

Large Scale Infrastructure for Health Data Analytics

Samantha Crossfield, Owen Johnson, Thomas Fleming
contact: s.crossfield@leeds.ac.uk
Leeds Institute for Data Analytics
University of Leeds
Leeds, UK

Abstract— The opportunities for data analytics to inform the science and practice of health care are growing. For example, in the UK, 65 million citizens have lifelong e-health records that can be used to examine patterns of disease, treatment and outcomes. Similarly the real-world impact of interventions such as new drugs can be evaluated in these records. Such approaches need to solve issues around information governance, confidentiality, understanding data and provenance and developing methods for big data. Here we describe a large-scale service that addresses the opportunities brought by the availability of large-scale e-health records. The service has been used to support 50 research projects in the UK across a wide range of scientific areas and can be seen as an exemplar for the developing field of health data analytics.

Keywords— health systems, analytics, infrastructure

I. INTRODUCTION

In the UK, 65 million residents have life-long primary care e-health records [1] with the largest single database system being TPP SystmOne, used by 5,000 NHS organizations to manage the care of 40 million individuals. This scale of e-health systems is new, as is the opportunity for data analytics at an equivalent scale to inform the science and practice of health care. Health data analytics has been largely run by siloed research teams, working independently to solve similar issues around information governance, data confidentiality, understanding systems and data, and developing new methods. Here we describe one approach, used in the UK, to meet the challenges of doing health data analytics at scale.

The Leeds Institute for Data Analytics (LIDA) is a £12m investment in data science at the University of Leeds, funded by UK research councils, in partnership with TPP SystmOne, Leeds Teaching Hospitals Trust (LTHT) and the national NHS Health and Social Care Information Centre which is based in Leeds. The partnership includes the ResearchOne data service which provides de-identified access to the e-health records of TPP SystmOne (currently 7 million opted in) and a similar service from the PPM hospital e-health record system at LTHT (2.5 million) which includes access to images, genomic and rich bio-informatics data. The investment aimed to improve the pace and quality of big data research through sharing knowledge, tools and equipment. The approach was to develop the Integrated Research Campus (IRC): a service built for secure and large-scale data analytics.

II. METHODS

The development of the IRC included designing infrastructure for large-scale data capture, storage and analysis (firewall-protected servers with 700 cores, 4 TB RAM and 2,000 TB storage) and secure links to data services such as ResearchOne and LTHT's PPM. Data is stored in secure zones and accessed by Virtual Research Environments (VRE) where each VRE is tailored to the data, security, software and processing needs of each research project and each team member [3]. Data is managed through 'privacy by design' information security processes based on template ethico-legal frameworks. Space for multi-disciplinary co-location includes a PowerWall data visualization suite, collaborative working spaces and data 'safe rooms'. Researchers are supported by a pool of experts in data science, information governance and data handling.

III. RESULTS

The IRC services source and de-identify health data and make it available for research. All operations are governed by an information security management system (ISO27001 certification pending). A cost remuneration scheme enables competitive costing for research bids. An IRC Data Team (currently 7) and interns (currently 10) provide expertise in data handling and information governance. LIDA associates (currently 150 associates including clinicians and researchers in computing, health and economics) have service access.

Since LIDA's launch in June 2015 over 50 projects have used IRC services. For example one study used IBM Watson Content Analytics to develop natural language processing algorithms to identify diagnoses from 50 million clinical reports. Another pseudonymously linked e-health records with environmental, genomic and tumor data to investigate skin cancer in order to spot cancer and improve treatment.

IV. SUMMARY

The IRC aids research on a large, cost-effective basis and this data is now used at scale by multiple research teams. LIDA offers a model for the developing field of health data analytics.

REFERENCES

- [1] T.H. Payne, D.E. Detmer, J.C. Wyatt, and I.E. Buchan, "National-scale clinical information exchange in the United Kingdom: lessons for the United States," *J. Am. Med. Inf. Assoc.*, vol 18, pp. 91-98, 2011.
- [2] R. Smith, J. Xu, S. Hima and O. Johnson, "GATEway to the cloud: Case study: A privacy-aware environment for electronic health records research," In: *Proceedings - IEEE 7th SOSE*, pp. 292-297, March 2013.

