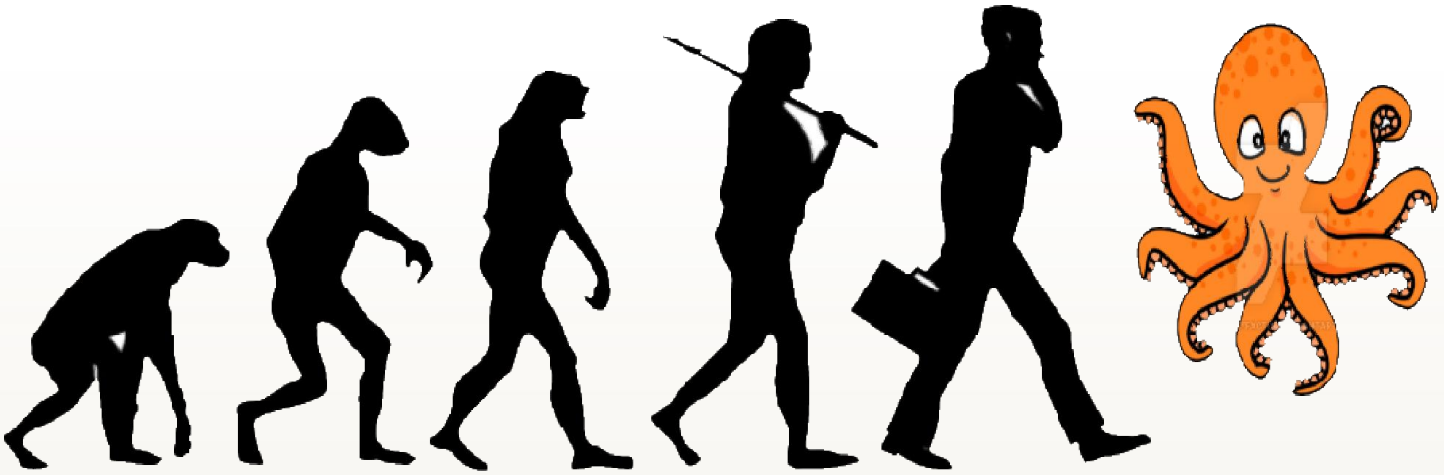
Taming the Octopus



Simon Ellwood-Thompson

previously a child, currently an adult

Swansea University think I am the CTO for the College of Medicine –
SAIL, CIPHER, FARR, ADRC, MS, CLIMB, DPUK

- **Simon Thompson** from CIPHER at the University of Swansea on their research infrastructure developments and projects, including working with the NHS as part of the Farr Institute, and as the Administrative Data Research Centre for Wales.

Agenda – Its all about acronyms

1. SAIL programme
2. SAIL Databank
3. CIPHER
4. FARR
5. NRDA
6. UKSeRP
7. NLP
8. ADRC
9. DPUK
10. MS
11. CLIMB
12. Joined up-ness

SAIL (Secure Anonymised Information Linkage)

- Initially we “liberated” 8 compute nodes from BlueC Super computer.

  IBM Power Series 7 series which we put IBM InfoSphere DB2 Warehouse edition (All data partitioned and DB transactions parallelized over all cluster nodes).

- Acquired 100TB SAN and 200TB tape backup.
- Partnership with Informing Health Care (IHC)
  Now called NWIS an NHS organisation

Lots of system design development
SAIL Project was born (2007)

SAIL Split File Principal

Supplier Data

ID	Name	Address	BP	Diag
56	Fred Bloggs	The Big house	120/80	G33..
78	Jim Jones	87 peterson rd	135/45	P123.
45	Harry Lucas	19 meirwen	125/75	G77..

Demographics + Link Key

File 1

ID	Name	Address
56	Fred Bloggs	The Big house
78	Jim Jones	87 peterson rd
45	Harry Lucas	19 meirwen

Linkage

Clinical (s) + Link Key

File 2

ID	BP	Diag
56	120/80	G33..
78	135/45	P123.
45	125/75	G77..

File 3

ID	ALF	Conf
56	65276573	88
78	32377722	97
45	27638236	95

Load into SAIL

ALF_E	BP	Diag
4252	120/80	G33..
7482	135/45	P123.
8436	125/75	G77..

Overview of SAIL Databank (up to now)

- Operational since 2007
- Repository model - no data ever leaves unless with patient consent
- The familiar "split file" approach
- Strong separation used - TTP in NHS organisation (NWIS), SAIL in university
- Only de-identified data held in university - linked via an "ALF_E"
- High levels of trust within Wales (Health, local gov and Welsh Government)
- Lots of UK-wide data linkage projects undertaken using SAIL

Overview of SAIL Databank (up to now)

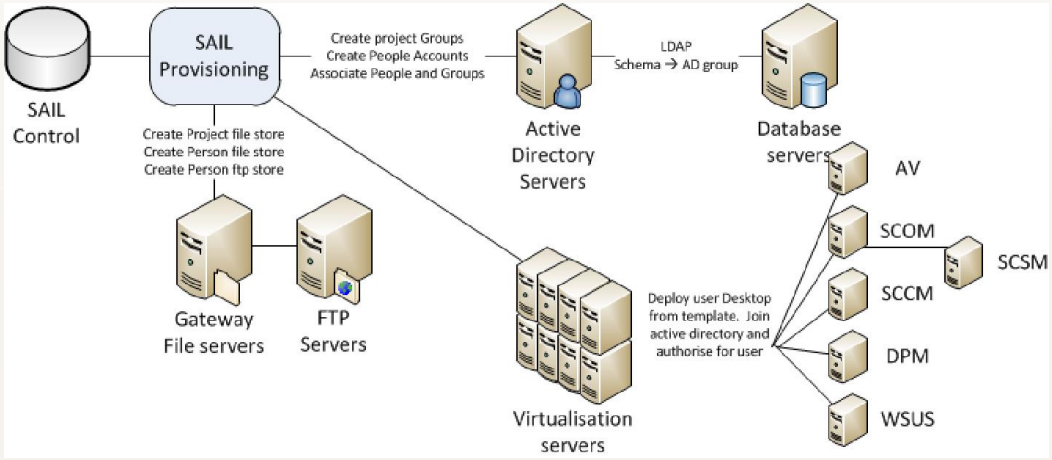
- High levels of automation - matching, anonymisation, data loading, etc.
- Matching against national "spine" (Welsh Demographics Service)
- Projects approved by independent Information Governance Review Panel
- User approval process - only experienced, backed up by enforceable agreement
- SAIL Gateway only route to data access. Tools, knowledge banks, etc.
- Two factor authentication used, encrypted connections, offsite access possible with approval
- Total separation between projects, project-specific IDs used
- IGRP resubmission required for project variation
- Projects archived and taken offline at completion

SAIL Established, time to expand

Everything integrated with Active Directory

Automate as much as possible

Partnership with HPC Wales – full technology refresh and doubling of capacity. Addition of Vmware View and Vmware cluster for application servers



Health Informatics Group
SAIL Databank

- Internationally renowned data linkage system
- Not just health - far more
- Privacy protection and information governance fundamental
- Built on IBM DB2
- Recent big investment in NLP (Watson) via Open Connections
- 300+ feeder systems across Wales
- >£5 million investment in high performance IT
- Industrial strength, reusable infrastructure.

- Secure, remote access from anywhere through VMs, given approvals
- Over 9 billion records for >5 million people
- Much data goes back 10-20 years
- All pre-linked data
- >300 users,
- >£90m projects from UKRCs
- 140+ approved SAIL projects, with 79 active today
- 100 staff in Swansea working on Health Informatics-related projects

SAIL Reach

OVER
370
USERS

78
ORGANISATIONS
WORLD-WIDE

SAIL

This stuff really is not easy !!

Drawing by

Martin Heaven, Senior Data Analyst
Age 53 and ¾

“inside by mind”

Health Informatics Group
MRC Farr Institute @CIPHER

- MRC-funded institute of health informatics
- £9.3 million investment at Swansea
- Four Centres across the UK
- Centre for Improvement in Population Health through E-records Research (**CIPHER**)
- (Swansea, Cardiff, Bristol, Uni of W Australia, Curtin, Ottawa)
- Focus on large scale studies

- **Aim:** provide the physical and electronic infrastructure to facilitate collaboration across the four nodes
  - **UK Secure e-Research Platform (UKSeRP)**
  - **National Research Data Appliances (NRDA)**
- New methods, public engagement, innovative governance, capacity building

FARR @ Swansea – Capital Investment

Additional capital investment, Our GOALS:-

1. UKSeRP: Offer our an expanded version of infrastructure as a service (IaS) to other major programmes (none-SAIL)
2. Data Appliance: Provide local capabilities to manage datasets so that dataset discover and availability become easier
3. Natural Language Processing

Context: Large amount of automation already developed but predicted

massive increase in workload without increase on staffing

FARR @ Swansea – Capital Investment

Additional capital investment without conditions

Due to large previous investment, no call on this resources

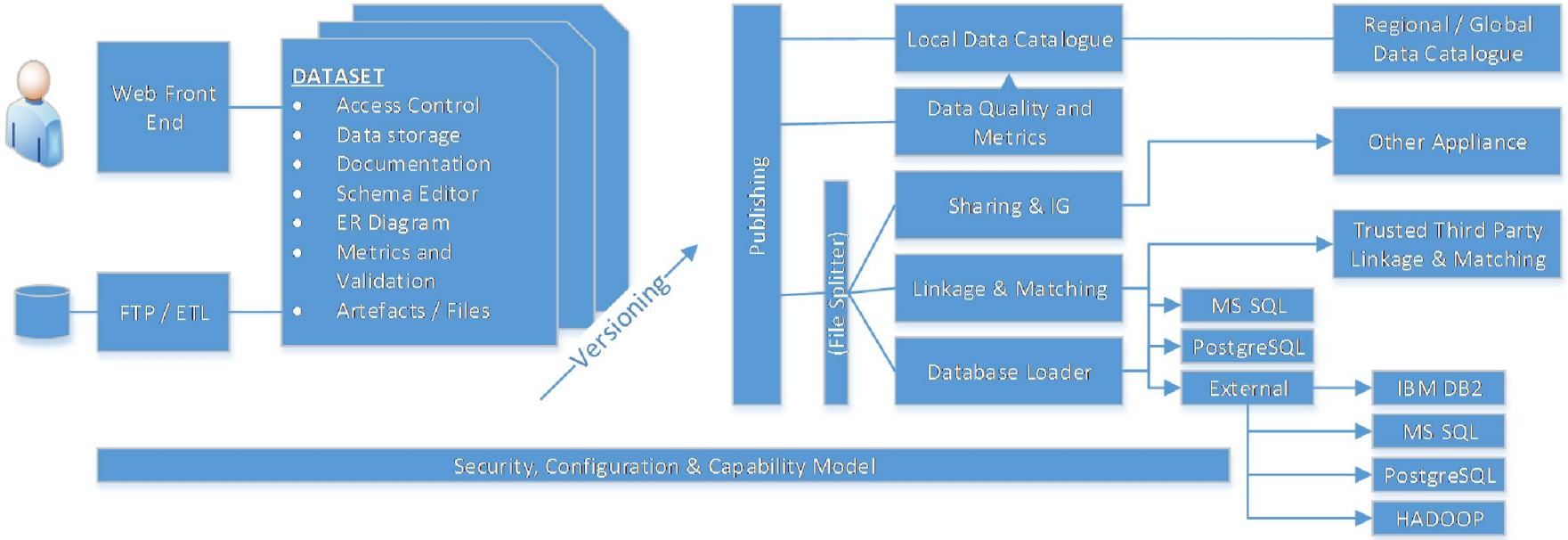
= Great opportunity to invest and take a chance !!!!!

= Chance to reflect on what we do / want to do

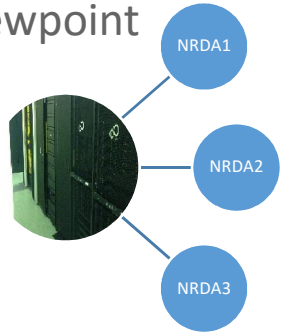
FARR @ Swansea – Capital Investment

NRDA

National Research Data Appliance (NRDA)

