Swansea University
Prifysgol Abertawe

Data platforms to support research, evaluation & practice

David V Ford
Professor of Health Informatics
School of Medicine, Swansea University

The Farr Institute @ Swansea

MRC Medical Research Council

SAIL

Administrative Data Research Centre Wales

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Outline

1. Swift overview of SAIL Databank as used in Wales
2. Everything we know in a box

SAIL

Research Data Appliances

The SAIL Databank

A national e-research platform for Wales

David Ford
Professor of Health Informatics
College of Medicine
Swansea University

SAIL: 5 Minute Overview

- SAIL= Secure Anonymised Information Linkage
- Databank of >9 billion recordings on the people of Wales
- Data on >5 million people
- 500+ feeder systems from Wales, inc >350 GP (75%)
- Much data goes back 10-20 years
- All pre-linked
- £5m+ investment in high performance I
- Strong privacy protection & IG
- Remote access from anywhere, given approvals
- Industrial strength, reusable infrastructure.

SAIL: Data resources

Core holdings:

- Annual District Birth Extract (ADBE)
- Annual District Death Extract (ADDE)
- Bowel Screening Wales (BSW)
- Breast Test Wales (BTW)
- Cervical Screening Wales (CSW)
- Congenital Anomaly Register and Information Service (CARIS)
- Emergency department Data Set (EDDS)
- National Community Child Health Database (NCCHD)
- Outpatient Dataset (OPD)
- Patient Episode Database for Wales (PEDW)
- Primary Care GP dataset (360 practices, 2m people, 75% of practices)
- Welsh Cancer Intelligence and Surveillance Unit (WCISU)
- Welsh Demographic Service (WDS)

Project-specific holdings:

- Clinical trials participants
- Conventional cohort participants
- Cross sectional survey participants
- Many others

Reference data:

- Data quality reports
- Extract histories
- Coding and mapping information
- Metadata
- Organisation codes

SAIL features

1. Built to consume (new data) swiftly 
2. Secure sharing for projects, based on views and a UKSERP instance (more later).
3. Total population coverage
4. Used for observational research, trial feasibility, outcome data for trials, extending cohorts, natural experiments . .
5. Lots of trial and cohort study participants embedded in SAIL

SAIL: 5 Minute Overview

1. Currently supporting externally funded projects with value >£90m
2. >100 Peer-reviewed publications to date
3. 140+ approved and active SAIL projects.
4. Over 300 registered users, with 150 active today from across the world
5. 100 staff in Swansea working on Health Informatics-related projects
6. Average 35 day turnaround from application to data
7. Applications open to all

SAIL: 5 Minute Overview

1. Now new UK Centre of Excellence in HI
   – CIPHER – Centre for Improvement in Population Health through E-records Research (£4.4m). Partners: Cardiff Bristol, Sussex, Australia, Canada

2. CIPHER now one of 4 UK-wide Farr Institute Centres (+£5m)

3. New award: CADRE – Centre for Administrative Research & Evaluation (£8.0m) ESRC, as part of the ESRC ADRN and its “Big data” initiative. Partners: Cardiff

SAIL: 5 Minute Overview

New Data Science building @ Swansea

- Security and privacy protection from the ground up
- Funded by MRC (Farr); ESRC (ADRN) and Welsh Government
- High security home to Farr@Swansea and CADRE
- Office space for NHS and other public sector staff and industry to work alongside university staff

MRC Medical Research Council

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Llywodraeth Cymru Welsh Government

Research Data Appliances & UKSERP

Everything we know, in a box

5 Minute Overview

Now in beta test:

1. UK Secure E-Research Platform (UKSeRP) - based on the SAIL Gateway – massively extended to provide a secure platform for data sharing across the UK – not just SAIL data (for Farr and ADRN). 270001 certified and IG Toolkit compliant
2. National Research Data Appliances: New concentrator technology to gift to NHS and other organisations, including automated matching, anonymisation, data management, metadata capture, data quality assessment, etc.
3. New focus on the capture and analysis of electronic free text data (on-board NLP in the Appliances)
4. Funded by MRC Farr institute and ESRC ADRN grants

Research Data Appliance

- Brings SAIL's capabilities onto combined hardware and software
- Shrink wrapped, ready to go.
- Any size from data stick to massive!
- Easy to use, low expertise barrier
- Multiple Appliances work together as a larger whole
- Purposes: concentrate data, make it research ready
- Provide utility to data owners and partners
- Provided free (limited numbers) to our collaborators

