ILLINOIS

State of the Art in Data Citation

Ruth Duerr

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois

Ronin Institute

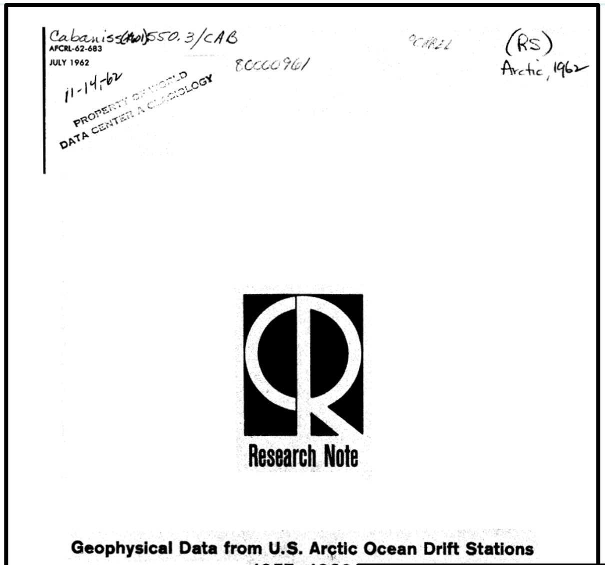
Overview

- Citing data in publications is a re-emerging practice that:
  - Encourages reproducibility of results
  - Promotes transparency of the research process, improving research standards and ensuring accountability
  - Provides credit to data producers and data publishers
  - Assists data repository and service providers in tracking usage to develop appropriate support mechanisms

- Assigning persistent identifiers is necessary to maintain long-term access to the cited data

Data was in the literature!

