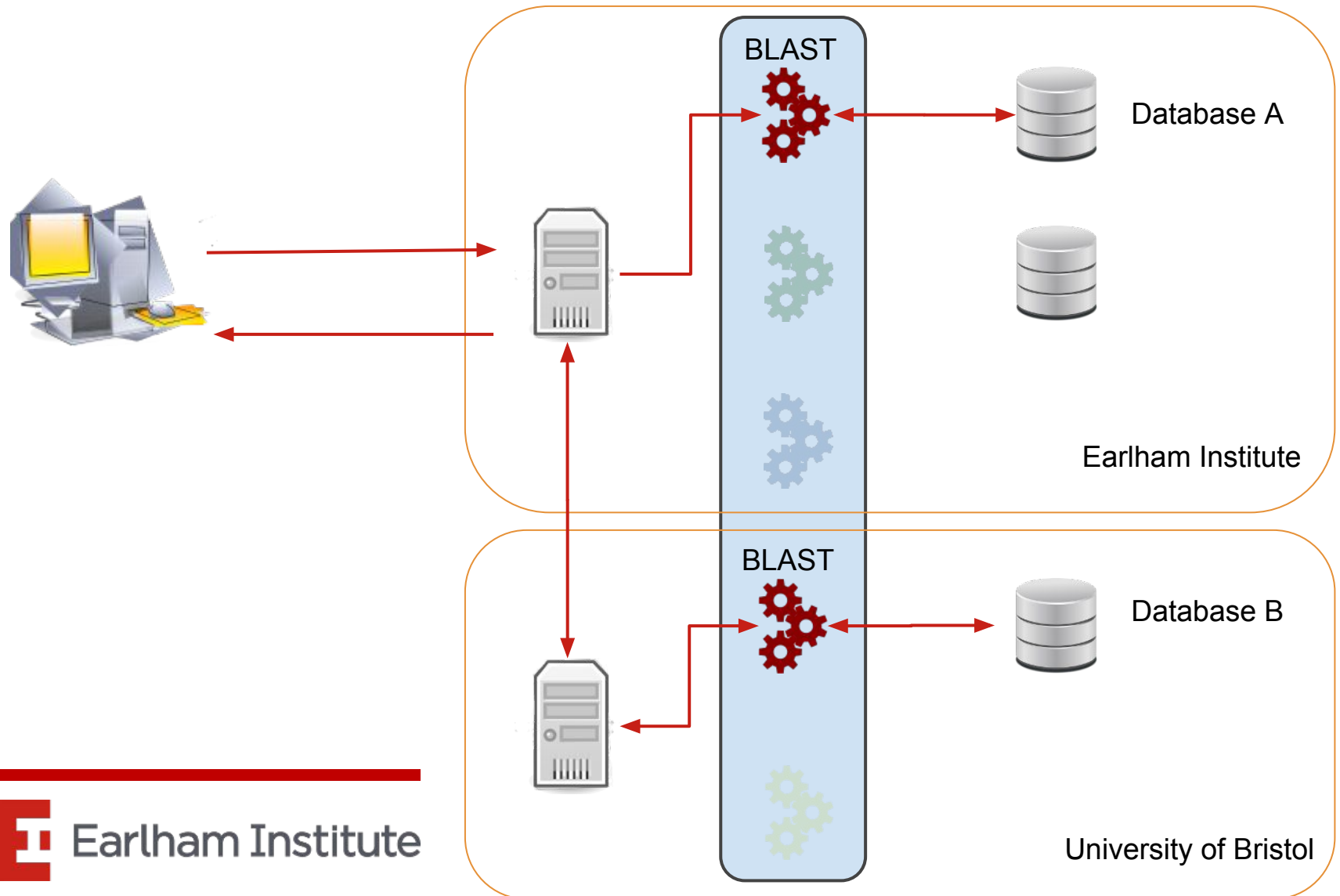
Duplicated Services...



