PERSISTENT IDENTIFIERS FOR THE UK: SOCIAL AND ECONOMIC DATA

MATTHEW WOOLLARD

UK **DATA ARCHIVE**
ECONOMIC AND SOCIAL DATA SERVICE
UNIVERSITY OF ESSEX

METADATA AND PERSISTENT IDENTIFIERS FOR
SOCIAL AND ECONOMIC DATA, 7-8 MAY 2012, BERLIN

UK • DATA ARCHIVE

WHY CITE DATA?

It’s a vital part of the scientific research process

- Acknowledges the researcher’s **sources**
- Gives data creators, authors and data curators **proper credit** when their work is reused
- Aids scientific **replication**
- Provides **permanent and reliable information** on data sources produced and used in research
- Facilitates data **resource discovery and access**
- Helps track the use and **impact** of data collections

OUR APPROACH TO CITATION

- Required by our **user agreement (End User Licence)** for many years:

8. To preserve at all times the confidentiality of information pertaining to individuals and/or households in the data collections where the information is not in the public domain. Not to use the data to attempt to obtain or derive information relating specifically to an identifiable individual or household, nor to claim to have obtained or derived such information. In addition, to preserve the confidentiality of information about, or supplied by, organisations recorded in the data collections. This includes the use or attempt to use the data collections to compromise or otherwise infringe the confidentiality of individuals, households or organisations.

9. To acknowledge, in any publication, whether printed, electronic or broadcast, based wholly or in part on the data collections, the original data creators, depositors or copyright holders, the service funders and the data service provider(s) in the form specified on the data distribution notes or in accompanying metadata received with the dataset or notified to me and without prejudice to paragraph 5 above to comply with any restrictions on my use of the data collections referred to or referenced therein or otherwise notified to me from time to time. To cite, in any publication, whether printed, electronic or broadcast, based wholly or in part on the data collections, the data collections used in the form specified on the data distribution notes or in accompanying metadata received with the dataset or notified to me.

10. To supply the relevant data service provider with the bibliographic details of any published work based wholly or in part on the data collections.

11. That the members of the Data Team may hold and process any personal data submitted by me for validation and statistical purposes, and for the purposes of the management of the service or for any other lawful purpose notified to me and to which I have consented under this Agreement in relation to a particular data collection, and they may also pass the information on to other parties such as: (i) depositors and distributors of material contained in or accessed via the data service

OUR APPROACH TO CITATION

- Should include **enough information** to ensure the exact version can be located

  University of Essex. Institute for Social and Economic Research and National Centre for Social Research, *Understanding Society: Wave 1, 2009-2010* [computer file]. 2nd Edition. Colchester, Essex: UK Data Archive [distributor], November 2011. SN: 6614.

- Different from an **acknowledgement**
  - A general statement giving credit to the source and distributor
  - Not an acceptable method of citing data

INTRODUCING **PERSISTENT IDENTIFERS**

- The idea of using **Persistent Identifiers (PI)** for data followed that for other research outputs
  - Persistence must mean enduring
  - Identifiers must be unique
- The digital ‘object’ should be **clearly defined** to ensure appropriate granularity
- PIs are also being applied to **people and places**
  - unique IDs for institutions and researcher, e.g. ORCID system

DEVELOPING OUR **METHODOLOGY**

- Our ‘data collections’ are not digital objects
- Need to capture changes made to data
  - Versioning data in a commonly understood manner
  - Rule-based but human mediated (in defining a ‘significant’ or ‘high impact’ change)
  - Use structured data so machine-actionable
- Integrate processes with:
  - Digital preservation activities
  - Current infrastructure / work flows
- Desire to ‘get it right first time’

CHANGES TO DATA

- Approx. **15% UKDA data collections** are altered within first year after first publication

- Some data collections are issued as **new editions**:
  - Changes to data/variables
  - Adding new ‘waves’ in a data series
  - Regrossing of a data series
  - Changes to documentation

RECORDING **SIGNIFICANT CHANGE**

- We have distinguished between major and minor changes to a data collection
- High impact vs. low impact
- Largely social science users want most recent data for for research, but information about earlier versions of data must be available … and we **should** be making earlier versions available … coming soon

