Citing Datasets and Linking Them to Publications

Alex Ball

DCC/UKOLN, University of Bath

8 December 2011

Funded by JISC

AWARENESS LEVEL

A Digital Curation Centre Briefing Paper

19th July 2011

Data Citation and Linking

By Alex Ball and Monica Duke, UKOLN, University of Bath

Introduction

On the surface, citing datasets is a trivially easy thing to do. Style manuals such as the *Publication Manual of the American Psychological Association* and the *Oxford Manual of Style* have provided sample citations for datasets since at least the early 2000s. The process of making datasets citable, however, is rather more difficult. In consequence of this and other factors, a culture of citing datasets has been slow to develop. Nevertheless, it is vital that researchers cite the datasets they use, if datasets are to be regarded as legitimate academic outputs in their own right.

Short-term Benefits and Long-term Value

There are several short-term benefits to making datasets citable, citing them in practice, and linking datasets to papers that make use of the data.

- If the authors of a scientific publication properly cite the data that underlies it, it is much easier for the reader to locate that data. This in turn makes it easier for the reader to validate and build on the publication's findings.
- Data citations ensure that data contributors receive proper credit when their work is reused by other researchers.
- If a dataset links back to the paper that describes its collection, a reader coming to the dataset direct can use that link to put it in context and understand the methodology used.
- If a dataset links to other papers that make use of it, these links can be used by the contributors and data publishers to demonstrate the impact of the data. Potential reusers might use these links to discover critiques of the data or to provide inspiration for how to use them.

Once a culture of data citation has been established, several other benefits are likely to become apparent.

- The publishing infrastructure that makes the data citable will also help to ensure they are available for reference and reuse long into the future.
- There will be less danger of rival researchers 'stealing' results from those who publish their data openly, as failure to give due credit would amount to plagiarism and thus be punishable.
- Services built around data citation will make it easier for researchers to discover relevant datasets.
- Data citations could be used to measure the impact of both individual datasets and their contributors.
- Researchers could gain professional recognition and rewards for published data in the same way as for more traditional publications.

Taking these points together, there would likely be an increase in the quantity and quality of data published, with all the benefits this implies for the transparency and rate of scientific research.

A Digital Curation Centre 'working level' guide

How to Cite Datasets and Link to Publications

Alex Ball (DCC) and Monica Duke (DCC)

Cite Datasets and Link to Publications

In this section

Curation Reference Manual
Curation Lifecycle
Policy and legal
Data Management
Case studies
Tools and applications
Briefing Papers
How-to Guides
- Appraise & Select Data
- Cite Datasets
- Develop a Data Plan
- License Research Data

Standards
Publications
External resources
Roles
Curation journals
Informatics research

This guide will help you create links between your academic publications and the underlying datasets, so that anyone viewing the

http://www.dcc.ac.uk/resources/how-guides/cite-datasets

interest researchers and principal investigators working on data-led research, as well as the data repositories with which they work.

By Alex Ball (DCC) and Monica Duke (DCC)

Published: 18th October 2011

Browse the guide below or download the PDF

Please cite as: Ball, A. & Duke, M. (2011). 'How to Cite Datasets and Link to Publications'. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides

Outline

Motivation

Elements of a data citation

Issues and challenges

Guidance for researchers

Guidance for data repositories

What's great about journal papers?

- Awareness raising
- Protection from plagiarism
- Verification of results
- Basis for future research
- Reward models
- Permanent access

What's great about journal papers?

- Awareness raising
- Protection from plagiarism
- ~~Verification of results~~
- ~~Basis for future research~~
- Reward models
- Permanent access

Data citations provide. . .

- Visibility for data
- Protection from plagiarism
- Possibility for verification of results
- Data on which to base future research
- Possibility for reward models
- Access

What's great about forward links?

Linking from resources to those that cite them

- helps gauge impact
- provides context
- reveals commentary and critique

Citation styles

Four data citation styles: which elements do they use?