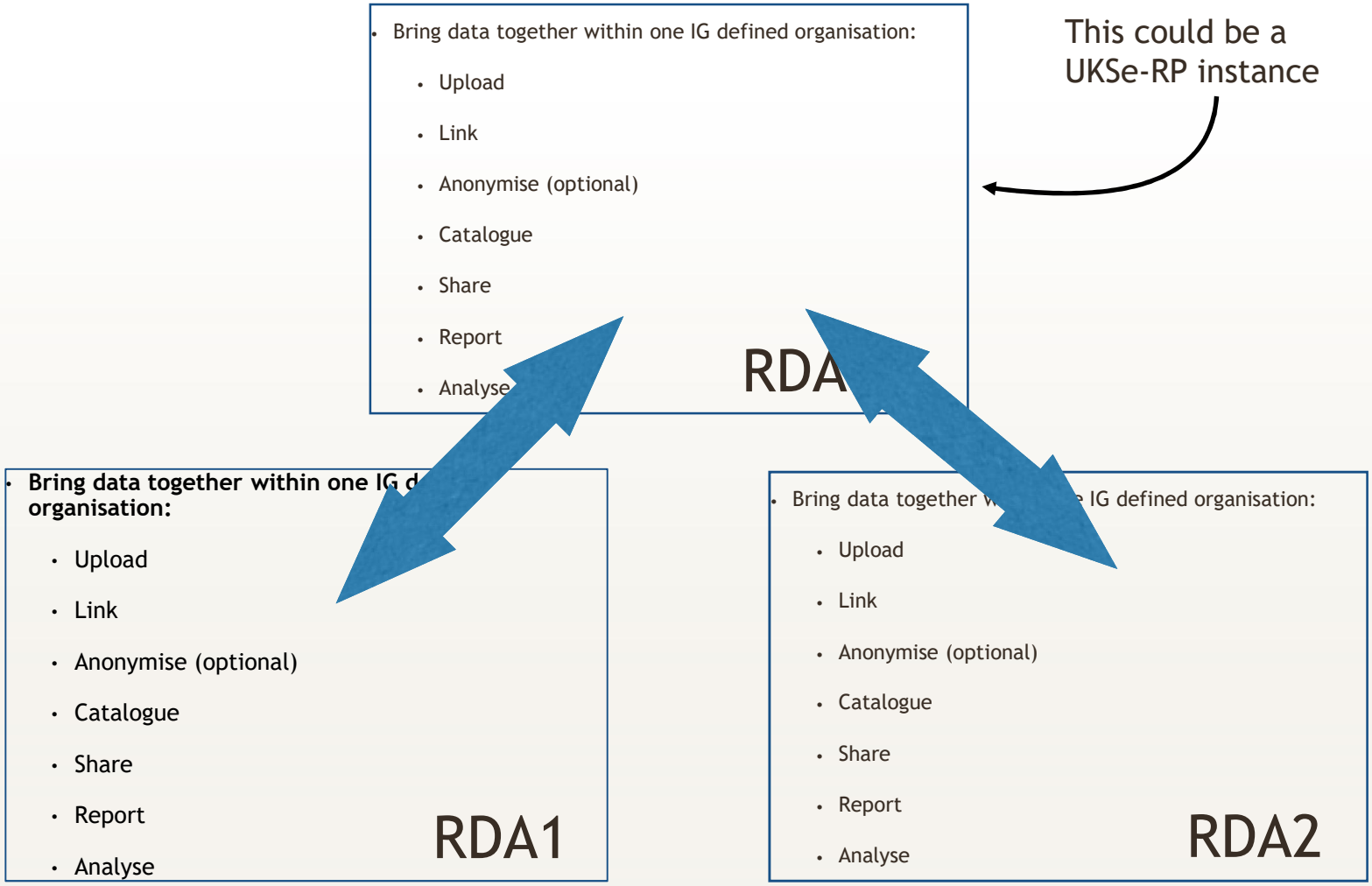


Simplistic Viewpoint

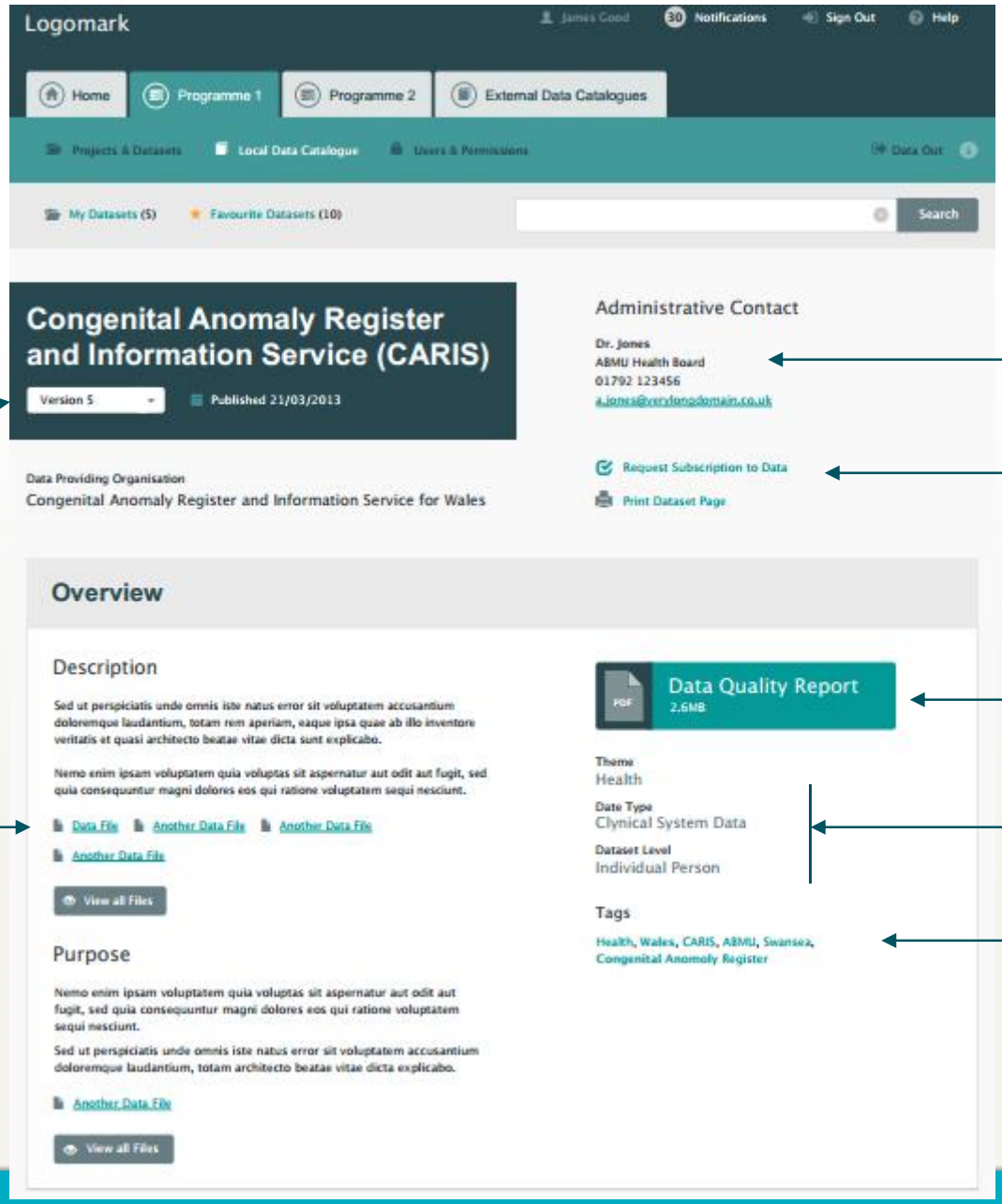


User interface for dataset management
Matching and Linkage
Data Loader
Data Quality
Data Catalogue
Pluggable architecture

A Simple RDA deployment model



A Dataset



Specific version & Date

All section attach files

Contact

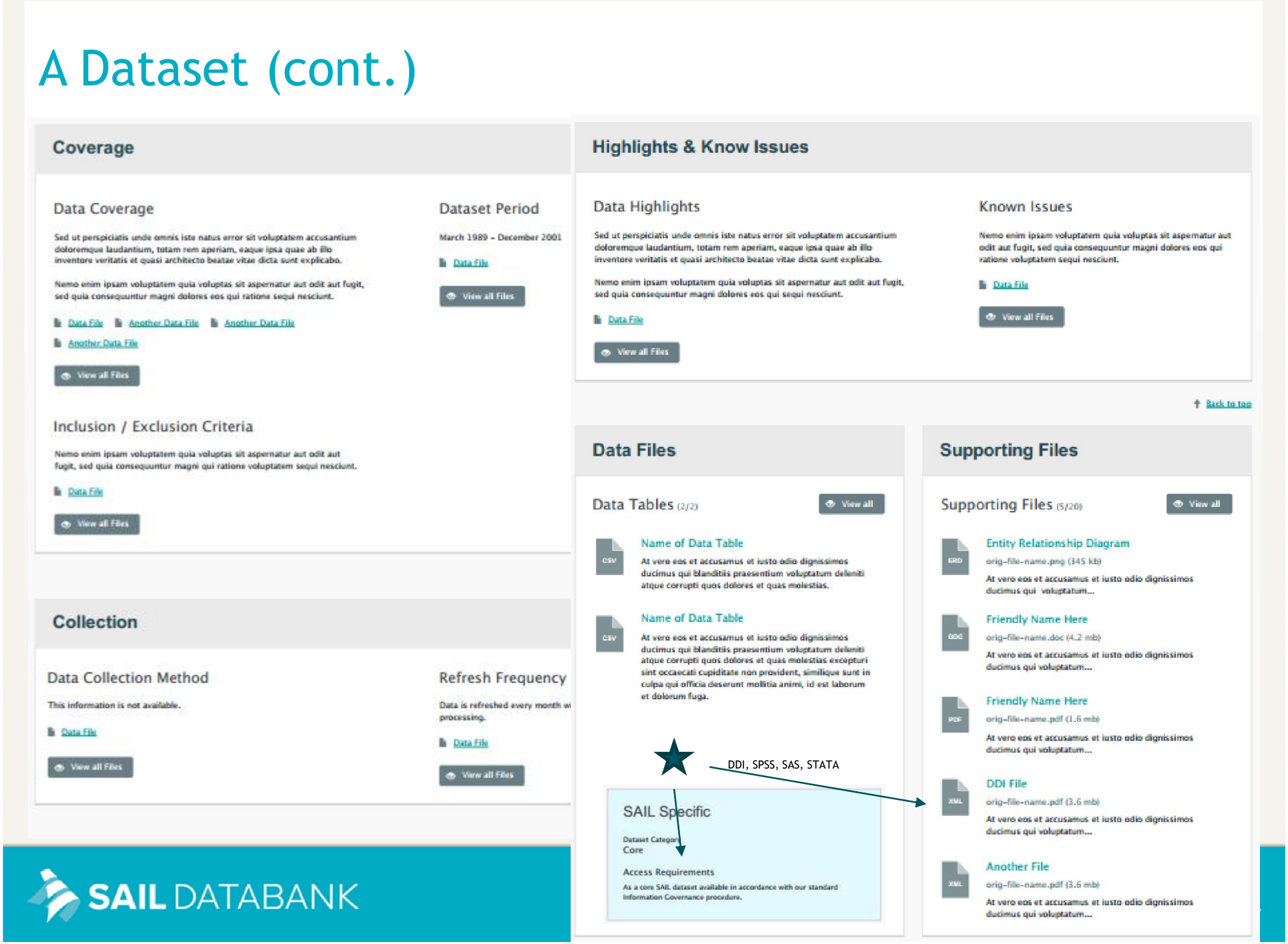
Request

VIMO

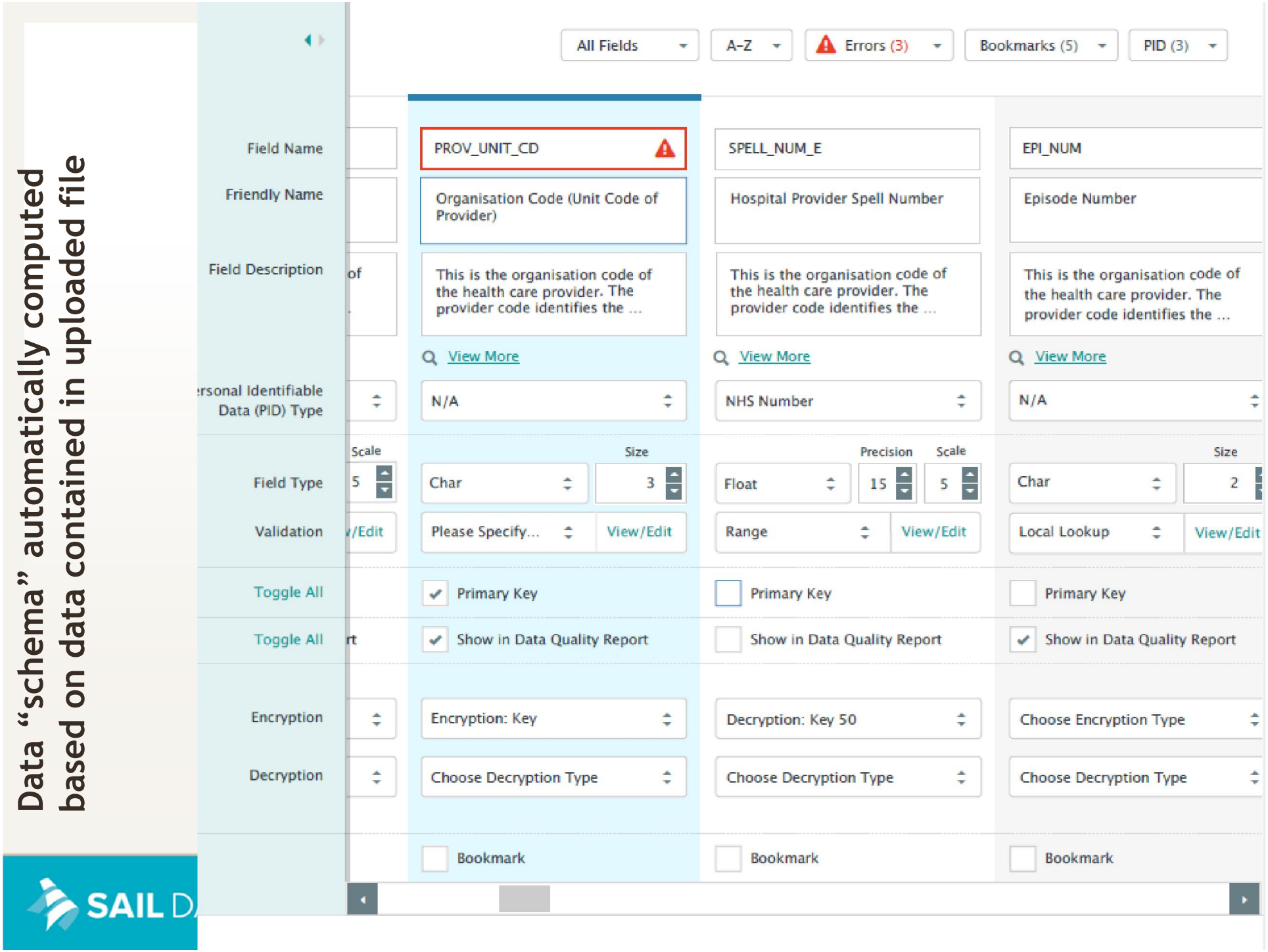
Theme / Type / Level

Tags

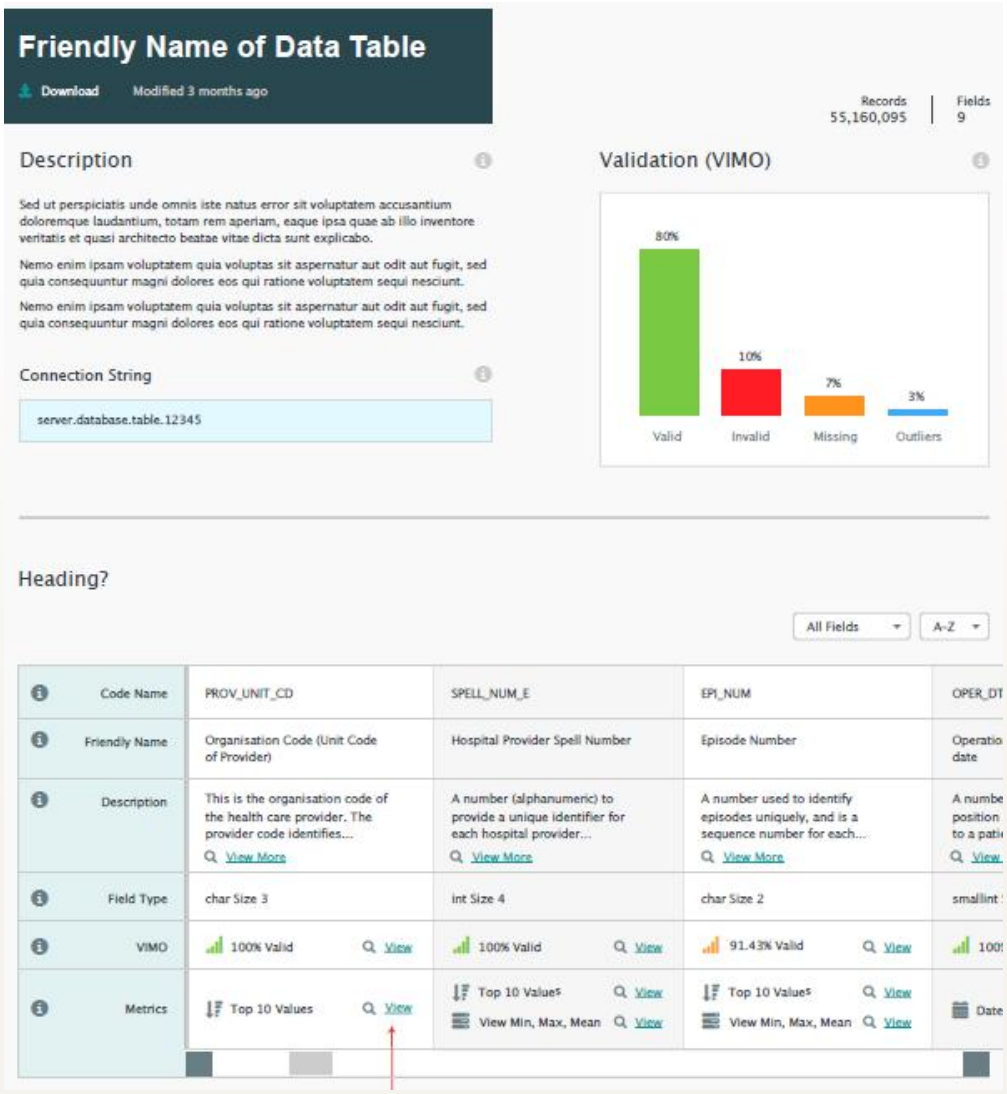
A Dataset (cont.)