MINOR CHANGES – **LOW IMPACT**

- Publication reference added
- Correction of spelling in variable labels
- Small changes in variable labels
- Removal of (erroneously supplied) admin variables
- Correction of spelling in metadata
- Minor changes in documentation
- New index terms
- Additional documentation added (non-fundamental)
- Change in access conditions

MAJOR CHANGES – **HIGH IMPACT**

- New variable added
- New labels/value codes added
- Weighting variables reconstructed
- Wrong data supplied (e.g., March not April)
- Mis-coded data (e.g., Don’t know/Refused confused)
- Change in format (file migration)
- Significant changes in documentation
- Change in access conditions

DEFINING AN **INSTANCE**

- Concept of an ***instance*** to denote a changed collection
- **Internal change** during ingest process (unreleased)
  ➔ new internal instance
- **Low impact** change (released)
  ➔ new external instance with unchanged PI
- **High impact** change (released)
  ➔ new external instance and new PI

ENTER DATACITE'S **DIGITAL OBJECT IDENTIFIER**

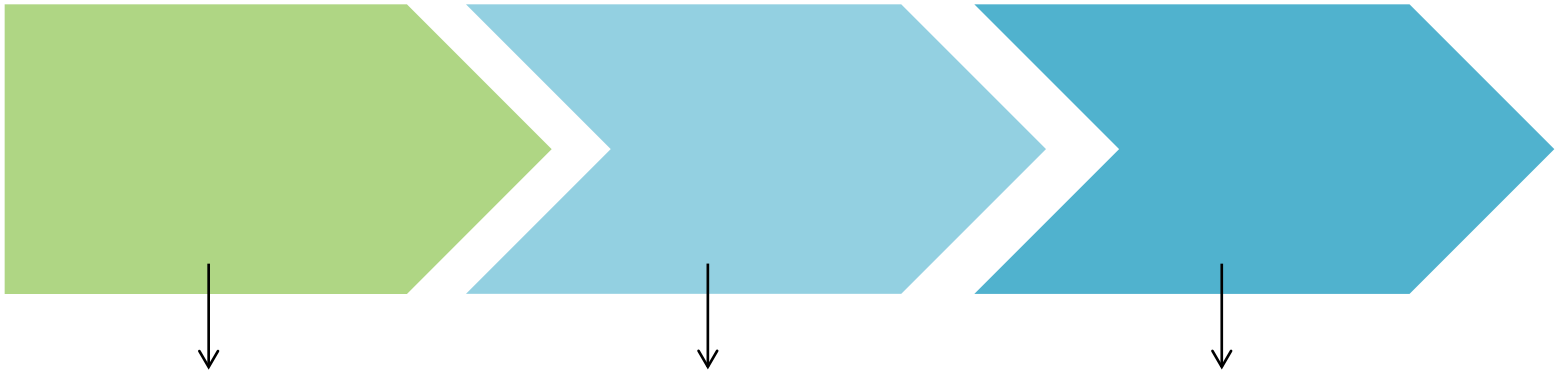
- In 2011 we started working with the British Library and DataCite to develop a permanent, reliable method of citing our data collections

- DataCite
  - Founded by organisations from six countries
  - Established a citation format for research data to provide an effective method of persistent identification
  - Works with data publishers, e.g. established data centres and institutional repositories

WHAT IS **A DOI?**

- The DOI framework is an **international and persistent standard** for identifying objects (or information about them) in a globally unique way
- A DOI is **a string of letters and numbers** that can be used to make resources directly available to anyone over the internet
- Subscribers can apply for DOIs which, when assigned, will be included in DataCite's DOI metadata store/resolver system
- There are already more than 1.3m (and growing) registered objects with DOIs

HOW DOIs ARE **CREATED**



The data publisher registers and obtains DOIs from a DataCite member, e.g. the British Library