Altman and King (2007): Dataverse

Lawrence et al. (2008): BADC

Green (2010): OECD

Starr and Gastl (2011): DataCite

Citation styles

Author

Altman and King (2007): Dataverse

- Sidney Verba.
  NORC [Producer];

Lawrence et al. (2008): BADC

- Iwi, A. and B. N. Lawrence

Green (2010): OECD

- OECD

Starr and Gastl (2011): DataCite

- Irino, T; Tada, R

Citation styles

Publication date

Altman and King (2007): Dataverse

- Sidney Verba. 1998.
  NORC [Producer];

Lawrence et al. (2008): BADC

- Iwi, A. and B. N. Lawrence (2004).

Green (2010): OECD

- OECD (2009),
  (Accessed on 14 September 2009)

Starr and Gastl (2011): DataCite

- Irino, T; Tada, R (2009):

Citation styles

Title

Altman and King (2007): Dataverse

- Sidney Verba. 1998. "U.S. and Russian Social and Political Participation Data," NORC [Producer];

Lawrence et al. (2008): BADC

- Iwi, A. and B. N. Lawrence (2004). A 500 year control run of HadCM3.

Green (2010): OECD

- OECD (2009), "Key short-term indicators", Main Economic Indicators (Accessed on 14 September 2009)

Starr and Gastl (2011): DataCite

- Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797.

Citation styles

Version

Altman and King (2007): Dataverse

- Sidney Verba. 1998. "U.S. and Russian Social and Political Participation Data," NORC [Producer];

Lawrence et al. (2008): BADC

- Iwi, A. and B. N. Lawrence (2004). A 500 year control run of HadCM3. Version 1.

Green (2010): OECD

- OECD (2009), "Key short-term indicators", Main Economic Indicators (Accessed on 14 September 2009)

Starr and Gastl (2011): DataCite

- Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. V.2.

Citation styles

Feature

Altman and King (2007): Dataverse

- Sidney Verba. 1998. "U.S. and Russian Social and Political Participation Data," NORC [Producer];

Lawrence et al. (2008): BADC

- Iwi, A. and B. N. Lawrence (2004). A 500 year control run of HadCM3. [GridSeries, http://ndg.nerc.ac.uk/csml2/GridSeries] Version 1.

Green (2010): OECD

- OECD (2009), "Key short-term indicators", Main Economic Indicators (Accessed on 14 September 2009)

Starr and Gastl (2011): DataCite

- Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. V.2.

Citation styles

Resource type

Altman and King (2007): Dataverse

- Sidney Verba. 1998. "U.S. and Russian Social and Political Participation Data," NORC [Producer]; data set [Type (DC)]

Lawrence et al. (2008): BADC

- Iwi, A. and B. N. Lawrence (2004). A 500 year control run of HadCM3. [GridSeries, http://ndg.nerc.ac.uk/csml2/GridSeries] Version 1.

Green (2010): OECD

- OECD (2009), "Key short-term indicators", Main Economic Indicators (database). (Accessed on 14 September 2009)

Starr and Gastl (2011): DataCite

- Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. V.2. Dataset.

Citation styles

Publisher

Altman and King (2007): Dataverse

- Sidney Verba. 1998. "U.S. and Russian Social and Political Participation Data," NORC [Producer]; data set [Type (DC)] ICPSR [Distributor].

Lawrence et al. (2008): BADC

- Iwi, A. and B. N. Lawrence (2004). A 500 year control run of HadCM3. [GridSeries, http://ndg.nerc.ac.uk/csml2/GridSeries] Version 1. BADC.

Green (2010): OECD

- OECD (2009), "Key short-term indicators", Main Economic Indicators (database). (Accessed on 14 September 2009)

Starr and Gastl (2011): DataCite

- Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. V.2. Geological Institute, University of Tokyo. Dataset.

Citation styles

Identifier

Altman and King (2007): Dataverse

- Sidney Verba. 1998. "U.S. and Russian Social and Political Participation Data," hdl:1902.4/00754 NORC [Producer]; data set [Type (DC)] ICPSR [Distributor].