... Get Amalgamated



Consolidate Services - Under the hood



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

... 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

[illegible]

Sbjct 1610 CTGTAGATGTGCACCTTGATGGTATCCTCGGCGATGAGCTCGAAGACGCAAACATCGAAC 1551

Query 61 TTCTCCAGATTGTTGCCGATCGAGAACTGGCTCCAGCCTCT 101

[illegible]

Sbjct 1550 TTCTCCAGATTGTTGCCGATCGAGAACTGGCTCCAGCCTCT 1510

Lambda	K	H
0.634	0.408	0.912

Gapped

Lambda	K	H
0.550	0.210	0.460



... To Run Another Service Automatically

The screenshot displays the Grassroots Client application. The 'Services' panel on the left lists several services, with 'BlastN service' and 'SamTools service' selected. The 'SamTools service' details panel on the right shows the configuration for a BLAST search using the 'Triticum aestivum CS42 cds' provided by Billy's Grassroots server 1. The 'Scaffold name' is 'TRIAE_CS42_6DS_TGACv1_542925_AA1732620.1' and the 'Max Line Length' is set to 60. The 'Run SamTools service' checkbox is checked.

The 'Results' panel at the bottom left shows the job details: 'Job: TRIAE_CS42_6DS_TGACv1_542925_AA1732620.1', 'Service: SamTools service', and 'Description: Triticum aestivum CS42 cds'. The 'Results' tab is active, showing the BLAST output for the job.

The BLAST output is displayed in the 'View' panel on the right, showing the sequence alignment results for the job. The output is a BLAST format text (FASTA) showing the alignment of the query sequence (TRIAE_CS42_6DS_TGACv1_542925_AA1732620.1) against the Triticum aestivum CS42 cds database. The alignment is highly similar, with a score of 100.0 and a p-value of 0.0.

```
>TRIAE_CS42_6DS_TGACv1_542925_AA1732620.1
ATGGCAGGGCTGAAGAAGGGGAAGGGGAAGGGGAGGGAGCGACGGGTGACCTTTGCACCGGAATC
CGGGGAGCGATGGCTCATTGTCGCCGAACCACTGACGAGAGCCGACCAAGATCTGATCGATCTCGGGGACG
GCCTGGACAGAGGACTTCGCCGTGCTCCGCCGAGAGAAAGCTCCCGGATCTCATGGTCACGGTAGGGATGAAG
AAGGTGAAGGGGAAGGGGAGGGGGAAGGATGAGCGACGGGTGACCTTTGCACCAAGAATCCGGGAGCGACGGGT
GTCAGCAACAAGTCAGAAGAGGAACATGAATGCAGCTGACTCATCAAAAAATTAGAGATTCATGAAATGGAACCT
CAAGGAAGCGGATTAGAGAAATTGATGCAGATGGCTCAAAAAAGGGCAACACAGCGGGTCTGGAGCTTACAAGTC
AGGCCAGGAAACCTTAAAGTACTAAAGATTCTCATTCTCAAAACACGTGTCCTCAAAGTTTGACCGCTGGCTGG
AATTCCCGGACACAAACAGGAACCTTACGGACATGGCTGTGAGGGTGAACGTGAAAGTGGAGGAGGGGGAAGT
AATGAAGTTTGGGCGAGAAATTCATCAGGGTATTCAAATACGGGGAGCGATTGAGAATCCACAATCATTGATCAGT
ACCTTCAGAATCAACGAAGTGGGTGCGTTTACTAAGAGGTCAAAGTGGGAATAAATGGCTTGTTGGAACCTTGCTTCA
GACACCGAAGGATTTTTCTTTAGACGTGGGTGGAAGGAATTTGGTATAGATCATTCCATTGGGGGAAGGAAACCTCTT
ATTATTTTGTATGATGGATACGGACAGTTCTCAGTTAATATATTTAATGGAATGTGTGTTGAGAAGCCATCAGCTCTT
CATGCTAAGCCTTCAAAGGATTTTAAAGAGAGAGTGATGAAGATGACAATGGAGTAGCTCCCCAAGAAGAGAATAA
TGGAACCAACAAGAAGCCGACTGGAGAAATTTACGCAGATGGCTCTACGTTGAAGAAGTGTTCTAATGCATCAGTTA
AGGGTAAGCAAAAAATGCCTGAATCTTTGGTTGGTACTTGTAAATGTGTCAGCAACAAGACAGAAGGAACAAGAAT
GCAGCTGACCCATCAGAAAGCTTGGAGGTTGCTGGAGCTTGCAGATCAGCGCGAACAAGCATTAAACTGTCATGA
AAGTTCAAAGGCAACATGCAATAGTATCAAAAGGTCGCCGGTAAGTGAAGGGCAGAAGAAGTACGCTCTTCAAAG
GGCAGAGATATTTACATCTAAGTACCTTCAACTCTGCAAGTAATTAAGGCGGCTAGTGCCCTACAACCTATTTTTCAT
GGTCATCCCATCTGAATTTGTCAGGGAGCATCTCCCCACACCAGCAAGAAGTTGACCTGTGGGATCCGCAAGCA
CGCCCTGGCAAGTTGACTACCTTACTGCAGCCATCGTTCCGCCGGCGCTTTGACGAGAGGCTGGAGCCAGTTCTCG
ATCGGCAACAATCTGGAGAAGTTCGATGTTTGCCTCTTCGAGCTCATCGCCGAGGATACCATCAAGGTGCACATCTA
CAGAGCTCAGCTCTGATAGCACAGTACCAACACCAAGAAGAACTTAGCGCTGCAATCCCAACTAA
```

