Best practices for infrastructure development and data curation: accelerating timely and appropriate access to key research datasets

Roundtable meeting convened by the Expert Advisory Group on Data Access (EAGDA)

Tuesday 11 October 2016, Room 4.16 SSRC Building, UK Data Archive, Colchester

Summary

- Stakeholders from funders, repositories, institutions and cohort studies met to discuss some of the key challenges for sustainable data management, curation and access, across different disciplines.
- Difficulties, lessons and developing best practice from across different disciplines and perspectives in the data lifecycle were shared.
- The benefits and drawbacks of discipline-specific repositories were extensively discussed. Better discoverability, interoperability and use across disciplines is underpinned by common standards, practices, vocabulary and processes: these need to be balanced with the recognition that standardising practices may over-generalise and inhibit innovation.
- It was agreed that opportunities to share experience and learn from one another was beneficial and more fora this sort of discussion would be valuable.
- Funders were challenged to consider how they could better support data curation and management in a longer-term, more joined up way.

Introduction – Matthew Woollard (UK Data Archive)

Matthew opened the meeting, outlining the broad work of the UKDA and highlighting several key issues from the repository perspective on how to make data infrastructures and data initiatives more integrated.

- Matthew outlined to core work of the UKDA and its datasets, including survey data; aggregate data (macro); historical data; and provision of the UK Data Service .
- The UKDA is currently looking at differences between Wellcome and ESRC funded researchers who create data in their attitudes towards sharing data.
- The key question for the day is: *how can we accelerate access to key data?* One of the challenges with seeking to accelerate access to data is the need to determine how the responsibility for curation, formatting etc should be distributed between data creators and repositories or archives.
- There has been significant progress towards determining what 'appropriate access' to data for secondary analysis should mean in practice. The UKDA uses the 'five safes' framework, which is solely based on disclosure risk of data.
    - The key criterion for this access mechanism is the balance between enabling access and ensuring confidentiality.
- Repositories face challenging times with funding, in part because five year funding cycles make long term planning extremely difficult. Business continuity plans need to be extensive to deal with anticipating changing funding scenarios and this is wasteful of time, expertise and resources.
    - It would be helpful for there to be funding mandates for ongoing data preservation and curation.
- The meeting has been convened to consider issues from broad stakeholder views, not just those of repositories and funders, with a view to understanding who is best placed to

make decisions about different aspects of the data curation and management process. There is a role of for data owners, creators, repositories and funders, and all need awareness of the legal and ethical challenges to data use:

- Data creators need to be more aware of what repositories do and how they work and to understand the data lifecycle. They may also need to be involved in decisions about access.
- Funders need to be able to guide the process of data curation and use, but are not always the best people to deal with issues or anticipate them.
- There also needs to be dialogue between those with legal responsibilities and those who are trying to promote maximise reuse of data.

Background to EAGDA – Tim Hubbard (KCL) & David Carr (Wellcome Trust)

- Tim and David introduced the Expert Advisory Group on Data Access (EAGDA), which is a group convened by Cancer Research UK, Economic and Social Research Council, Medical Research Council and the Wellcome Trust since 2012. It provides strategic advice to its funders on emerging ethical, legal and scientific issues for data access, for human genetic and cohort studies.
- David outlined the report on '*Establishing Incentives*'. Based on 35 interviews with researchers and data managers about the barriers and challenges for better sharing of research data, several key findings emerged, including:
  - Costs of data sharing are not anticipated at the start of the project and not reviewed by funders
  - No rewards for data sharing
  - Infrastructure doesn't exist or is not user friendly – barriers need to be lowered for researchers to share their data
- One of the key recommendations from the report was to *"Ensure key data repositories serving the data community have adequate funding to meet the long-term costs of data preservation, and develop user-friendly services that reduce the burden on researchers as far as possible."* It was this recommendation that led to the convening of this meeting.