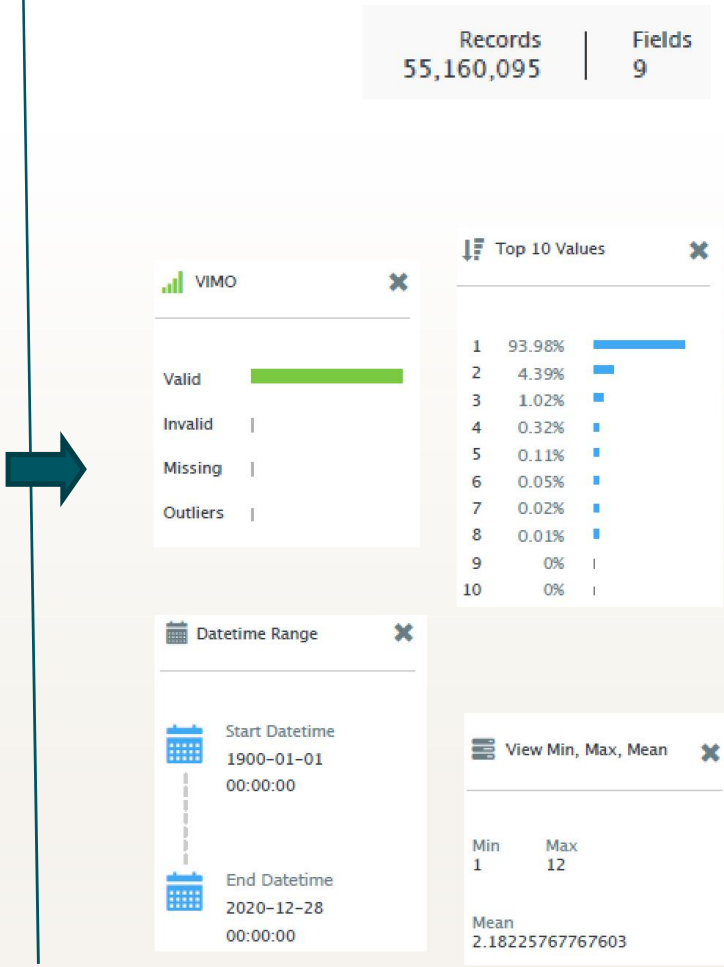


Data "schema" automatically computed based on data contained in uploaded file

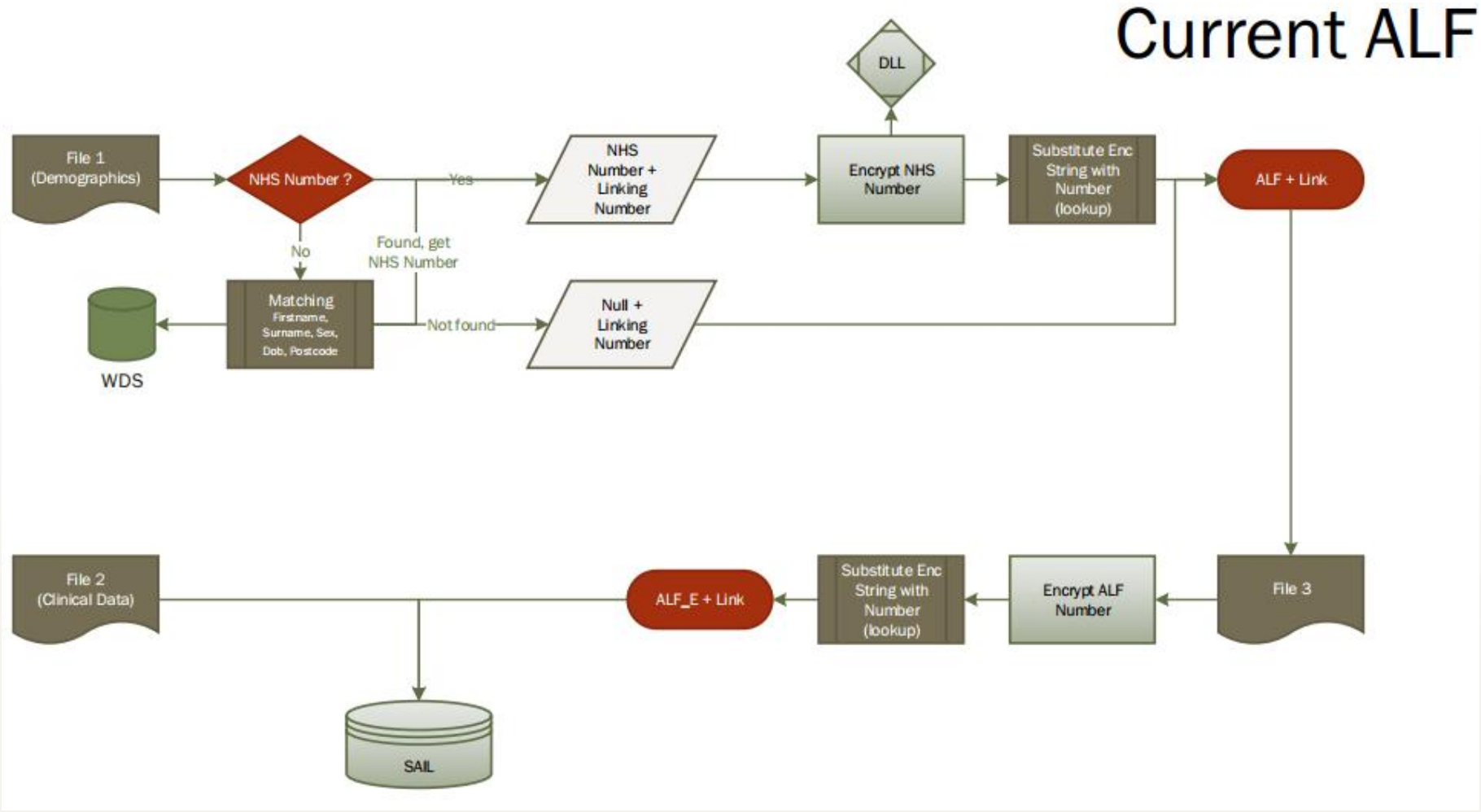


Data Catalogue – a specific table



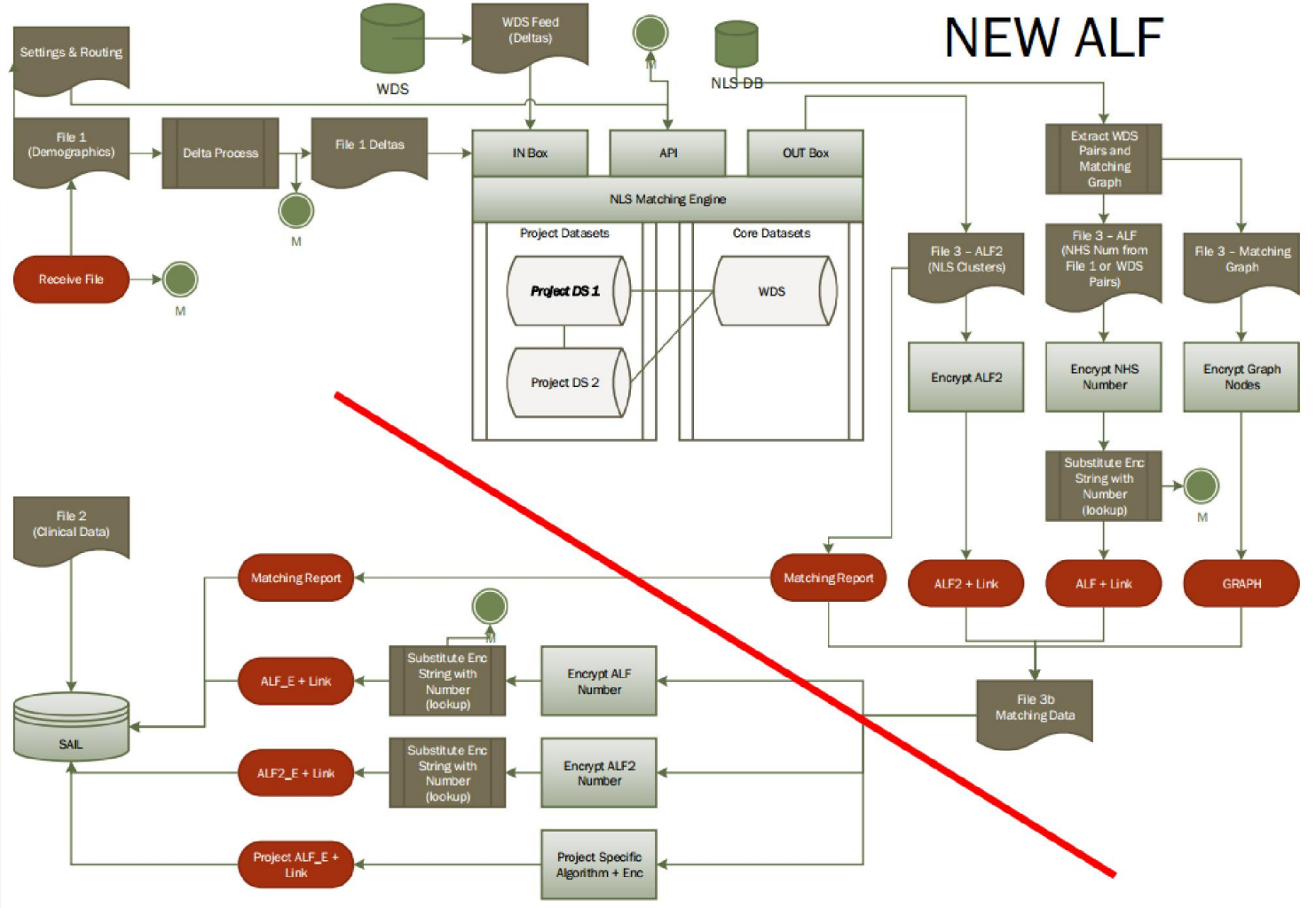


SAIL Split File Principal



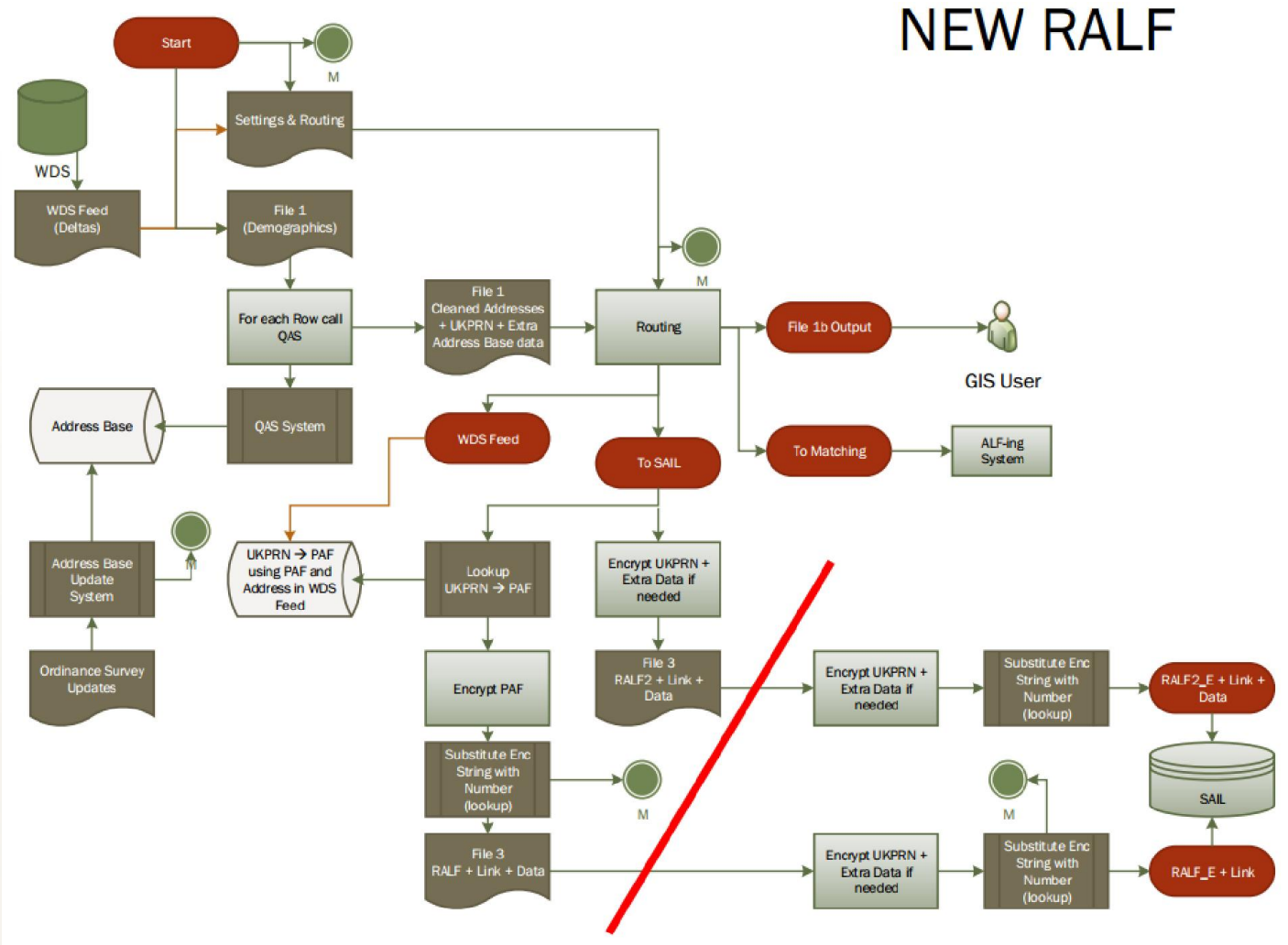
Additional Project level encryption of ALF_E → PALF_E

New ALF Process



RALF – Residential ALF.

Relevant as now we can pipeline address cleaning to improve matching

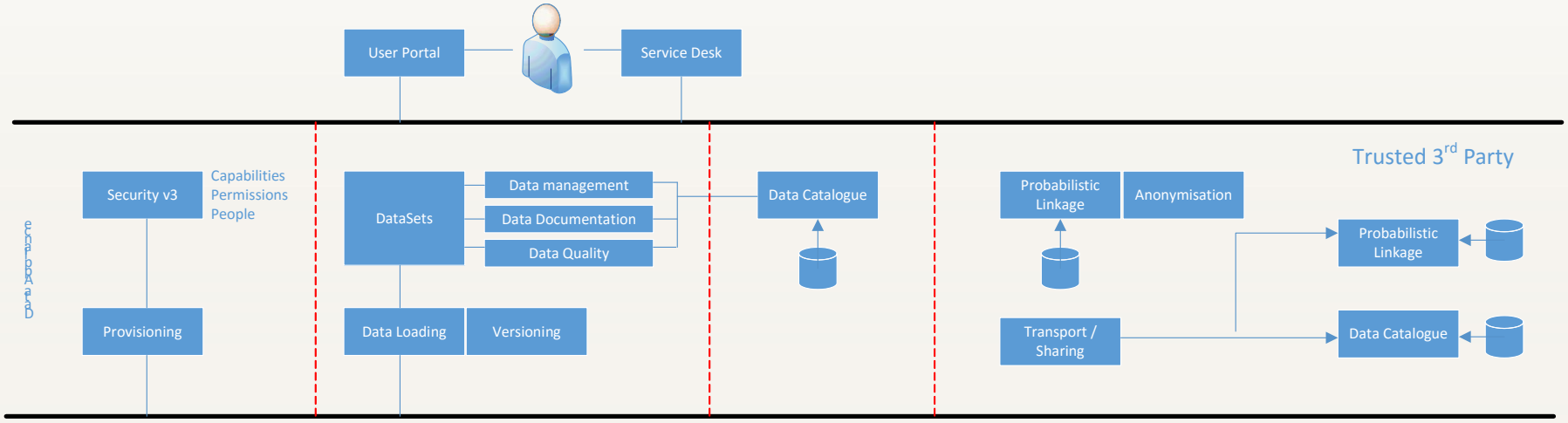


1000 foot view of new linkage capability

- Create matching pools.
- All datasets fed into a pool will be linked to all other datasets in pool where possible.
- A matching pool can be linked to a “core pool” (latest version any other pool), remove dependence of spine while keeping spine.
- Linkage expressed as a graph of nodes with confidence weighting between nodes
- Clusters identified and numbered – Encrypted id is new ALF
- Current ALF maintained for backward compatibility.
- Pools retain data to enable linkage to any future data added, NRDA has option to remove pool after linkage.

Key areas delivered by NRDA

- Security and Access control
- User portal
- Dataset management
- Data Catalogue
- Probabilistic linkage
- Anonymisation
- Sharing and transportation
- Optional Modules



3 initial configurations – plug and play single cable

Small (Development / Demo)

- Single servers everything on. 4 cores, 6gb
- Web site, Workflow engine, Modules, SQL Express, MongoDB, RabbitMQ

Medium (Single Physical Server)

- **Single HyperV server, multiple v-servers for different roles. Dual 10 core CPU (40 virtual cores), 160GB memory, 6TB Disk**
- **SQL Express replaced by SQL server Standard 2012**
- **10 special versions having extra modules for CliniThink NLP**

Large (Four Physical Clustered Servers)

- Dual Server HyperV server, Dual 10 core CPU (40 virtual cores), 160GB memory, shared 24TB Disk
- Dual SQL server 2012 Enterprise, Dual 8 core, 96GB memory, 22 x 300gb local disk, SQL 2012 Server AG Cluster

---

Software : Custom software in C# .net 4 / MVC 4. RabbitMQ, MongoDB, RavenDB for Large version, GoodSync FTP Replication

Costs for medium version : Hardware plus licencing for Microsoft Windows Server 2012 Enterprise & Microsoft SQL server 2012 standard

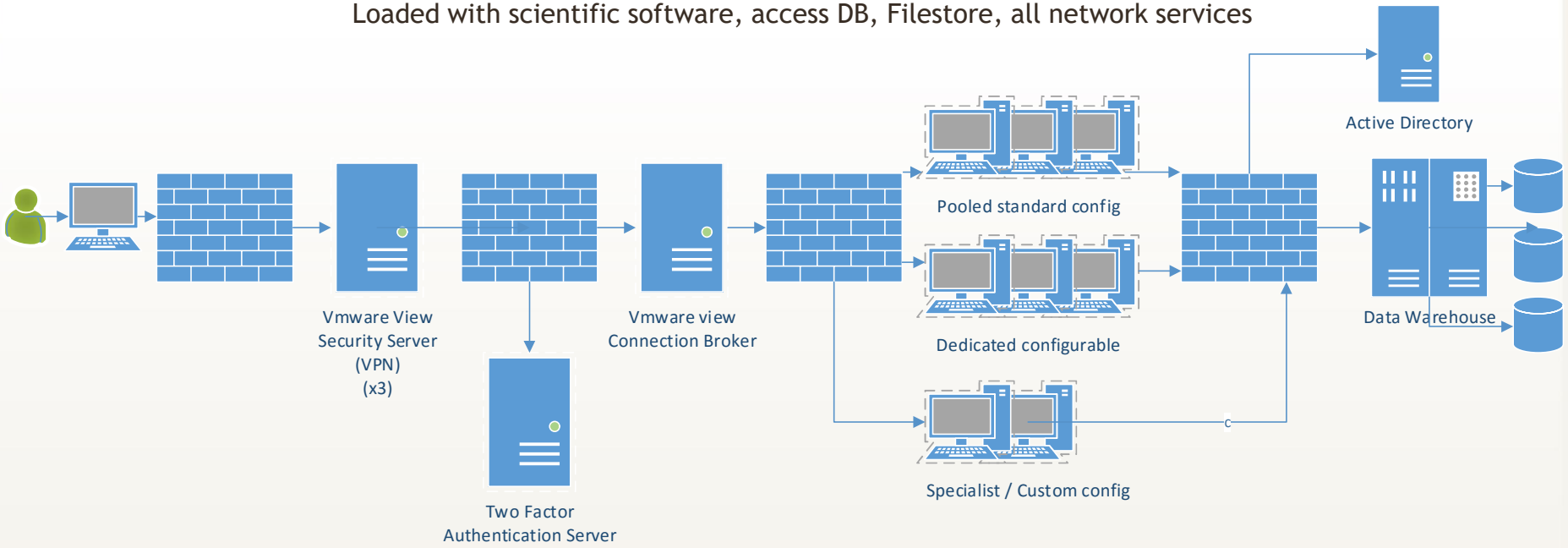
FARR @ Swansea – Capital Investment

UKSeRP

Access Portal

End User Experience – Remote desktop to Windows 7 / Ubuntu shared/dedicated system