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 Internet¹
- Sequence Ontology
 - A collaborative ontology project for the definition of sequence features used in biological sequence annotation²
- EDAM Ontology
 - A simple ontology of well established, familiar concepts that are prevalent within bioinformatics³
- FALDO
 - Lightweight interval-based genomic feature descriptors⁴

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

Grassroots Client

File Connect Tools

All Services Run by search

Services

☐ iRods search service

iRods search service
A service to search the metadata within iRods
Provided by [Ei test iRODS server](#), Grassroots running on Ei test iRODS server.

☒ Data objects metadata

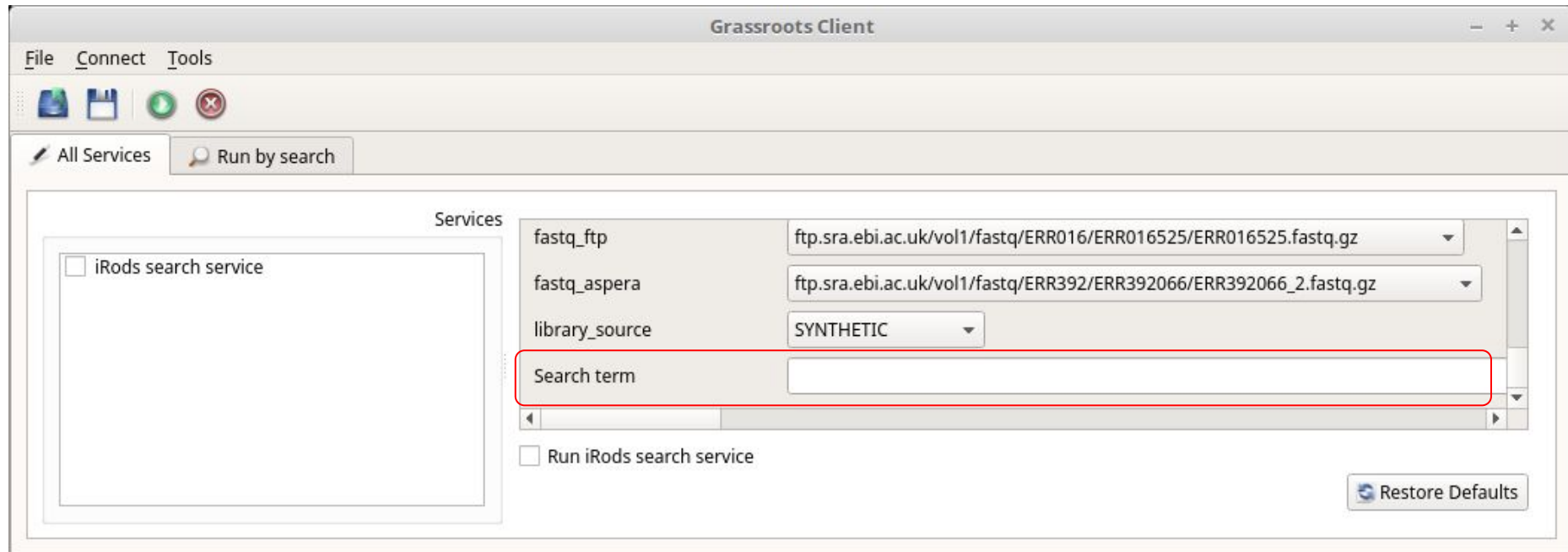
base_count	3942264700
broker_name	ArrayExpress
center_name	China Agricultural University
experiment_accession	ERX1220768
experiment_alias	KTC32_Kronos652_140528_I1135_FCC4K1CACXX_L5
experiment_title	Illumina HiSeq 2000 sequencing; Exome capture of T. turgidum cv Kronos: EMS mutant K
fastq_aspera	fasp.sra.ebi.ac.uk:/vol1/fastq/ERR125/ERR125556/ERR125556.fastq.gz
fastq_bytes	1775512424
fastq_ftp	ftp.sra.ebi.ac.uk/vol1/fastq/ERR120/007/ERR1201757/ERR1201757_1.fastq.gz
fastq_galaxy	ftp.sra.ebi.ac.uk/vol1/fastq/ERR120/007/ERR1201757/ERR1201757_1.fastq.gz
fastq_md5	0640df44c967c3c998d7aa890597865b
first_public	2010-06-16
instrument_model	454 GS FLX
instrument_platform	ILLUMINA
irods_path	/eiZone/public/reads/triticum_aestivum/SRP056412/SRS895375/SRR1958769_1.fastq.gz
last_updated	2012-07-19
library_layout	PAIRED
library_name	KTC32_Kronos652_140528_I1135_FCC4K1CACXX_L5