Session 1: Current challenges for repositories

European Genome Phenome Archive (EGA) – Paul Flicek

Paul provided a brief overview of the European Bioinformatics Institute (EBI) and the EGA within it. The EBI has three main roles: providing archive resources, often in collaboration internationally, of which the EGA is one; value added resources such as Emsembl (genomics) and Uniprot (proteins); and conducting research.

- The peer archive for the EGA is dbGaP in the US, which is the mandated archive for NIH funded genomics researchers. EGA shares metadata with them and there are plans to share metadata with a Japanese archive in future so that people can find data even if they can't access it through EGA. Usual data types are genome sequence and basic phenotype traits.
- Access to EGA data is controlled by individual Data Access Committees, not the EGA itself. EGA deals only with the mechanics and infrastructure for enabling data access, not the legal and ethical complexities of data sharing.
- Since 2012, the EGA has worked with a regulatory centre in Barcelona to share the load of managing access, and this is coordinated by Elixir to share the infrastructure and

assist collaboration across Europe. When the EGA was initially created there was no clear way of sharing data.

- Data sharing has increased massively over the past few years. There are nearly 2000 studies in the EGA now, with more than 2.5PB of data, and a 30% growth in data requestors in 2015. The EGA processes around 200 helpdesk requests per month. Many requests are from commercial organisations, but they are all processed by study DACs.
- People mostly share their data with the EAG because of journal requirements. For deposition, a DAC needs to be created for the dataset to be accepted by the EGA.

Challenges for data curation and management from the EGA perspective:

- Doubling times (for processing and hard drive cost) need to be around 18 months to be sustainable in terms of costs. The EGA's doubling time is currently around 12 months, which is just manageable.
- There is not a good controlled vocabulary across datasets, which makes finding data much harder for users.
- There's a heterogeneous landscape for access mechanisms, because of the sheer number of DACs (over 300). The EGA does not review consent agreements or legal terms for access, so there may be significant variation across studies.
- Because there is no coordination, DACs need to go through a manual authorisation process to check data requesters' credentials, and to check requests against the terms of participants' consent.

Steps the EBI is taking to improve the data discovery and access process:

- The Beacon Network, from the Global Alliance for Genomics and Health, is seeking to make discovery easier, by encouraging repositories and archives to implement simple code in their databases that can respond to a query "does dataset contain x variant?", with a simple yes, no or can't tell because of controlled access.
- EGA is also considering how to develop consent form codes that can automatically define what the data can be used for. This is a difficult balance; if they are too granular they will cease to be useful but they need to be fine-grained enough for data requestors to identify if the dataset they want can be used for their purposes.
- EBI are also working towards a federated EGA in different countries, which would aim to resolve jurisdictional issues for data sharing across borders.

UK Data Service – Louise Corti

Louise provided an overview of the UK Data Service (UKDS) and its principles of enabling data to be open where possible but closed where necessary. It seeks to balance being a 'data police' with being supportive as well: fulfilling the needs of open science and transparency but also ensuring protections and ethical and legal obligations are met.

- The UKDS provides three types of data access: open, safeguarded and controlled. 90% of datasets are 'safeguarded', which typically means users have to sign and End User Licence and are then able to download of access data online.
- The UKDA provides outreach and training support for data users, including webinars and videos for the research community on good data management and planning.
- ESRC data policy has been optimised since the 90s, and includes a penalty for non-compliance with data sharing requirements of withholding final grant payments.