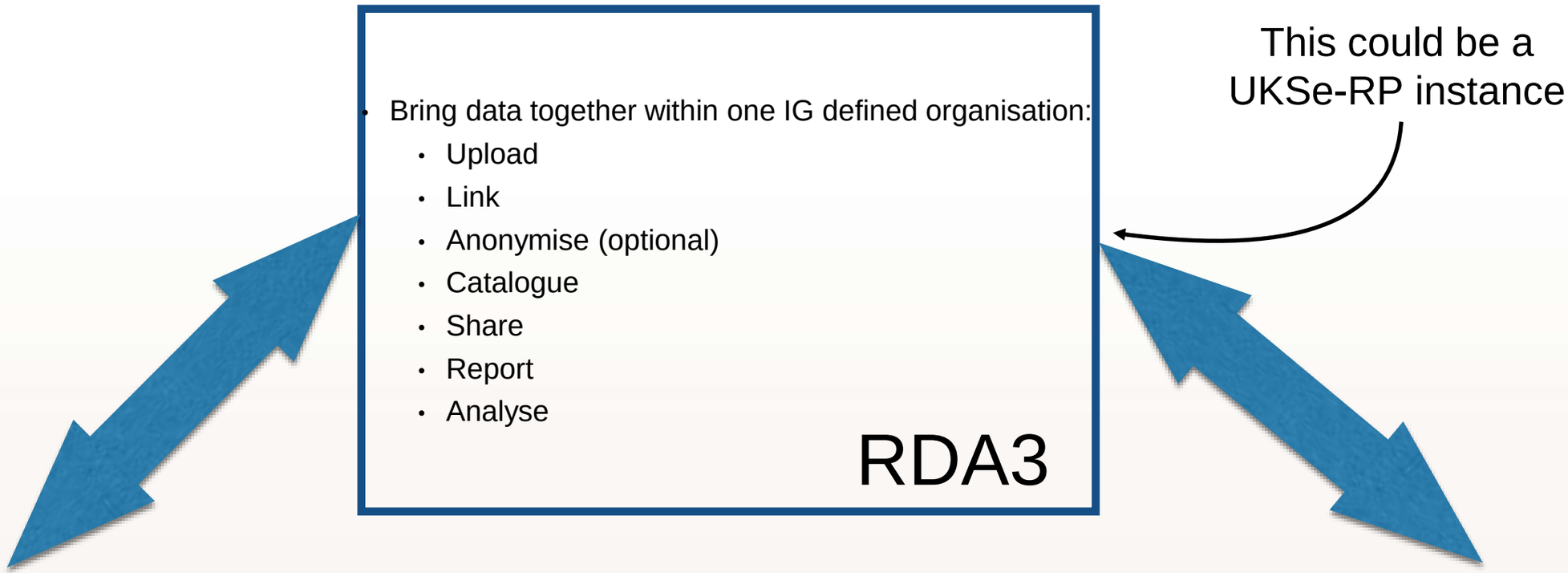
RDA Usecase

Bring data together within one IG defined organisation:

- Upload
- Link
- Anonymise (optional)
- Measure
- Catalogue
- Share
- Pool
- Manage & Report
- Analyse

All under strong IG control

A Simple RDA deployment model



Bring data together within one IG defined organisation:

- Upload
- Link
- Anonymise (optional)
- Catalogue
- Share
- Report
- Analyse

RDA1

Bring data together within one IG defined organisation:

- Upload
- Link
- Anonymise (optional)
- Catalogue
- Share
- Report
- Analyse

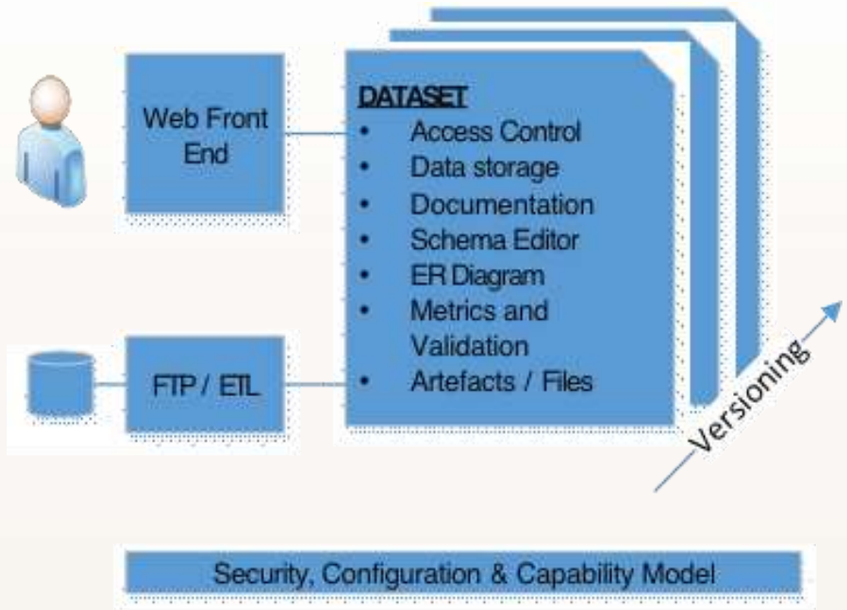
RDA2

Research Data Appliances

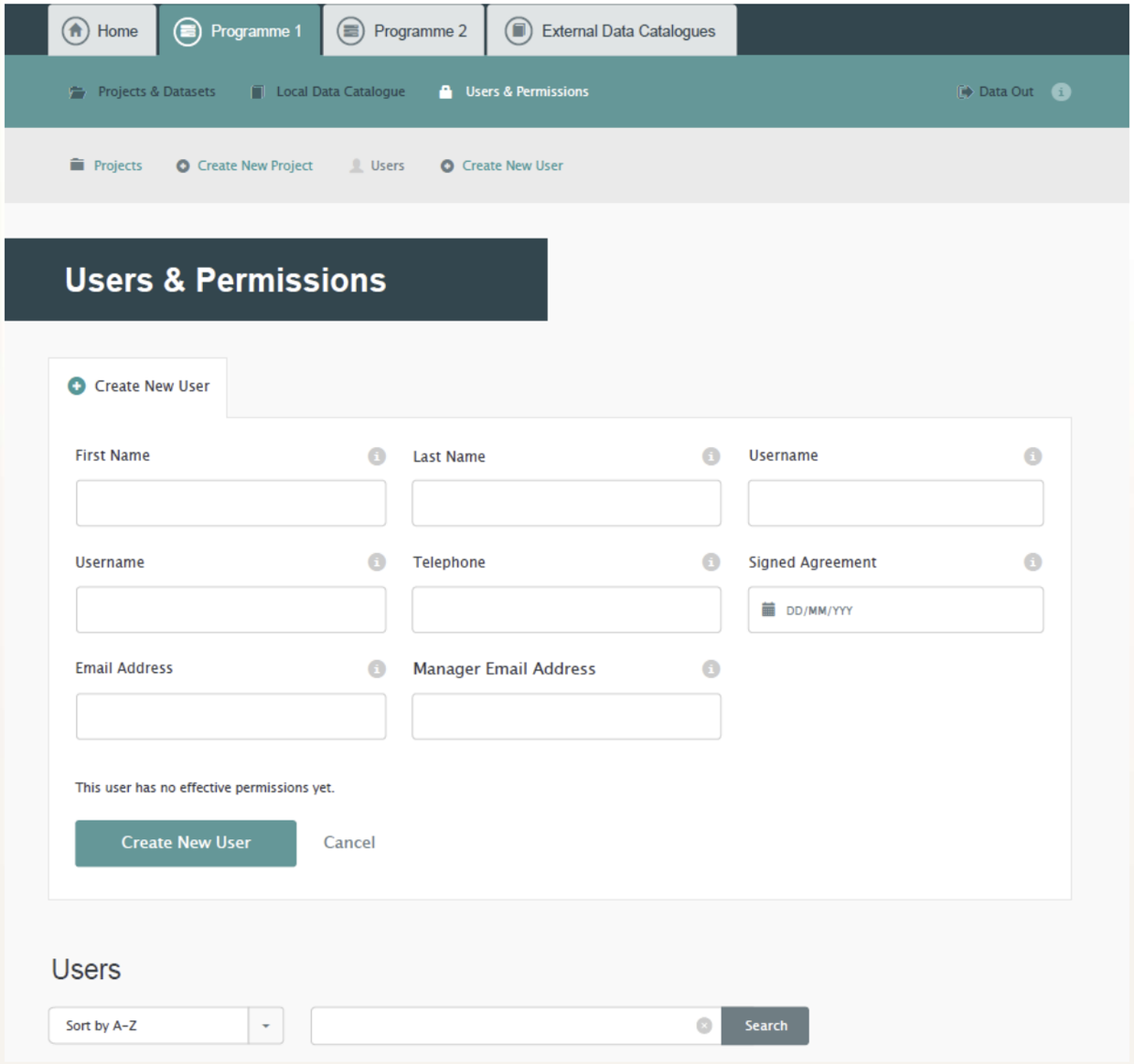
Concepts
(Presented at SHIP 2013)

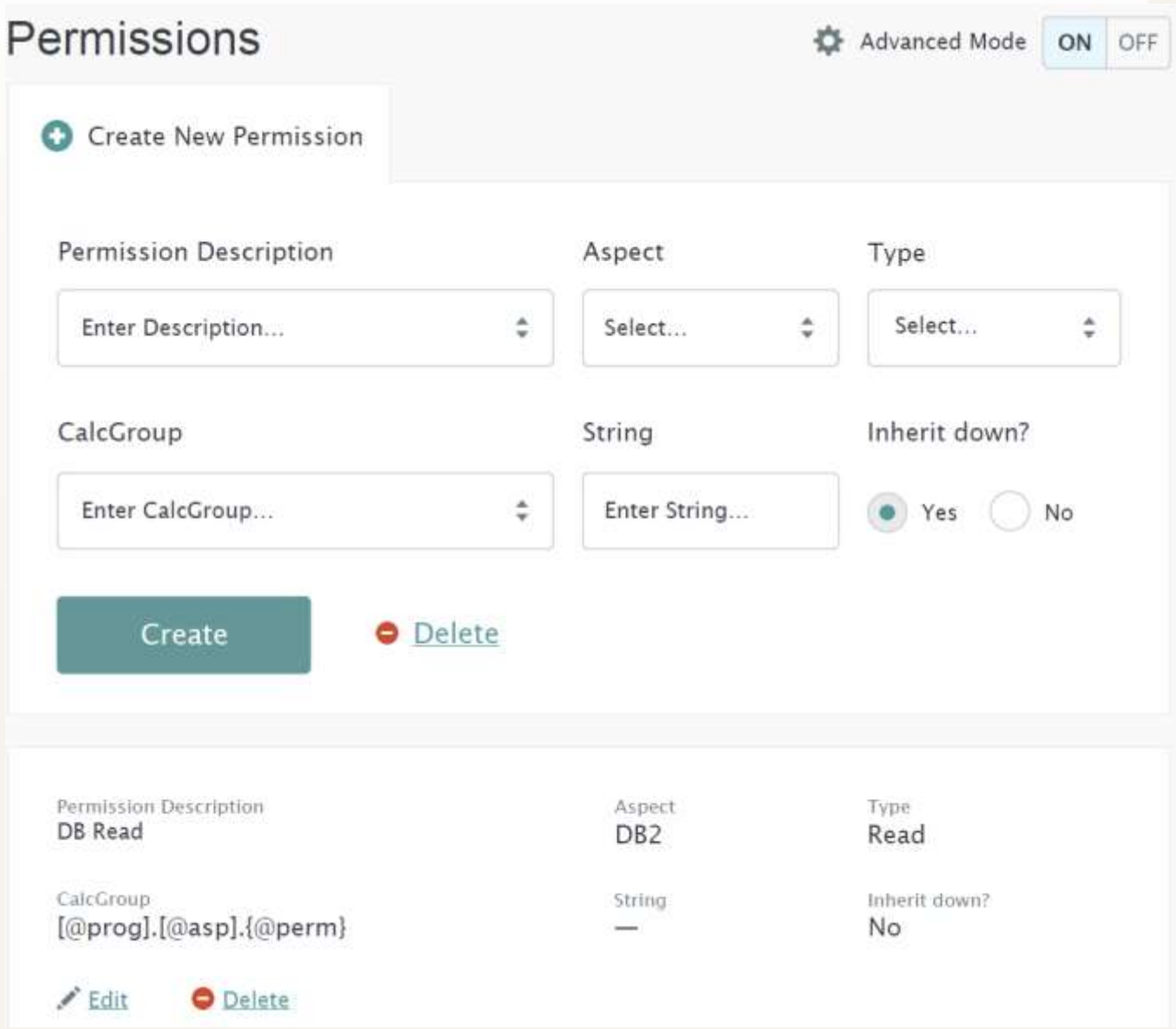
Benefits to health partners:

- Data concentrator with file store and database capabilities
- On board ETL
- Very easy to use (low technical barriers)
- Built on best privacy practice principles
- On-board high quality probabilistic identity matching system
- Integrated automated Metadata Catalogue and data quality reporting
- Full IG control locally - local data for local reuse.
- Robust standardised anonymisation (as required)
- NLP facilities option for unstructured (free text) data stores



Self management





Congenital Anomaly Register and Information Service (CARIS)

Unpublished - Version 5 (Draft) | View Log

Dataset Description | Data Files ⚠ | Supporting Files | Entity Relationship Diagram | Share Settings | Publish to Catalogue

Dataset Description

Overview

Give your Dataset a unique name, so others can understand what the data is about.

Dataset Name (Mandatory) 8
Congenital Anomaly Register and Information Service (CARIS)

Theme 11
Deprivation

Data Providing Organisation (Mandatory) 9
Congenital Anomaly Register and Information Service (CARIS)

Data Type (Mandatory) 12
Research Data (survey, questionnaire etc.)

Description (Mandatory) 10
Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo.

Dataset Level (Mandatory) 13
Other, please specify...
Some other text here

Artifact Name | Artifact Name | Artifact Name

Link a file

Purpose 14
Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo.

Link a file

Tags 15
tag, another tag, a really long tag,

Congenital Anomaly Register and Information Service (CARIS)

Unpublished - Version 5 (Draft) | View Log

Dataset Description | Data Files ⚠ | Supporting Files | Entity Relationship Diagram | Share Settings | Publish to Catalogue

Data Files (3 files)

Upload new data file | Remote upload data file

CSV

Friendly Name of Data Table (NEW)

Original-file-name-01.csv

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo.

Data Details | Data Import Settings

Options

Include in Data Quality Report

Status: ⚠
Errors - Not ready to publish

Data Details

Name (Mandatory) 2
Friendly Name of Data Table

Table Name 3
friendly-name-of-data-table

Description (Mandatory) 4
Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo.

Date Format 5
DD/MM/YYYY

Date / Time Format 6
DD/MM/YYYY 00:00:00

Distribution Column 7
PROV_UNIT_CD

Save and Close

Congenital Anomaly Register and Information Service (CARIS)

Unpublished - Version 5 (Draft) | View Log

Data Files (3 files)

Upload new data file | Remote upload data file

CSV

Friendly Name of Data Table (NEW)

Original-file-name-01.csv

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo.

Data Details | Data Import Settings

Origin: My Dataset

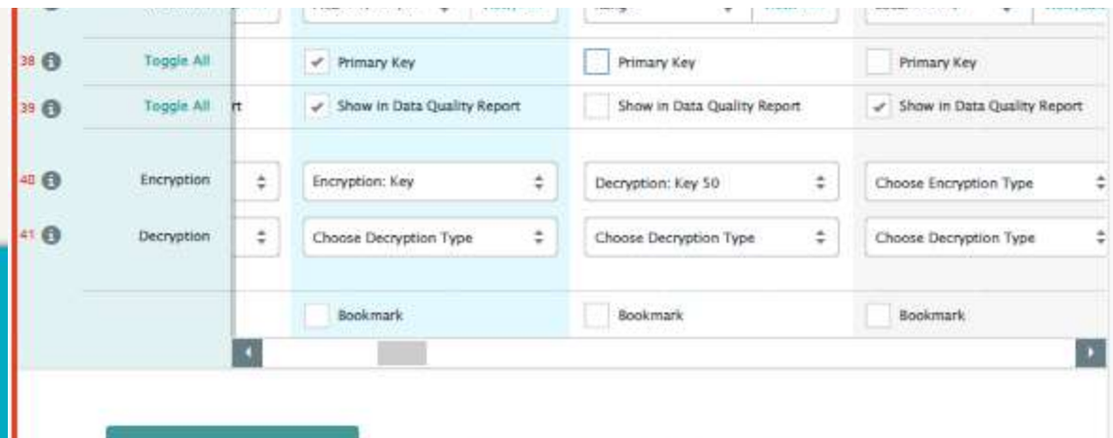
Options

Include in Data Quality Report

Status: Errors - Not ready to publish

Data Import Settings

29	30	31
Personal Identifiable Data (PID)	Import Settings	Import Settings from Local Data Catalogue
Choose template	Choose existing settings	View list of settings