Loaded with scientific software, access DB, Filestore, all network services

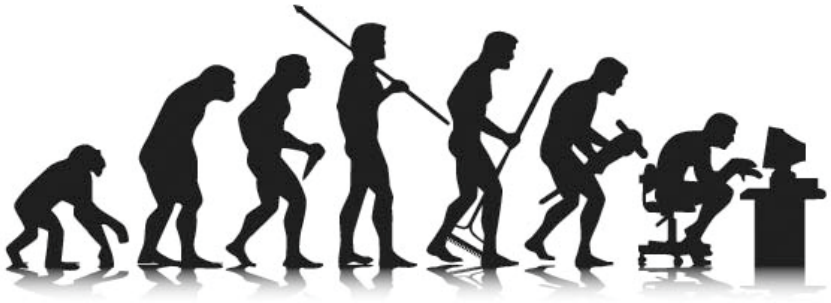


Admin / Cohort Experience – Access to NRDA:-

control and load datasets,

control users, permissions & access

UKSeRP – Expanding



Additional SQL server 2014 AV cluster

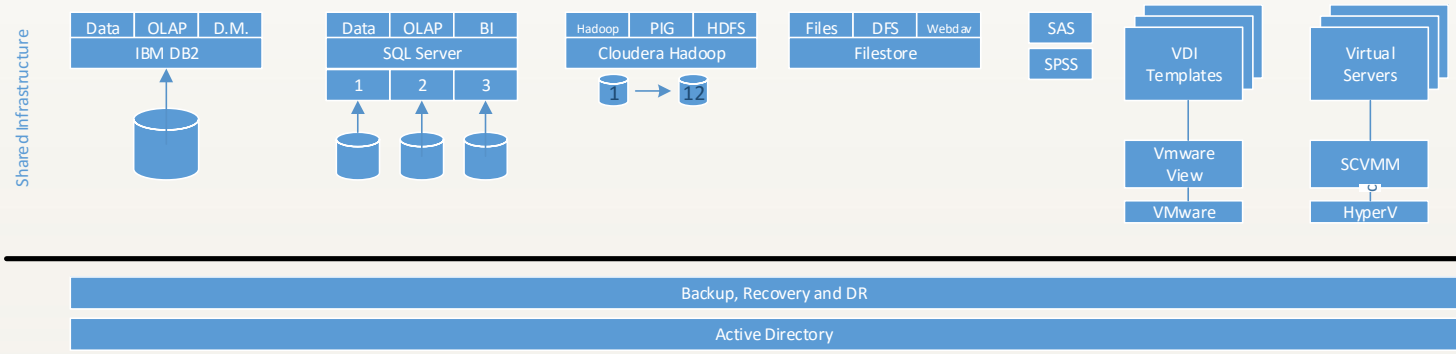
Additional HADOOP Cloudera Cluster

Additional PostgreSQL Cluster

Additional Remote Access servers (double)

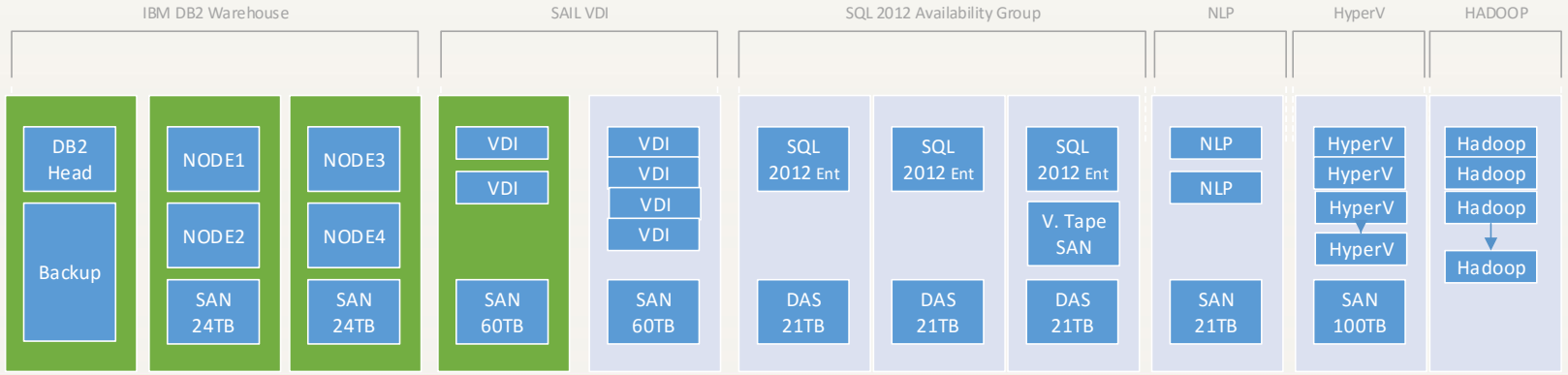
Additional Virtualisation stack

Additional Natural Language Processing Infrastructure

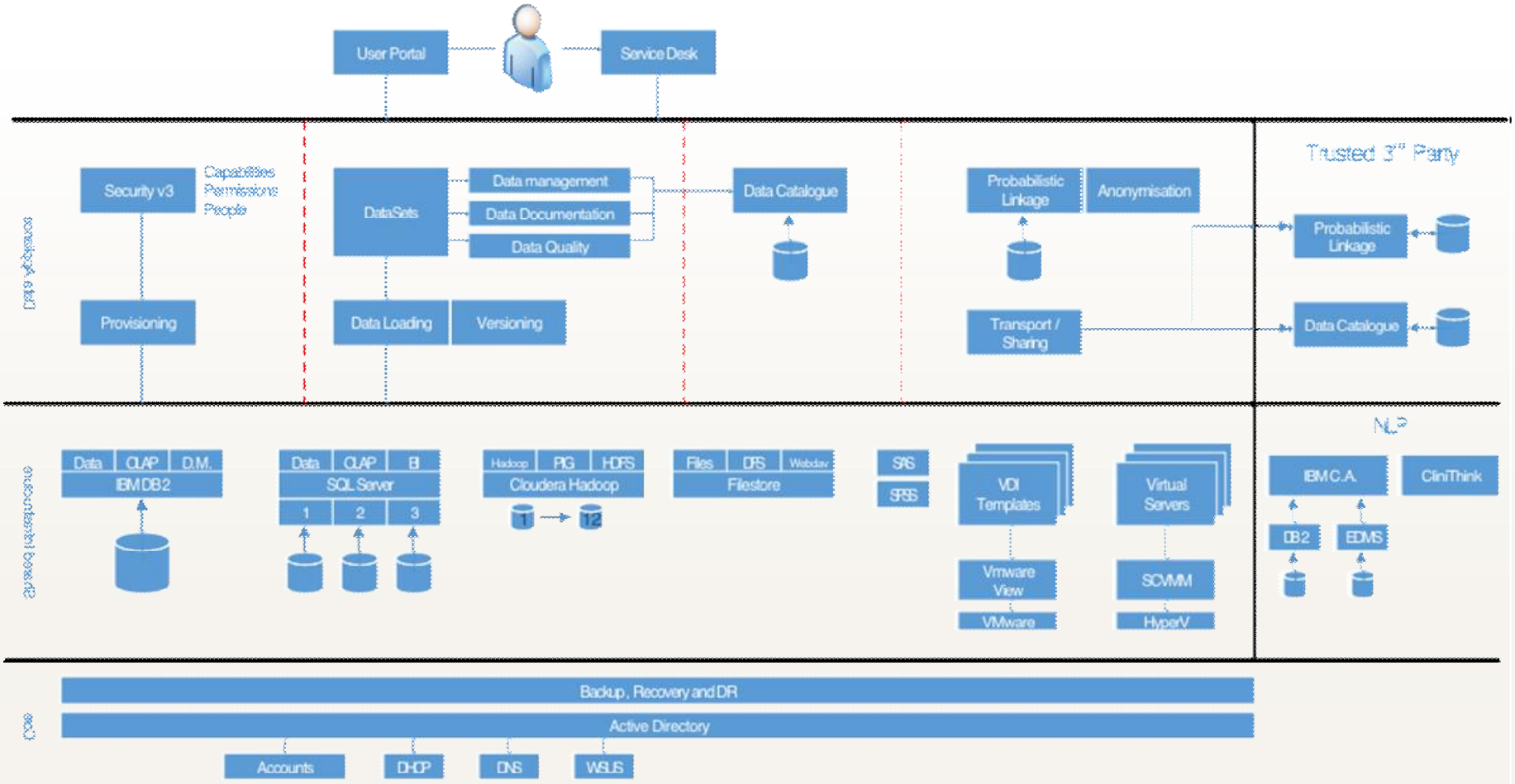


FARR – UKSeRP (quick overview)

- Expand Technical Platform
  - Double SAIL Gateway and increase power of each desktop
  - Add software e.g. SAS & BI tools
  - Add SQL server 2014 3 node AG Cluster
  - Add HADOOP cluster – big data
  - HyperV clusters
  - IBM ICA / NLP
  - Additional 10 racks
- Additional management requirements
  - Now three database platforms
  - Selective delegated management and control
  - Multiple configuration and security models



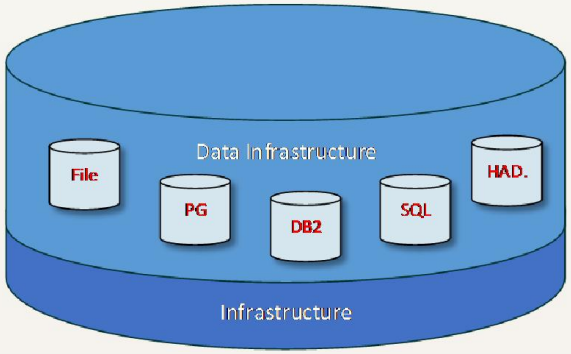
UKSeRP uses NRDA



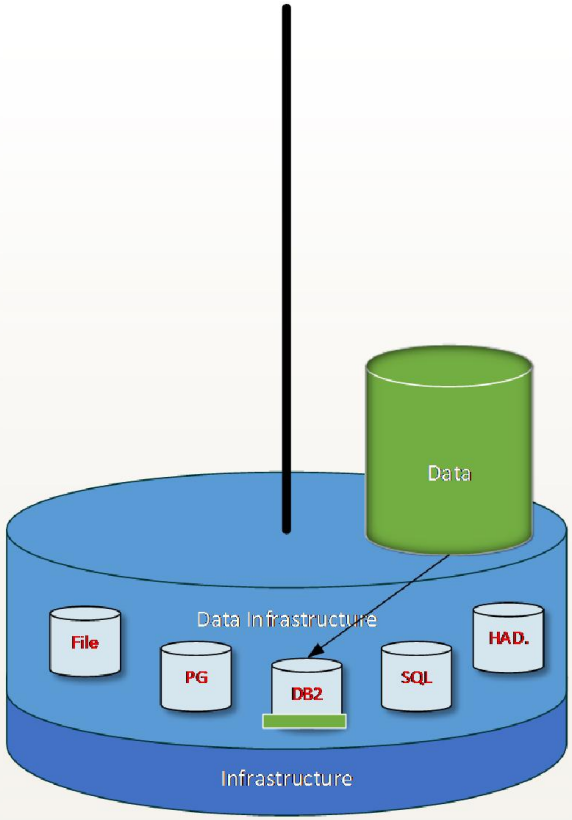
UKSeRP allow others to build on this infrastructure

Better ROI for funders

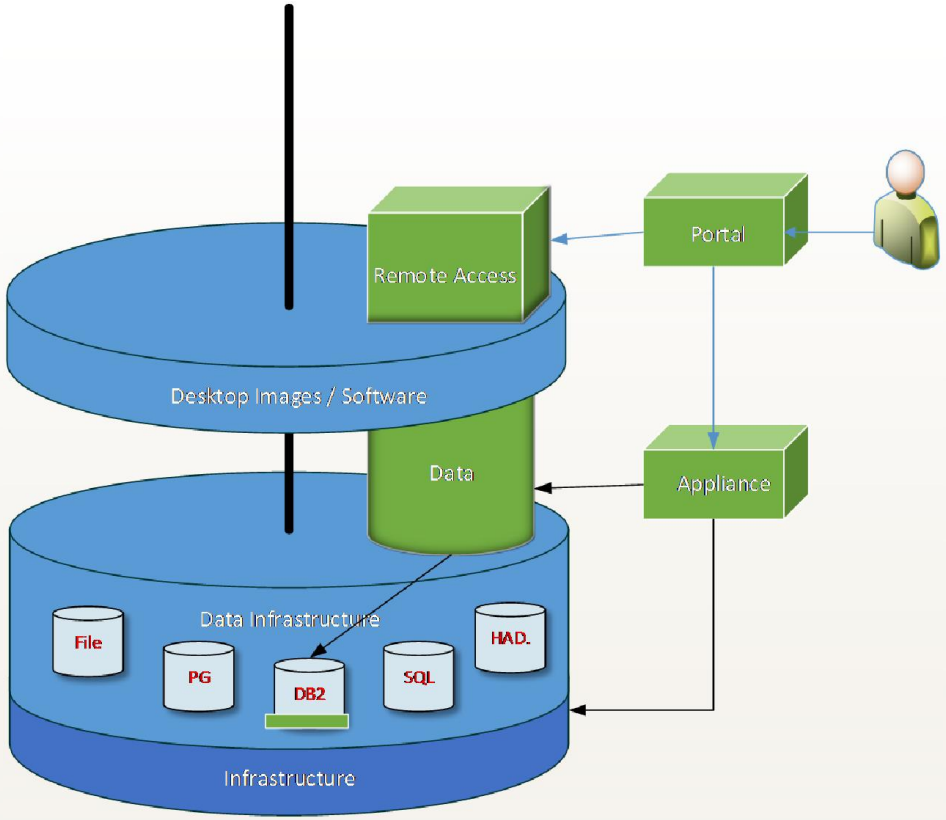
- Researchers can focus on doing research
- IT specialist run the IT
- Better performance from combined infrastructure



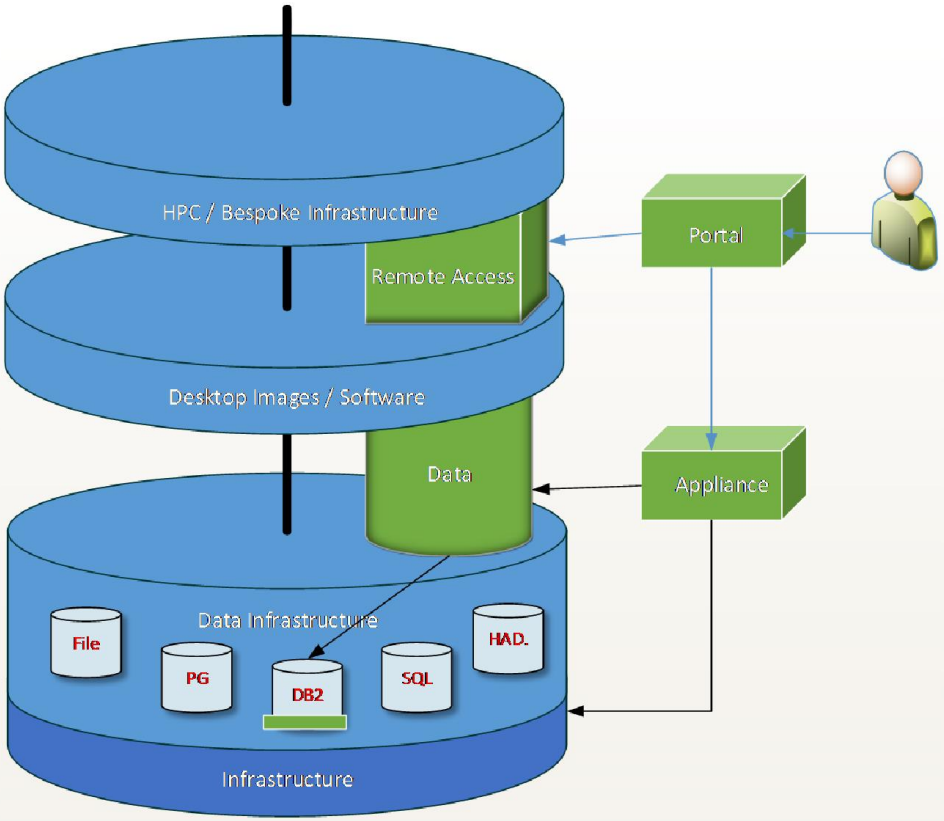
Programme bring data in and choose platform