- There is also an ESRC requirement for data management plans to be created at the start of a project, although there remain questions about their role, and whether they should be live, published documents to help set out researchers' intentions for their data.
- Louise noted that that data management plans are important but there are too many formats across funders and they are often not properly peer reviewed. There are several intervention points across the data life cycle at which at which data management should be thought about and planned for.
- The UKDS does not host a lot of biosocial data but services could be extended to host cohort studies as well. Researchers working with these types of data could go through SURE (Safe Use of Research Environments) training (approved by ONS, HMRC, NHS).
- UKDS also has a dedicated collection of scientific data journal datasets and is the host repository for *Nature* Data papers. This highlights the need for researchers to better use existing repositories for their data: *Nature* has criteria for trusted repositories across the disciplinary spectrum that researchers can use to identify appropriate hosts.
- UKDS is seeking to extend dataset use through creating different versions of same dataset, with different access conditions from open through to controlled.
- Data depositors have to allow data to be accessed via a standard data sharing agreement or license rather than creating a bespoke set of access conditions. They must also use standard metadata and vocabulary to use UKDS as a repository.
- There are still legacy licenses with older datasets that have specific conditions attached to them, and more could be done to harmonise especially across different cohort studies in terms of governance, access conditions etc.

Challenges for improving accessibility and use of data:

- Funders invest in data collection but not necessarily long term management and data access. We need to anticipate the big data future and the benefits that different disciplinary perspectives could bring to existing datasets.
- There continue to be many siloes, which reduces accessibility. The community could work towards a common data policy underpinned by the principles of: there is a duty to share; researchers should be resourced to do so; and there are penalties for not trying.

Discussion

Audience question: The EGA has open data, and data that is heavily DAC controlled: there are only two choices. Is there a way to have a light touch registration/regulation for some EGA data, for example like the UKDS 'safeguarded' category of data with standard licenses? And could this harmonisation extend to common metadata, vocabulary and access conditions? This would be more efficient than always needing a DAC.

Paul Flicek: Many data owners don't trust a registered user access model and want to have oversight and control over who gets access to what data. The EGA accepts the political reality of accepting data from 60+ countries, with different legal frameworks in place. It has been a deliberate decision to allow this variance, even though it is inefficient.

The EGA is proposing a registered data access tier, which would need trusted methodologies to authenticate users, understands their use patterns, and attempt to persuade DACs to accept that this is good enough.

Audience follow up: Perhaps there could be moves to ensure data requestors are trained, for example through SURE training, and that they will work according to the five safes principles. There's a need to show how trust in such a system can build up over time (e.g. through comparisons with ONS 'approved researcher' status. This sort of system, with automated authentication, could cut down the work of DACS to check credentials and streamline a lot of the work to reduce the administrative burden.

Session 2: Perspectives from the user community

ALSPAC and University of Bristol – Lynn Molloy, Olly Butters & Stephen Gray

Lynn and Olly provided an overview of ALSPAC, the Avon Longitudinal Study of Parents and Children, which is a cohort study begun in the 1990s involving over 14,000 families in the Bristol area.

- ALSPAC gathers phenotypic, genetic, environmental and linkage data and has its own DAC which meets weekly. It deals with 15-20 proposals per month
- It operates a cost recovery model whereby data requestors pay for access, with the average cost currently £3200 to access a dataset from ALSPAC. Since introducing this model in April 2014 there has not been a decrease in access requests.
- ALSPAC provides an extensive support service for data requestors, through a 'data buddy' system. The cohort datasets are complex and applicants often need support to help with the processes and technicalities.
- The ALSPAC data are held in different places: UKDA, in consortia: STELAR, CPRD. Non-standard data is managed locally. There is a data dictionary but this is a 200MB pdf that is not machine readable or easily searchable at present. Data storage is file based and most processing is done in house or by the PI/collaborators.
- The ALSPAC team are keen to integrate the data pipeline and make it more automated. They are also considering a new model for data release, aiming to make as much data as possible accessible for free via an online portal.

There are several challenges for ALSPAC:

- Cost recovery is not ideal and costs are rising every 12-18 months. This could create a barrier to accessing data and is not sustainable.
- The data management infrastructure has been built up over time as a series of 'bolt ons' to the main scientific project, which means it is not ideally structured.
- Funder requirements, journal requirements and expectations change constantly, making long term planning difficult. It also makes it hard to acquire funding for retrospective data curation and re-curation: many datasets could be made more valuable through re-duration as data collection trends change and new possibilities for linkage emerge, but there are no resources to support this.
- Standard tabular data could be shared via a public repository (e.g. UKDA) to minimise the DAC's workload, but resources are needed to prepare the data for deposition.