SAIL DATABANK

Institute of Life Science
College of Medicine

Swansea University
Prifysgol Abertawe

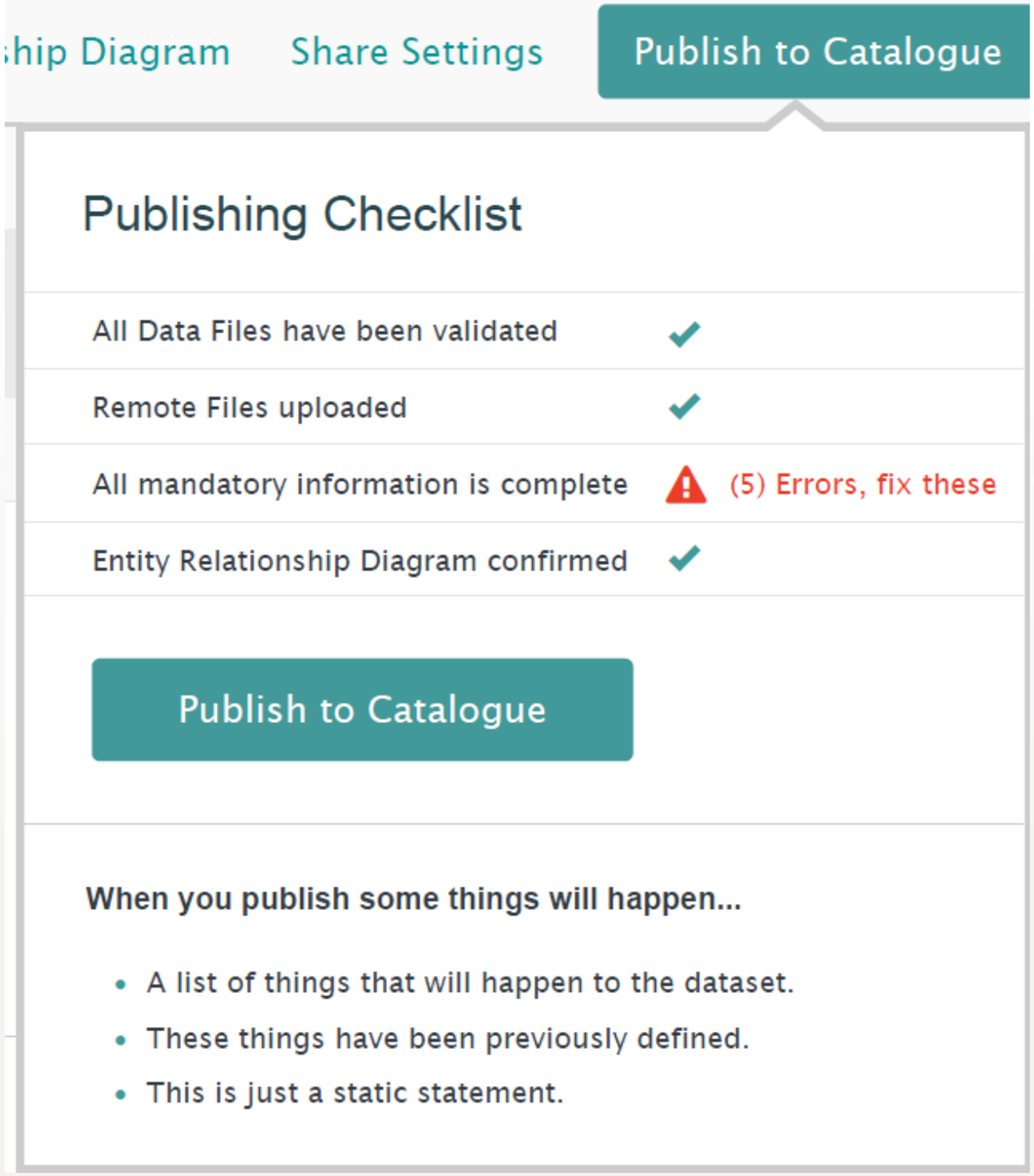
Data “schema” automatically computed based on data contained in uploaded file