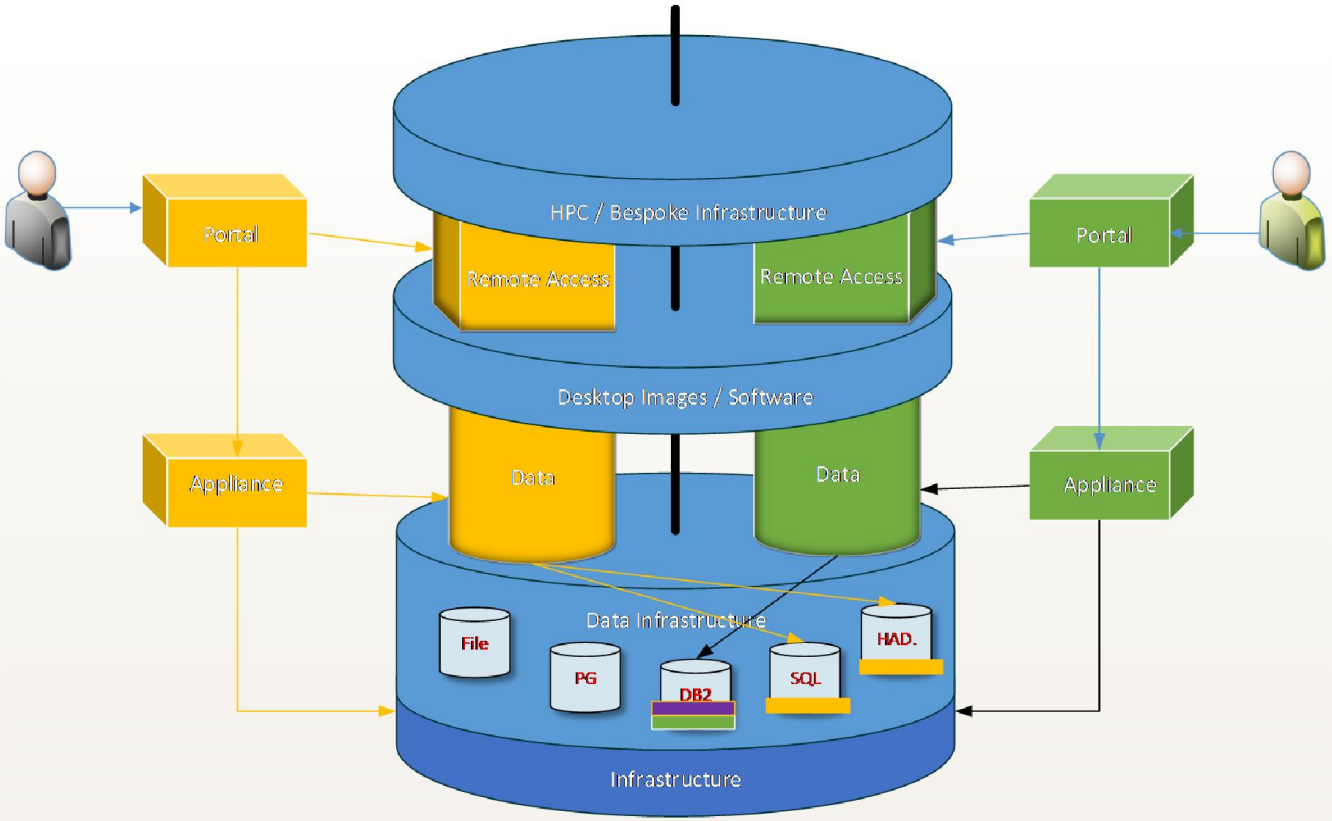
Provide remote access, automate data loading and management, governance, control



Bring in any extra/bespoke infrastructure



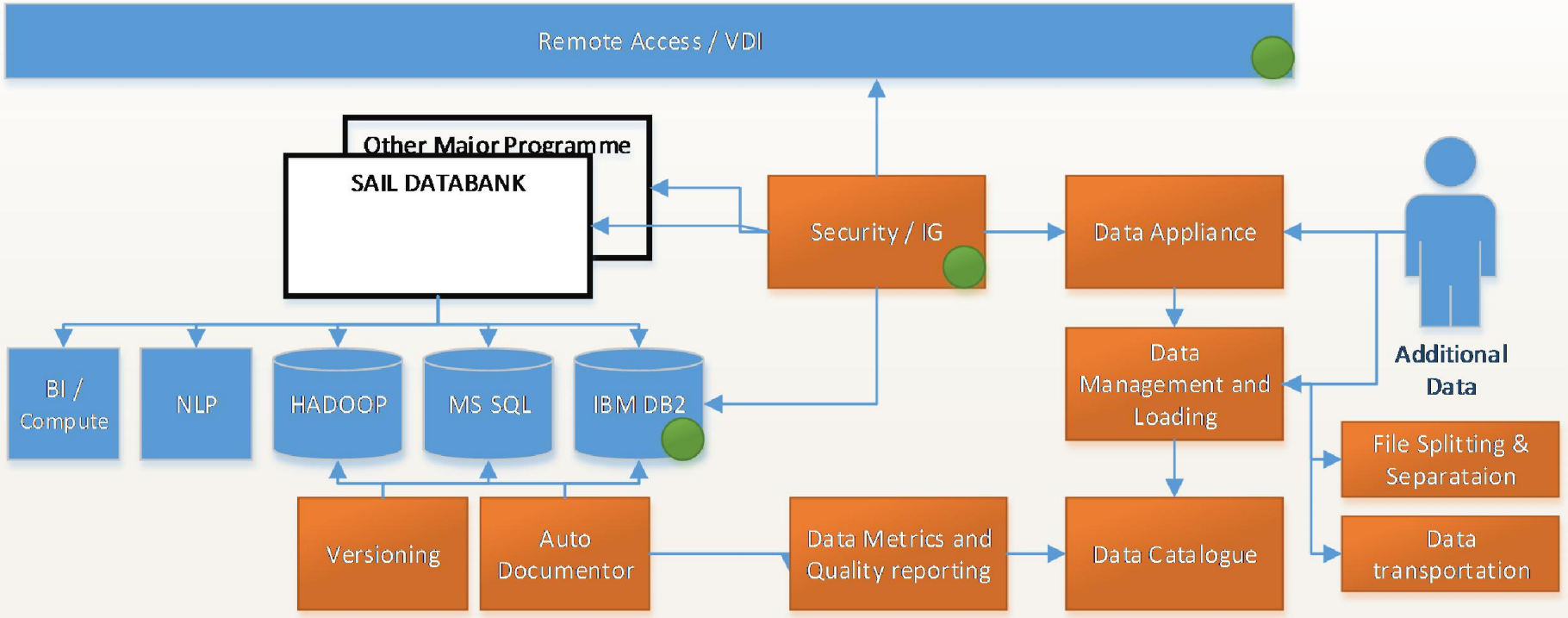
Environment can be shared



Why such a disruptive technology (9 months to build !!)

SAIL Databank has/will/is becoming an instance of UKSeRP

and fully dependant on Data Appliance



UKseRP is born

SAIL

CIPHER

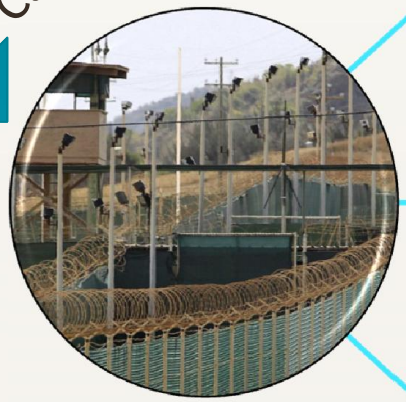
Farr

Additional Technical Capabilities

SAIL

ADRC

DPUK



IaaS

Significant Infrastructure,
Policies and Procedures,
Information Governance,
Knowledge & Reputation

UK Secure e-Research Platform (UKSeRP)

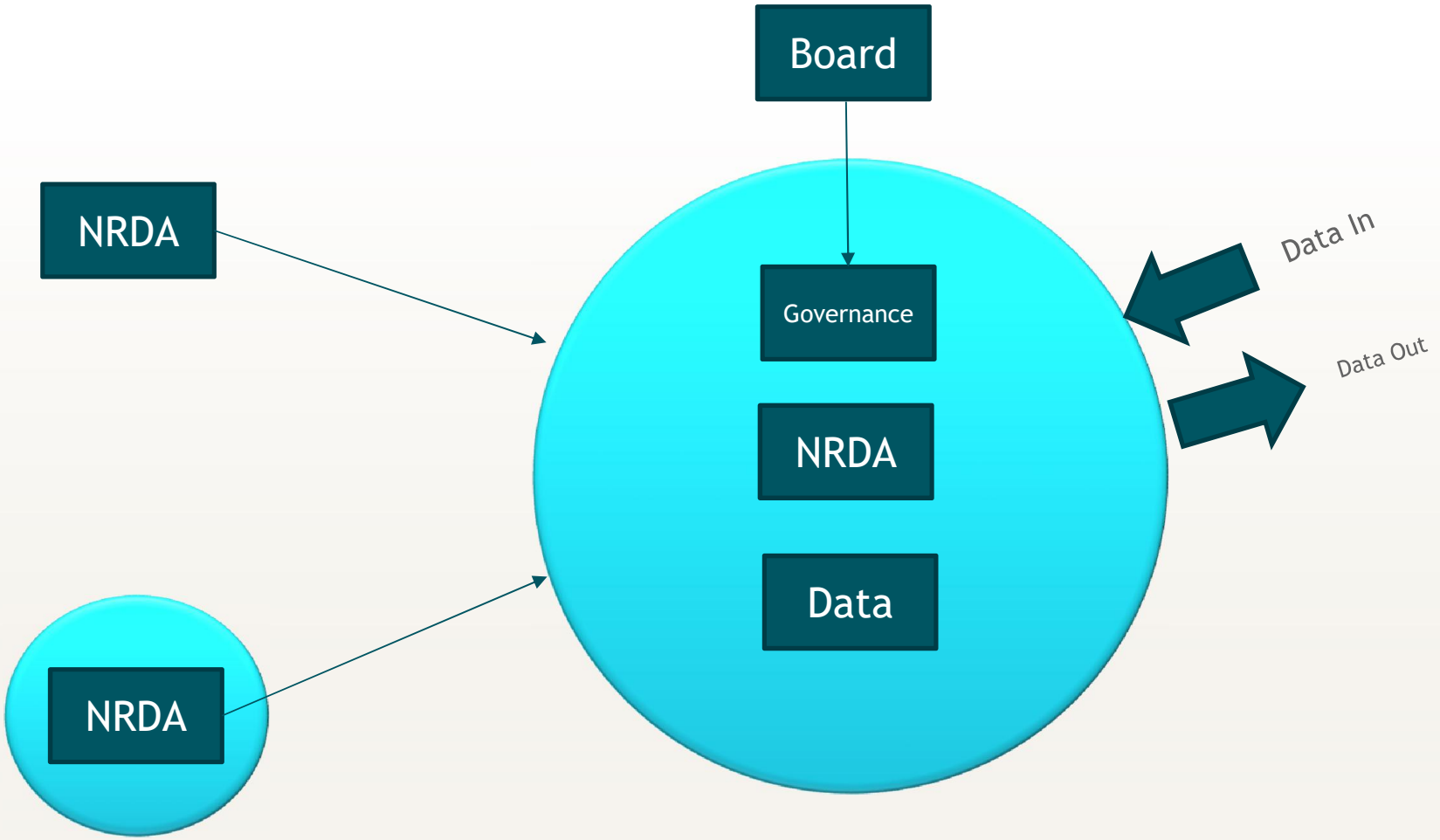
- Large scale data and compute platform
  - Performance and scale

- A remote access analytics platform
  - Best practice: data management, security, governance
  - Suite of standard and bespoke data analytical tools
  - Accessible across UK and internationally

- Leaves data ownership with the cohorts/programme
  - Each ‘controls’ slice of UKSeRP
  - Devolved account and access control
  - Information governance remains with cohorts/programme
  - Brings together data for DPUK across cohorts

- Enables researchers to focus on the science Reduce Costs, Reduce Risk

Wider Governance Models (deviate from SAIL's chosen model)



Natural Language Processing (NLP)

Two Different Approaches

- 10 NRDA will carry CliniThink with a licence of 1 million documents

  across all appliances

- Large capital investments in a small IBM Content Analytics (Watson)
  - Redaction gives an undefinable confidence ! So system will run more in the NHS and in the University

As we effective scale up changes are required :-

- Service management (ITLv3)
- ISO 27001:2013
- BIL4 server room and BIL3 building & Pan Government Accreditation
- GCHQ Approval – Security Clearance SC (my team)
- Technical Responses
  - 3 Data centers moving to 10GBe
  - 4 different vender firewalls (min of 2 at perimeter)
  - DR Solution
  - Full system Center deployment
  - IPS / IDS / Pen Testing
  - SEIM implement NSA guidelines for AD
  - Secure Building – taking control of door access system (swipe in/out) and digital CCTV georeplicated
  - Configuration managemet. Build Servers + Octopusdeploy. Linux – Anisible, Windows – Puppet / MS DSC / SCCM

Health Informatics Group
ESRC ADRC - Wales

- ADRN is a UK-wide partnership:
  - Universities
  - Government
  - National statistics authorities
  - Third sector
  - Funders
  - Researchers
- ADRC-W one of four centres
- £8m investment from ESRC
- Part of the focus on governmental data sharing
- Information assurance and privacy protection

- Secure environment for research
- Using SAIL infrastructure in Wales, with UK SeRP
- Aimed at UK social researchers
- Help accredited researchers carry out social and economic research
- Help to using linked, de-identified administrative data – information which is routinely collected by government organisations.

ADRC

Use of previous investment in systems / knowledge / development

Very similar to FARR at the 1000 foot view, lots of differences in detail

Lots of time on perfecting the design of the wheel – must be a better design than square ???

These dataset have not be shared at scale before – lots of nervousness

Not a repository model

Compile dataset – Do research – Publish – Destroy

Data is transitory and specific to a project

Data Linkage

New linkage capabilities in NRDA required

Possible Encryption at source with linkage

based on encrypted demographics

Security

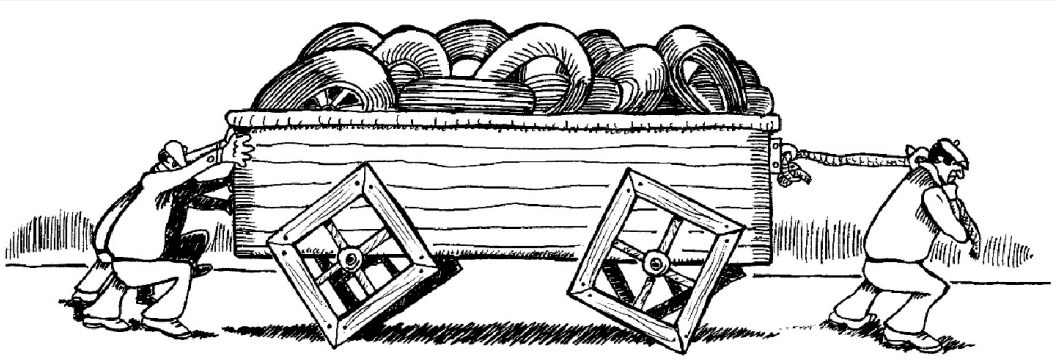
Much higher security requirements –

All researcher must have Safe Researcher Training / Cert

System Admin / Developers – Security Cleared

Safe Setting – linking to Cardiff, Bristol

NRDA and new linkage



SAIL DATABANK

Data Science Building

- New building solely for MRC / ESRC
- Whole building considerably more secure/controlled than any existing buildings.
- A £8 million 2900 sqm purposed-designed building.
- Focus on health and administrative data research unleashing the potential of large scale data.
- Six floors including offices, conference and meeting rooms, high security IT research labs and server room.
- High Performance Computing architecture to support safe and secured sharing of large scale data.
- Adjacent to ILS1/ILS2 and Singleton Hospital sites

Dementia Platform - DPUK

What is the Dementias Platform UK?