Stephen provided an overview of *data.bris*, the University of Bristol's institutional online repository, which has its own access conditions and levels.

- *Data.bris* is a repository for researchers who have no discipline-specific place to deposit data. Some data are confidential, and some are potentially commercially sensitive.

- The repository takes static datasets from any discipline, and issues a DOI for them. It plans to maintain deposits for at least 10 years.
- There are three access levels: open (licence, CC); restricted (pre-approved for sharing, minimal checking); controlled data (decision by UoB DAC).

European Prospective Investigation of Cancer (EPIC) and the University of Cambridge – Robert Luben & Marta Teperek

Robert introduced the EPIC study, which is looking at the effects of diet on cancer, and highlighted some key challenges for data management. It has two UK cohorts (Norfolk and Oxford) and around 50,000 participants Europe-wide between 1993-7. There are 10,000 items of data for each person.

- Data access requests are discussed by the study Steering Committee. The study has an online data dictionary for people to find variables.
- The ability to collaborate in the UK may be restricted if NHS Digital change their approach of disseminating anonymous datasets to researchers, which would be highly problematic for cross-European working.
- Data sharing can be inhibited by the need to set up collaborative research agreements between institutions, which can create lengthy delays. The University doesn't necessarily understand the type of data that's being created and the right balance of protections that ae required for it: it is hard to develop the right balance institutionally between enabling access and protecting data.
- There is an issue of legacy consents, given that the study was conducted in the early 1990s: consents were very generic and more recent standards of consent are more complex. This can create issues in determining what kinds of use would or would not fall under the terms of the original consents.

Marta provided background to the University of Cambridge institutional repository, which has been in place since 2005. It covers six schools, which are separate entities, and 130 departments and research institutes. It receives 20-30,000 site visits per month.

- The university structure means that control over policies and protocols is highly devolved. This makes it difficult to roll out harmonised data policies across the university and to support the wide range of different disciplines.
- The service does provide education and training, for example through workshops on research data management, but these are difficult to scale.
- The repository has an outdated technical infrastructure. There is a pressing need to do more than just archive: active curation is necessary.
- There is no capacity to support managed access to data – any such requests have to be pushed out to other services such as the UKDS.
- The repository is seeking now to implement 3 different levels of access: open, safeguarded and controlled (with a DAC).
- Some disciplines wouldn’t have established procedures for managing sensitive personal data (e.g. engineering) but could be supported if they do generate this sort of data.
- Sustainability is a challenge as there is limited central resource to support the repository.
- The repository would like to help researchers to make datasets discoverable, for example by connecting UK institutions together to help make metadata discoverable and available and to flag how and where to access data.

Marta outlined what is needed within the community to help support data curation and management:

- Education: more widely established training provisions on data management and sharing, right from grant planning stage (writing data management plans and creating consent forms).
- Better communication between institutions, repositories, funders and the research community: unless you're embedded in research practice and understand researcher needs they won't listen or shift their behaviour towards a more proactive approach to data management. Institutions are well-placed to play this advocacy role for their researchers.

Session 3 – Towards a sustainable infrastructure for cohort and longitudinal studies

The meeting broke into small groups to discuss three key questions:

- What are the main gaps and challenges in providing infrastructure to support cohort and longitudinal studies?
- How could existing repositories develop to meet these needs?
- How could we break down silos and move toward a joined up infrastructure that meets the needs of data generators and data users?

Observations:

There has been significant progress over the last decade or so: what was previously impossible is now merely extremely difficult. The next step will be making the transition to making data access, curation and management easy.

