Identity and Identifiers

Contents

V1: Why is identification important?

V2: *What* are we identifying?

V3: *How* do we identify?

V4: A practical example: XML *canonicalization*

V1: Why is identification important?

(A short introduction, referring to earlier discussions)

Identification Problems in Data Curation (again)

Identification problems

Archiving: Is this dataset already in the archive?

Preservation: Was the information preserved in the new file format?

Security: Has this dataset been tampered with?

Authentication: Is this the data we think it is?

Reproducibility: Does this XML file have the same information as that JSON file?

Provenance: Were these datasets derived from the same data?

Conversions: Does the converted file have the same data as the original?

Identifiers – what are they for?

Identifiers. . .

Enable *discovery and reuse* of relevant data sets

Support management of data sets
including *version control, correction, conversion*, etc.

Support workflow and provenance tracking

Promote transparency and reproducibility

Give credit to data produces

And more.

Identifiers, how they are used

An *identifier is often the one word answer to questions like:*

Which data set was the input for your analysis?

Which data set was the output of your analysis?

Is there a data set that has the temperatures by zip code?

Is this the JSON version of that XML data set??

Which version was corrected and anonymized?

Sure, you could say answer those questions by things like saying:

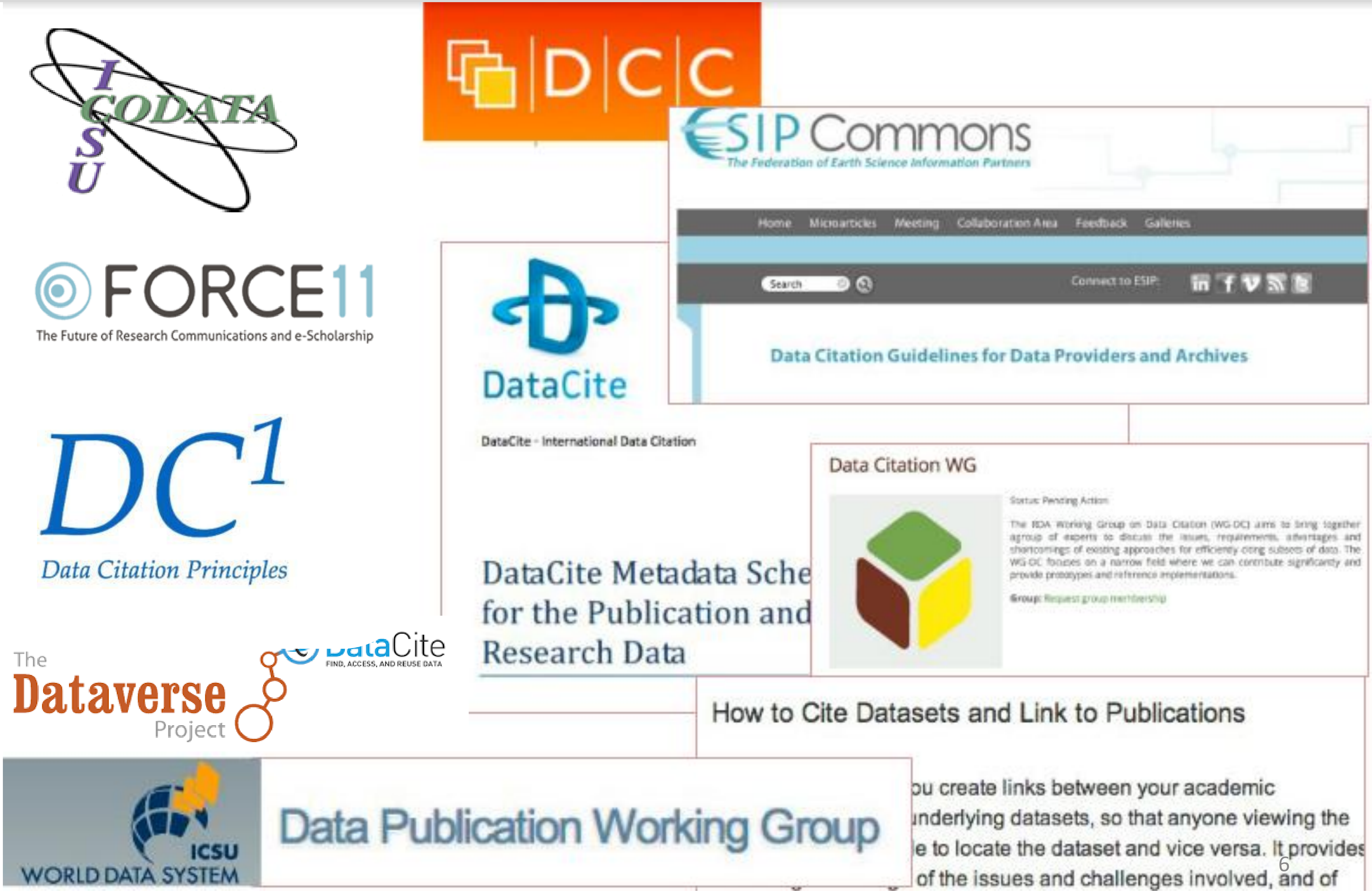
It’s this one, I think
The one on the red USB drive

Bill has it, or maybe Daphne

NewMikeVersionSecondFinalV3.zip

But in the long run answers of this sort are neither reliable nor efficient.

Surf these up!



I CODATA S U

DCC

ESIP Commons

The Federation of Earth Science Information Partners

Home Microarticles Meeting Collaboration Area Feedback Galleries

Search

Connect to ESIP:

Data Citation Guidelines for Data Providers and Archives

FORCE11

The Future of Research Communications and e-Scholarship

DataCite

DataCite - International Data Citation

DC1

Data Citation Principles

Data Citation WG

Status: Pending Action

The RDA Working Group on Data Citation (WG-DC) aims to bring together agroup of experts to discuss the issues, requirements, advantages and shortcomings of existing approaches for efficiently citing subsets of data. The WG-DC focuses on a narrow field where we can contribute significantly and provide prototypes and reference implementations.

Group: Request group membership

DataCite Metadata Sche
for the Publication and
Research Data

DataCite

FIND, ACCESS, AND REUSE DATA

The Dataverse Project

How to Cite Datasets and Link to Publications

ou create links between your academic
nderlying datasets, so that anyone viewing the
e to locate the dataset and vice versa. It provides
of the issues and challenges involved, and of

ICSU WORLD DATA SYSTEM

Data Publication Working Group

Readings

=> Persistent Identifiers, Fixity and Checksums, in *The Digital Preservation Handbook*, Digital Preservation Coalition, 2017.

=> On the utility of identification schemes for digital earth science data: An assessment and recommendations. Duerr, R. E., Downs, R. R., Tilmes, C., Barkstrom, B., Lenhardt, W. C., Glassy, J., Bermudez, L. E., & Slaughter, P. (2011). *Earth Science Informatics*.

Clifford A. Lynch, "Canonicalization: A Fundamental Tool To Facilitate Preservation and Management of Digital Information," *D-lib Magazine*, 5:9 (September 1999).

Canonical XML. Version 1.0. W3C Recommendation, John Boyer, March 2001. Latest version: http://www.w3.org/TR/xml-c14n.