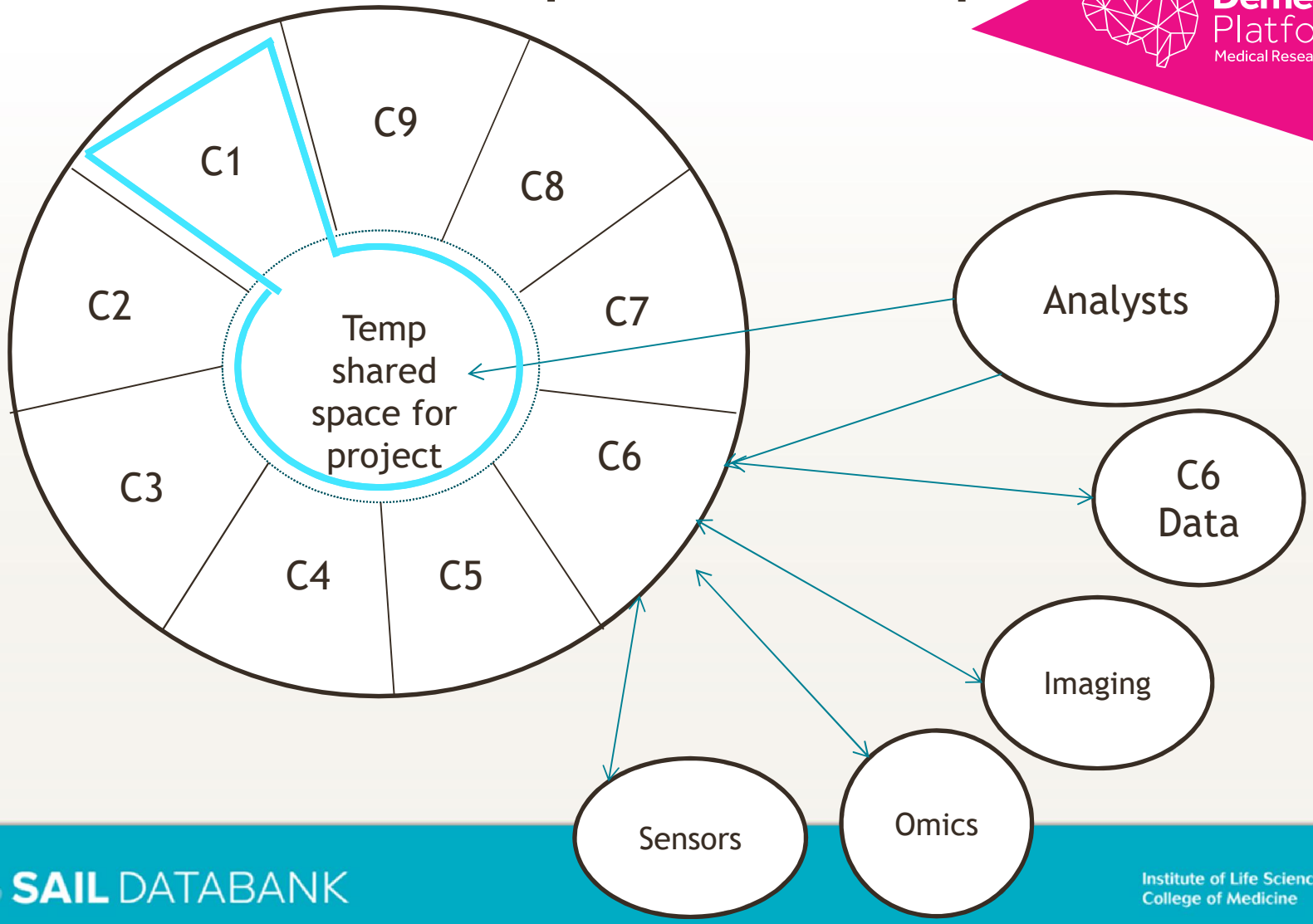
DPUK is a UK powerhouse for dementias research designed to fast-track scientific understanding, treatments and the prevention of the disease.

Established by the Medical Research Council (MRC) in June 2014, it is a £53 million collaboration between universities and drug companies to transform the best dementia research into the best treatments as quickly as possible.

It combines the power of different types of population study to compare healthy people with people at all stages of dementias and can be used to look at how this is affected by other conditions that people may have.

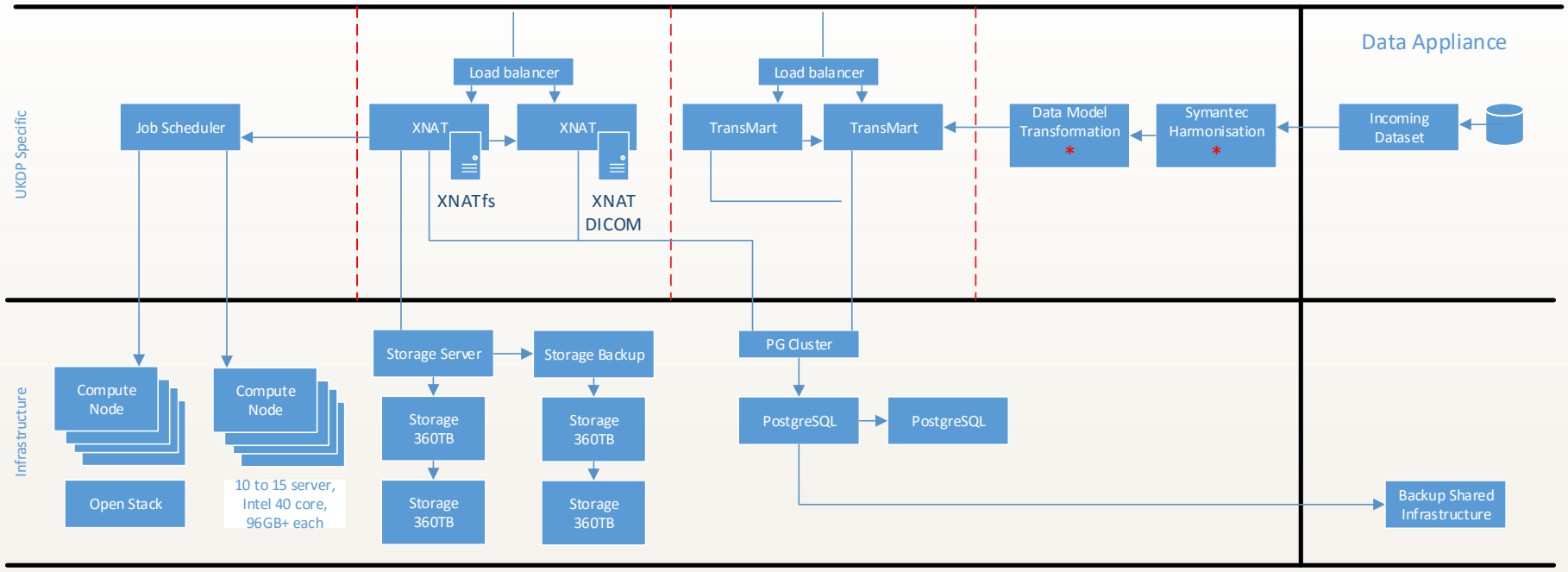
DPUK - expand

WP2: Shared Space Concept

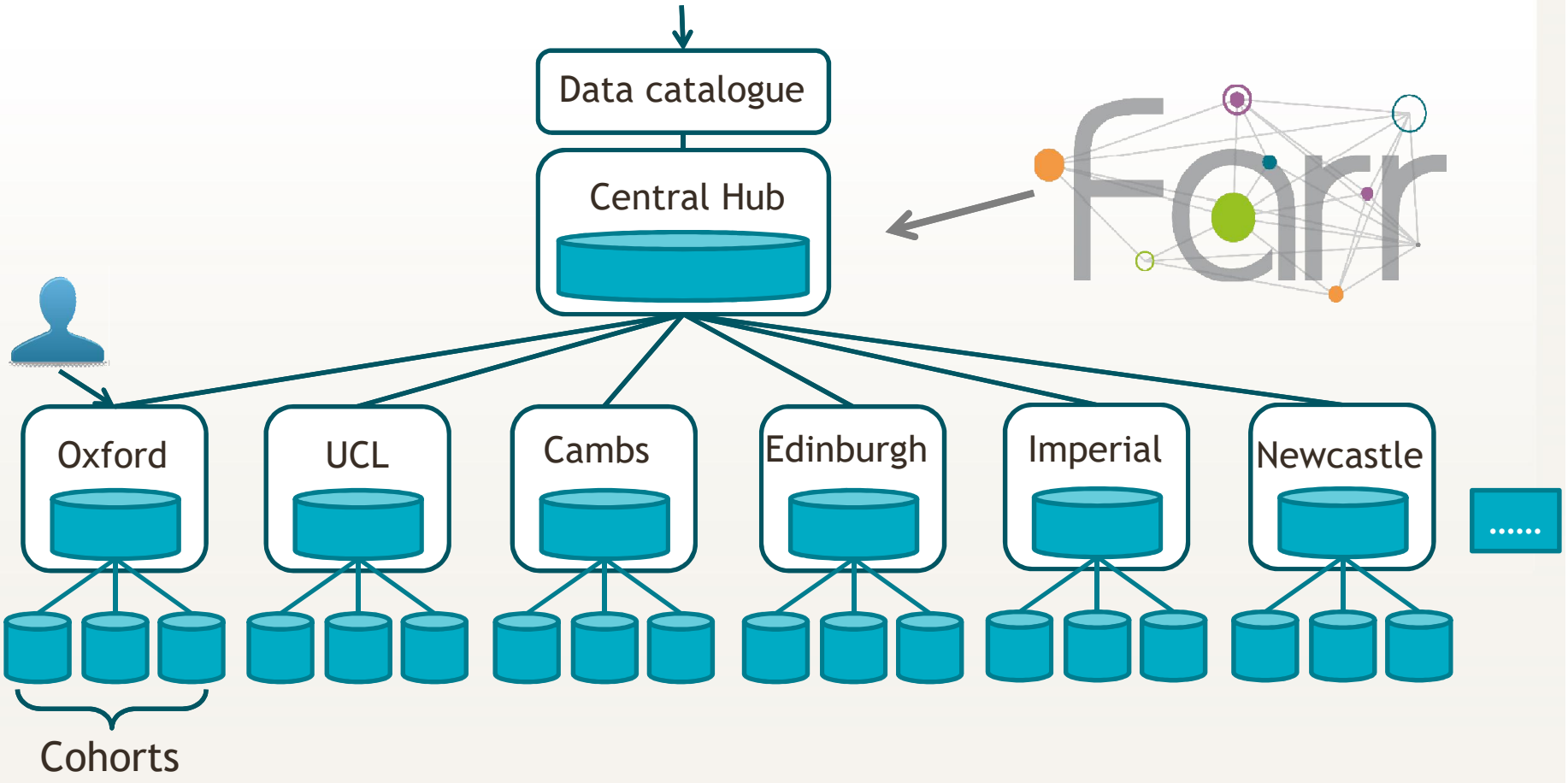


DPUK – Dementia Platform UK

Image storage, HPC Cluster, Transmart, EMIF



DPUK – Imaging WP (recently awarded central hub contract)



Health Informatics Group
UK Multiple Sclerosis Register

- Ground-breaking study designed to increase our understanding of living with MS
- Unique disease register of people with MS in the UK
- 3 data sources:
  - Clinic e-systems
  - Routine data
  - Direct from PwMS (or PROMS)
- Data protection and privacy
- Informed consent

- Secure web portal with mobile compatibility
- Regular, direct contact with PwMS
- Regular Quality of Life (QoL) measurement & reporting
- "Gamification"
- High buy-in:
  - 12,390 MS patients (out of a projected 100,000 across the UK) have signed up

MRC | Cloud Infrastructure for Microbial Bioinformatics

- £4m investment from MRC
- Swansea, Cardiff, Birmingham, Warwick Universities
- Large in CPU or Memory servers as host servers
- OpenStack VM Stack

- Capacity >1000 virtual research servers
- Compute cloud for academics
- 2,880 CPU cores
- 4PB storage (2.8PB usable)

https://www.climb.ac.uk

CLIMB – plug into the side

CLIMB is a collaboration between Birmingham, Cardiff, Swansea and Warwick Universities, to develop and deploy a world leading cyber infrastructure for microbial bioinformatics.

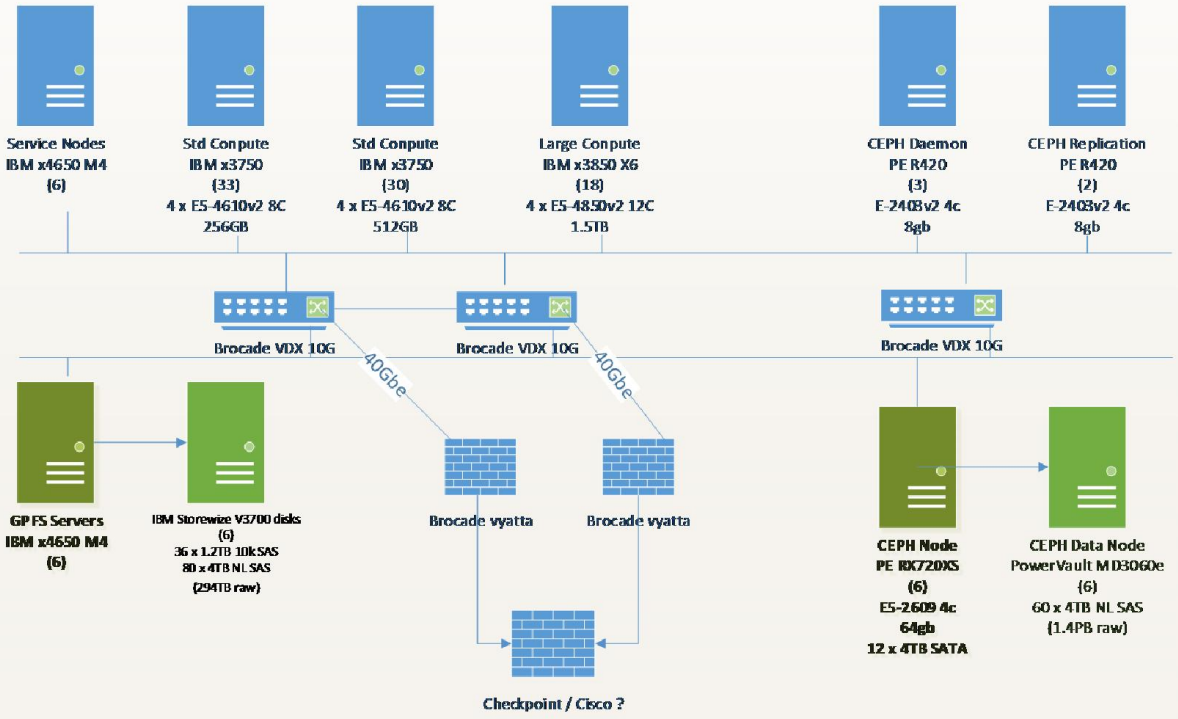
Provision of high memory or high CPU count VM “servers” for research groups

At Swansea:-

- 2880 vCPU
- 50 TB Memory
- 400 TB SAN
- 1.4 PB storage

Cluster: 4 Identical Sites

Goal : 1000 VM’s



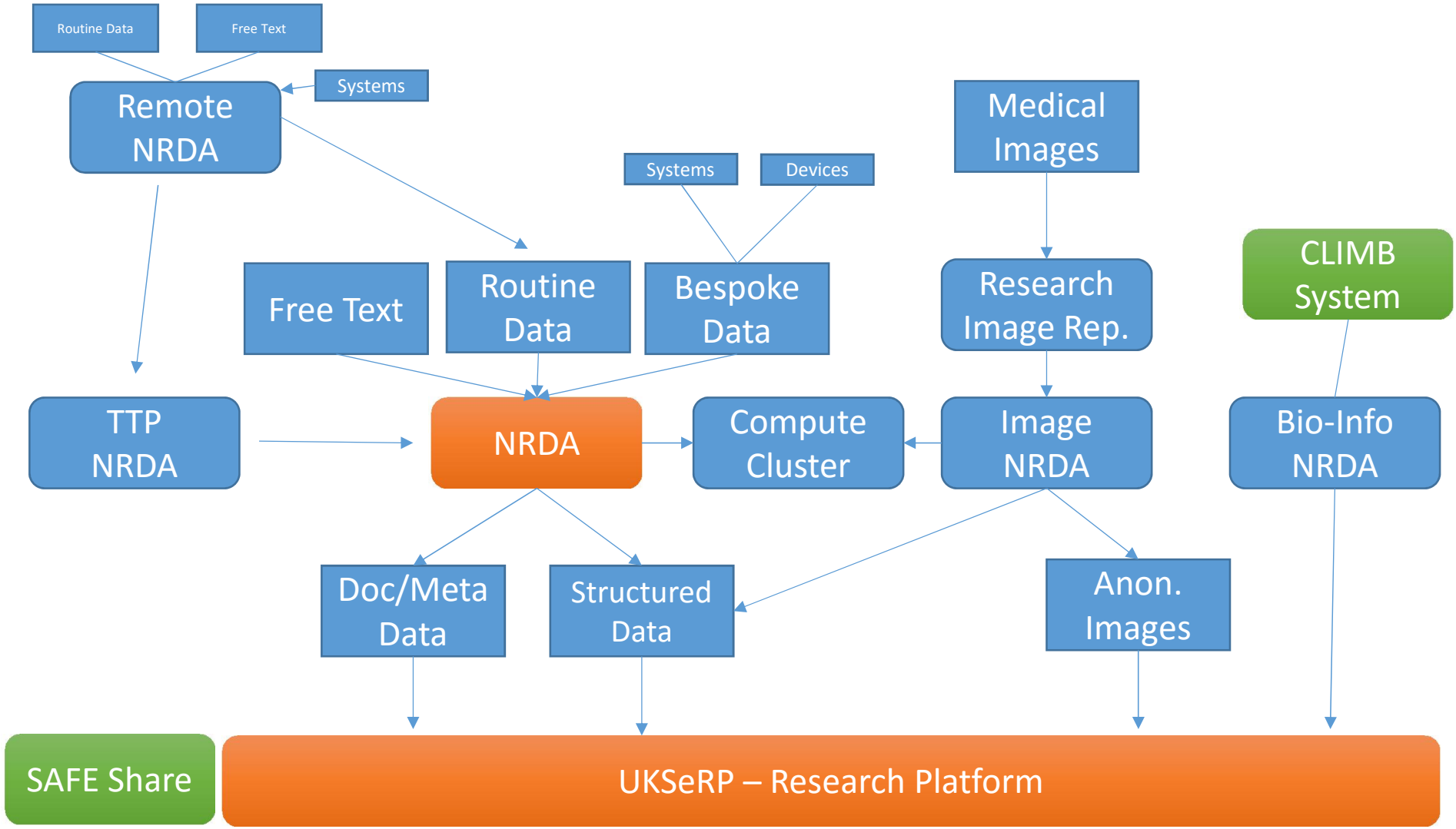
UKSeRP so far

Production	Planning	Discussion	Opportunity
SAIL	ALSPAC	E. Oz	CLIMB
ADRC	BioBank	M. CA	GEL
DPUK	PH/NHS	W. CA app	
MS			