In Books and Technical Reports



Cabaniss(#01)550.3/CAB
AFCRL-62-683
JULY 1962
11-14-62

8CCC0961

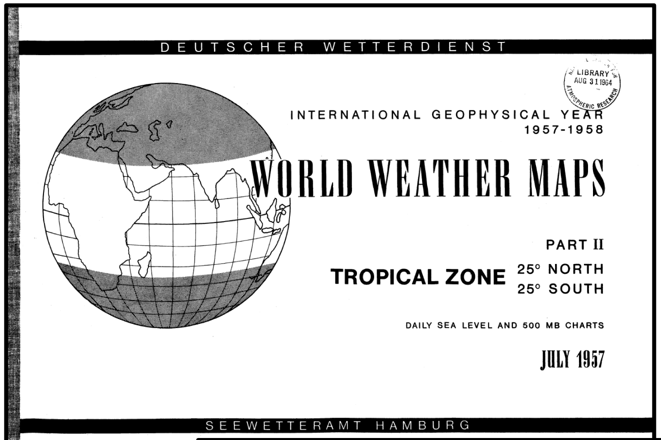
9CHREL

(RS)
Arctic, 1962

Research Note

Geophysical Data from U.S. Arctic Ocean Drift Stations
1957—1960

G. H. CABANISS
EDITOR



DEUTSCHER WETTERDIENST

INTERNATIONAL GEOPHYSICAL YEAR
1957-1958

WORLD WEATHER MAPS

PART II

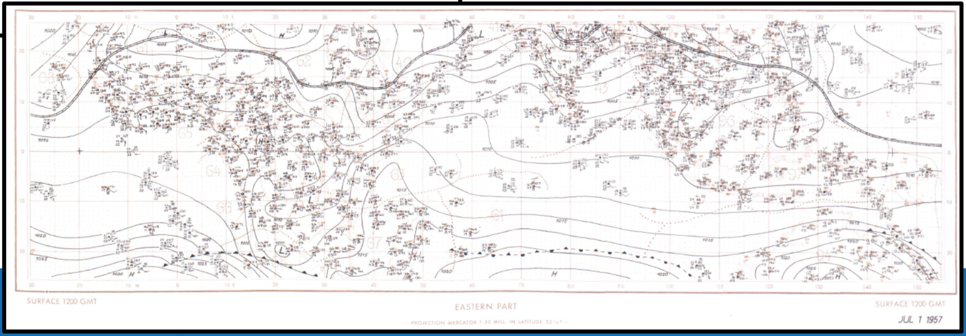
TROPICAL ZONE 25° NORTH
25° SOUTH

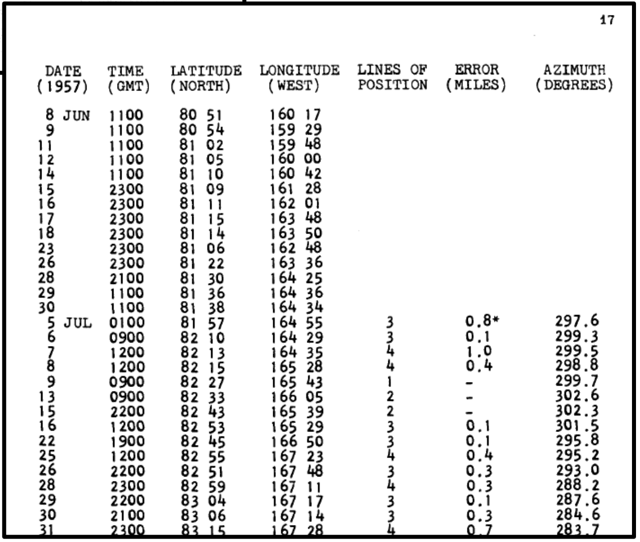
DAILY SEA LEVEL AND 500 MB CHARTS

JULY 1957

SEEWETTERAMT HAMBURG

DATA
90022



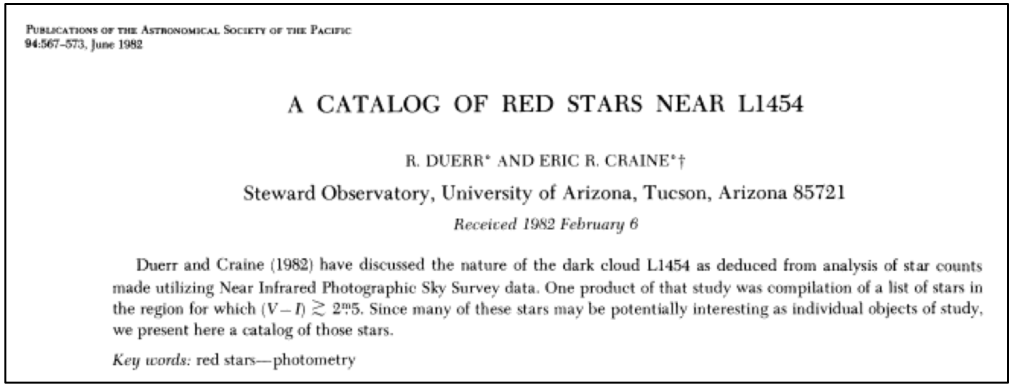


17

DATE (1957)		TIME (GMT)	LATITUDE (NORTH)	LONGITUDE (WEST)	LINES OF POSITION	ERROR (MILES)	AZIMUTH (DEGREES)
8	JUN	1100	80 51	160 17			
9		1100	80 54	159 29			
11		1100	81 02	159 48			
12		1100	81 05	160 00			
14		1100	81 10	160 42			
15		2300	81 09	161 28			
16		2300	81 11	162 01			
17		2300	81 15	163 48			
18		2300	81 14	163 50			
23		2300	81 06	162 48			
26		2300	81 22	163 36			
28		2100	81 30	164 25			
29		1100	81 36	164 36			
30		1100	81 38	164 34			
5	JUL	0100	81 57	164 55	3	0.8*	297.6
6		0900	82 10	164 29	3	0.1	299.3
7		1200	82 13	164 35	4	1.0	299.5
8		1200	82 15	165 28	4	0.4	298.8
9		0900	82 27	165 43	1	-	299.7
13		0900	82 33	166 05	2	-	302.6
15		2200	82 43	165 39	2	-	302.3
16		1200	82 53	165 29	3	0.1	301.5
22		1900	82 45	166 50	3	0.1	295.8
25		1200	82 55	167 23	4	0.4	295.2
26		2200	82 51	167 48	3	0.3	293.0
28		2300	82 59	167 11	4	0.3	288.2
29		2200	83 04	167 17	3	0.1	287.6
30		2100	83 06	167 14	3	0.3	284.6
31		2300	83 15	167 28	4	0.7	283.7

Data was in the literature!