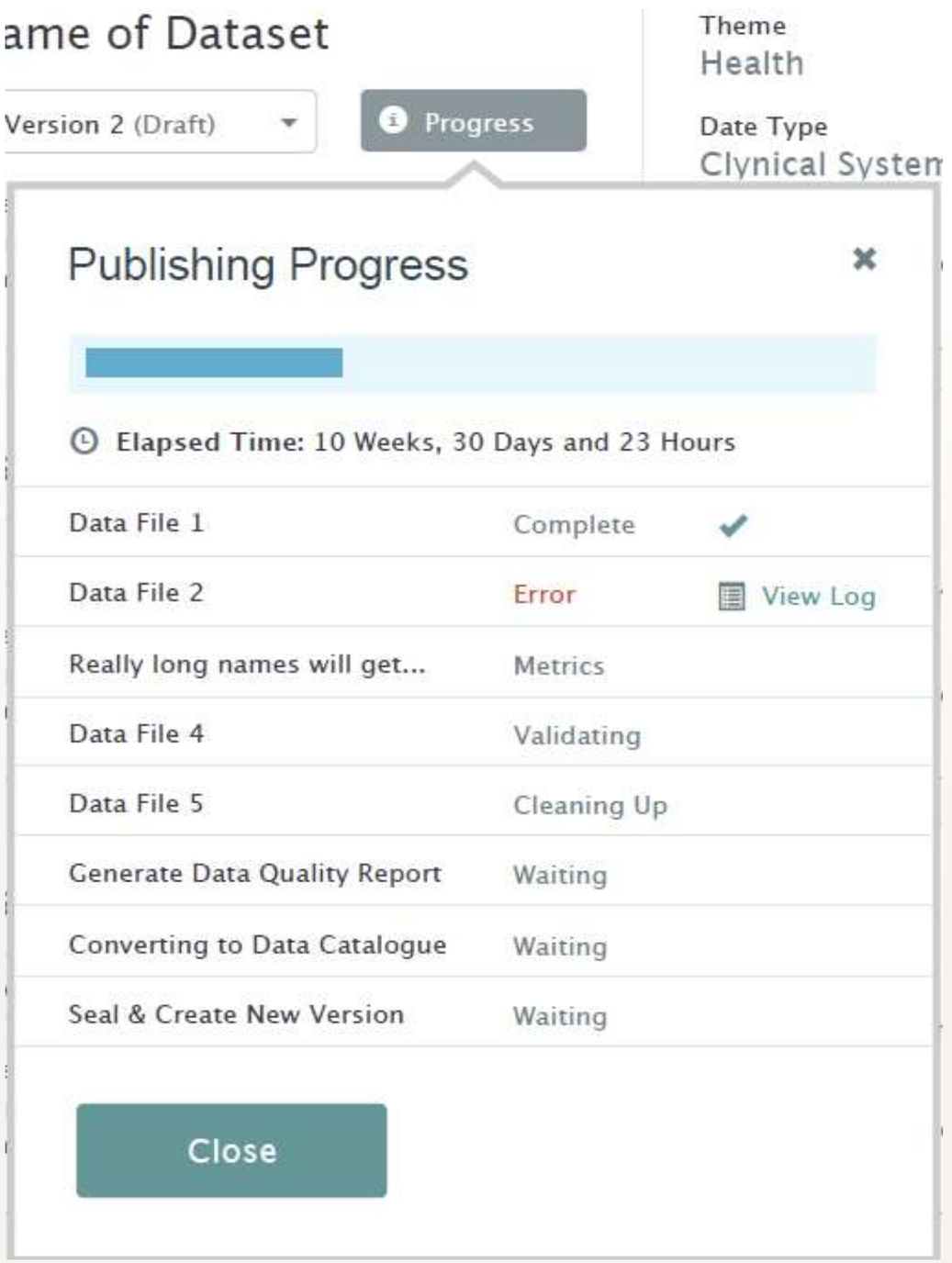
All Fields | A-Z | Errors (3) | Bookmarks (5) | PID (3)

Field Name	PROV_UNIT_CD	SPELL_NUM_E	EPI_NUM
Friendly Name	Organisation Code (Unit Code of Provider)	Hospital Provider Spell Number	Episode Number
Field Description	This is the organisation code of the health care provider. The provider code identifies the ...	This is the organisation code of the health care provider. The provider code identifies the ...	This is the organisation code of the health care provider. The provider code identifies the ...
	View More	View More	View More
Personal Identifiable Data (PID) Type	N/A	NHS Number	N/A
	Scale / Size	Precision / Scale	Size
Field Type	Char 3	Float 15 5	Char 2
Validation	Please Specify... View/Edit	Range View/Edit	Local Lookup View/Edit
Toggle All	Primary Key	Primary Key	Primary Key
Toggle All	Show in Data Quality Report	Show in Data Quality Report	Show in Data Quality Report
Encryption	Encryption: Key	Decryption: Key 50	Choose Encryption Type
Decryption	Choose Decryption Type	Choose Decryption Type	Choose Decryption Type
	Bookmark	Bookmark	Bookmark