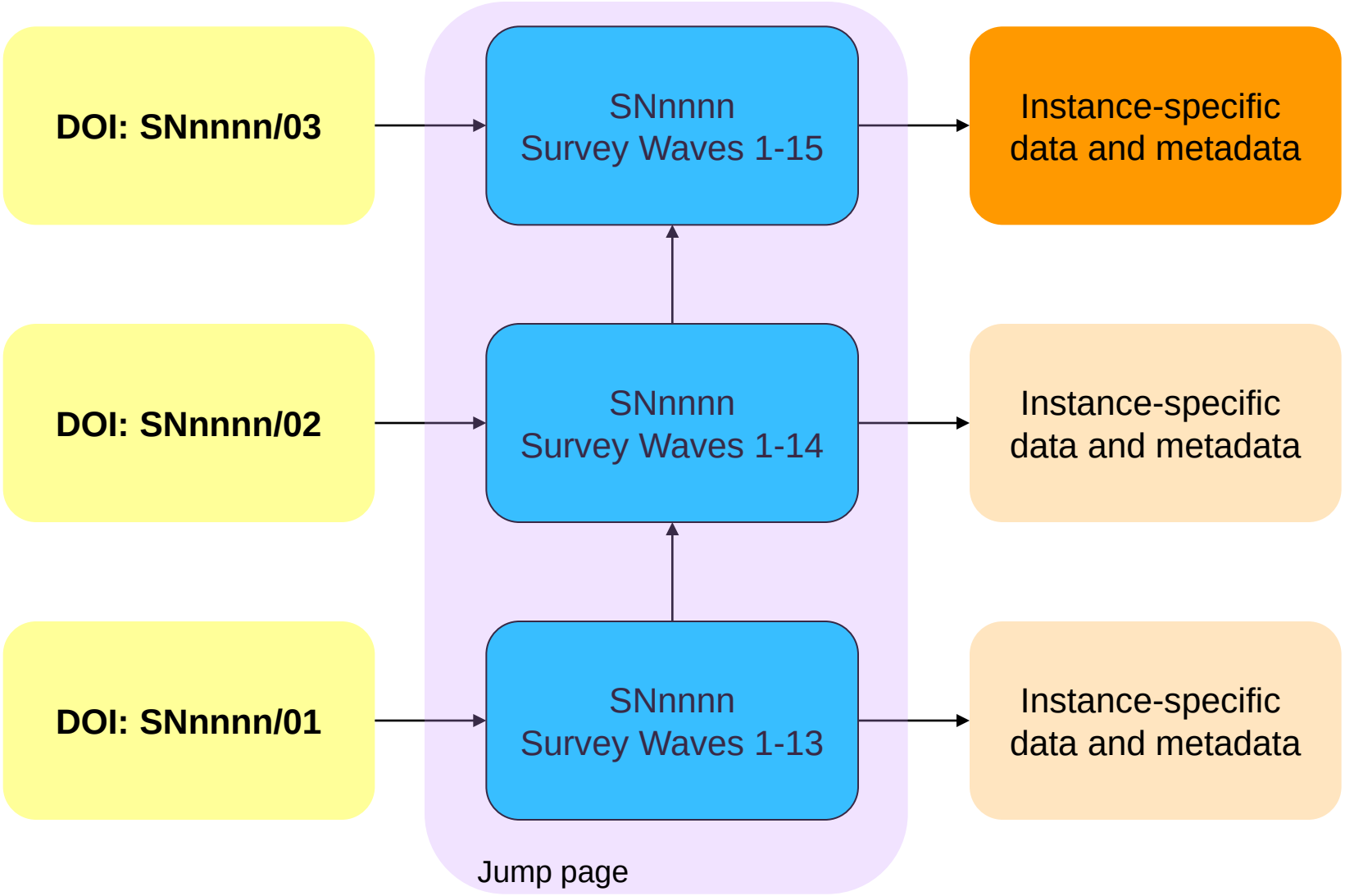
We mint and update DOIs with our metadata management infrastructure

We use DataCite' s application programming interface (API) to mint DOIs

WHAT WAS **OUR SOLUTION?**

- **Original solution**
  - DOI allocated to core metadata (title, etc.) relating to a data collection
  - Problem: even titles can change
- **Final solution**
  - DOI allocated to metadata relating to *each external* instance (metadata record) of a data collection
    - DOIs resolve to “jump” page pointing to all external instances
    - New DOI = High Impact change, with explicit logging

THIS IS WHAT THAT **LOOKS LIKE**



Understanding Society: Wave 1, 2009-2010 and Wave 2, Year 1 (Interim Release), 2010

A new Digital Object Identifier (DOI) is assigned to the data collection each time there is a major change to data, c
metadata. The new DOI will resolve to an updated version of this page containing a log of changes to this data coll
of its first DOI. The DOI system supports resource discovery and simplifies citation for users of data collections. Da
directly through increased visibility of their work.