Lawrence et al. (2008): BADC

- Iwi, A. and B. N. Lawrence (2004). A 500 year control run of HadCM3. [GridSeries, http://ndg.nerc.ac.uk/csml2/GridSeries] Version 1. BADC. urn:badc.nerc.ac.uk_coapec500yr

Green (2010): OECD

- OECD (2009), "Key short-term indicators", Main Economic Indicators (database). doi: 10.1787/data-00039-en (Accessed on 14 September 2009)

Starr and Gastl (2011): DataCite

- Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. V.2. Geological Institute, University of Tokyo. Dataset. doi:10.1594/PANGAEA.726855.

Citation styles

Location

Altman and King (2007): Dataverse

- Sidney Verba. 1998. "U.S. and Russian Social and Political Participation Data," hdl:1902.4/00754 NORC [Producer]; data set [Type (DC)] ICPSR [Distributor].

Lawrence et al. (2008): BADC

- Iwi, A. and B. N. Lawrence (2004). A 500 year control run of HadCM3. [GridSeries, http://ndg.nerc.ac.uk/csml2/GridSeries] Version 1. BADC. urn:badc.nerc.ac.uk_coapec500yr [Available from http://badc.nerc.ac.uk/data/coapec500yr].

Green (2010): OECD

- OECD (2009), "Key short-term indicators", Main Economic Indicators (database). doi: 10.1787/data-00039-en http://dx.doi.org/10.1787/data-00039-en (Accessed on 14 September 2009)

Starr and Gastl (2011): DataCite

- Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. V.2. Geological Institute, University of Tokyo. Dataset. doi:10.1594/PANGAEA.726855. http://dx.doi.org/10.1594/PANGAEA.726855

Citation styles

Unique Numeric Fingerprint

Altman and King (2007): Dataverse

- Sidney Verba. 1998. "U.S. and Russian Social and Political Participation Data," hdl:1902.4/00754 `UNF:3:ZNQRI14053UZq389x0Bffg?==` NORC [Producer]; data set [Type (DC)] ICPSR [Distributor].

Lawrence et al. (2008): BADC

- Iwi, A. and B. N. Lawrence (2004). A 500 year control run of HadCM3. [GridSeries, http://ndg.nerc.ac.uk/csml2/GridSeries] Version 1. BADC. urn:badc.nerc.ac.uk_coapec500yr [Available from http://badc.nerc.ac.uk/data/coapec500yr].

Green (2010): OECD

- OECD (2009), "Key short-term indicators", Main Economic Indicators (database). doi: 10.1787/data-00039-en http://dx.doi.org/10.1787/data-00039-en (Accessed on 14 September 2009)

Starr and Gastl (2011): DataCite

- Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. V.2. Geological Institute, University of Tokyo. Dataset. doi:10.1594/PANGAEA.726855. http://dx.doi.org/10.1594/PANGAEA.726855