and Journals



PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF THE PACIFIC
94:567–573, June 1982

A CATALOG OF RED STARS NEAR L1454

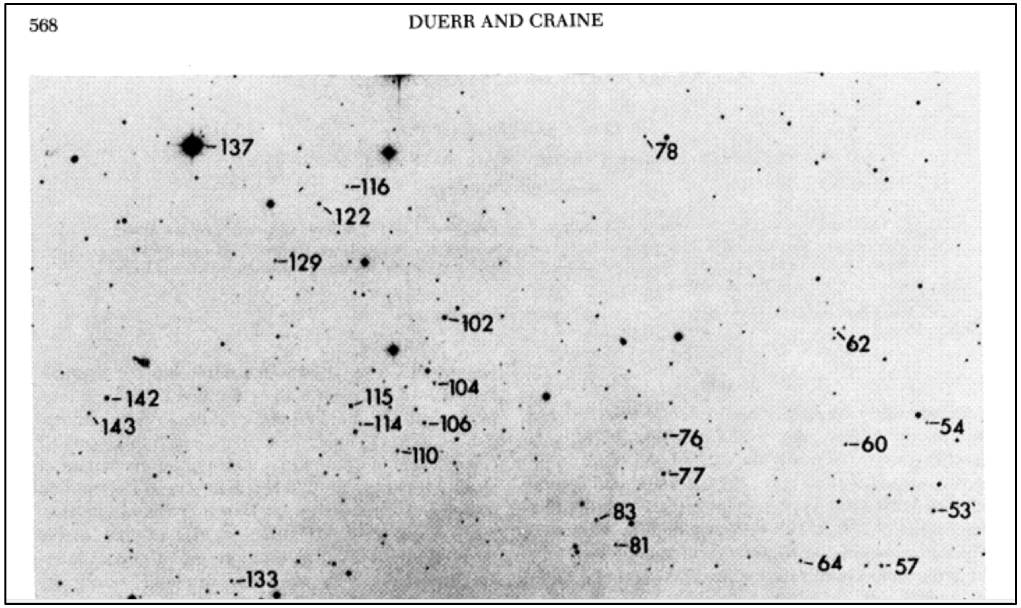
R. DUERR* AND ERIC R. CRAINE*†

Steward Observatory, University of Arizona, Tucson, Arizona 85721

Received 1982 February 6

Duerr and Craine (1982) have discussed the nature of the dark cloud L1454 as deduced from analysis of star counts made utilizing Near Infrared Photographic Sky Survey data. One product of that study was compilation of a list of stars in the region for which $(V-I) \gtrsim 2^{m}5$. Since many of these stars may be potentially interesting as individual objects of study, we present here a catalog of those stars.

