WP 2

Data Portal
Start date: 1 Aug 2014 **Completion date:** 30 Jun 2020

Overall work package objectives:
The integration of data across cohorts and single portal data access. This involves the collection, integration, curation and analysis of the substantial volumes of data obtained during the course of project. The final goal is to integrate the cohort data and UK-wide data linkage into a DPUK informatics portal.

1. UK-wide record linkage, understanding data capability
2. Integration of cohort research data and metadata
3. Informatics portal established
4. Informatics portal dissemination

Deliverables	Milestones	Milestone deadline	Work package dependencies	Person(s) responsible
Objective 1:				
D1.1 Prioritising of cohorts for data transfer and facilitation of data transfer	M1.1.1 Readiness of data for transfer established for priority cohorts: Airwave, BDR, Biobank, Cam-CAN, CamPaIGN, CAPS, MRC CFAS, CFAS II, Cygnus, ELSA, GS: SFHS, ICICLE-PD, LBC1936, OPDC Discovery, PICNICS, PREVENT, Whitehall II	M1.1.1 Complete	None	CO
	M1.1.2 Facilitation of data transfer initiated	M1.1.2 Complete		
D1.2 Accurate and up-to-date information available to researchers	M1.2.1 Comparison tool created.	M1.2.1 Complete	None	JB
D1.3 Analysis of cohort metadata for DPUK studies	M1.3.1 Primary metadata study to analyse the spread of data available within DPUK	M1.3.1 Complete	None	Oxford analyst
	M1.3.2 Comparative study with other big data consortia	M1.3.2 Complete		
	M1.3.3 Further analysis of cohort metadata to establish data capability of DPUK with specific attention on variable collection	M1.3.3 Complete		
Objective 2:				
D2.1 Utilise Gap Analysis and contact with Cohort IT Leads to identify cohorts with fewer difficulties Identify cost effective solutions with early adopters Migration of data initiated	M2.1.1 Support options and cost recovery policy in place	M2.1.1 Complete	None	CO, JB, ST
D2.2 Migration of data completed – ongoing development as cohort partners are constantly joining and some cohorts have bespoke sharing permutations	M2.2.1 First data set migrated	M2.2.1 Complete	None	CO, JB, ST
	M2.2.2 Five prioritised cohorts' data on the platform	M2.2.2 Complete		
	M2.2.3 12 Prioritised cohorts' data on the platform	M2.2.3 Complete		
	M2.2.4 16 Prioritised cohorts' data on the platform – change of prioritisation based on cohort circumstances	M2.2.4 Complete		

	M2.2.5 20 cohorts data on the platform	M2.2.5 Complete		
	M2.2.6 26 cohorts data on the platform	M2.2.6 Complete		
	M2.2.7 32 cohorts data on the platform	M2.2.7 Complete		
D2.3 Integrate cohorts fully into DPUK	M2.3.1 Meet with cohorts to explain latest on DPUK and facilitate cohort data transfer	M2.3.1 Complete	None	CO, JB, ST
	M2.3.2 Engage at four cohort events for 2016 and 2017 – forums, seminars, conferences	M2.3.2 Complete		
	M2.3.3 Dissemination of completed DPUK literature for cohort consumption	M2.3.3 Complete		
	M2.3.4 Open up consortium services to cohort groups and individual cohorts for rapid scientific engagement	M2.3.4 Complete		
	M2.3.5 Cohort data services to be initiated	M2.3.5 Complete		
D2.4 Support and facilitate DPUK scientific analysis studies	M2.4.1 5 studies supported and ongoing in Data Portal	M2.4.1 Complete	None	CO, JB, ST SB, MN
	M2.4.2 10 studies supported and ongoing in Data Portal	M2.4.2 Complete		
	M2.4.3 15 studies supported and ongoing in Data Portal	M2.4.3 Complete		
	M2.4.4 35 studies supported and ongoing in Data Portal	M2.4.4 Complete		
	M2.4.5 45 studies supported and ongoing in Data Portal	M2.4.5 Complete		
	M2.4.6 55 studies supported and ongoing in Data Portal	M2.4.6 Complete		
	M2.4.7 First study publications to be facilitated	M2.4.7 Complete		
	M2.4.8 65 studies supported and ongoing in Data Portal	M2.4.8 Complete		
	M.2.4.9 100 studies supported and ongoing in Data Portal	M2.4.9 Jun 2019		
Objective 3:				
D3.1 Establish DPUK UKSeRP instance and governance structure	M3.1.1 DPUK UKSeRP instance and governance structure in place	M3.1.1 Complete	None	JB, ST
D3.2 Establish DPUK data portal front-end	M3.2.1 Create DPUK portal website	M3.2.1 Complete		JB, ST, CO
	M3.2.2 Build expression of interest and full application pages for study applications	M3.2.2 Complete		
	M3.2.4 Create policy documents for DPUK portal and procedures for each product (data access, data management)	M3.2.3 Complete		
	M3.2.4 Create policy documents for DPUK portal and procedures for each product (data access, data management)	M3.2.4 Complete		
	M3.2.5 Launch application phases and documentation	M3.2.5 Complete		
	M3.2.6 Fully operational study application system	M3.2.6 Complete		
	M3.2.7 Active links to analysis platform and all other DPUK platforms (imaging, genetic etc)	M3.2.7 Complete		
	M3.3.1 First data set tested on UKSeRP infrastructure	M3.3.1 Complete		JB, ST, CO