Gaps and challenges

1. *Balancing the need for harmonisation with recognising the specifics of different disciplines and data types*

- There is a constant tension between the drive to harmonise policies, processes and language to make things simpler and enable more/better sharing, and the recognition that there are cogent reasons for differences in approach, e.g. if data is international and needs to accommodate different jurisdictions (even within the UK's devolved nations).
- Across disciplines, there are mostly three categories of data used, corresponding roughly to open, safeguarded and controlled, but these are not consistent with one another (e.g. does 'open' require a licence/signed agreement?)
- The way that data controllers make decisions about disclosure risk judgements is not based on unifying principles. The issue of anonymisation and risks of re-identification can be a red herring in determining access levels: the UKDA is good at flipping the argument to focus on the importance of the research environment and how the data can be used safely, which may be an appropriate approach for other disciplines.
- There are significant disciplinary differences in cultures about controls over data and the expectations attached to data use, for example, insistence on collaboration or only sharing within an established consortium.
- There is no consensus on the best set up for data management for different studies. Cohorts have evolved relatively independently, piecing together data management approaches as they've developed. There's little incentive for them to align, and

significant conversion costs to change their established practices and procedures. Initiatives such as CLOSER are helping address this, but they are limited in scale.
- The number and range of repositories also highlights the difficulties of establishing the right balance: a large number may make discovery more difficult, but too few covering a broad range of data might mean metadata is not granular enough, which could make specific datasets harder to find.

2. Sustainable funding

- Hand-to-mouth, short term funding makes it very hard to build a good infrastructure to support data curation and management. 'Bolt on' approaches are often inefficient in the long term, not scalable and make it difficult to plan for the future, especially as data generation is increasing rapidly. Longer term funding commitments are essential.
- Cost recovery may not be an appropriate mechanism of studies/repositories, because it may make the barriers to access too high or entail too high a resource commitment to fulfil requests. What is an appropriate alternative?
- The provision of a 'safe settings' environment for the UKDS creates additional costs of purchasing licences for software (programmes like SPSS, STATA, Windows license) – this cost is estimated to around £100k/year. This could be calculated in the data access charge but may be prohibitive to users.
- Funding for infrastructures could be separate from funding for research. At the moment, funding for most infrastructure support (with the exception of UKDA/UKDS) is considered alongside research applications as part of the same grant award, and thus is often considered insignificant or of less importance than a novel science proposal.
- Repositories may need new solutions that could be efficient in the long run but require capital investment in equipment and skills to manage.

3. Skills

- Resources need to be invested in providing training in data management and science. Incentives for staff retention and maintenance of skills for data management are essential.
- Stakeholders need to recognise that the challenges for data management are about more than the bare infrastructure required: people, development and support are critical.
- There's value in ensuring that data managers and researchers mutually understand one another's perspectives, experiences and expertise. It's helpful if data managers have knowledge of research and research skills.
- There are vital administrative and data management skills that are needed as part of good data curation, but these are not 'valued' skills or career paths at present. Service provision is not necessarily an attractive path for career progression.
- Standard training on data management skills should be a core part of ongoing researchers' skills development.

4. Future thinking

- Historical and legacy consents will continue to be a major challenge and there is not a consistent approach to how these are managed.
- Current methods for curation and management will also generate problems now and in future: static datasets becoming dynamic; increased linkage possibilities; improved data

science analysis and interpretation techniques; and the sheer quantity of data will make good curation and management all the more important.

How could existing repositories develop?

1. *Share good practice on specific issues*

- There are many examples of good practice that could be adapted and shared across domains, for example approved researcher status. Repositories could take a leading role in establishing best practice that works across disciplines.
- Social science and biomedical repositories haven't typically worked well together, but there may be particular issues on which to join up, e.g. on data discovery, sharing metadata and metadata schema.
- In terms of skills sharing, there is a great deal repositories can learn from each other all the way through the data life cycle and opportunities are needed to facilitate these conversations. Most are hard pressed to be able to spend time on networking.
- Federated models enable control to be retained for individual studies or datasets, but this requires common standards across datasets, e.g. those being established through the data documentation initiative (DDI).