10.5255/UKDA-SN-6614-3

Citation:
University of Essex. Institute for Social and Economic Research and National Centre for Social Research, *Understanding Society: Wave 1, 2009-2010 and Wave 2, Year 1 (Interim Release), 2010* [computer file]. *3rd Edition.* Colchester, Essex: UK Data Archive [distributor], February 2012. SN: 6614, http://dx.doi.org/10.5255/UKDA-SN-6614-3

Change log:
20 February 2012: For the third edition (February 2012) data and materials for the first year of Wave 2 were added to the study.

^

10.5255/UKDA-SN-6614-2

Citation:
University of Essex. Institute for Social and Economic Research and National Centre for Social Research, *Understanding Society: Wave 1, 2009-2010* [computer file]. *2nd Edition.* Colchester, Essex: UK Data Archive [distributor], November 2011. SN: 6614, http://dx.doi.org/10.5255/UKDA-SN-6614-2

Change log:
23 November 2011: For the second edition (November 2011), materials for the second year of Wave 1 were added to the study, which now comprises the full set of Wave 1 data and documentation.

Minor changes

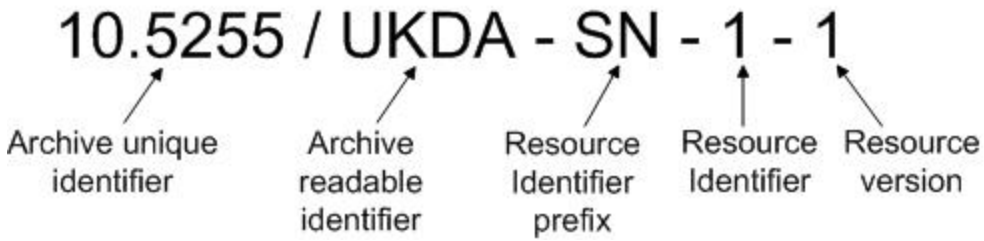
CREATING AND UPDATING DOIs

- **New** catalogue record
  - mint new DOI through DataCite
  - update DOI change log
  - create new citation file

- **Update** catalogue record
  - enter high/low impact changes
  - create/update DOI through DataCite
  - update DOI change log
  - create new citation file where high impact change

DOI **FORMAT AND VERSIONING**

- Archive readable identifier
- Resource type identifier
- Resource identifier
- Resource version
- http://dx.doi.org/10.5255/UKDA-SN-1-1



