Joint Declaration of Data Citation Principles: Implementation and Compliance in the Dataverse Repository

Mercè Crosas, Ph.D.

Twitter: @mercecrosas

Director of Data Science

Institute for Quantitative Social Science, Harvard University

NISO Virtual Conference, April 23, 2014

A brief History of Data Citation

Standards in Scholarly Citation:
author/creator, title, dates,
publisher or distributor of the work

1906
Chicago Manual of Style

1960
First scientific digital data archives

1977 – 1998
ASBR ("Data File" type)
MARC (machine readable catalog)

1999-2014
Data Repositories (NESSTAR, Dataverse, Dryad, Figshare)
DOI services(DataCite)

Altman M., Crosas M., 2014, "The Evolution of Data Citation: From Principles to Implementation" IASSIST Quarterly, *In Press*

The Making of the Principles

- Decades of research and practices in data citation
- Consolidated to a single set of Principles
- By a synthesis group representing 25+ organizations
- Driven by the premise that:

"sound, reproducible scholarship rests upon a foundation of robust, accessible data"

and

"data should be considered legitimate, citable products of research"

Joint Declaration of Data Citation Principles

1. **Importance**
2. **Credit and Attribution**
3. **Evidence**
4. **Unique Identification**
5. **Access**
6. **Persistence**
7. **Specificity and Verifiability**
8. **Interoperability and flexibility**

Full Principles: https://www.force11.org/datacitation

Endorsement: https://www.force11.org/datacitation/endorsements

Joint Declaration of Data Citation Principles

1. Importance

Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.

Joint Declaration of Data Citation Principles

2. Credit and Attribution

Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.

Joint Declaration of Data Citation Principles

3. Evidence

In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.

Joint Declaration of Data Citation Principles

4. Unique Identification

A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.

Joint Declaration of Data Citation Principles

5. Access

Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.

Joint Declaration of Data Citation Principles

6. Persistence

Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe.

Joint Declaration of Data Citation Principles

7. Specificity and Verifiability

Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific time slice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.

Joint Declaration of Data Citation Principles

8. Interoperability and flexibility

Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.

About Dataverse

- A software framework to **build data repositories**.
- Provides a **preservation and archival** infrastructure,
  … while researchers **share, keep control of and get recognition for their data** through a web interface.
- Harvard Dataverse is open to all researchers and disciplines.
- It contains more than 50,000 data sets.
- Other large Dataverse instances throughout the world:
  ODUM at UNC, Dutch Universities, Scholar Portal, Fudan University.
- Dataverse 4.0 (June 2014) brings an entirely new UI and improved data publishing workflows.

Data Citation Implementation in Dataverse

The Dataverse generates a Data Citation for each deposited data set compliant with the Principles:

Authors, Year, Dataset Title, DOI, Data Repository, UNF, version

Example:

Logan Vidal, 2013, "ANES data coding ", http://dx.doi.org/10.7910/DVN/23274 Harvard Dataverse, UNF:5:0fdUNzmCsyeqrVKtgUG74A==, V8

Compliant with Principle 2

Authors, Year, Dataset Title, DOI, Data Repository, UNF, version

Principle 2:

Credit and Attribution: …facilitate giving scholarly credit and … attribution to all contributors to the data, …

Compliant with Principles 4, 5, 6

Resolves to landing page with access to metadata, docs, code and data

↓

Authors, Year, Dataset Title, DOI, Data Repository, UNF, version

↑

Principles 4, 5, 6

Unique Identification: …machine actionable, globally unique, and widely used by a community …
Access: … access to the data themselves and to such associated metadata, documentation, code, and other materials …
Persistence: … even beyond the lifespan of the data they describe.

Landing Page Example: Metadata

ANES DATA CODING

doi:10.7910/DVN/23274UNF:5:0fdUNzmCsyeqrVKtgUG74A==
Version: 8 – Released: Tue Apr 22 12:38:15 EDT 2014

CATALOGING INFORMATION | Data & Analysis | Comments (0) | Versions

If you use these data, please add the following citation to your scholarly references. Why cite?