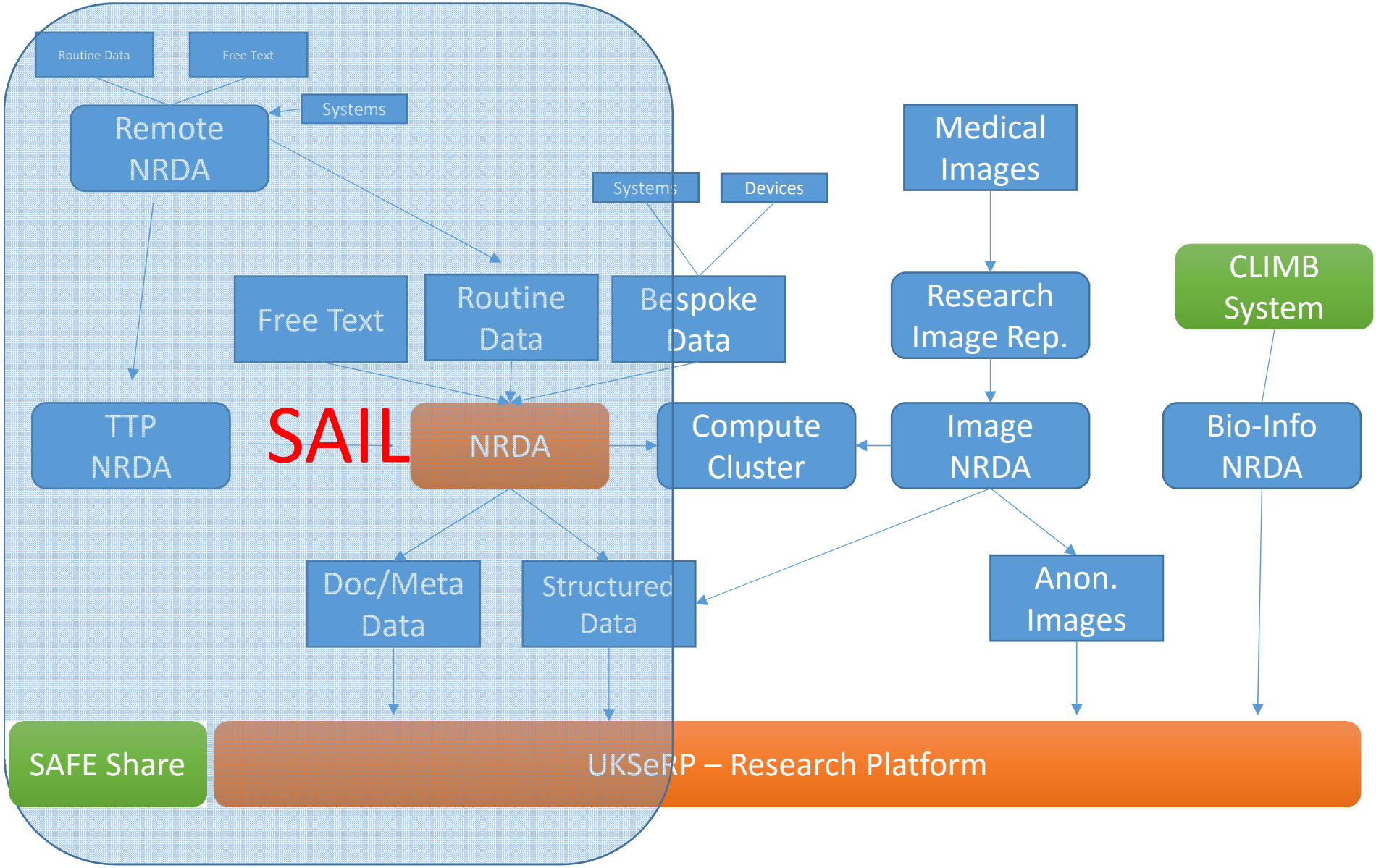
SAIL – Bring in – Data controller
ADRC – Bring in – Project Use – Destroy
DPUK – Bring in – Data Owner control access
MS - Bring in – mix of controller and devolved

Ind. Gov. Board

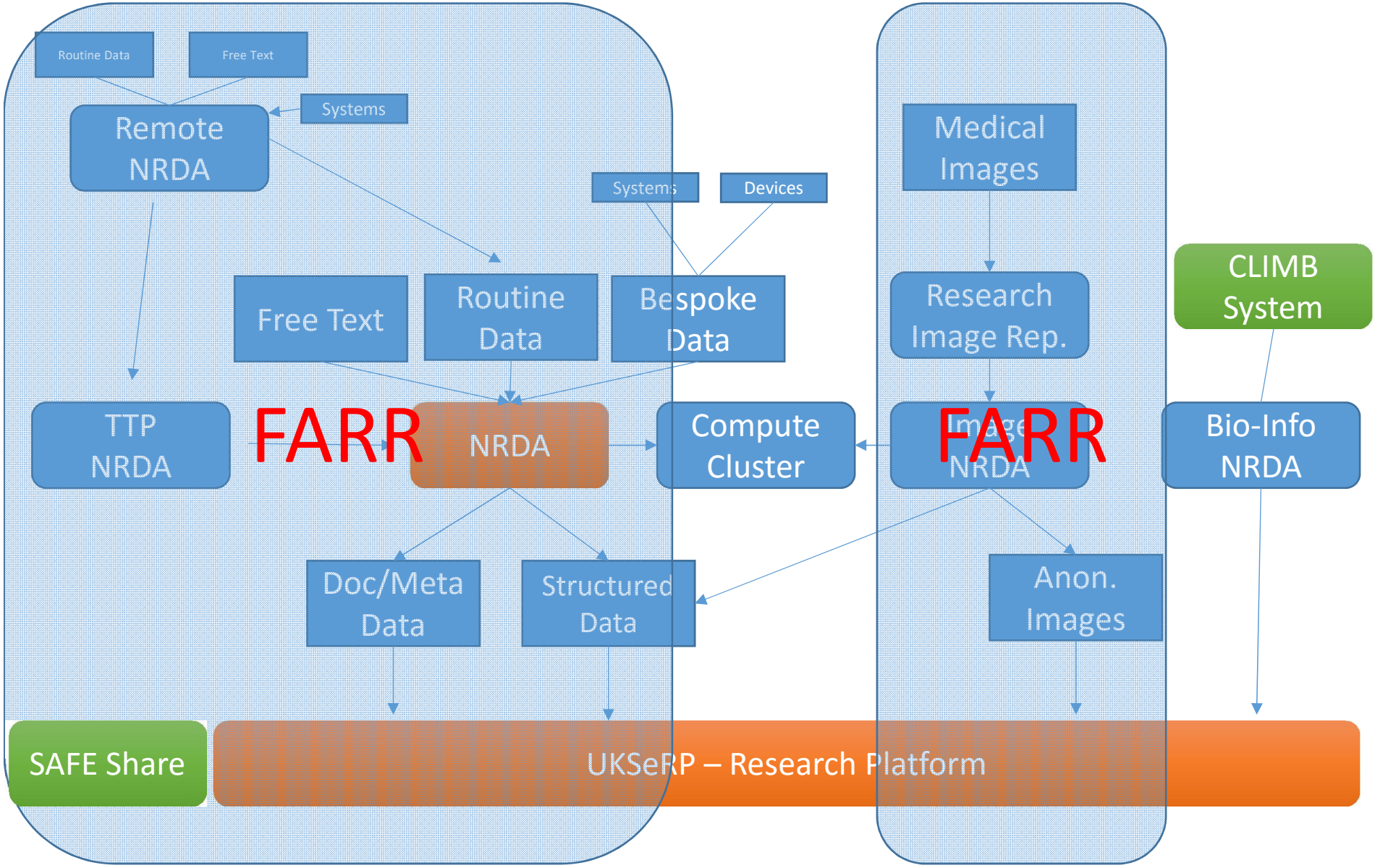
The joining up of efforts and re-use is absolutely critical



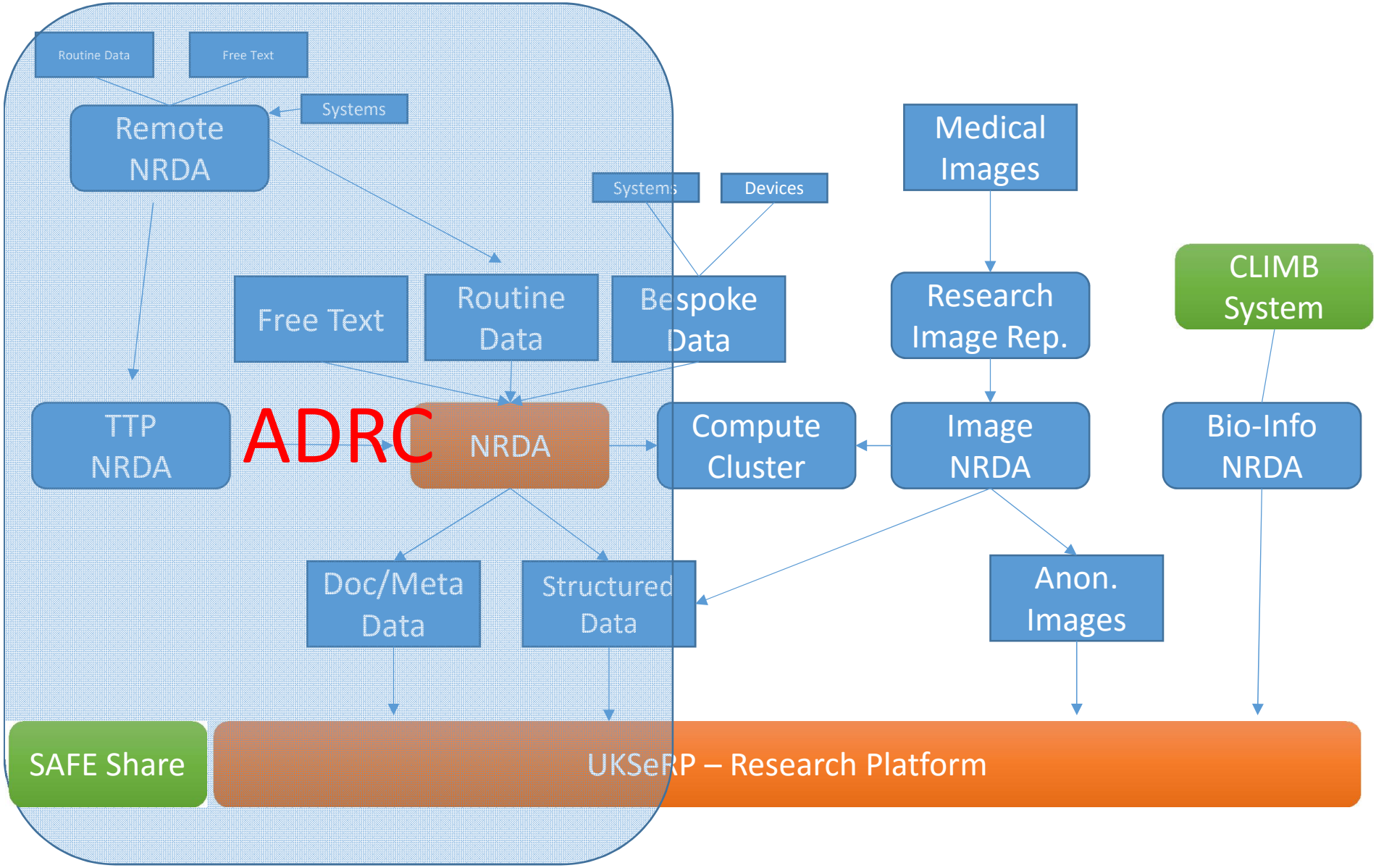
The joining up of efforts and re-use is absolutely critical



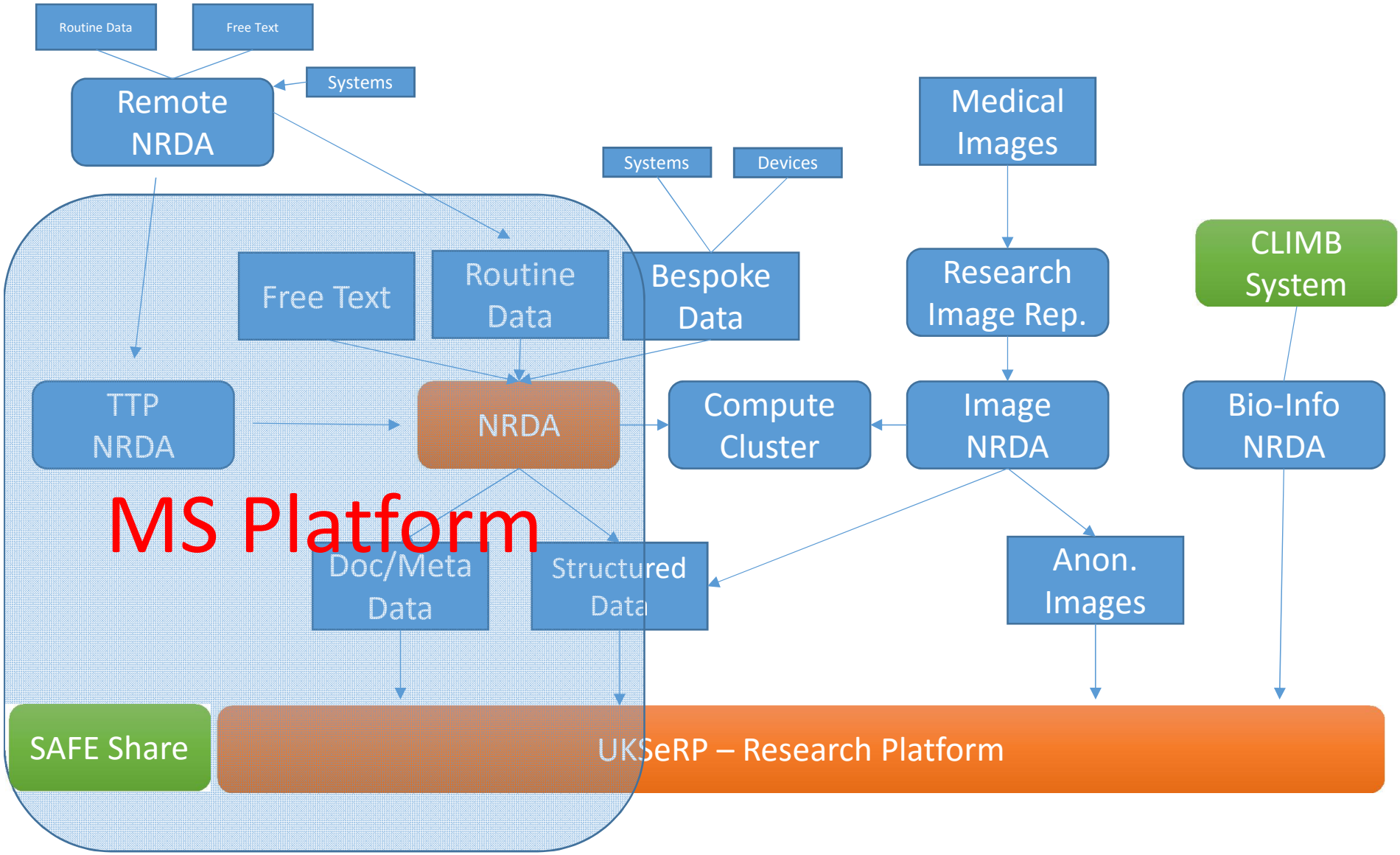
The joining up of efforts and re-use is absolutely critical