☐ Run iRods search service

Restore Defaults

iRODS - Metadata search service

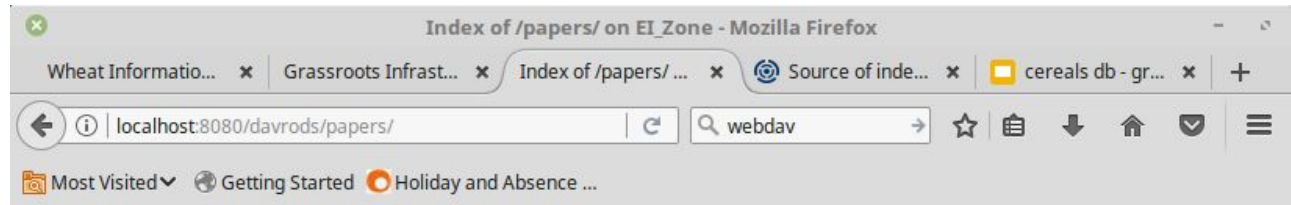
- Global keyword search across all metadata keys and values



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



Index of /papers/ on EI_Zone

[^ Parent collection](#)

Name	Size	Owner	Last modified
CerealsDB 3.0.pdf	1.5M	irods	2017-02-24 12:25
PhysRevE.88.042701.pdf	1.0M	irods	2017-02-24 12:25
SNPSplit.pdf	1.5M	irods	2017-02-26 18:38