Key words: red stars—photometry



568 DUERR AND CRAINE

Losing the data citation tradition

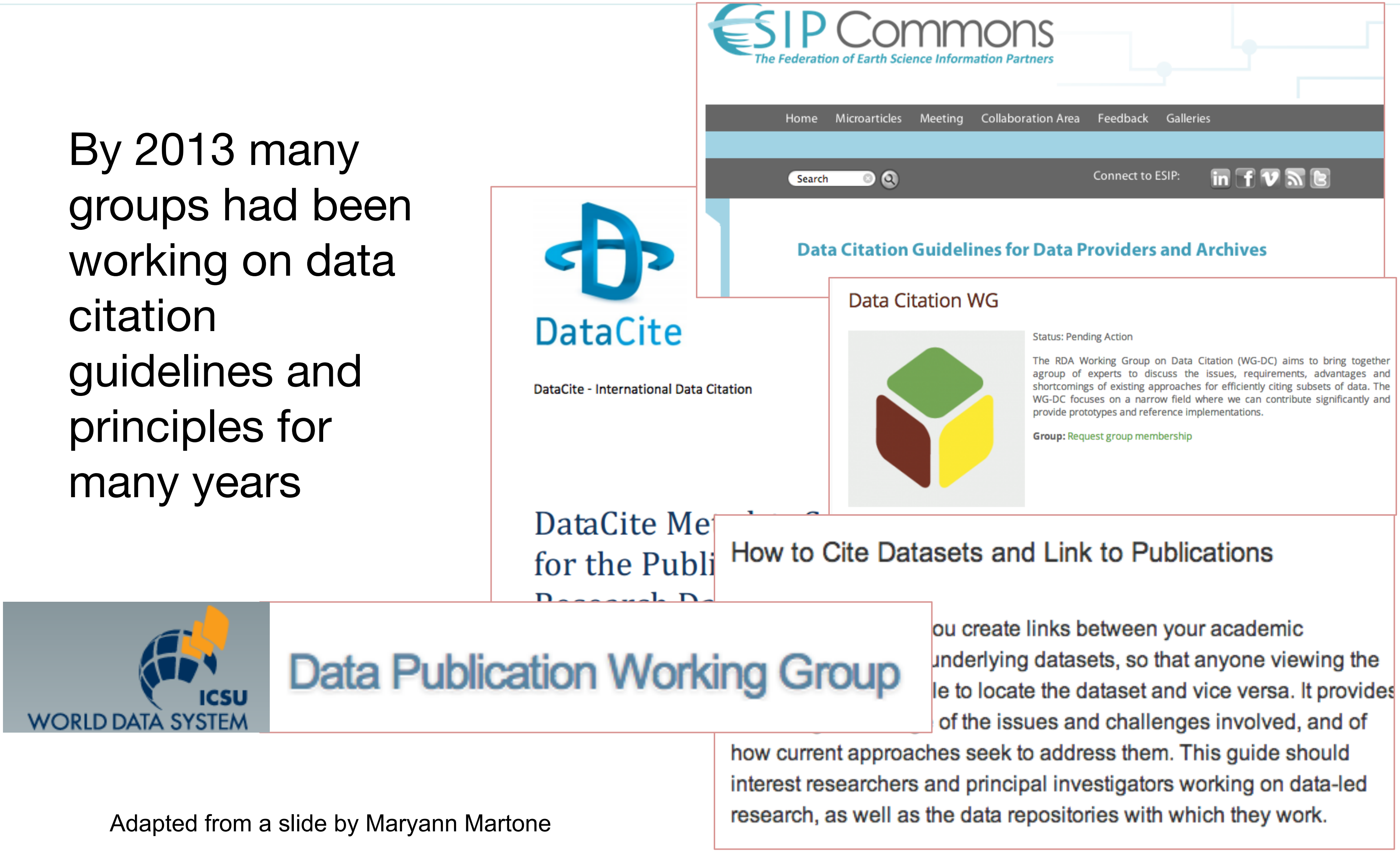
- That started changing with the advent of digital data and media other than paper
  - At first because the publications were still paper
    - Why would you want to make your data less accessible?
    - Now how do you represent a multi-dimensional data set in a two-dimensional medium? What about audio? Videos?
    -
    -
    -
  - Later because often the data was voluminous

Data Repositories started as Data Libraries

- World Data Centers set up during the International Geophysical Year 1957/8
- Social science repositories started up about that same time
- Many repositories transitioned to dealing with digital data in the last half of the twentieth century
- Many have been promoting data citation for decades

Data citation guidelines and principles

By 2013 many groups had been working on data citation guidelines and principles for many years



ESIP Commons

The Federation of Earth Science Information Partners

Data Citation Guidelines for Data Providers and Archives

DataCite

DataCite - International Data Citation

DataCite Me

for the Publi

Research Da

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

How to Cite Datasets and Link to Publications

ou create links between your academic underlying datasets, so that anyone viewing the le to locate the dataset and vice versa. It provides of the issues and challenges involved, and of how current approaches seek to address them. This guide should interest researchers and principal investigators working on data-led research, as well as the data repositories with which they work.

ICSU WORLD DATA SYSTEM

Data Publication Working Group

Adapted from a slide by Maryann Martone

Paul Uhlir "...a plea to come together"

Photo: Flickr

Joint Declaration of Data Citation Principles

- ***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.
- ***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.
- ***Evidence***: In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.
- ***Unique Identification:*** A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.