High impact change

10.5255/UKDA-SN-1-1

High impact change
Increments major version

10.5255/UKDA-SN-1-2

Low impact change

10.5255/UKDA-SN-1-1

Metadata is updated with minor version

Metadata Version	Created
1	2011-09-20T13:47:21.000Z
0	2011-09-20T13:27:26.000Z

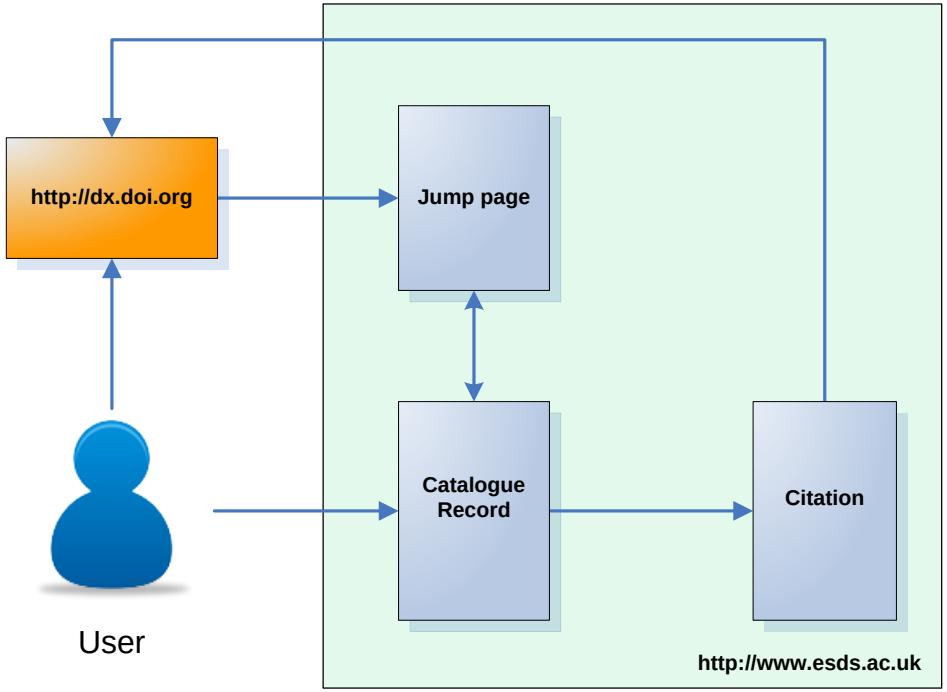

10.5255/UKDA-SN-6614-3
http://data.datacite.org/text/html/10.5255/UKDA-SN-6614-3
Reader
Google
DataCite Content Service Beta
DataCite
doi:10.5255/UKDA-SN-6614-3
This page represents DataCite's metadata for doi:10.5255/UKDA-SN-6614-3.
For a landing page of this dataset please follow http://dx.doi.org/10.5255/UKDA-SN-6614-3
Citation
University of Essex. Institute for Social and Economic Research National Centre for Social Research; (2010): Understanding Society: Wave 1, 2009-2010 and Wave 2, Year 1 (Interim Release), 2010; UK Data Archive, University of Essex. http://dx.doi.org/10.5255/UKDA-SN-6614-3 RIS BibTeX
Descriptions
Abstract
Understanding Society, or the United Kingdom Household Longitudinal Study (UKHLS), is conducted by the Institute for Social and Economic Research (ISER), at the University of Essex. The survey research organisation is National Centre for Social Research (NatCen), and in Northern Ireland, the Central Survey Unit of the Northern Ireland Statistics and Research Agency (NISRA).