iRODS - Davrods customisation

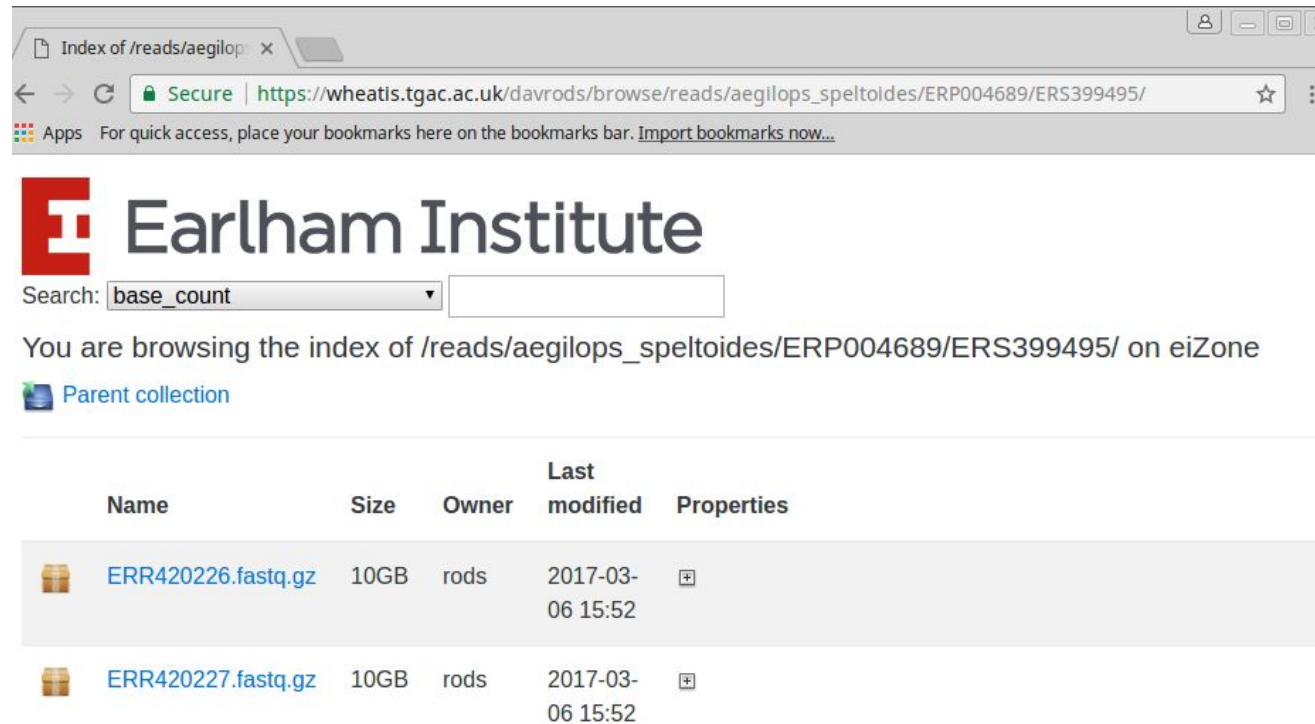
Open source project available at <https://github.com/billyfish/davrods>





- 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



Name	Size	Owner	Last modified	Properties
 ERR420226.fastq.gz	10GB	rods	2017-03-06 15:52	
 ERR420227.fastq.gz	10GB	rods	2017-03-06 15:52	

Listing generated by mod_davrods, © 2016 by Utrecht University and © 2017 by the Earlham Institute.
Filetype icons are taken from the Amiga Image Storage System © 2004 - 2016 by Martin Mason Merz.

iRODS - Davrods metadata enhancements

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

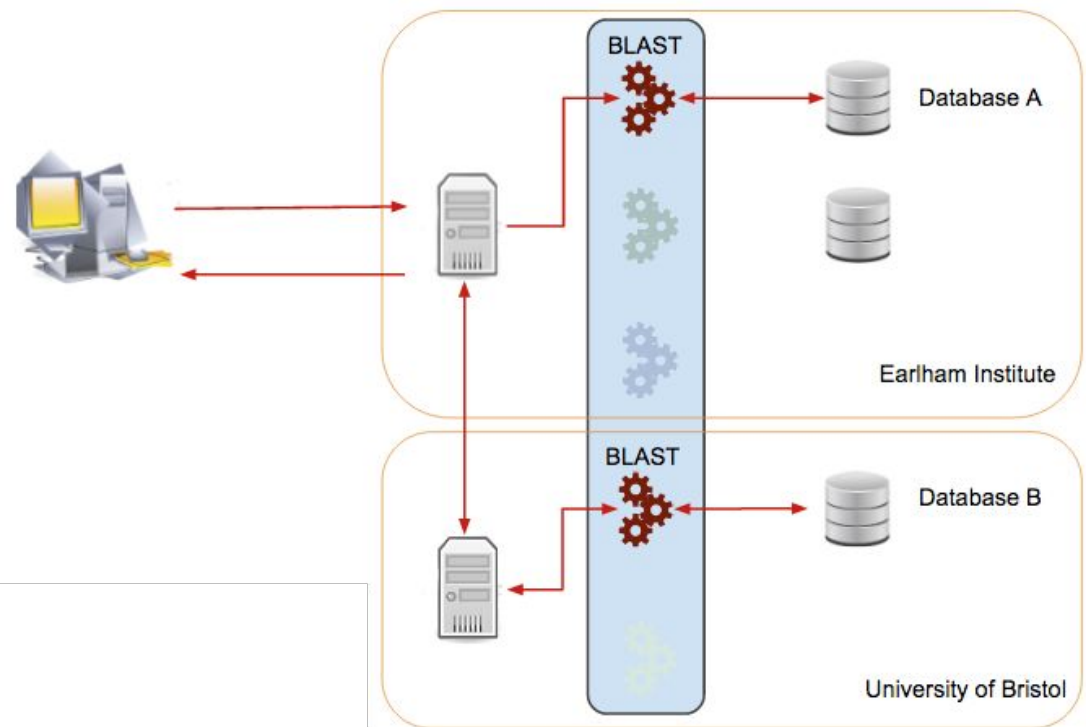
The screenshot shows a web browser window with the following details:

- Browser Tabs:** Mail - Simon.Tyrrell@e..., iRODS UGM - grassroot..., Index of /reads/aegilop...
- Address Bar:** Secure | https://wheatis.tgac.ac.uk/davrods/browse/reads/aegilops_speltoides/ERP004689/ERS399495/
- Page Header:** Earlham Institute logo and name.
- Search:** A dropdown menu showing 'base_count' and an adjacent input field.
- Text:** 'You are browsing the index of /reads/aegilops_speltoides/ERP004689/ERS399495/ on eiZone' and a link for 'Parent collection'.
- Table:** A table with columns: Name, Size, Owner, Last modified, and Properties.

Name	Size	Owner	Last modified	Properties
ERR420226.fastq.gz	10GB	rods	2017-03-06 15:52	<div><div>base_count: 12415278200</div><div>center_name: TGAC</div><div>experiment_accession: ERX386526</div><div>experiment_alias: speltoides</div><div>experiment_title: Illumina HiSeq 2000 paired end sequencing; mRNA-Seq for Associative Transcriptomics in hexaploid wheat</div><div>fastq_aspera: fasp.sra.ebi.ac.uk/vol1/fastq/ERR420/ERR420226/ERR420226.fastq.gz</div><div>fastq_bytes: 11154573984</div><div>fastq_ftp: ftp.sra.ebi.ac.uk/vol1/fastq/ERR420/ERR420226/ERR420226.fastq.gz</div><div>fastq_galaxy: ftp.sra.ebi.ac.uk/vol1/fastq/ERR420/ERR420226/ERR420226.fastq.gz</div><div>fastq_md5: 3bc8c40f34faa2e6ffb5ac173edc261b</div><div>first_public: 2014-01-31</div><div>instrument_model: Illumina HiSeq 2000</div></div>