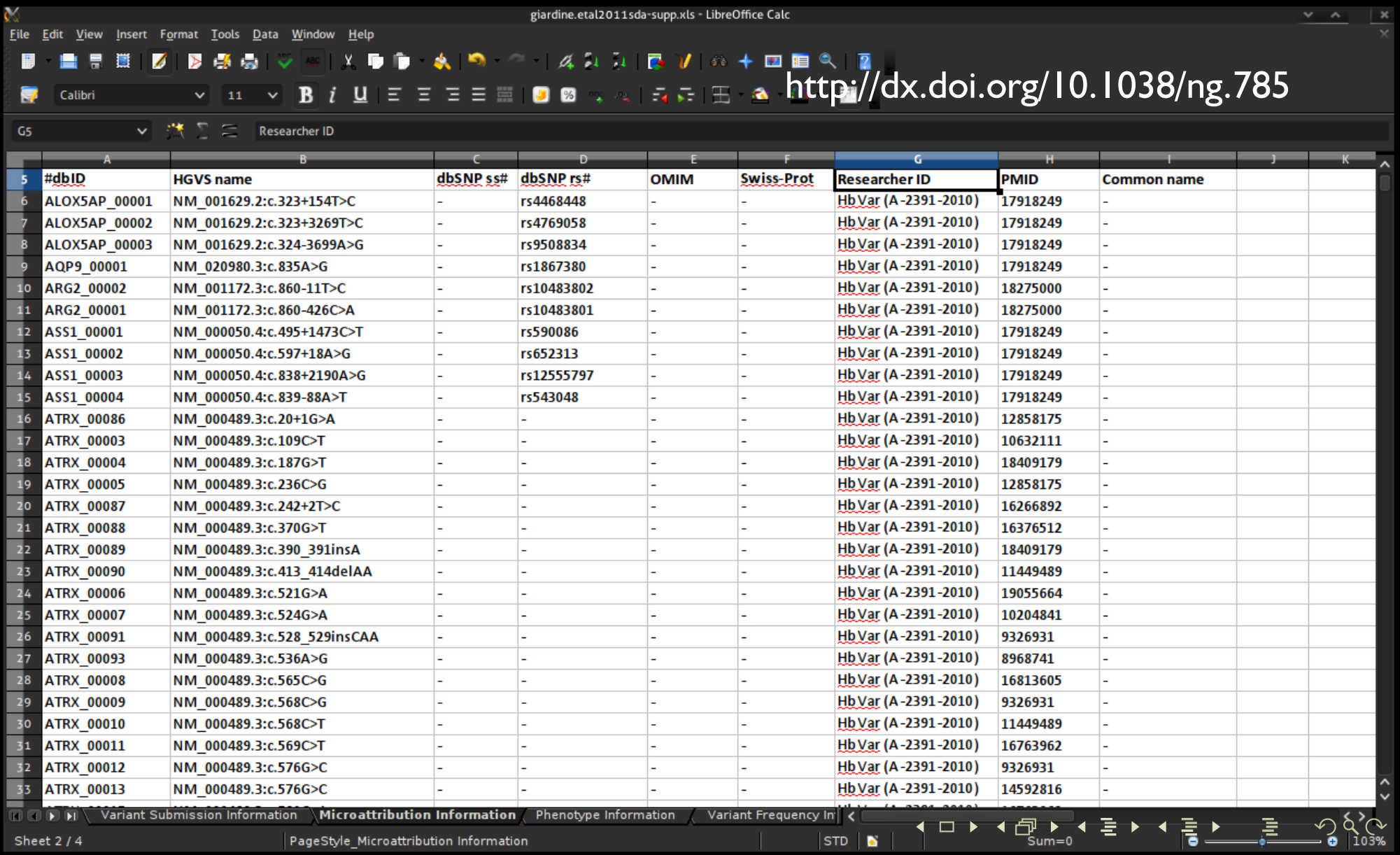
Key citation elements

- Author
- Publication date
- Title
- Location

Key citation elements

- Author
- Publication date
- Title
- Location (= identifier)