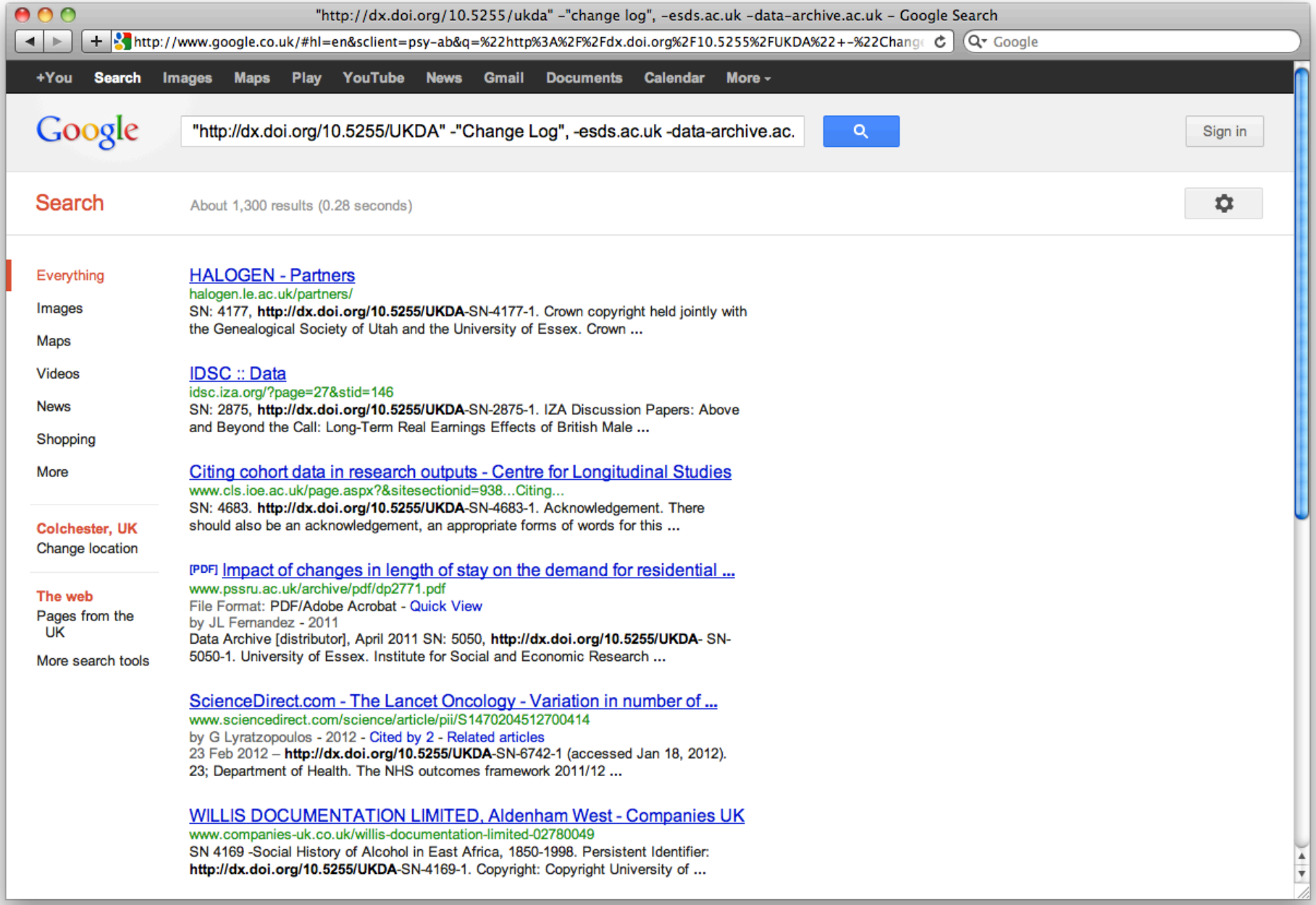
 As a multi-topic household survey, the purpose of Understanding Society is to understand social and economic change in Britain at the household and individual levels. It is anticipated that over time the study will permit examination of short- and long-term effects of social and economic change, including policy interventions, on the general well-being of the UK population. The study has a strong emphasis on domains of family and social ties, work, financial resources, and health. Further information about the survey may be found in the documentation, and on the Understanding Society web site.

 The study is an annual survey of each adult member of a nationally representative sample. The same individuals are re-interviewed in each wave. If individuals leave their household, all adult members of their new household are interviewed. Each wave is collected over 24 months, such that the first wave of data collection started in January 2009 and finished in January 2011. Data collection takes place using computer assisted personal interviewing (CAPI). One person completes the household

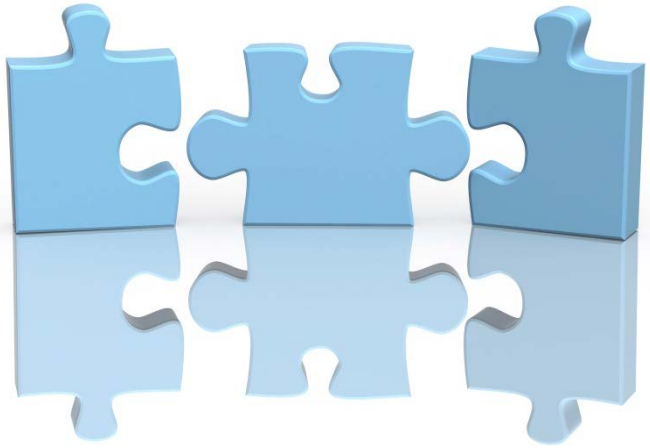
USING DOIs

- DOI links to an anchor on the jump page
- Citation file or the most recent DOI address
- Previous citations and data available on request only