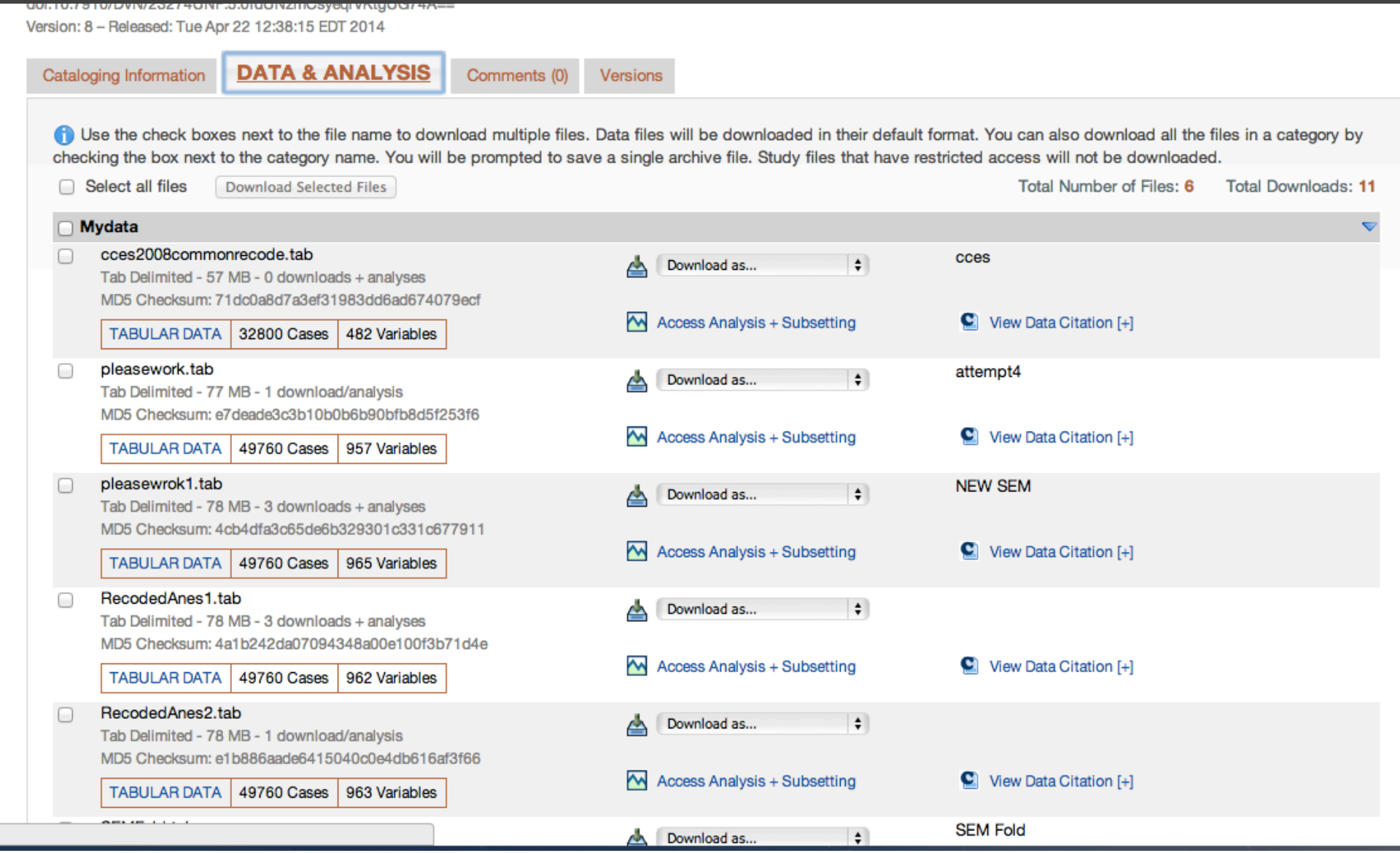
Data Citation	Logan Vidal, 2013, "ANES data coding ", http://dx.doi.org/10.7910/DVN/23274 UNF:5:0fdUNzmCsyeqrVKtgUG74A== Harvard Dataverse Network [Distributor] V8 [Version]
Citation Format	Print
Publications	ANES 1952-2008