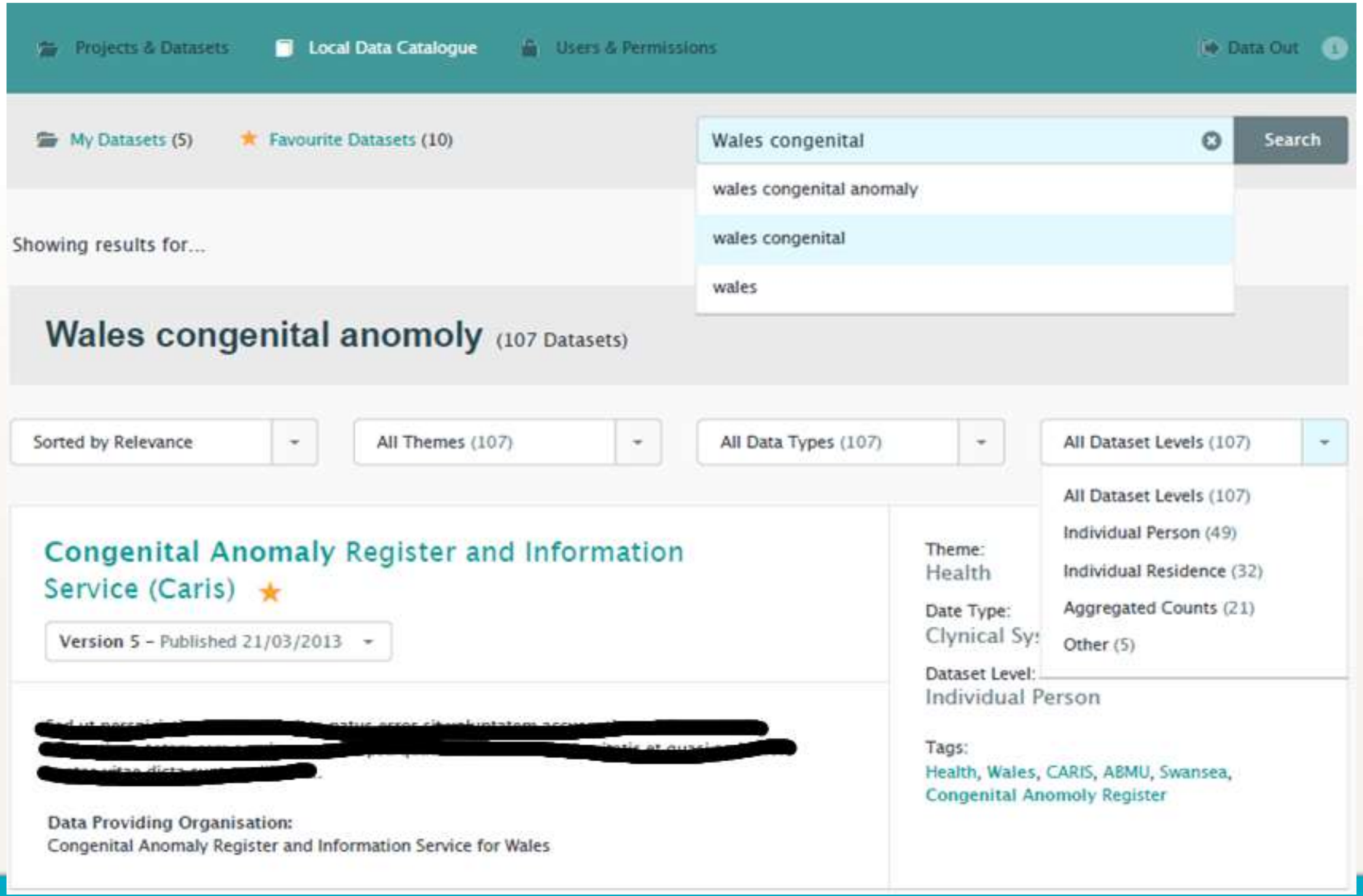
SAIL D

Publish Dataset - Depend on Configuration/Capabilities. Data will now be available



