British Library, *F1000 Research*, *Good Form and Spectacle*,
London Metropolitan University, Multi-centre Advisory Group

AMASING AIMS

- Develop the text-DataSHIELD analysis packages (BL application)
- Plan the pilot DataSHIELD implementation with F1000 Research
- Scope user interface (non-command line)

14/7/2015

AMASED: Access Methods for Analysing Sensitive Data

Research data spring

University of BRISTOL

DataSHIELD
Secure Bioscience Collaboration

Team

- Who have you worked with?
  - DataSHIELD team, British Library, F1000 Research, London Metropolitan Uni (meeting pending)
- Who got involved over the past 3 months?
  - Feedback from wider community
    - Prof. David Zeitlyn (Social Anthropology, University of Oxford) re: UK Data Service
    - Dr Adam Crymble (Digital History, University of Herts)
    - Others users: Wellcome Trust, data journals, research libraries

Demand outstrips resource to supply!

Valuation

- Amazing and useful
  - Simultaneous emergence of serious real-world problem (how to "share" confidential data) & flexible/affordable technical solution (DataSHIELD)
  - Basic tech/stats methods proven: next enhance usability, implementation in practice
- Love and commitment
  - Open-source, free to obtain software fundamental to all components
  - **Numerous varied applications based on common foundation**
  - Growing user & developer community
- Sustainability
  - Now, core funds still needed for:
    - Known tech, governance & support challenges; grow strengthen & organise user/dev communities; scope business models for future user/dev support
  - Going forward: could provide viable, flexible & **affordable** national service providing access to academic data with tailored privacy protection
    - Jisc, funders, journals, universities, individual developers
    - Basic infrastructure cheap; improved modular functionality, "interesting" to program, governance evolve naturally from other (inter)national initiatives

Progress

» Scope challenges of implementing DataSHIELD within F1000 Research

Results: Seroprevalence for HIV 0.47% (95% CI 0.30-0.76) (17/3,599), and syphilis 0.39% (95% CI 0.23-0.65) (14/3,592), were low. Syphilis was significantly lower in refugees (0.07% 95% CI 0.01-0.38) (1/1,469), than in

F1000Research F1000Research 2015, 3:123 Last updated: 02 JUL 2015

RESEARCH NOTE

REVISED **Audit of antenatal screening for syphilis and HIV in migrant and refugee women on the Thai-Myanmar border: a descriptive study [v2; ref status: approved 1, approved with reservations 1, http://f1000r.es/5dn]**

Rose McGready[1-3], *et al*

```
> ds.table1D("hiv")
$counts
        hiv
0      3582
1        17
Total 3599
```

```
> ds.glm("hiv~1",family="binomial")

Family: binomial
Link function: logit


$coefficients
             Estimate Std. Error   z-value
(Intercept) -5.350463  0.2431105 -22.00836
                    p-value low0.95CI.LP high0.95CI.LP
(Intercept) 2.394913e-107     -5.826951     -4.873975
                      P_OR low0.95CI.P_OR high0.95CI.P_OR
(Intercept) 0.004723534     0.002938389      0.007584949
```

SCOPING EXERCISE

Key challenges identified to be met in next phase

- Where to site data and who should "manage" it?
- What are optimal rules for disclosure protection and how contextual should those rules be? – *e.g.* syphilis tabulation
- Modify functions to augment disclosure control – *e.g.* ds.glm
- Modify functions to augment utility – *e.g.* enhanced missing data handling; how manuscript-specific?
- Add new functions to augment utility – *e.g.* survival analysis, genome-wide association data; how manuscript-specific?
- Implementation of formal governance mechanism

TECHNICAL SOLUTION

Standard multi-site DataSHIELD reconfigured for single-site application

- Works in principle

```
> ds.table2D("D$syphillis","D$Refugee_migrant")
WARNING: Invalid contingency table from 'study1' !
         Only total values are returned in the output table(s).
$counts
$counts$`pooled-D$syphillis(row)|D$Refugee_migrant(col)`
          1    2
1        NA   NA
2        NA   NA
Total 1469 2123
      Total
1        14
2      3578
Total  3592
```

DataSHIELD
Secure Bioscience Collaboration

Progress

» Deploy DataSHIELD to analyse unrestricted digitised books

AMASED / bl2

Dictionary | Summary | Values | Permissions

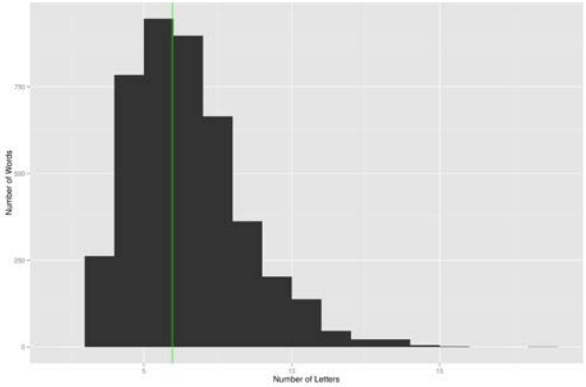
Properties

Name	bl2
Entity Type	Participant

Variables

+ Add Variable

Name	Label	Value Type
id		text
content		text
hpos		integer
vpos		integer
width		integer
height		integer
wc		decimal
cc		integer