Data Citation Synthesis Group. (2014). Joint Declaration of Data Citation Principles.
http://force11.org/datacitation

Ronin Institute

Joint Declaration of Data Citation Principles

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

Data Citation Synthesis Group. (2014). Joint Declaration of Data Citation Principles.
http://force11.org/datacitation

Ronin Institute

Data Citation Implementer's Group

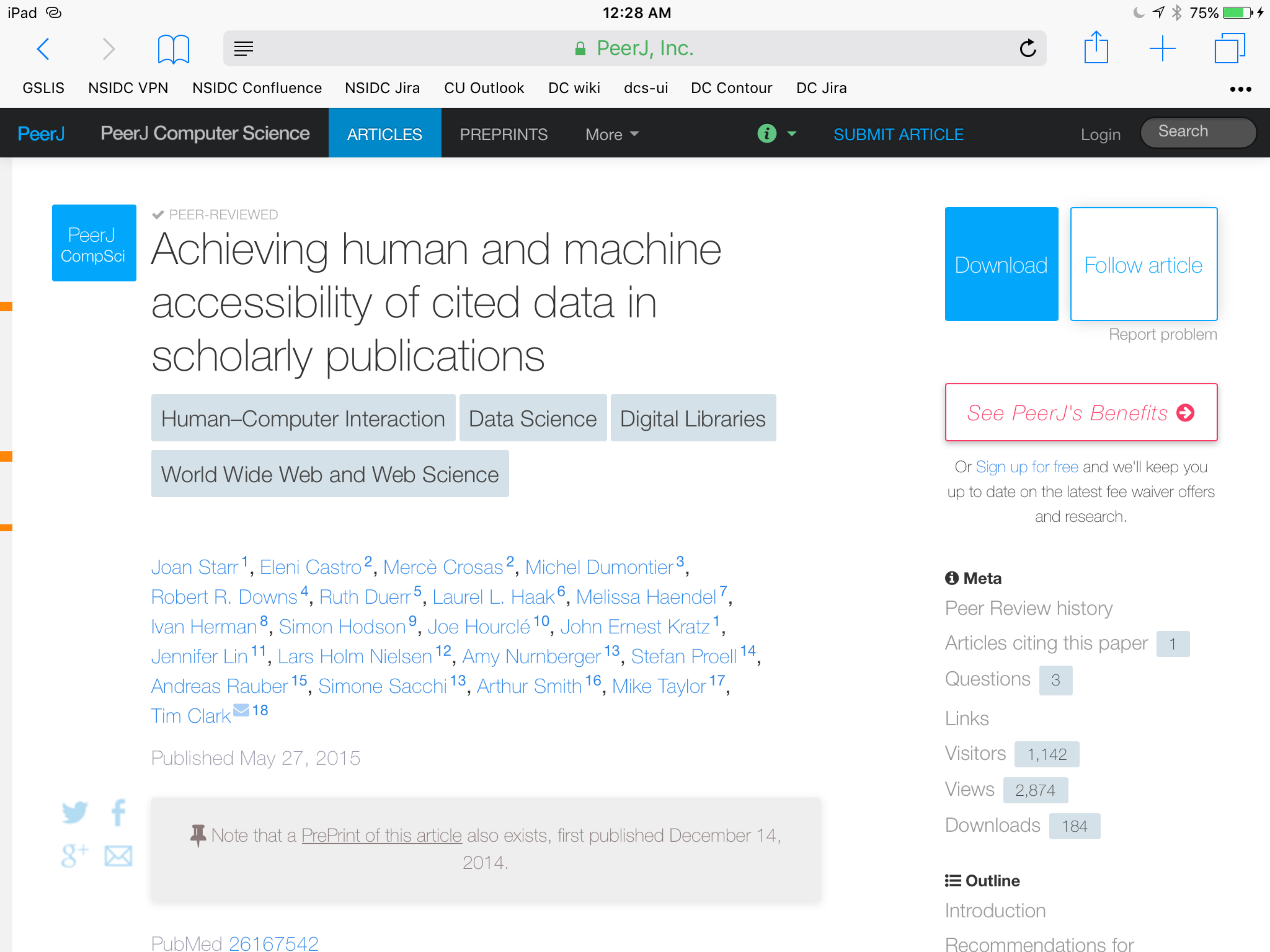
- Work in 4 areas:
  - NISO JATS.
  - Identifiers and associated metadata.
  - Common repository interfaces.
  - Putting together and analyzing some exemplar journal workflows with suggestions on how the editorial process can deal with data citations, to provide context and analysis of commonality for the other tasks.