"http://dx.doi.org/10.5255/ukda" -"change log", -esds.ac.uk -data-archive.ac.uk - Google Search
http://www.google.co.uk/#hl=en&sclient=psy-ab&q=%22http%3A%2F%2Fdx.doi.org%2F10.5255%2FUKDA%22+-%22Chang
Google
+You Search Images Maps Play YouTube News Gmail Documents Calendar More
Google
"http://dx.doi.org/10.5255/UKDA" -"Change Log", -esds.ac.uk -data-archive.ac.
Sign in
Search
About 1,300 results (0.28 seconds)
Everything
Images
Maps
Videos
News
Shopping
More
Colchester, UK
Change location
The web
Pages from the UK
More search tools
HALOGEN - Partners
halogen.le.ac.uk/partners/
SN: 4177, http://dx.doi.org/10.5255/UKDA-SN-4177-1. Crown copyright held jointly with the Genealogical Society of Utah and the University of Essex. Crown ...
IDSC :: Data
idsc.iza.org/?page=27&stid=146
SN: 2875, http://dx.doi.org/10.5255/UKDA-SN-2875-1. IZA Discussion Papers: Above and Beyond the Call: Long-Term Real Earnings Effects of British Male ...
Citing cohort data in research outputs - Centre for Longitudinal Studies
www.cls.ioe.ac.uk/page.aspx?&sitesectionid=938...Citing...
SN: 4683. http://dx.doi.org/10.5255/UKDA-SN-4683-1. Acknowledgement. There should also be an acknowledgement, an appropriate forms of words for this ...
[PDF] Impact of changes in length of stay on the demand for residential ...
www.pssru.ac.uk/archive/pdf/dp2771.pdf
File Format: PDF/Adobe Acrobat - Quick View
by JL Fernandez - 2011
Data Archive [distributor], April 2011 SN: 5050, http://dx.doi.org/10.5255/UKDA- SN-5050-1. University of Essex. Institute for Social and Economic Research ...
ScienceDirect.com - The Lancet Oncology - Variation in number of ...
www.sciencedirect.com/science/article/pii/S1470204512700414
by G Lyratzopoulos - 2012 - Cited by 2 - Related articles
23 Feb 2012 – http://dx.doi.org/10.5255/UKDA-SN-6742-1 (accessed Jan 18, 2012). 23; Department of Health. The NHS outcomes framework 2011/12 ...
WILLIS DOCUMENTATION LIMITED, Aldenham West - Companies UK
www.companies-uk.co.uk/willis-documentation-limited-02780049
SN 4169 -Social History of Alcohol in East Africa, 1850-1998. Persistent Identifier: http://dx.doi.org/10.5255/UKDA-SN-4169-1. Copyright: Copyright University of ...