- Written tools to ingested BL dataset into our data warehouse (Opal) and reshape data for analysis
- Successful non-disclosive, unrestricted text analysis using standard R packages

Progress / Next Phase

- Scope challenges of integration of data cleaning tool and DataSHIELD
  - Because we have progressed further on F1000 and BL goals than hoped we have moved this goal to phase 2 goal 5
- Success Indicators: Defined, realistic methodology for integrating data cleaning tool and DataSHIELD

Next phase

- Goal 1 Set up advisory group (researchers, text miners, collators of digital collections, DataSHIELD developers)
  - Deliverables: Establish advisory group who will
    - Identify relevant analytical techniques
    - Identify data restrictions on digitised books
    - Create workflow to prevent data disclosure (statistical and computational methods)
  - Success Indicators: Generalised methodology for preventing disclosure of restricted data from digitised books

Next phase

- Goal 2: Develop proof of concept implementing findings of Goal 1
  - Deliverables:
    - Implement DataSHIELD methodology scoped in Phase 1 Goal 2and Phase 2 Goal 2 for application to open digitised books
    - Build proof of concept DataSHIELD text analysis package of shortlisted analytical functions
  - Success Indicators: Demonstrate remote non-disclosive text analysis using DataSHIELD methodologies

Next phase

- Goal 3: Develop proof of concept for the remote analysis of F1000 Research paper data
- Objectives:
  - Adapt existing DataSHIELD infrastructure based on findings scoped in Phase 1 Goal 3
  - Replicate an F1000 Research paper analysis
  - Liaise with F1000 Research to identify a plan for pilot implementation
  - Deliverables:
    - Build proof of concept DataSHIELD for application in data publishing
    - Create implementation plan for pilot
  - Success Indicators:
    - Demonstrate remote non-disclosive analysis of research paper data can be replicated using DataSHIELD and key challenges have been scoped and a forward plan developed to deal with them:
      - Technical aspects of IT infrastructure; analytic flexibility; disclosure control; management/siting of data; governance; QC; business model and sustainability

Next phase

- Goal 4: Scope user interface for the software
  - Objectives: Interaction design team to liaise with DataSHIELD developers, users and the Advisory Group to scope interface
  - Deliverables: Project report outlining scoping findings and suggested interface
  - Success Indicators: Model for design and implementation of user interface

Funding

		Cost including Inflation		Cost excluding Inflation	
		YEAR 1	TOTAL	YEAR 1	TOTAL
		£	£	£	£
Staff Costs	**FTE**				
Dr Rebecca Wilson					
Basic Salary		4,720	**4,720**	4,720	**4,720**
National Insurance		390	**390**	390	**390**
Superannuation		755	**755**	755	**755**
Total	0.116	5,865	**5,865**	5,865	**5,865**
Scheme					
Basic Salary		9,649	**9,649**	9,401	**9,401**
National Insurance		763	**763**	738	**738**
Superannuation		1,544	**1,544**	1,504	**1,504**
Total	0.267	11,956	**11,956**	11,643	**11,643**
Total Staff Costs [1]		**17,821**	**17,821**	**17,508**	**17,508**
Non Staff Costs					
Travel & Subsistence		5,000	**5,000**	5,000	**5,000**
Fees (Int & Ext)		17,000	**17,000**	17,000	**17,000**
Total Non Staff Costs [2]		**22,000**	**22,000**	**22,000**	**22,000**
Facility Costs					
Total Facility Costs [3]		**0**	**0**	**0**	**0**
Estate Costs					
Estate Costs		2,983	**2,983**	2,983	**2,983**
Infra Lab Technician Costs		0	**0**	0	**0**
Total Estate Costs [4]		**2,983**	**2,983**	**2,983**	**2,983**
Indirect Costs					
Indirect Costs [5]		**16,602**	**16,602**	**16,602**	**16,602**
TOTAL COSTS [6=1+2+3+4+5]		**59,406**	**59,406**	**59,093**	**59,093**

Direct Costs	£39,821.00	
Total fec	£59,406.00	
Jisc	£44,554.50	75%
University of Bristol	£14,851.50	25%

Not for the pitch, but please fill in

- Contact person Dr Becca Wilson, University of Bristol
- Social media presence
  - @Data2Knowledge
  - @drbeccawilson
  - #d2kDataSHIELD
  - www.datashield.ac.uk