Attributing datasets to many contributors



	A	B	C	D	E	F	G	H	I	J	K
5	#dbID	HGVS name	dbSNP ss#	dbSNP rs#	OMIM	Swiss-Prot	Researcher ID	PMID	Common name		
6	ALOX5AP_00001	NM_001629.2:c.323+154T>C	-	rs4468448	-	-	Hb Var (A-2391-2010)	17918249	-		
7	ALOX5AP_00002	NM_001629.2:c.323+3269T>C	-	rs4769058	-	-	Hb Var (A-2391-2010)	17918249	-		
8	ALOX5AP_00003	NM_001629.2:c.324-3699A>G	-	rs9508834	-	-	Hb Var (A-2391-2010)	17918249	-		
9	AQP9_00001	NM_020980.3:c.835A>G	-	rs1867380	-	-	Hb Var (A-2391-2010)	17918249	-		
10	ARG2_00002	NM_001172.3:c.860-11T>C	-	rs10483802	-	-	Hb Var (A-2391-2010)	18275000	-		
11	ARG2_00001	NM_001172.3:c.860-426C>A	-	rs10483801	-	-	Hb Var (A-2391-2010)	18275000	-		
12	ASS1_00001	NM_000050.4:c.495+1473C>T	-	rs590086	-	-	Hb Var (A-2391-2010)	17918249	-		
13	ASS1_00002	NM_000050.4:c.597+18A>G	-	rs652313	-	-	Hb Var (A-2391-2010)	17918249	-		
14	ASS1_00003	NM_000050.4:c.838+2190A>G	-	rs12555797	-	-	Hb Var (A-2391-2010)	17918249	-		
15	ASS1_00004	NM_000050.4:c.839-88A>T	-	rs543048	-	-	Hb Var (A-2391-2010)	17918249	-		
16	ATRX_00086	NM_000489.3:c.20+1G>A	-	-	-	-	Hb Var (A-2391-2010)	12858175	-		
17	ATRX_00003	NM_000489.3:c.109C>T	-	-	-	-	Hb Var (A-2391-2010)	10632111	-		
18	ATRX_00004	NM_000489.3:c.187G>T	-	-	-	-	Hb Var (A-2391-2010)	18409179	-		
19	ATRX_00005	NM_000489.3:c.236C>G	-	-	-	-	Hb Var (A-2391-2010)	12858175	-		
20	ATRX_00087	NM_000489.3:c.242+2T>C	-	-	-	-	Hb Var (A-2391-2010)	16266892	-		
21	ATRX_00088	NM_000489.3:c.370G>T	-	-	-	-	Hb Var (A-2391-2010)	16376512	-		
22	ATRX_00089	NM_000489.3:c.390_391insA	-	-	-	-	Hb Var (A-2391-2010)	18409179	-		
23	ATRX_00090	NM_000489.3:c.413_414delAA	-	-	-	-	Hb Var (A-2391-2010)	11449489	-		
24	ATRX_00006	NM_000489.3:c.521G>A	-	-	-	-	Hb Var (A-2391-2010)	19055664	-		
25	ATRX_00007	NM_000489.3:c.524G>A	-	-	-	-	Hb Var (A-2391-2010)	10204841	-		
26	ATRX_00091	NM_000489.3:c.528_529insCAA	-	-	-	-	Hb Var (A-2391-2010)	9326931	-		
27	ATRX_00093	NM_000489.3:c.536A>G	-	-	-	-	Hb Var (A-2391-2010)	8968741	-		
28	ATRX_00008	NM_000489.3:c.565C>G	-	-	-	-	Hb Var (A-2391-2010)	16813605	-		
29	ATRX_00009	NM_000489.3:c.568C>G	-	-	-	-	Hb Var (A-2391-2010)	9326931	-		
30	ATRX_00010	NM_000489.3:c.568C>T	-	-	-	-	Hb Var (A-2391-2010)	11449489	-		
31	ATRX_00011	NM_000489.3:c.569C>T	-	-	-	-	Hb Var (A-2391-2010)	16763962	-		
32	ATRX_00012	NM_000489.3:c.576G>C	-	-	-	-	Hb Var (A-2391-2010)	9326931	-		
33	ATRX_00013	NM_000489.3:c.576G>C	-	-	-	-	Hb Var (A-2391-2010)	14592816	-		