D3.3 Pilot testing and provision of cohorts to the analysis platform	M3.3.2 First data set available for analysis	M3.3.2 Complete		
D3.4 Instance of EMIF fingerprinter, DPUK fingerprinter, EMIF variable/participant selection tools and TranSmart to be housed on platform - UNDER REVIEW	M3.4.1 EMIF fingerprinter installed	M3.4.1 Complete		JB, ST, CO
	M3.4.2 DPUK fingerprinter installed and populated	M3.4.2 Complete		
	M3.4.3 EMIF VST/PST tools installed	M3.4.3 Complete		
	M3.4.4 TranSmart installed	M3.4.4 Closed		
D3.5 Oxford/EMIF staff will be given access to the raw data within the portal uploaded by Swansea to harmonise to the TranSmart data model – NOT ACTIVE	M3.5.1 TranSmart installed	M3.5.1 Closed		JB, ST, CO
D3.6 Instances of static and interactive metadata tooling to be installed – Swansea build and install	M3.6.1 Project proposal raised to management and accepted based on data scoping exercise and initial tooling design feedback	M3.6.1 Complete		ST, CO
	M3.6.2 Install of primary components for tooling deployment	M3.6.2 Complete		
	M3.6.3 Deployment of static 3-tiers of information on DPUK Portal website (Cohort Landing Page, Cohort Overview, Data Dictionaries)	M3.6.3 Complete		
	M3.6.4 Swansea metadata tooling to be generated and demonstrated intra-DPUK	M3.6.4 Complete		
	M3.6.5 Swansea tooling to be made live	M3.6.5 Complete		
	M3.6.6 Continual development of tooling according to market research	M3.6.6 Complete		
	M3.6.7 Complete redesign and rework of Cohort Matrix and Cohort Directory – interactive and journey through from matrix to directory	M3.6.1 Complete		
	M3.6.8 Version 1 visualisation tooling installed	M3.6.2 Complete		
	M3.6.9 Install of GAAIN server tooling	M3.6.3 Complete		
D3.7 Link to genetics platform	M3.7.1 Genetics software installed in Swansea	M3.7.1 Complete		CO, ST (Swansea) JF (Digital Health Labs) GM (Cardiff)
	M3.7.2 Genetics platform remotely linked to Swansea infrastructure	M3.7.2 Complete		
	M3.7.3 First genetic data provisioned to platform	M3.7.3 Complete		
	M3.7.4 Genetic data analysis to begin	M3.7.4 Complete		
D3.8 Link to imaging platform	M3.8.1 Imaging hub established in Swansea	M3.8.1 Complete		CO, ST (Swansea) CM, MS (Oxford)
	M3.8.2 Imaging nodes to be established across the UK	M3.8.2 Complete		
	M3.8.3 Imaging data to fully link with central hub	M3.8.3 Ongoing		
	M3.8.4 Imaging data to be received from nodes	M3.8.4 Complete		
	M3.8.5 Imaging hub to provision access for centralised data analysis	M3.8.5 Complete		