Implications of NISO-JATS support for data citation

- Enabling the citation of data to be treated with the same “respect” as article citations
- Journals empowered to structure the citation of data in machine-actionable form …
- … ultimately supporting development of new applications and processes
- Agreements on implementation best practice will become important as uptake grows (Data Citation Principles!)

FORCE11 The Future of Research Communications and e-Scholarship

OpenAIRE



✔ PEER-REVIEWED

Achieving human and machine accessibility of cited data in scholarly publications

Human–Computer Interaction | Data Science | Digital Libraries | World Wide Web and Web Science

Joan Starr[1], Eleni Castro[2], Mercè Crosas[2], Michel Dumontier[3], Robert R. Downs[4], Ruth Duerr[5], Laurel L. Haak[6], Melissa Haendel[7], Ivan Herman[8], Simon Hodson[9], Joe Hourclé[10], John Ernest Kratz[1], Jennifer Lin[11], Lars Holm Nielsen[12], Amy Nurnberger[13], Stefan Proell[14], Andreas Rauber[15], Simone Sacchi[13], Arthur Smith[16], Mike Taylor[17], Tim Clark✉[18]

Published May 27, 2015

Note that a PrePrint of this article also exists, first published December 14, 2014.

PubMed 26167542

Or Sign up for free and we'll keep you up to date on the latest fee waiver offers and research.