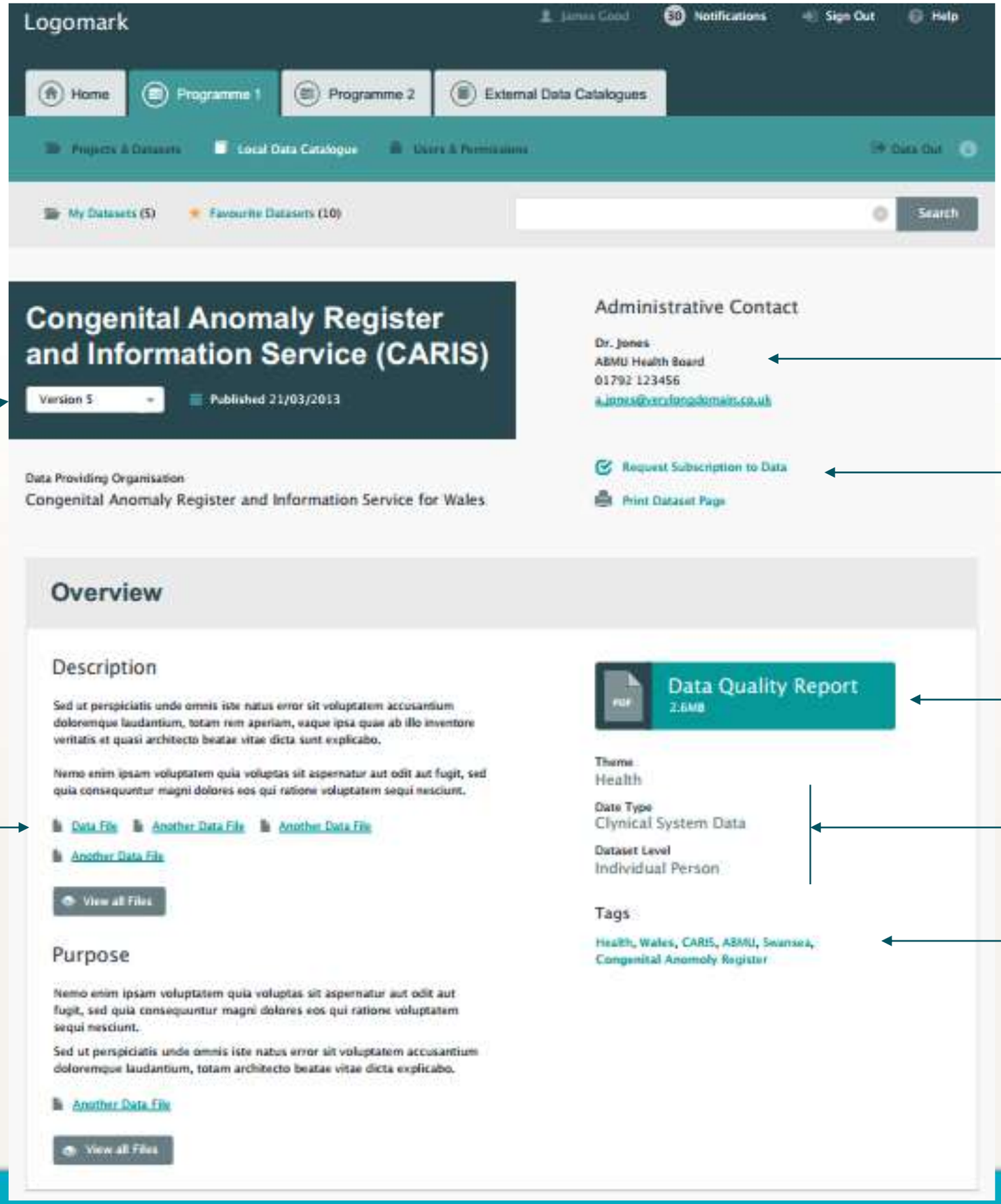


Data Catalogue – Key Component



Additional points following previous sessions:
All DA carry a DC, DS can inherit from other DS DC entries, DC related to Programme/Security domain. DC's replicate to Regional/Global DC.
Road map: DC used to define and create DS

A Dataset



Specific version & Date

Contact

Request

VIMO

All section attach files

Theme / Type / Level

Tags

A Dataset (cont.)

Coverage

Data Coverage

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo.

Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione sequi nesciunt.

Data File Another Data File Another Data File

Another Data File

View all Files

Dataset Period

March 1989 – December 2001

Data File

View all Files

Inclusion / Exclusion Criteria

Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni qui ratione voluptatem sequi nesciunt.

Data File

View all Files

Collection

Data Collection Method

This information is not available.

Data File

View all Files

Refresh Frequency

Data is refreshed every month w processing.

Data File

View all Files

Highlights & Know Issues

Data Highlights

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo.

Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui sequi nesciunt.

Data File

View all Files

Known Issues

Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt.

Data File

View all Files

Back to top

Data Files

Data Tables (2/2)

View all

Name of Data Table

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias.

Name of Data Table

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga.

DDI, SPSS, SAS, STATA

SAIL Specific

Dataset Category
Core

Access Requirements

As a core SAIL dataset available in accordance with our standard Information Governance procedure.

Supporting Files

Supporting Files (5/20)

View all

Entity Relationship Diagram
orig-file-name.png (345 kb)
At vero eos et accusamus et iusto odio dignissimos ducimus qui voluptatum...

Friendly Name Here
orig-file-name.doc (4.2 mb)
At vero eos et accusamus et iusto odio dignissimos ducimus qui voluptatum...

Friendly Name Here
orig-file-name.pdf (1.6 mb)
At vero eos et accusamus et iusto odio dignissimos ducimus qui voluptatum...

DDI File
orig-file-name.pdf (3.6 mb)
At vero eos et accusamus et iusto odio dignissimos ducimus qui voluptatum...

Another File
orig-file-name.pdf (3.6 mb)
At vero eos et accusamus et iusto odio dignissimos ducimus qui voluptatum...

SAIL DATABANK

Data Catalogue – a specific table

Friendly Name of Data Table

Download Modified 3 months ago

Records 55,160,095 | Fields 9

Description

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo.

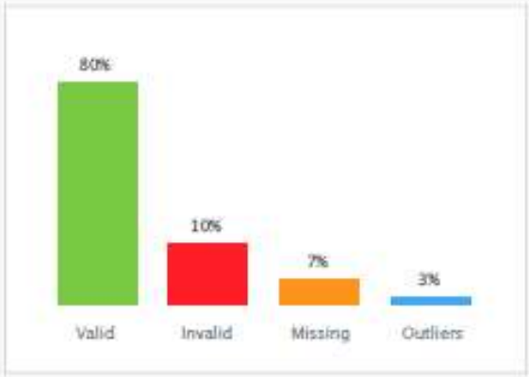
Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt.

Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt.

Connection String

server.database.table.12345

Validation (VIMO)



Heading?

All Fields | A-Z

	Code Name	PROV_UNIT_CD	SPELL_NUM_E	EPI_NUM	OPER_DT
	Friendly Name	Organisation Code (Unit Code of Provider)	Hospital Provider Spell Number	Episode Number	Operatio date
	Description	This is the organisation code of the health care provider. The provider code identifies... View More	A number (alphanumeric) to provide a unique identifier for each hospital provider... View More	A number used to identify episodes uniquely, and is a sequence number for each... View More	A numbe position to a pati View
	Field Type	char Size 3	int Size 4	char Size 2	smallint
	VIMO	100% Valid View	100% Valid View	91.43% Valid View	100
	Metrics	Top 10 Values View	Top 10 Values View View Min, Max, Mean View	Top 10 Values View View Min, Max, Mean View	Date

Records 55,160,095 | Fields 9

VIMO

Valid
Invalid
Missing
Outliers

Top 10 Values

1	93.98%	
2	4.39%	
3	1.02%	
4	0.32%	
5	0.11%	
6	0.05%	
7	0.02%	
8	0.01%	
9	0%	
10	0%	

Datetime Range

Start Datetime
1900-01-01
00:00:00

End Datetime
2020-12-28
00:00:00

View Min, Max, Mean

Min 1
Max 12

Mean
2.18225767767603

Key features of relevance

- Appliances federate with each other to create a sharing network
- Network can be hierarchical or peer-to-peer
- Full IG controls available at every point
- Metadata builds automatically and publishes to a global catalogue
- Data quality measurement automated
- NLP address rich, free text datasets, converting them to SNOMEDCT
- High quality identity reconciliation automatic, de-identification optional
- UKSERP provides scalable, performant analytics platform, with full IG controls

The Farr Institute of Health Informatics Research

Questions?