2. *Seek a balance on specialising versus generalising*

- Existing repositories could seek to develop to accommodate large datasets that include data beyond their usual realm of expertise, for example to give rich cohort data a natural home. This could also help repositories build up skills sets in different data types.
- In some cases, harmonisation is not appropriate: discipline specific repositories remain important to retain the right skills to do discipline specific research and support.

3. *Consider cost efficiencies*

- Given the prohibitive costs of software licences for safe settings, repositories could institute a policy of researchers needing to bring their own software licenses e.g. for SPSS and Windows. The national safe haven in Scotland is doing this.
- CLOSER as a research project could be mined to look at benefits of cross-working between the disciplines. It has allowed pump priming in cross study activities that would have been difficult to do on an individual basis.

How could we break down silos?

- Better exchange of good practice and all stakeholders learning from each other's experiences and expertise.
  - If possible, shared principles should be developed, with a common language that applies across different data types.
  - Standardise data access criteria: wherever possible these should be shared among data controllers to ensure transparency and consistency.
  - There needs to be a common understanding of assessing disclosure risk and clarity on the legal gateways for accessing data.
- Some parts of the data access process could be automated and shared across disciplines, e.g., data requestor authentication.
- If funders are going to mandate data sharing, they need a coherent, consistent, compelling policy that applies across disciplines:

- The Research Council structure doesn't help as it tends to reinforce silos. Work across funding councils and other funders would be beneficial on specific areas, not just high-level policies.
- Join up across funders is required across range of activities and disciplines so we don't end up reinventing bespoke solutions to generic problems.

Session 4 – The Way Ahead: Actions for Funders

David Carr (Wellcome Trust)
Wellcome is strongly advancing the Open Research agenda:

- Considering dedicated funding to support initiatives that seek to make data more shareable, especially outside of the typical five year funding cycle;
- Looking to be more strategic, e.g. on clinical trials' data access;
- Identifying how we could harmonise policies and processes;
- Recognising the skills development issue and the difficulties of establishing the roles of funders versus institutions on this.

Rhoswyn Walker (MRC)
The MRC recognises that the costs and the values of data sharing are often buried and difficult to identify:

- It is hard to demonstrate value to users to develop community buy in and support.
- There's a need to break down the 'us versus them' culture that is an obstacle to data sharing, especially in the clinical world, and ensure the benefits flow back to the providers of data.
- The establishment of UKRI should facilitate better cross-council collaboration: data is primed to be one of the major beneficiaries of this joining up, and HEFCE is being brought into conversations about data needs for institutions.

Mark Thorley (NERC)
NERC is supporting a network to manage environmental data, costing £12mn, but data access has a low profile and there is insufficient recognition for the infrastructure.

- Domain specific repositories are a critical part of long term research infrastructures and have the skilled people to manage them.
- Funders need to put their collective will into getting more value out of the data generation they fund, and rise to the challenge, for example, by not funding a cohort unless there is a clear strategy for sharing data.
- Cyber security is a major issue and the integrity of research datasets needs to be preserved. In sensitive areas, a breach could really damage research and people's perception of it.

Q&A
Q: What could studies do better to help themselves with funders?
A: The roles of all of the team in the research project should be clearly articulated and valued, especially the data managers. Applicants should also be very clear if they need costs for data infrastructure or management, especially when funding is needed for data management to augment or maximise the value of the research being proposed.

Q: How should we weigh up the value of centralised solutions versus small, specialised repositories?

A: Funders need evidence of costs and benefits from data repositories to make appropriate decisions. Joined-up big infrastructure projects can be high-risk and might prevent innovation: it's a challenge the balance the benefits of centralising (easier discoverability, standard terms, consistent access conditions) with its drawbacks (potentially stifling innovation; over-generalising leading to lack of specialist expertise with complex data; poor flexibility; risks of unsustainable funding).