Samantha Crossfield
s.crossfield@leeds.ac.uk
lida.leeds.ac.uk

Presented with grateful acknowledgement of support from
The MRC Medical Bioinformatics Centre at Leeds; MR/L01629X

Large Scale Infrastructure for Health Data Analytics

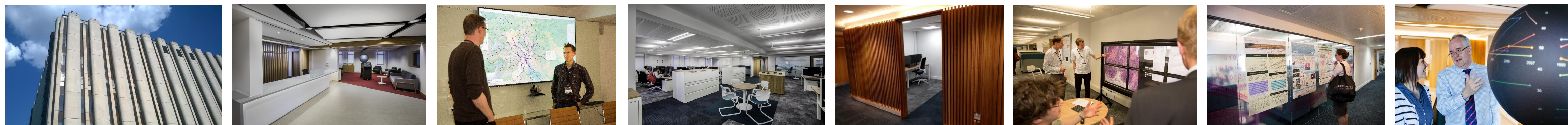
Samantha Crossfield, Owen Johnson, Tom Fleming

LEEDS INSTITUTE FOR DATA ANALYTICS

Funders: MR/L01629X; ES/L011891/1



UNIVERSITY OF LEEDS



Outline

Growth in e-health systems brings opportunity for data analytics to inform health research on a larger scale [1]. In the UK, 65 million residents have lifelong e-health records that can be examined for patterns of disease and to evaluate interventions in the real world [2]. To date, projects in health data analytics are often run by silo-ed research teams, independently solving similar issues around information governance, data confidentiality, understanding systems and data, and developing new methods.

We describe a way that uses large-scale infrastructure to address the opportunities for data analytics at scale in the UK. It has supported 50 projects in a range of scientific areas and can be seen as an exemplar for the developing field of data analytics.

Context

The UK population (n=65 million) has life-long primary care records with the largest single database system (n=40m) being TPP's SystmOne (www.tpp-uk.com), which has research access via ResearchOne (currently 7m opted in).

UK hospital systems also capture rich genomic, imaging and bio-informatics data. At Leeds Teaching Hospitals NHS Trust (LTHT) for example this covers 3 million patients in systems linked to their Patient Pathway Manager (PPM).

The Leeds Institute for Data Analytics (LIDA) is a £12m investment in data science at the University of Leeds, funded by UK research councils, to develop a partnership with TPP, LTHT and the NHS Health and Social Care Information Centre (HSCIC) (Table 1).

	TPP SystmOne	LTHT systems
EHR source	5,000 UK NHS organisations <i>incl. GP, community, prisons, social care</i>	Leeds Teaching Hospitals NHS Trust, Leeds, UK
Period	Life-long records	Episodes of care, 1996 onwards
Patients	40 million	3 million
Data examples	Demographics, diagnosis, referral, pathology, prescription, vaccinations	Diagnosis, prescription, genomic, procedure, laboratory, vital signs

Table 1. Examples of large scale systems in the LIDA partnership

Aim and Objectives

LIDA aims to improve the pace and quality of big data research through sharing knowledge, tools and equipment.

Its objectives are to:

- Enable large scale data analytics research
- Protect and increase revenue and intellectual property
- Increase the impact and visibility of research
- Improve the researcher experience

LIDA require a large-scale infrastructure to meet this. Therefore the University of Leeds developed the **Integrated Research Campus (IRC): a service for secure and large-scale data analytics.**