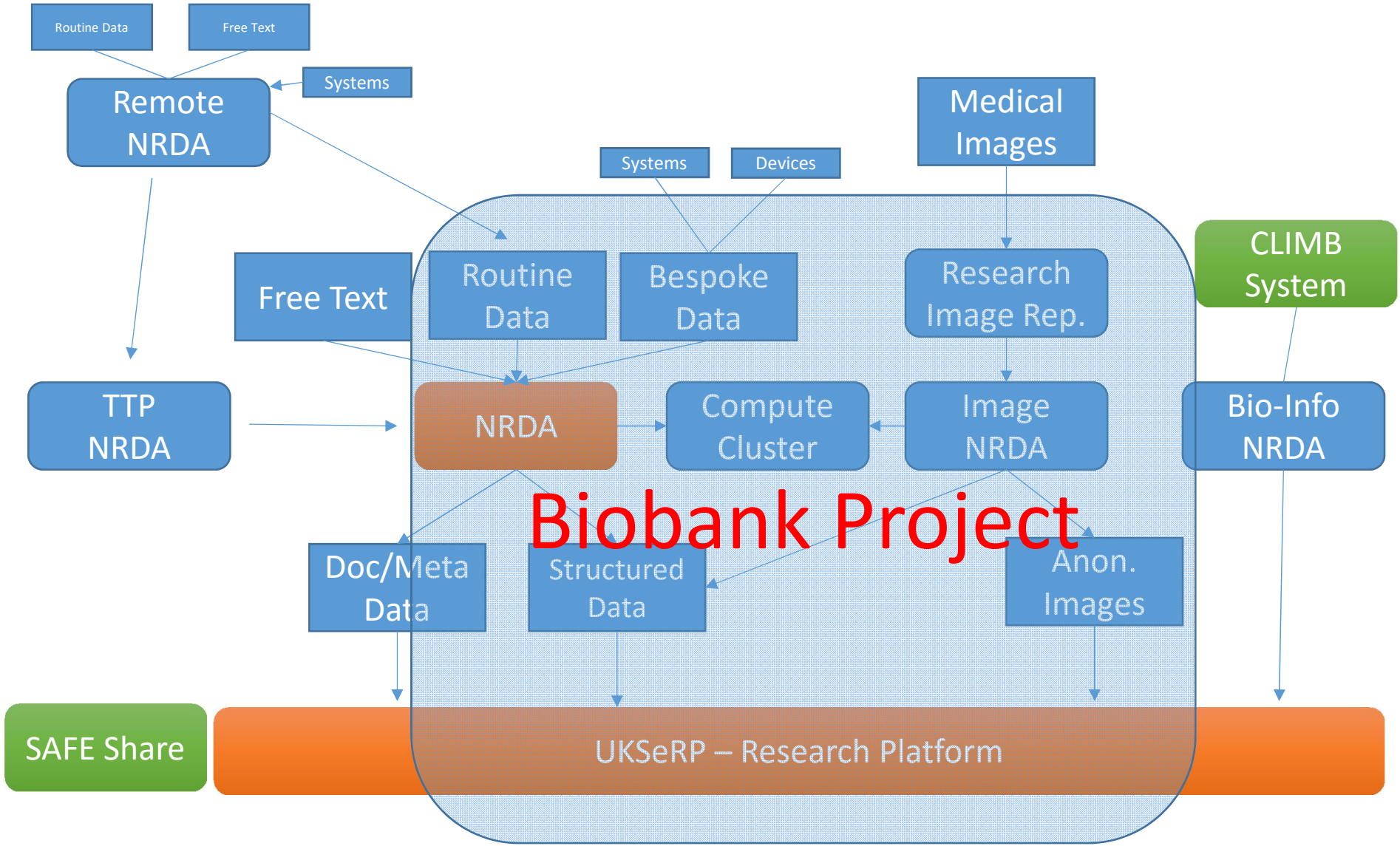
The joining up of efforts and re-use is absolutely critical



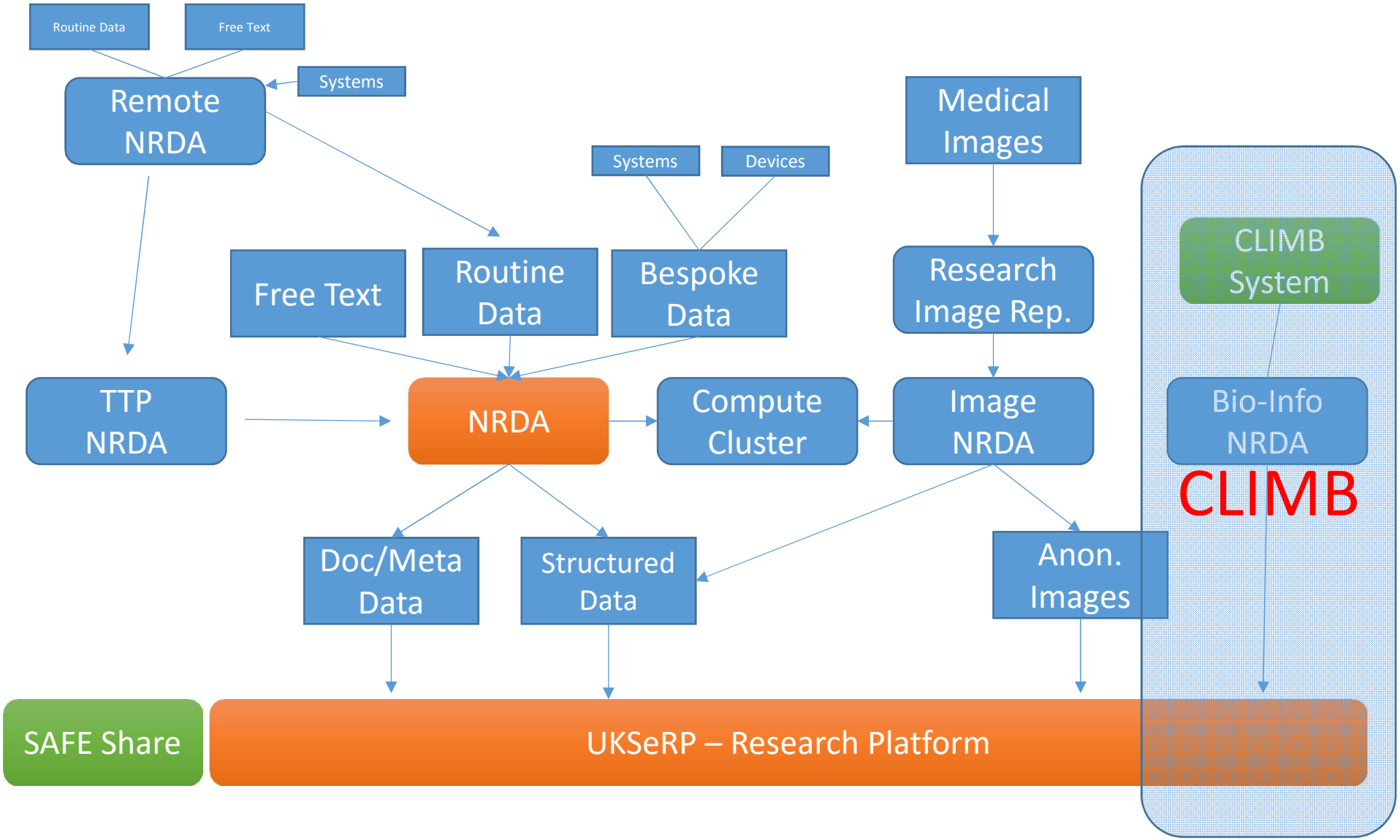
The joining up of efforts and re-use is absolutely critical



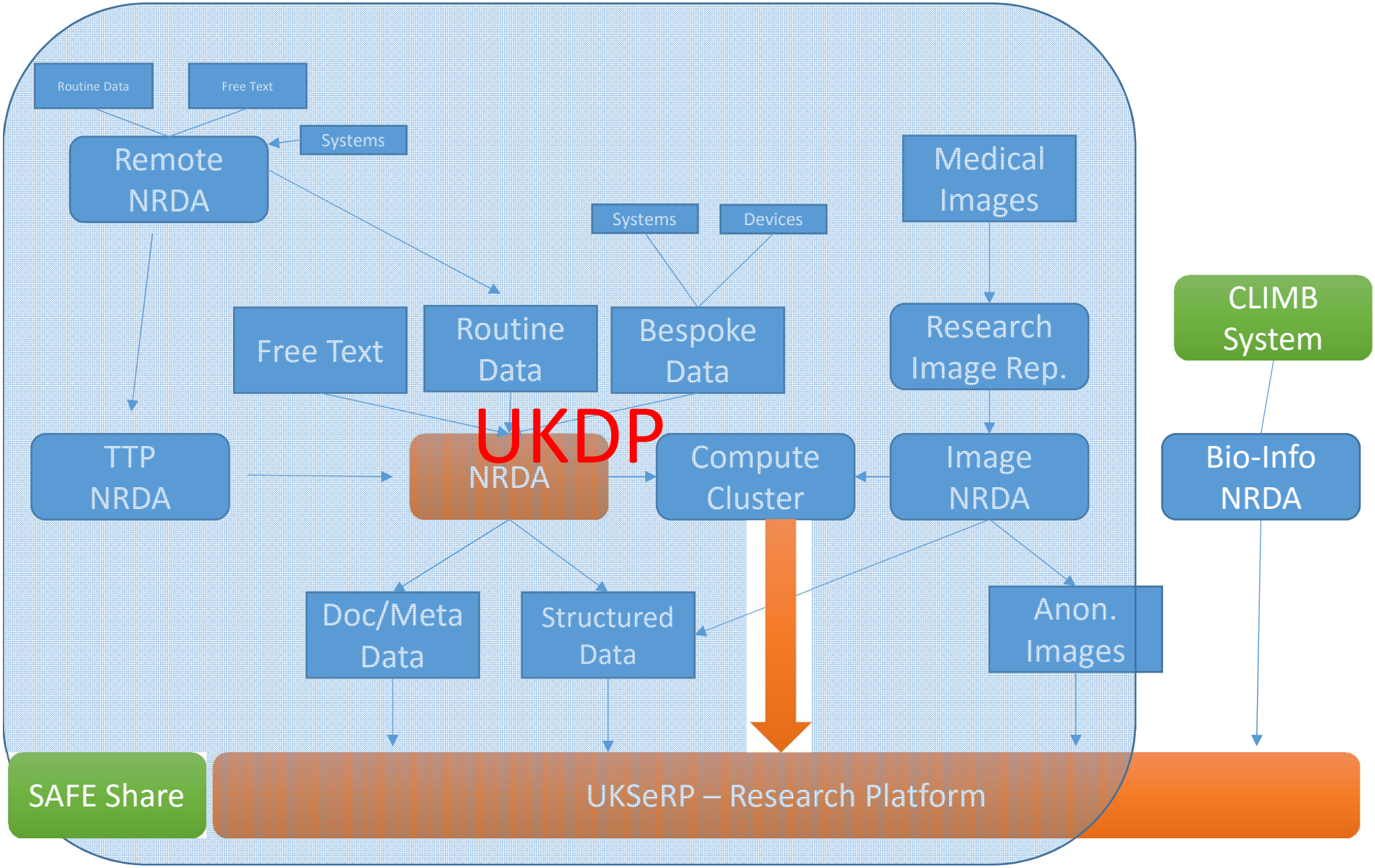
The joining up of efforts and re-use is absolutely critical



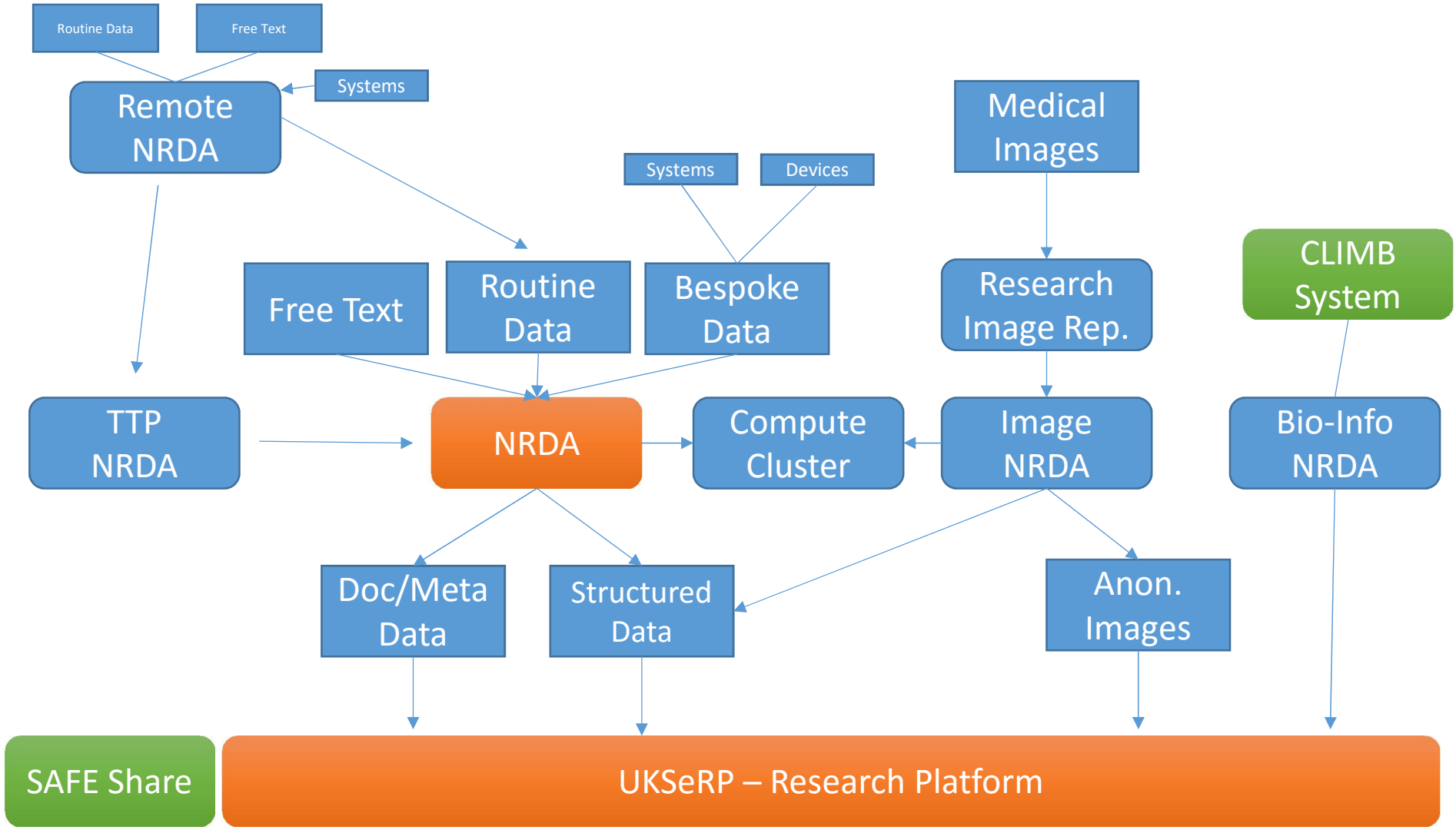
The joining up of efforts and re-use is absolutely critical



The joining up of efforts and re-use is absolutely critical



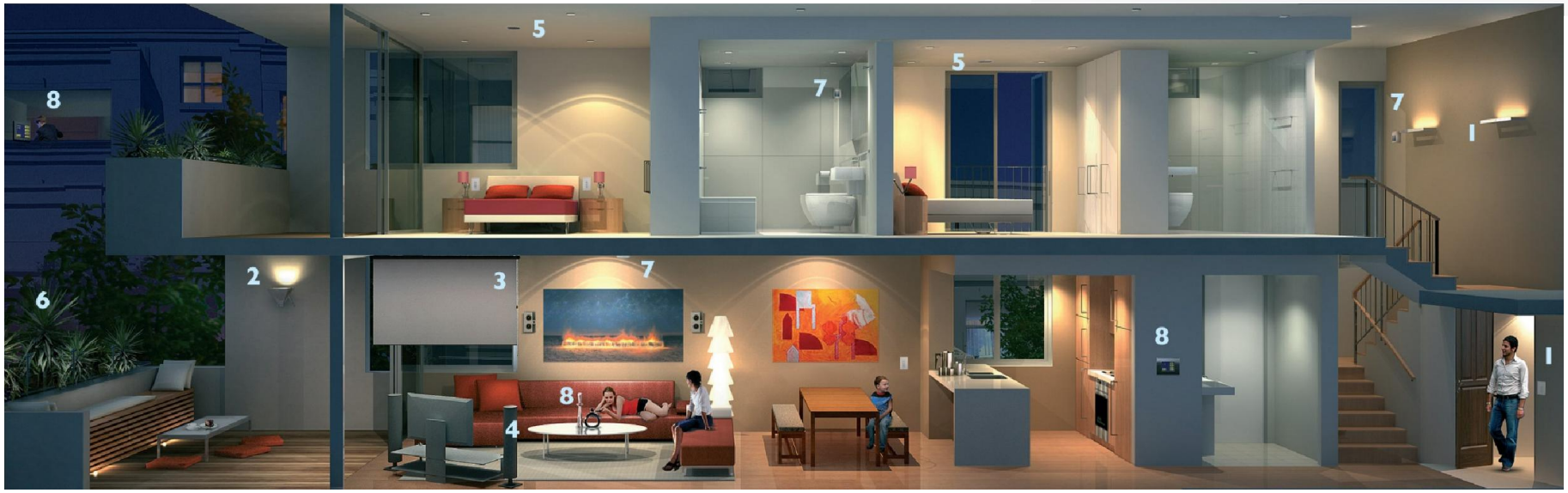
The joining up of efforts and re-use is absolutely critical



Building on opportunities

1. Need to act like an Entrepreneur, see the generic opportunity and grow capability. Allow enthusiasm to dominate ☺
2. Hard to achieve give university / NHS environment
3. Can get complicated – must deliver 100% what funders want, it's a question of how ☺

Example:-

- DPUK UKSeRP stack, decided to put XNAT in as was interesting and might be useful (just virtual machines anyway)
- Enthusiasm to understand this world and demonstrable competence to support large architectures, awarded the DPUK central imaging hub contract. Now have larger infrastructure.
- ILS2 / NHS Wales need research imaging capability – resell this capability to these efforts, use resources to expand infrastructure. Cost to University and NHS considerably lowered. Opened up opportunities to get involved in this type of research. MS project how has imaging capability which it could not afford previously

The tour of my world ends here !!