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

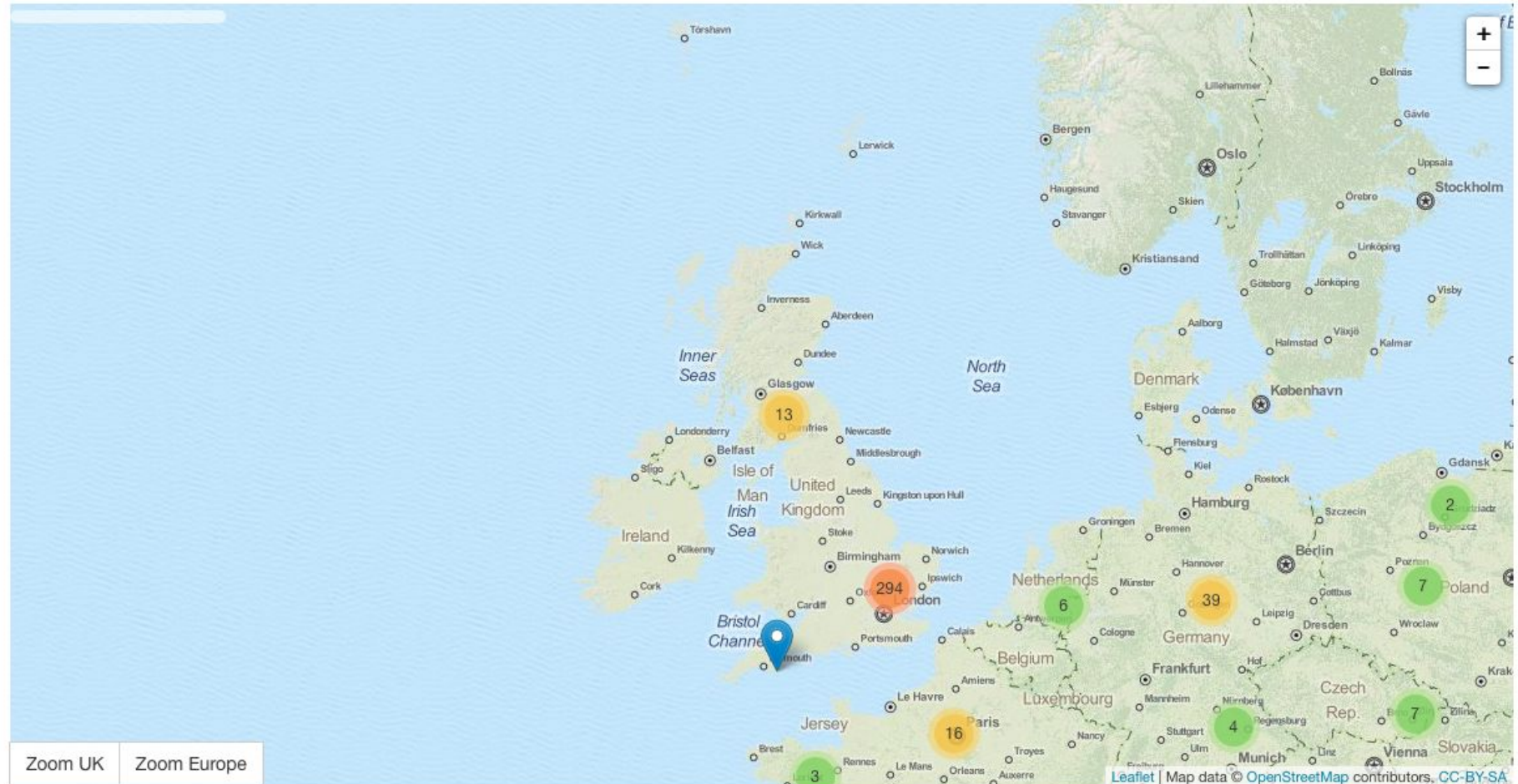


Field Pathogenomics Service

- 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

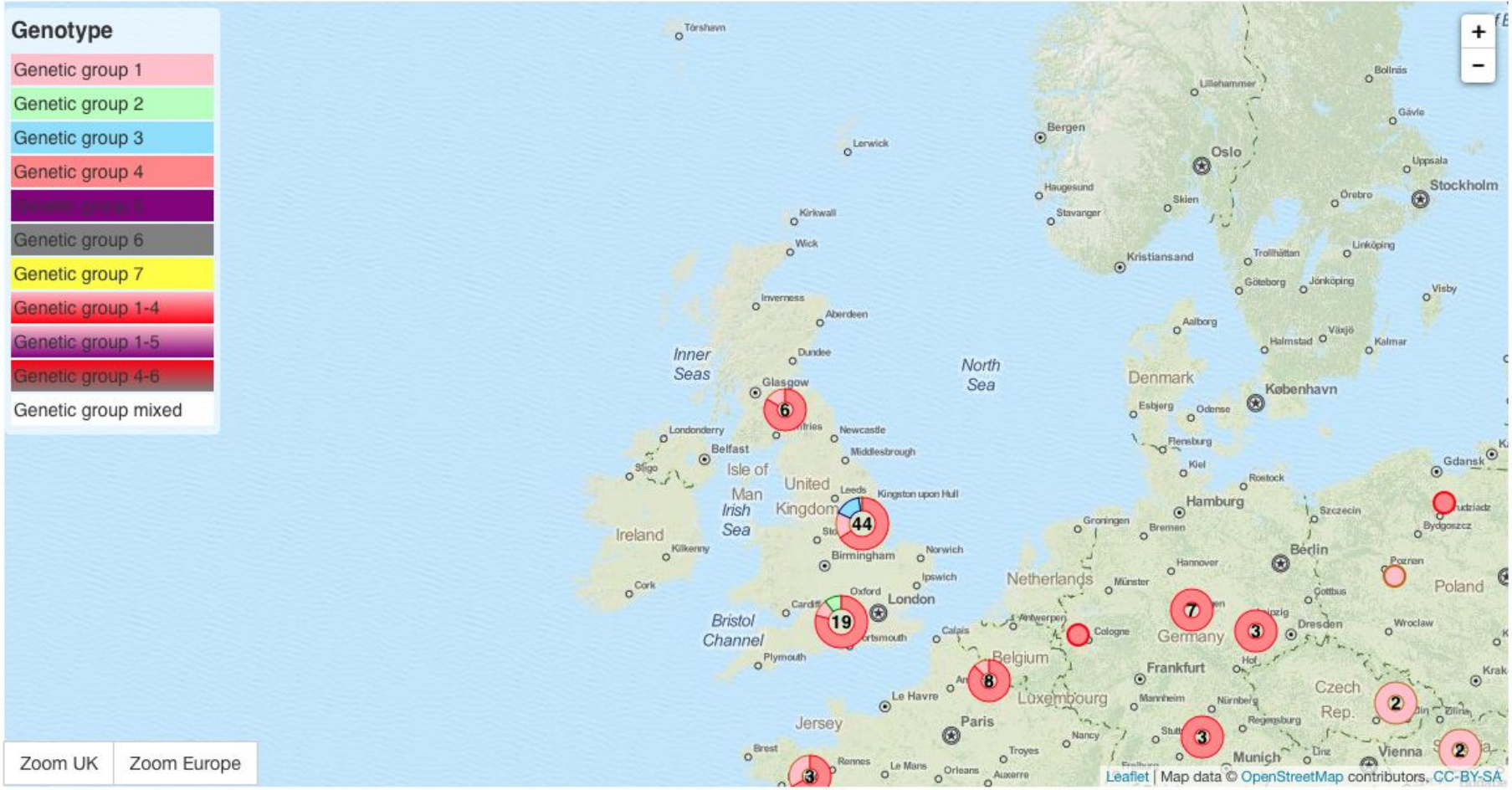
Field Pathogenomics - Samples view

Yellow Rust Map



Field Pathogenomics - Genotype view

Yellow Rust Map



Field Pathogenomics - table

Yellow Rust Map

Secure | <https://wheatis.tgac.ac.uk/yellowrust-map/>

Zoom UK Zoom Europe

Phenotype Data Genotype View Normal View 2013-01-01 2017-05-30 Select I UKCPVS Only ALL Request File

Show 10 entries Search:

ID	Country	UKCPVS ID	Rust Type	Collector	Date	Host	Phenotype	Genotype	Variety	Location	Verified	File request
13.0001	GB	13/01	Yellow Rust	Menka Hopma	2013-05-16	Wheat			Oakley	Welbourn	Verified	<input type="checkbox"/>
13.0002	GB	13/02	Yellow Rust	Catherine Johnson	2013-05-13	Wheat			Oakley	Childs Ercall	Verified	<input type="checkbox"/>