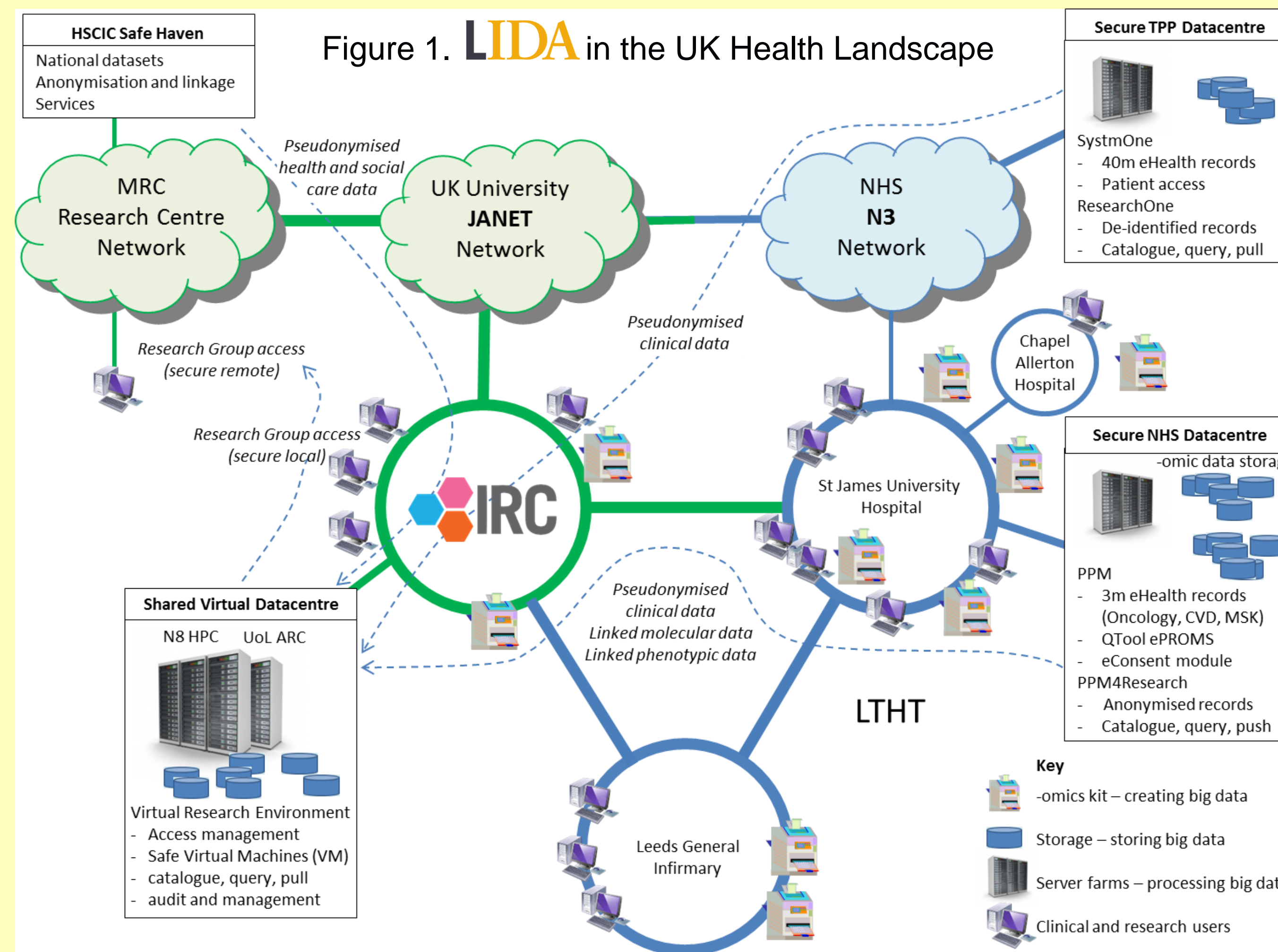


Figure 2. IRC Virtual Research Environment for isolated projects with remote data access