	M3.8.6 Imaging interoperability with general access procedure finalised	M3.8.6 Complete		
	M3.8.7 DPUK Common ID model concept trialled	M3.8.7 Complete		
	M3.8.8 DPUK Common ID model concept established	M3.8.8 Complete		
	M3.8.9 DPUK Imaging Desktop resource built	M3.8.9 Complete		
D3.9 Link to wearables platform	M3.9.1 Wearables platform software/hardware to be linked to Swansea infrastructure – ON HOLD	M3.9.1 Closed		JA, JC (Manchester) CO, ST (Swansea)
	M3.9.2 Wearables data to be tested via remote link	M3.9.2 Complete		
	M3.9.3 Wearables data to be used as part of a DPUK study using data portal	M3.9.3 Complete		
D3.10 Link to EMIF-AD cohort platform – NOT ACTIVE	M3.10.1 EMIF-AD PST/VST tools to be installed on DPUK environment	M3.10.1 Complete		CO, ST (Swansea), JC (Manchester), MVS, NH (Janssen)
	M3.10.2 EMIF-AD PST/VST tools to be integrated with DPUK-based data	M3.10.2 Complete		
	M3.10.3 EMIF-AD to be hosted in Swansea and linkable to DPUK via infrastructure and cover remote access for non-hosted data	M3.10.3 Closed		
D3.11 Interoperability of platforms displaying multi-modal data analysis in the Portal	M3.11.1 First multi-modal study ongoing in the Portal (Combination of genetics/imaging and clinical data)	M3.11.1 Complete	None	CO, ST, MS
	M3.11.2 Creation of genetics/omics desktop template	M3.11.2 Complete		
	M3.11.3 Creation of imaging desktop template	M3.11.3 Complete		
	M3.11.4 Link to HDR UK phenomics initiative created	M3.11.Complete		
Objective 4:				
D4.1 Promote the work of DPUK and specifically the informatics portal at UK-wide events, such as conferences, seminars and via UK media	M4.1.1 Posters and representation at DPUK Conference 2016	M4.1.1 Complete		CO, ST, RL (Swansea), JG, AM, SB, BS (Oxford)
	M4.1.2 Attendance at UK Biobank Conference 2016	M4.1.2 Complete		
	M4.1.3 Posters, presentations and lead infrastructure staff representation at IPDLN Conference 2016	M4.1.3 Complete		
	M4.1.4 Attend and represent at Informatics for Health 2017	M4.1.4 Complete		
	M4.1.5 Lead position at DPUK Conference 2017	M4.1.5 Complete		
	M4.1.6 Attend and represent at AAIC 2017	M4.1.6 Complete		
	M4.1.7 Attend and represent at ARUK 2018	M4.1.7 Complete		
	M4.1.8 Attend and represent as key speakers and as leads at DPUK Conference 2018	M4.1.8 Complete		
	M4.1.9 Attend and represent at ADR 2018	M4.1.9 Complete		
	M4.1.10 Attend and represent at IPDLN 2018	M4.1.10 Complete		
	M4.1.11 Attend collaborative meetings with OBI, GAAIN, and CPAD – December 2018	M4.1.11 Complete		

	M4.1.12 Attend second round of collaborative GAAIN/CPAD/OBI meetings – up to end of April 2019	M4.1.12 Complete		

Updates on delivery against milestones since last report

- M.2.4.9 100 studies supported and ongoing in Data Portal

135 data access requests have been submitted:
75 approvals

- 57 approved with analysis underway
- 2 approved but awaiting cohort data upload/share
- 3 awaiting DAA return
- 8 approved by at least 1 cohort but awaiting others
- 5 studies completed with no output produced