Meta

Peer Review history

Articles citing this paper 1

Questions 3

Links

Visitors 1,142

Views 2,874

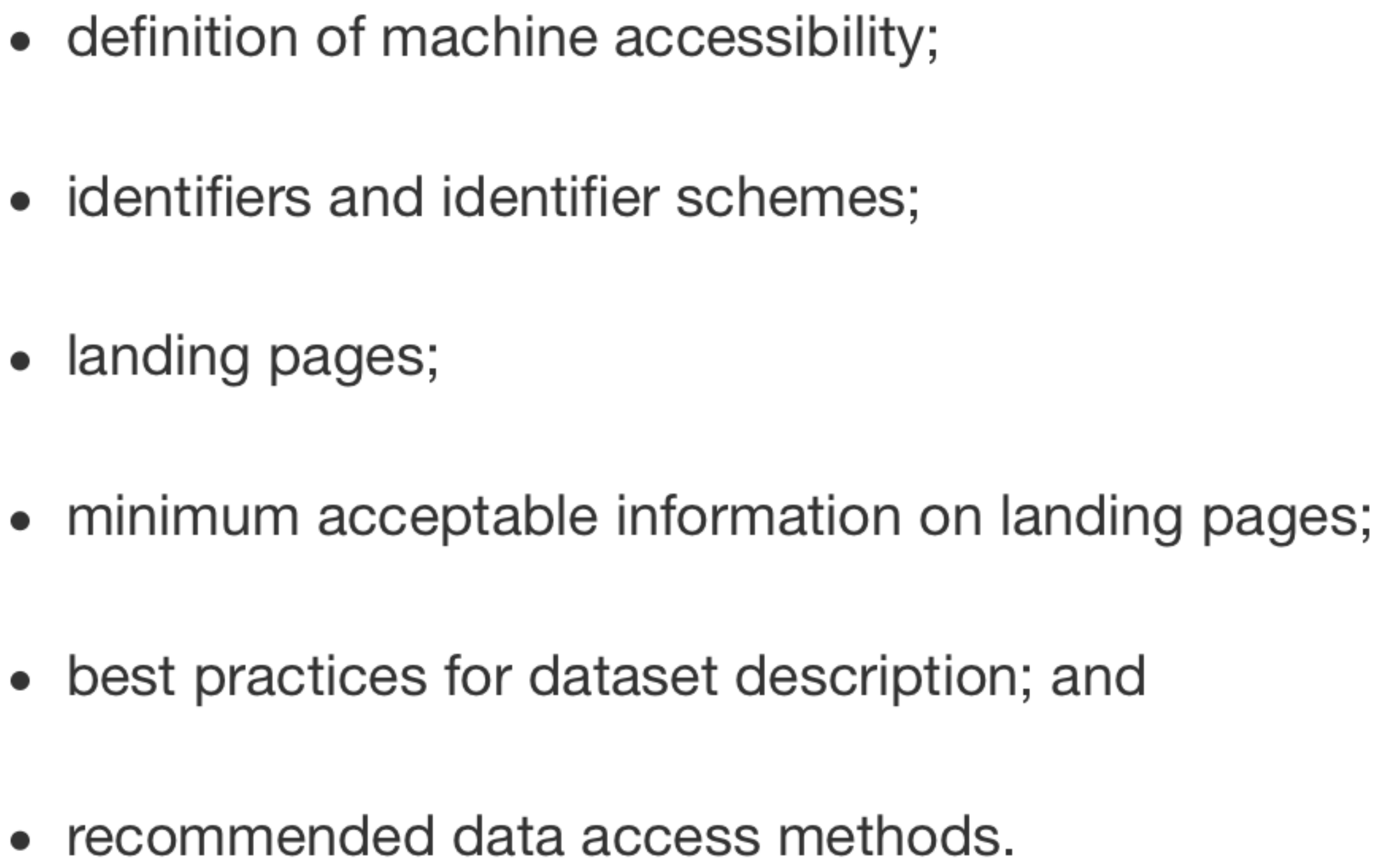
Downloads 184

Outline

Introduction

Recommendations for

Recommendations



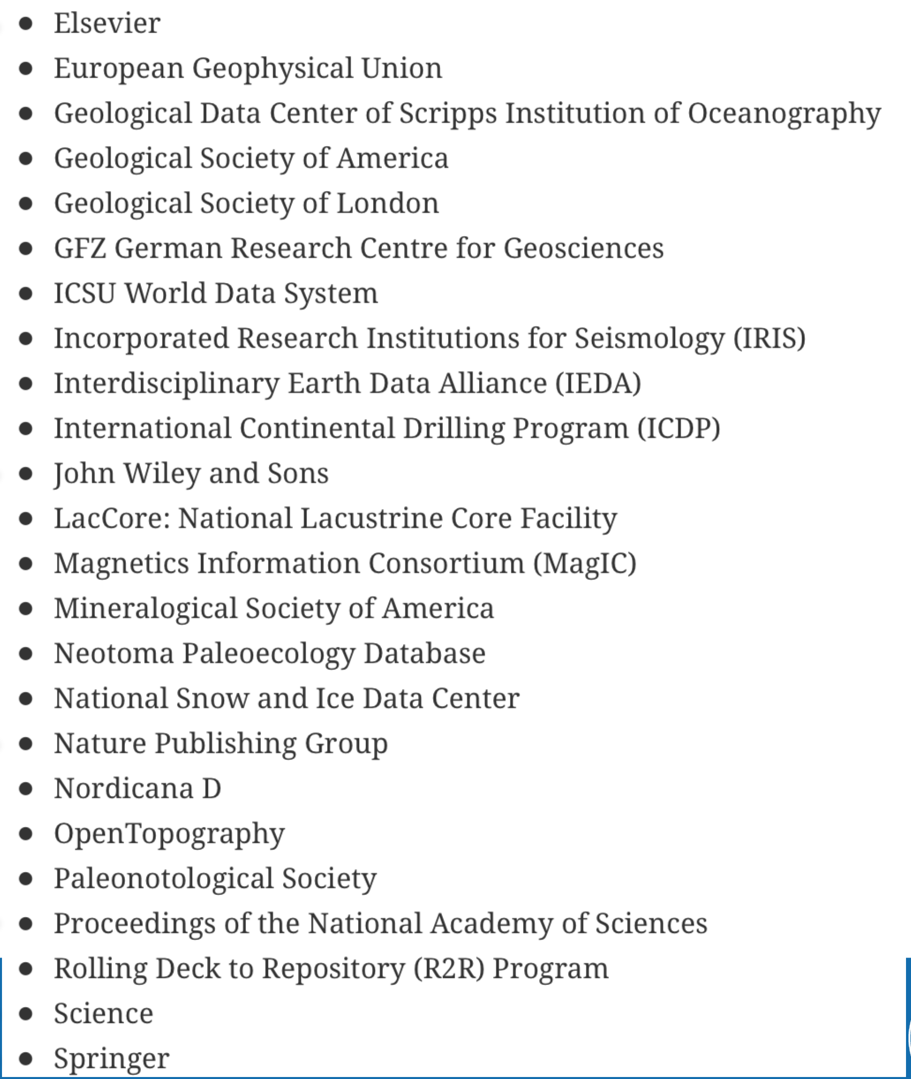
- definition of machine accessibility;
- identifiers and identifier schemes;
- landing pages;
- minimum acceptable information on landing pages;
- best practices for dataset description; and
- recommended data access methods.