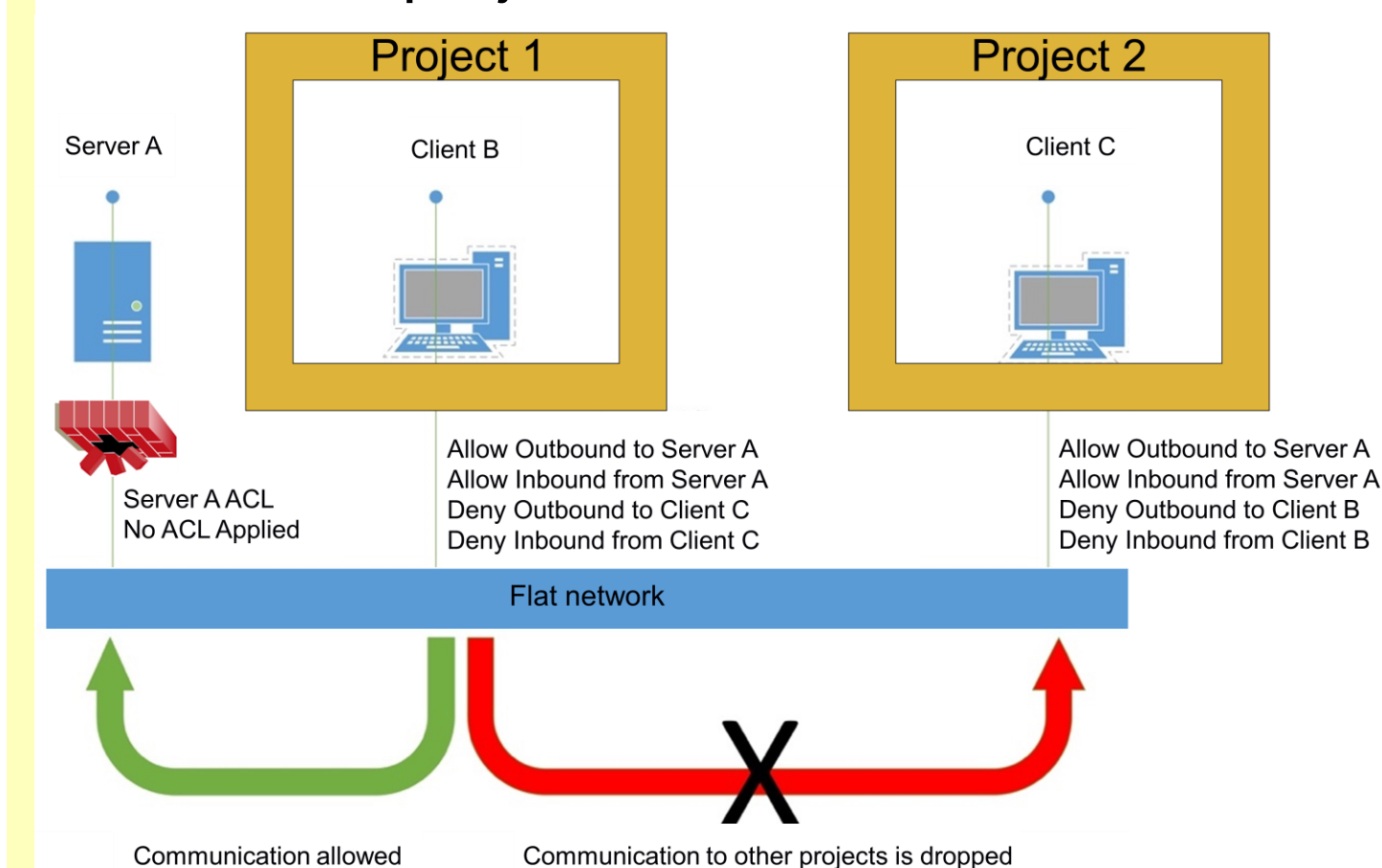


Figure 3. IRC Team Roles

- Data Services Manager**
- Ensure the IRC provides new services as required
 - Own and promote the infrastructure
 - Ensure value for money via re-use and exploiting new opportunities
 - Operationally responsible for execution of Standard Operating Procedures
 - Manage interactions with data providers and University IT
 - Technical problem management
- Data Services (currently 6)**
- Data linkage and risk profiling
 - Load and transform data
 - Responsible for information governance
 - Deliver bespoke software
 - Liaise with data providers
 - Produce datasets for research and publication
 - Assist researchers with data / compute issues
 - Support researchers through the whole project lifecycle
 - Manage areas of system administration
-

The Integrated Research Campus

- **Computer network** for large-scale data capture, storage and analysis (firewall-protected servers with 700 cores, 4 TB RAM and 2,000 TB storage), plus High Performance Computing (HPC)
- **Links to data services** such as ResearchOne and LTHT's PPM and development of their services, including a Leeds Data Warehouse to merge LTHT systems (Fig 1)
- **'Privacy by design'** information security system (ISO27001 pending)
- **Processes**, templates and guides for common research tasks and tools
- **Data access** via Virtual Research Environments (VRE) tailored to the data, security, software and processing needs of each project and each team member (Fig 2) [3]
- **IRC Data Team** provides support in data science, governance and data handling (Fig 3)
- **Space** for multi-disciplinary co-location includes a PowerWall data visualization suite, collaborative working areas and data 'safe rooms'
- **Cost remuneration** scheme for quick, competitive costing in bids

Outcome

From LIDA's launch in June 2015 it has provided IRC access to 150 associates including clinicians and researchers, in 50 projects, supported by 10 data science interns. For example:

- One study used IBM Watson Content Analytics to develop natural language processing algorithms to identify diagnoses from 50 million clinical reports
- Another pseudonymously links e-health records with environmental, genomic and tumor data in order to identify skin cancer and improve its treatment

The IRC aids research on a large, cost-effective basis and this data is used at scale by multiple research teams. LIDA offers a model for the developing field of data analytics.

References

1. A.B. Jensen, et al., 2014. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications*, 5: 1-11
2. O. Johnson, et al., 2014. Electronic health records in the UK and USA. *The Lancet*, 384 (9947): 954
3. R. Smith, et al., 2013. "GATEway to the cloud: Case study: A privacy-aware environment for electronic health records research," In: *Proceedings - IEEE 7th SOSE2013*.