13.0003	GB	13/03	Yellow Rust	Amelia Hubbard	2013-05-20	Wheat			Torch	Cambridge	Verified	<input type="checkbox"/>
13.0004	GB	13/04	Yellow Rust	Amelia Hubbard	2013-05-20	Wheat			Oakley	Cambridge	Verified	<input type="checkbox"/>
13.0005	GB	13/05	Yellow Rust	Jane Evans	2013-05-16	Wheat			Oakley	Circencester	Verified	<input type="checkbox"/>
13.0006	GB	13/06	Yellow Rust	Peter Burgis	2013-05-17	Wheat			Torch	Croft	Verified	<input type="checkbox"/>
13.0007	GB	13/07	Yellow Rust	Peter Burgis	2013-05-17	Wheat			Claire	Croft	Verified	<input type="checkbox"/>
13.0008	GB	13/08	Yellow Rust	Menka Hopma	2013-05-29	Wheat			Victo	Caythorpe	Verified	<input type="checkbox"/>
13.0009	GB	13/09	Yellow Rust	Menka Hopma	2013-05-29	Wheat	13/09	3	Oakley	Caythorpe	Verified	<input type="checkbox"/>
13.0010	GB	13/10	Yellow Rust	Jane Evans	2013-05-31	Wheat			Oakley	Circencester	Verified	<input type="checkbox"/>

Showing 1 to 10 of 747 entries

Previous 1 2 3 4 5 ... 75 Next

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/>