

'Big Data' and the social sciences – a perspective from the ESRC

Peter Elias



What do we mean by ‘Big data’

- ▶ Electronic data generated from research infrastructures (e.g. astronomy, particle physics, micro-biology, *etc.*)
- ▶ Electronic data not designed for research but with potential research value which records transactions, communications, physical movements (e.g. customer databases, service delivery records, internet search activity, *etc.*)

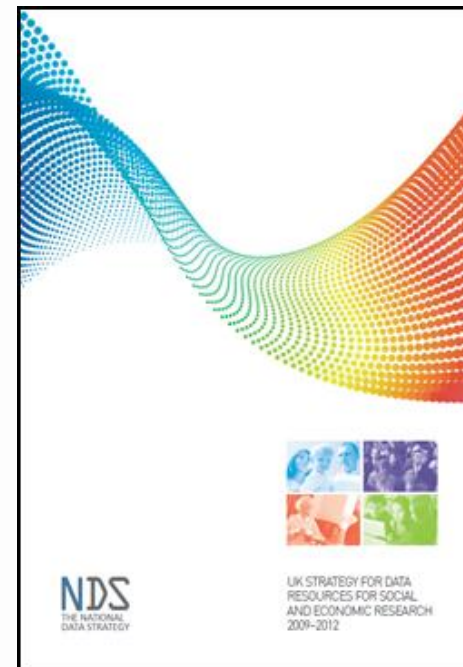
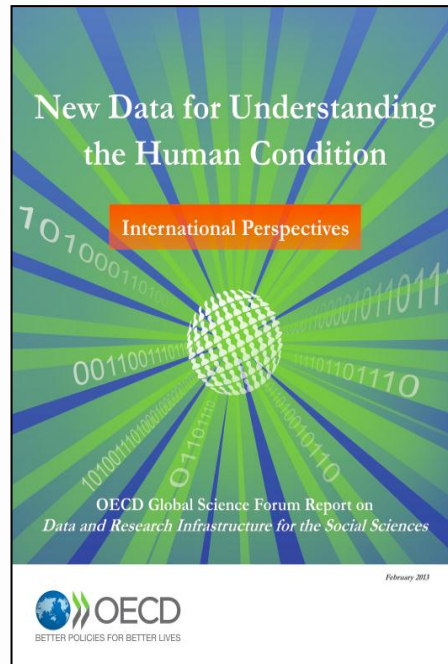
Why are such data of interest to social scientists?

- ▶ Their large scale means that events which are relatively unusual can be captured.
- ▶ Re-using existing data is cheaper than creating data for research
- ▶ Some types of data are ‘international’ – they are not generated via national processes
- ▶ They can often be combined with data collected by more traditional methods, to facilitate ‘triangulation’ and/or to enhance such data

What problems do researchers face when attempting to utilise 'Big data'

- ▶ Access problems
- ▶ Understanding their utility as research resources
- ▶ Data linkage and sharing for reuse may be difficult
- ▶ New skill sets are required (data science)
- ▶ Data curation needs have to be rethought

What has the ESRC been doing to promote a 'big data' research agenda?



The data landscape and ESRC

- ▶ UK Strategy for Data Resources for Economic and Social Research
 - First published in 2007, then 2009, and refreshed for 2013
- ▶ Integration of data services via the UK Data Service
 - Access to census and other datasets
 - Safe access to sensitive data
- ▶ The Administrative Data Liaison Service (ADLS)
 - Intermediary between academics and data holding organisations
- ▶ E-science and new forms of data
 - E-health centres & Digital Social Research

The data landscape and ESRC

Big data are identified in three main areas:

- ▶ Administrative data (from public bodies)
- ▶ Social media data
- ▶ Private sector data

Three related initiatives have been taken to progress the development of such data as research resources:

- ▶ Administrative Data Research Network
- ▶ Centre for International Social Media Research
- ▶ Business Datasafe

A vertical decorative bar on the left side of the slide, composed of horizontal segments in red, blue, and grey.

The Administrative Data Research Network

Some of the problems currently faced by researchers

- ▶ Inconsistent access conditions
- ▶ Severe time delays in granting access or refusal
- ▶ Lack of information about selection and/or linking of administrative datasets
- ▶ Restricted access to datasets – especially for addressing the counterfactual
- ▶ Data controllers making inconsistent decisions about appropriateness of data for research
- ▶ Research permitted then publication denied

What progress has been made in terms of research use of administrative data?

- ▶ Scottish Longitudinal Study (SLS); Scottish Health Informatics Programme (SHIP)
- ▶ Welsh Secure Anonymised Information Linkage (SAIL)
- ▶ Four new e-Health Centres (London, Manchester, Dundee, Swansea) funded by a consortium of 10 funders

ADT key recommendations

- ▶ (1) An **Administrative Data Research Centre (ADRC)** should be established in each of the countries in the UK
 - To commission and undertake data linkage and make the linked data available for analysis, thereby creating new resources for a growing research agenda
 - A UK **Governing Board** to provide the governance structure for the ADRCs
 - Information about the ADRCs, including access arrangements, to be managed through an **Information Gateway** (now the **Administrative Data Service**)