Granularity

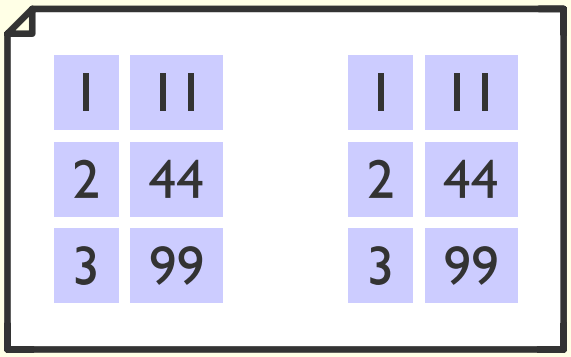
- Data points

Granularity

1	11
2	44
3	99

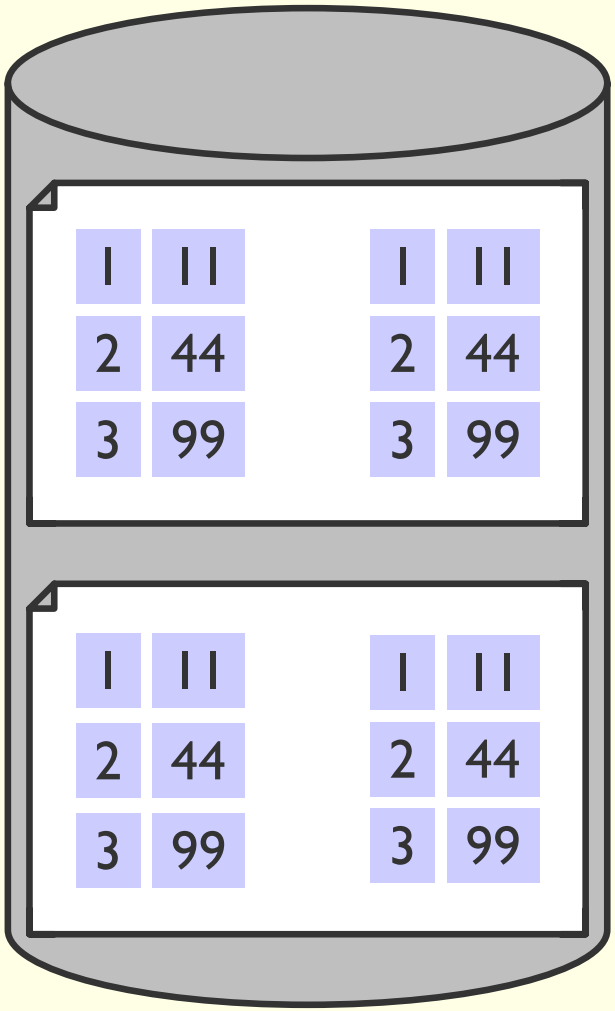
- Data points
- Data tables

Granularity



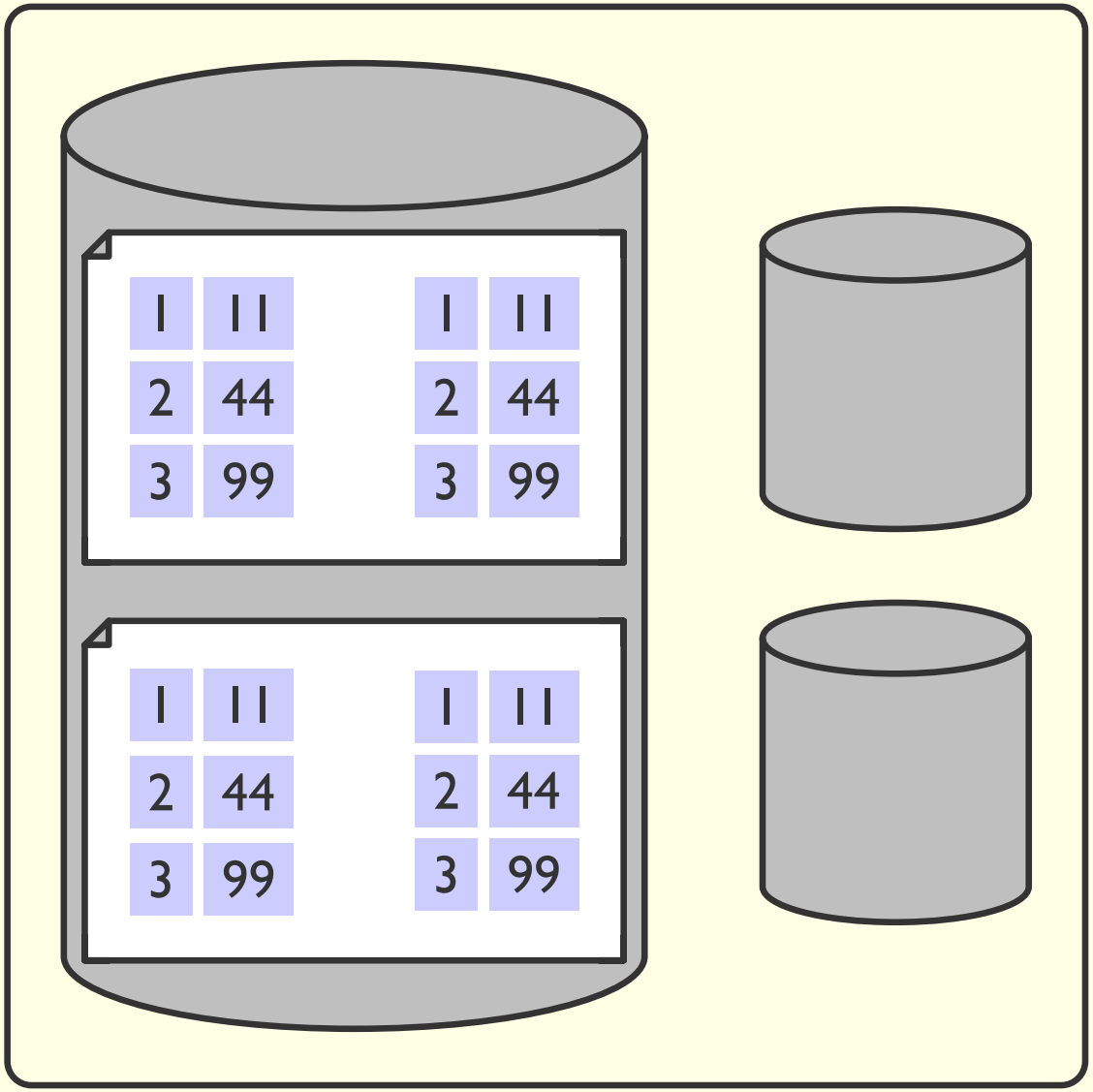
- Data points
- Data tables
- Data files