Data Citation Details

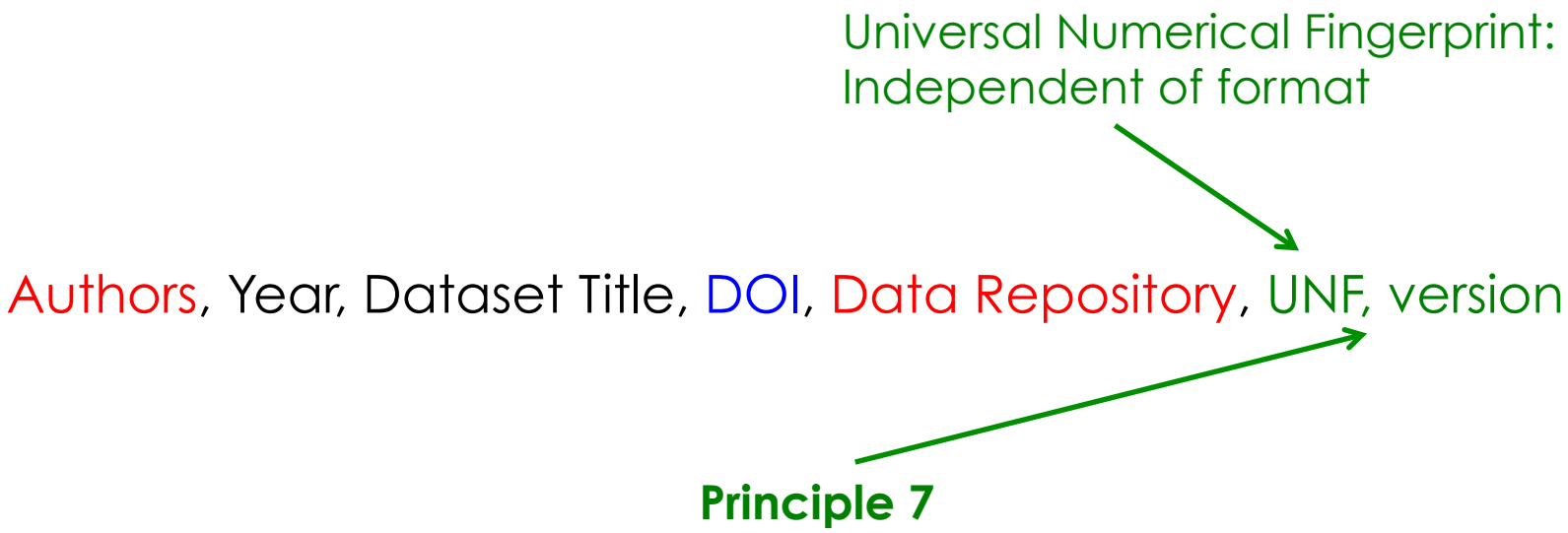
Title	ANES data coding
Study Global ID	doi:10.7910/DVN/23274
Authors	Logan Vidal (University of Wisconsin)
Producer	ANES, Michigan
Production Date	November 07, 2013
Funding Agency	University of Michigan
Distributor	Harvard Dataverse Network (POWERED BY THE Dataverse Network PROJECT)
Contact	Logan Vidal, lvidal@wisc.edu
Distribution Date	2013
Deposit Date	November 08, 2013
Original Dataverse	LVidal Data Dataverse

Description and Scope

Landing Page Example: Data, Code & Docs



Compliant with Principle 7



Specificity and Verifiability: …provenance and fixity sufficient to facilitate verifying that the specific time slice, version and/or granular portion of data …

Example of version History

ANES DATA CODING

doi:10.7910/DVN/23274UNF:5:0fdUNzmCsyeqrVKtgUG74A==
Version: 8 – Released: Tue Apr 22 12:38:15 EDT 2014