- Basic infrastructure (secure, modern storage platform) should be shared between all the repositories and costs shared across funders. This could be cost efficient.
    - Matthew Woollard described the transition of the UKDS's technical infrastructure and services into a single location (Essex) to illustrate the costs and effort required to do something like this. 4.5 years into the process and they are still in the transition, despite social science data being relatively well organised and structured in terms of data and metadata standards. He doubted this might effectively happen for discipline-specific repositories.
    - A strong and powerful central body is needed to dictate standards at the outset; it's much harder to bring people together once software, vocabulary and processes are already established.
- Instead of joined up infrastructures that are too closely tied together, the emphasis could be on interoperability, so that where there is value in joining up this can be done but it's not an imperative if it doesn't best serve the interests of the science.
    - Social science is generally good at describing their data; this coherence could be created/introduced across domains e.g. in genomics, which is usually poorly described. Joining up could be done at this low level, without which you can't have a decentralised interoperable infrastructure.
- A key issue for interoperability is that the social and medical sciences are incredibly diverse: datasets might need different metadata, mapping on to different ontologies. Would this make data visible and understandable to those outside the specific technical domains, or would it cease to be useful and meaningful?

Q: How do we continue to talk across disciplines?

A: We need the right fora to mix with different people and build mutually beneficial relationships: people will only reach out if there's an incentive to do so.

- METADAC is useful in this regard as it has created a system for consolidated access governance that bridges disciplines. Cross-discipline groups and committees are a valuable way of helping break down disciplinary silos.
- Joint funding programmes across funders are another good way to maintain these discussions: ADRN, Farr Institute, CLOSER etc.
- There are also several examples of initiatives focusing on specific aspects of data management across disciplines e.g. the Software Sustainability Institute.
- Lessons can be learned from other fields, e.g. databases have developed in model organism genomics completely independently and totally interoperable.

Q: How should the right support be provided for researchers?

A: This is a role for institutions and for discipline-specific experts. No online training and webinars will ever replace dedicated support that researchers need. Researchers need someone to be there to answer queries. They want to know they can speak with a person

who understands the data and cares about data quality and integrity to support their science. This is the role of institutions and discipline-specific experts.

Joining up training is one way to build relationships and collaborations over time and demonstrate the value of interaction: the SURE training was generated from interactions of 4 data services. It took 2 years to produce joint training from the starting point of having individual training courses, but the process produced a better training course than anyone had individually at the outset.

Meeting Participants

Natalie	Banner	Wellcome Trust
Olly	Butters	ALSPAC
David	Carr	Wellcome Trust
Louise	Corti	UK Data Archive
Philip	Curran	1946 NSHD
Simon	Doran	Institute of Cancer Research
Mark	Elliot	University of Manchester
Paul	Flicek	European Bioinformatics Institute
Alissa	Goodman	1958 Child Development Study
Stephen	Gray	data.bris
Tim	Hubbard	King's College London
Jon	Johnson	Centre for Longitudinal Studies
Robert	Luben	EPIC Cambridge
Katy	McNeill	UK Data Archive
Lynn	Molloy	ALSPAC
Jacky	Pallas	UCL Research Data Services
Alison	Park	CLOSER
Mark	Parsons	Edinburgh Parallel Computing Centre
Steve	Pavis	NHS Scotland
Robin	Rice	Edinburgh Data Share
Jonathan	Sellors	UK Biobank
Andrew	Steptoe	English Longitudinal Study of Ageing
Marta	Teperek	Cambridge Data Repository
Mark	Thorley	NERC
Neil	Walker	University of Cambridge
Rhoswyn	Walker	MRC
Matthew	Woollard	UK Data Archive
Melanie	Wright	University of Essex