9 studies awaiting cohort approvals
23 declined by cohorts or at initial scrutiny
15 withdrawn by applicants
13 awaiting resubmission

- M3.6.9 Install of GAAIN server tooling - COMPLETE

  The GAAIN Interrogator has been installed and trialled using 2 cohorts' data as part of the ADDI project
- M3.8.8 DPUK Common ID model concept established – COMPLETE

  The common ID model has been built and deployed to standardise IDs across MRC NSHD data in the imaging, genotype, and phenotype modalities. The project is still running but this work was undertaken between April and July.
- M3.8.9 DPUK Imaging Desktop resource built – COMPLETE

  Bespoke desktop specifications have been deployed for imaging analysis as part of the ADDI project (July-September 2019)
- M3.11.4 Link to HDR UK multi-omics initiative created – COMPLETE (amended slightly)

  Swansea has applied to lead a work package of the HDR UK multi-omics implementation project, with DPUK as a key part of the infrastructure which will inform the national strategy on computational architecture for multi-omic data science.
- M4.1.12 Attend second round of collaborative GAAIN/CPAD/OBI meetings – up to end of April 2019 – COMPLETE

  Collaborative calls were attended and now GAAIN and CPAD collaboration is continuing under the ADDI project, with OBI collaboration occurring as part of a network of neuroscience initiatives across the world working on standardising data curation and visualisation.

Team members <u>funded</u> (full or part-time) by DPUK	**Dependencies to and from other work packages, networks and themes**
Ronan Lyons – Principal Investigator, Chris Orton – Data Project Manager, Justin Biddle – IT Developer, Laura North – Data Analyst, Mark Newbury – Administrator **Team members involved with the project but <u>not</u> funded by DPUK** Simon Thompson – Chief Technical Officer, Swansea University Medical School	None

Summary of plan to deliver on outstanding work (with dates)

- All deliverables have been completed, or closed due to changing priorities as per previous reports.
- The key plan for the Portal is to mature all active operations for DPUK studies, collaborative science, and multi-modal data analysis.
- The focus is now changing to the development of the Portal for in-depth multi-omic and imaging analysis, alongside supporting large scale epidemiological studies.

- Other development includes the next phase of our Cohort Explorer visualisation tool, and developing our cohort directory for enhanced, detailed data discovery of our cohort studies.
- DPUK is continuing to grow its network of cohort studies and ongoing research studies, hosting two more datathons before the close of 2019, and continuing to be a key platform in the Alzheimer's Disease Data Initiative project.

Risks	**Mitigation**
1) Size of immediate audience for the Portal is beginning to outweigh the technical setup	1) Partnership with HDR UK and new UKSeRP system upgrades coming on board will ease pressure on the system to provide services to a hugely growing user base
2) No-cost extension period is tight for funding to make rapid/large-scale developments	2) Partnerships as above and working with externally funded projects (ADDI) will allow us to develop areas of greatest importance, rather than remain on a development plateau.

Outcomes

- DPUK Data Portal Analysis Environment
- DPUK Data Portal website and tools – https://portal.dementiasplatform.uk
- DPUK Data Portal: Data Resource Profile – https://www.biorxiv.org/content/10.1101/582155v1

Project narrative

WP2 has completed the initial objectives.

Our key focus is to now globalise the Portal user base by working with international partners and the ADDI project, as well as continuing to develop our multi-modal data analysis capability. In partnership with HDR UK and the ADDI project, the development of multi-modal data analysis will both be fostered within cohort studies and using real-world evidence data.

To date, the Portal has been successful in drawing in a diverse audience in epidemiology and is beginning to collaborate with bioinformaticians to further our omics/imaging portfolio.

Criticisms of the Portal still focus on data access times, which we are working to mitigate with our data providing community as a distinct activity. We are looking at providing more options for researchers in terms of computational capability for both standard statistical research, and multi-omic, multi-modal data science, in order to be competitive and market leading in the dementia research space. System upgrades in partnership with HDR UK are currently being finalised, which should allow more flexible and operational processes for a wider audience of researchers into 2020.