Capital funding

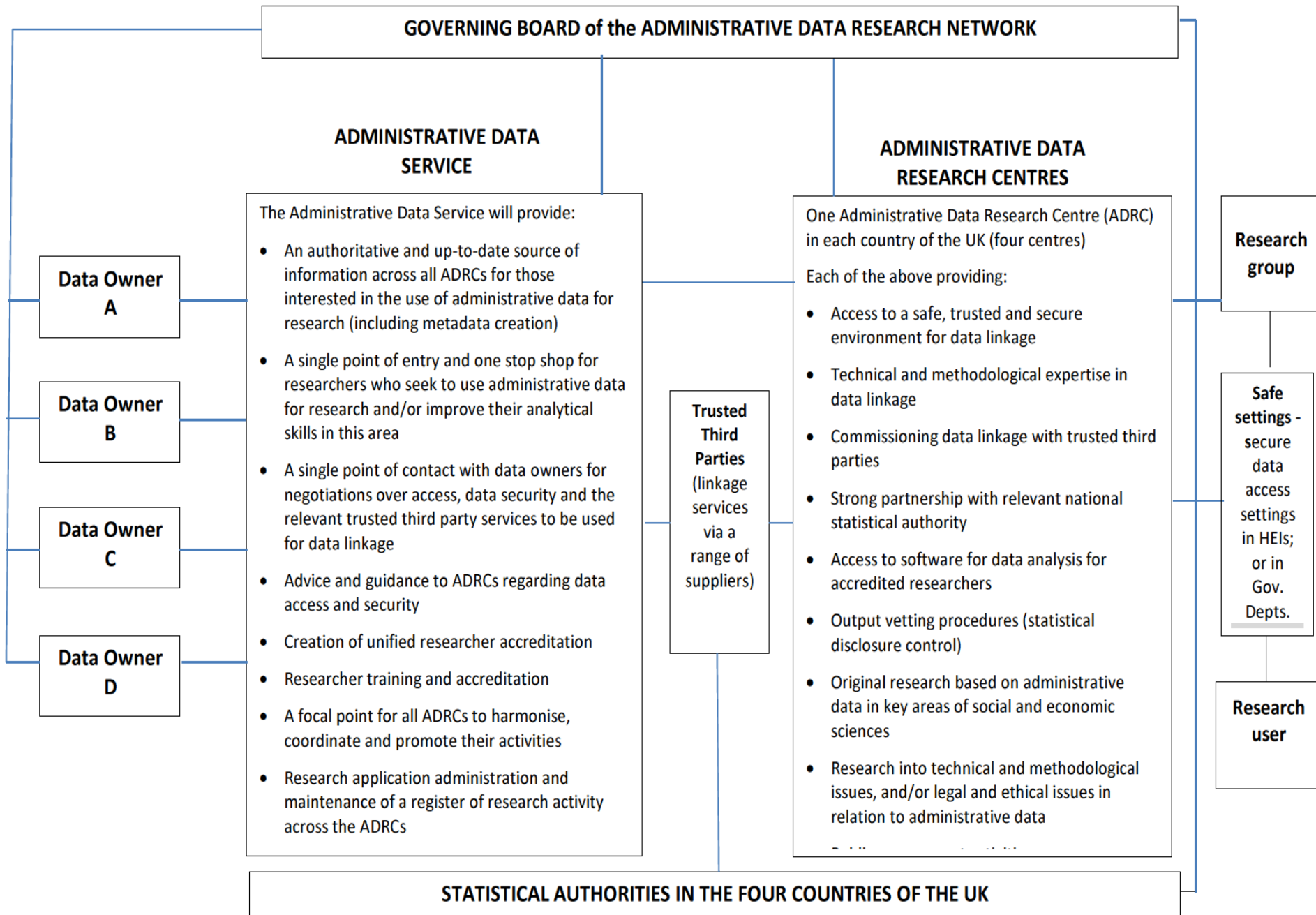
- ▶ Chancellor's Autumn Statement included £600m for science, research and innovation, £484m for RCUK
- ▶ Funds to support the development of innovative technologies across eight areas, including 'big data'
- ▶ Draws from RCUK Strategic Framework for Capital Investment (published Nov 2012)
- ▶ ESRC – earmarked £64m to support packages of activity within the 'big data' theme

ADRN elements: An overview

- ▶ Four Administrative Data Research Centres (ADRCs) in each country : England, Wales, Scotland and Northern Ireland
- ▶ An Administrative Data Service for the network
- ▶ The UK Governing Board for the network
- ▶ Researcher Training and Accreditation
- ▶ Safe Settings
- ▶ Public Engagement Strategy
- ▶ Key Partnerships with National Statistical Authorities

Figure 1

Functions of the Administrative Data Research Network (ADRN)



The AD Service



AD Service – key functions

As the source of information and coordination across the ADRN, the AD Service will be the:

- ▶ Single point of entry into and a front line service for researchers
- ▶ Principal point of contact with data owners and ADRCs for data negotiations
- ▶ Central point of expertise for advice and guidance for ADRCs on data access, safe settings, researcher training and accreditation
- ▶ Focal point for all ADRCs to harmonise, coordinate and promote their activities

Researcher training and accreditation

- ▶ ADT recommendations for a single UK-wide accreditation framework built on best practice
 - Researcher Training and Accreditation
 - Accredited access facilities
 - Approved research projects
- ▶ AD Service to lead on the collaboration with data owners for a uniform system of researcher accreditation, and to coordinate training provision
- ▶ Map existing provision - ADLS review

The ADRCs



ADRCs – key functions

Four Administrative Data Research Centres (ADRCs) - Scotland, Northern Ireland, England and Wales - each building on existing investment and providing:

- ▶ State-of-the-art facilities
- ▶ Data management
- ▶ Linked administrative data and research
- ▶ Training, capacity building, and public engagement

Central role for the national statistical authorities

Legislation

- ▶ The Taskforce recommended that primary legislation should be sought to provide a generic legal gateway for research and statistical purposes
- ▶ Given the complexities and pressures on the legislative timetable, a **dual track approach** was proposed.
- ▶ The establishment of the ADRN will enable improvement in research access to administrative data and lay the foundations ahead of any new legislation