Cataloging Information | Data & Analysis | Comments (0) | **VERSIONS**

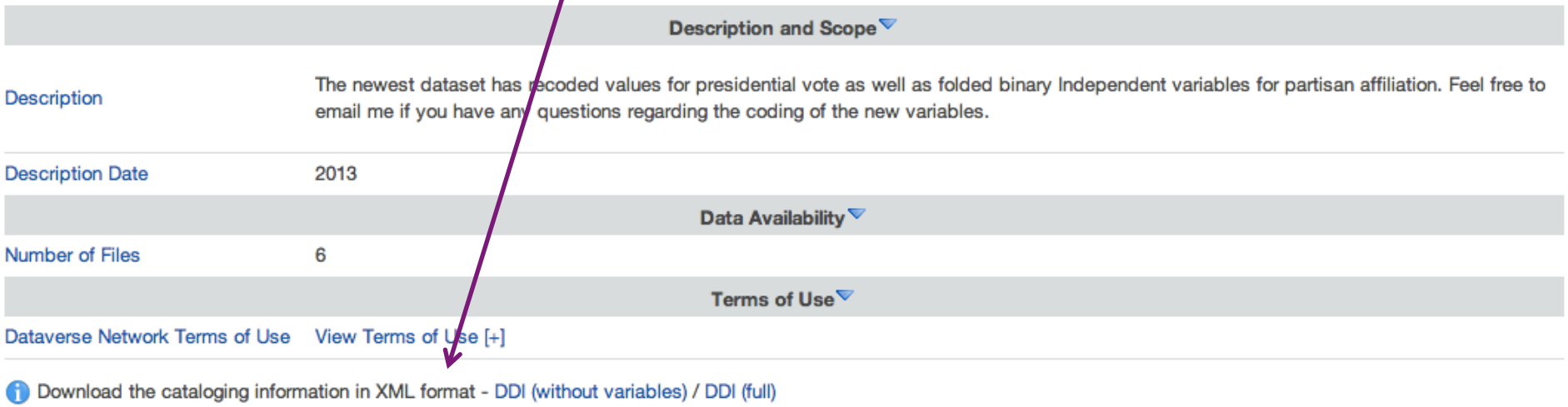
Version History

Difference	Version	Status	Comments	Released	Contributors
☐	8	Released		Tue Apr 22 12:38:15 EDT 2014	Logan Vidal
☐	7	Archived	Sem2	Fri Feb 14 15:18:11 EST 2014	Logan Vidal
☐	6	Archived	New SEM	Tue Feb 11 14:18:45 EST 2014	Logan Vidal
☐	5	Archived		Tue Feb 11 12:38:14 EST 2014	Logan Vidal
☐	4	Archived	Update	Fri Nov 15 12:57:51 EST 2013	Logan Vidal
☐	3	Archived		Fri Nov 15 11:34:00 EST 2013	Logan Vidal
☐	2	Archived		Fri Nov 15 11:07:57 EST 2013	Logan Vidal
☐	1	Archived	This my data with a simple coding of presidential vote as turnout.	Fri Nov 08 13:09:58 EST 2013	Logan Vidal

Compliant with Principle 8

Principle 8: Interoperability and flexibility:

Dataverse exports all citation metadata in XML, JSON formats



Description and Scope	
Description	The newest dataset has recoded values for presidential vote as well as folded binary Independent variables for partisan affiliation. Feel free to email me if you have any questions regarding the coding of the new variables.
Description Date	2013
Data Availability	
Number of Files	6
Terms of Use	
Dataverse Network Terms of Use	View Terms of Use [+]

Download the cataloging information in XML format - DDI (without variables) / DDI (full)

Implementation Suggestions for Publishers

- Upgrade data citation to references section **[Principle 1: Importance]**
- In article, cite data by claim **[Principle 3: Evidence]**
- Provide guidelines for authors based on Principles, but customized to each journal **[Principle 8: Interoperability and Flexibility]**
- Interoperate with, or recommend, trusted Data Repositories compliant with the Principles
- Build tools to access machine-readable metadata from datasets

Want to be involved?

Join the **Data Citation Implementation group**:

https://www.force11.org/datacitationimplementation

Remaining Challenges

- Challenges of Provenance: what is the chain of ownership and transformations to the data?
- Challenges of Identity: what should be cited? at what level of granularity and versioning for large, dynamic datasets?
- Challenges of Attribution: How do you support attribution for hundreds/thousands contributors?

Altman M., Crosas M., 2014, "The Evolution of Data Citation: From Principles to Implementation" IASSIST Quarterly, *In Press*