Granularity



- Data points
- Data tables
- Data files
- Datasets

Granularity



- Data points
- Data tables
- Data files
- Datasets
- Data collections

Granularity

- Cite datasets at the finest level that is appropriate and for which an identifier is provided.
- If that is not fine enough, provide details of the subset of data you are using at the point in the text where you make the citation.

Placement of data citations

- Special data resources section?
- Acknowledgements?
- Accession codes?
- Reference list?

Placement of data citations

- Special data resources section?
- Acknowledgements?
- Accession codes?
- Reference list?
- Alongside or independent of a reference to the related article?

Placement of data citations

- Include the citation in the reference list.
- When your data collection paper is published, notify the repository holding the dataset.
- When you publish a paper in which you reuse a prior dataset, notify the repository holding that dataset.

Dynamic datasets

Two types:

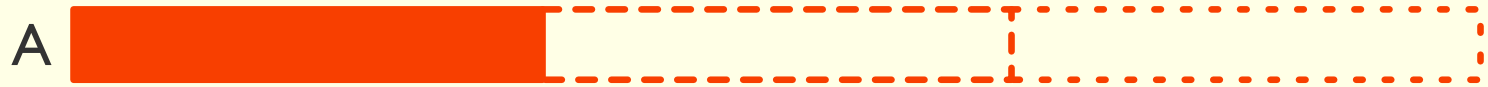
- Revised datasets

- Expanding datasets

Dynamic datasets

Three strategies:

1. Differentiate versions by access date rather than ID



2. Take time slices

3. Take snapshots

Guidance for researchers publishing a paper

- Deposit any data you have collected and used as evidence.
- Ask for a persistent ID/URL for your deposited data.
- When your data collection paper is published, notify the repository holding the dataset.

Guidance for researchers citing a prior dataset

- Use the data citation style required by the editor/publisher.
- If no style is specified, use a standard data citation style, adapted to match the style for textual publications.
- Default to writing IDs in the form of URLs if possible.
- Include the citation in the reference list.
- Cite datasets at the finest level that is appropriate and for which an identifier is provided.
- If that is not fine enough, provide details of the subset of data you are using at the point in the text where you make the citation.
- Cite the exact version of the dataset you need.
- When your paper is published, notify the repository holding the dataset you used.

Guidance for data repositories

- Provide persistent IDs for the datasets you host.
  - The ID should remain unique.
  - The ID should always point to the same version.
  - The ID should resolve to a URL.
  - The URL should locate the dataset's landing page.
- The explanatory metadata should not change for a dataset with a persistent ID.
- IDs should only be assigned once no further changes are expected.
- With dynamic datasets, provide IDs for snapshots or time slices.
- Provide sample citations on dataset landing pages.
- Link from landing pages to publications citing the dataset.

Thank you for your attention

DCC Website: http://www.dcc.ac.uk/
Alex Ball: http://www.ukoln.ac.uk/ukoln/staff/a.ball/