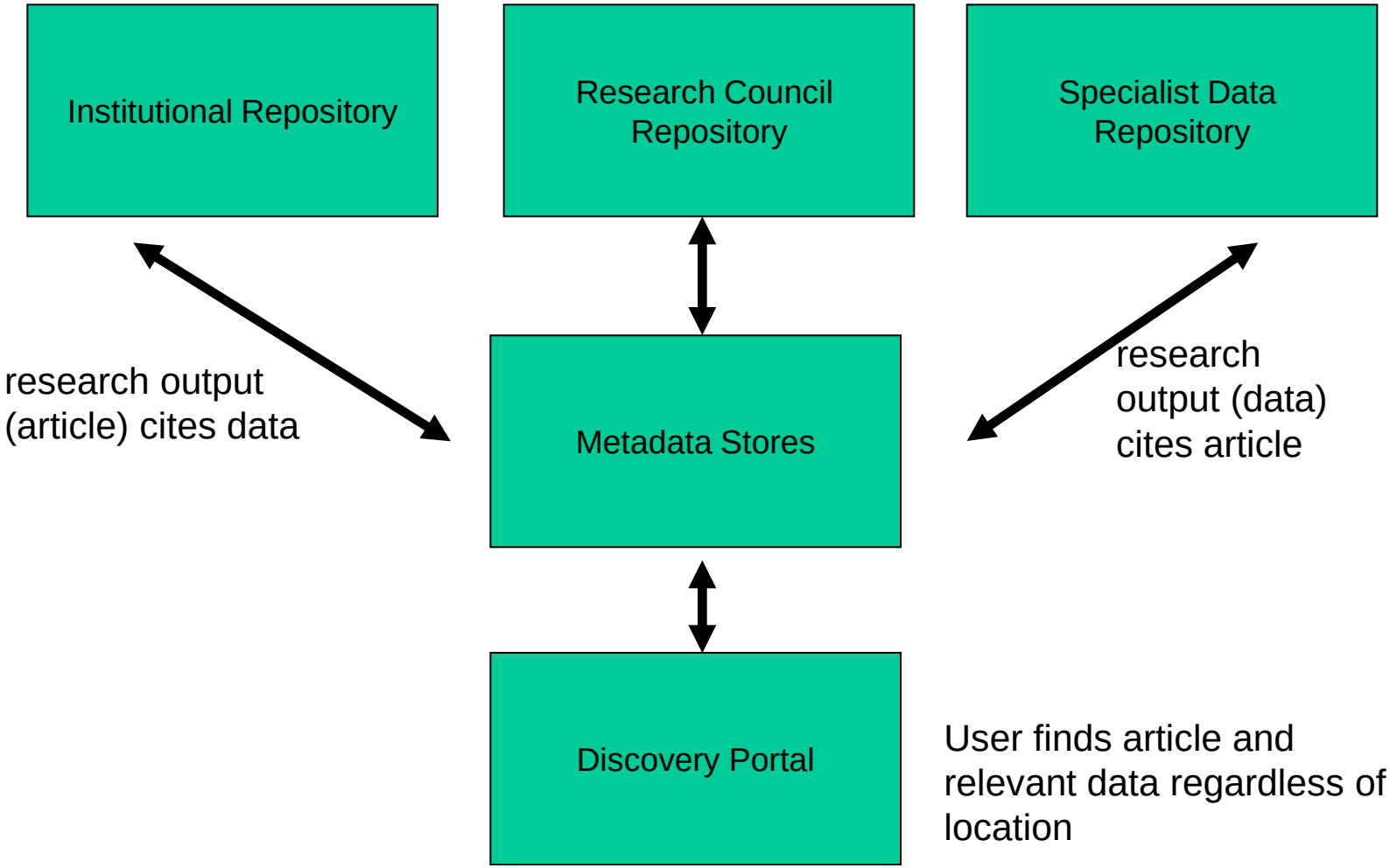
CHALLENGES FOR THE FUTURE



- Citing **parts (fragments)** of data collections
  - single files
  - subsets of quantitative data files
  - extracts of textual data

- Creating **relationships between** different objects
  - research outputs (articles) and research inputs (data)
  - research outputs (data) and research outputs (data)
  - any output and researcher/institution/funding information

LINKING RESOURCES VIA METADATA



RAISING AWARENESS IN THE SOCIAL SCIENCES

- ESRC funding for short-term project
- Aim to educate and inform best practice in citing research data
- Audiences
  - Professional organisations
  - Academic publishers and journal editors
  - Researchers and postgraduate students
- Key activities
  - Data citation principles for social sciences
  - Personal communications
  - Events with BL DataCite, JISC and PI community
  - Outreach through Doctoral Training Centres

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

What you need to know about **DATA CITATION**

ACKNOWLEDGEMENTS

- John Shepherdson, ESDS/UKDA
- Louise Corti, ESDS/UKDA
- Sharon Bolton, ESDS/UKDA

- Susan Noble, ESDS/MIMAS

CONTACT

UK **DATA ARCHIVE**
UNIVERSITY OF ESSEX
WIVENHOE PARK
COLCHESTER
ESSEX CO4 3SQ

T +44 (0)1206 872001
E comms@data-archive.ac.uk
www.data-archive.ac.uk

Economic and Social Data Service
University of Essex
Wivenhoe Park
Colchester
Essex CO4 3SQ

T +44 (0)1206 872001
E publicity@esds.ac.uk
www.esds.ac.uk