Research Data Alliance Working Groups

- Data bibliometrics
- Data services
- Data Workflows in conjunction with Force 11 group
- Cost recovery for data centers
- Dynamic data citation

Coalition for Publishing Data in the Earth Sciences (COPDESS)

- Statement of Commitment from Earth and Space Science Publishers and Data Facilities



- Elsevier
- European Geophysical Union
- Geological Data Center of Scripps Institution of Oceanography
- Geological Society of America
- Geological Society of London
- GFZ German Research Centre for Geosciences
- ICSU World Data System
- Incorporated Research Institutions for Seismology (IRIS)
- Interdisciplinary Earth Data Alliance (IEDA)
- International Continental Drilling Program (ICDP)
- John Wiley and Sons
- LacCore: National Lacustrine Core Facility
- Magnetics Information Consortium (MagIC)
- Mineralogical Society of America
- Neotoma Paleoecology Database
- National Snow and Ice Data Center
- Nature Publishing Group
- Nordicana D
- OpenTopography
- Paleontological Society
- Proceedings of the National Academy of Sciences
- Rolling Deck to Repository (R2R) Program
- Science
- Springer

Coalition for Publishing Data in the Earth Sciences (COPDESS)

- Data management policies
- Index of data facilities
- Released common verbiage for authors, editors, and reviewers for a wide variety of journals
- Extending Re3data.org schema in the Earth Sciences to allow detailed identification of what repositories a journal accepts as a reasonable place to put data

Making Dynamic Data Citeable

- Building blocks of supporting dynamic data citation:
  - Uniquely identifiable data records (for unique sort)
  - Versioned data, marking changes as insertion/deletion
  - Timestamps on data insertion / deletions
  - “Query language” for constructing subsets
- Add modules:
  - Persistent query store: queries, timestamp, hash, metadata including creator of subset
  - Query rewriting module
  - PID assignment to queries
  - Landing page design, citation text
- Stable across data source migrations (e.g. diff. DBMS), scalable, machine-actionable

S. Pröll, A. Rauber. **Scalable Data Citation in Dynamic Large Databases: Model and Reference Implementation.** In IEEE Intl. Conf. on Big Data 2013 (v BigData2013), 2013
http://www.ifs.tuwien.ac.at/~andi/publications/pdf/pro_ieeebigdata13.pdf

Dynamic Data Citation – Deployment

- Researcher uses workbench to identify subset of data
- Upon executing selection („download“) user gets
  - Data (package, access API, …)
  - PID (e.g. DOI) (Query is time-stamped and stored)
  - Hash value computed over the data for local storage
  - Recommended citation text (e.g. BibTeX)
- PID resolves to landing page
  - Provides detailed metadata, link to parent data set, subset,…
  - Option to retrieve **original data** OR **current version** OR **changes**
- Upon activating PID associated with a data citation
  - Query is re-executed against time-stamped and versioned DB
  - Results as above are returned

Earth Science View of Citation

ESIP has had guidelines for citation of dynamic data for many years

Doe, J. and R. Roe. 2001, updated daily. The FOO Gridded Time Series Data Set. Version 3.2. Oct. 2007- Sep. 2008, 84°N, 75°W; 44°N, 10°W. The FOO Data Center. http://dx.doi.org/10.xxxx/notfoo.547983. Accessed 1 May 2011.

The question is can a reproducible subset identifier be generated to replace the red bit.

Chemistry View of Citation

Excerpt of text from the body of an article that cites PubChem records and a Molecular Libraries chemical probe:

"We searched the PubChem BioAssay database for the biological activity and found one assay, AID: 2299 (1) from the Scripps Research Institute Molecular Screening Center, which reported the identification and development of chemical probe ML114 (2), a potent small molecule inhibitor against Retinoblastoma binding protein 9 (RBBP9). The chemical structure information for this probe is available in the PubChem Substance and Compound database through the substance identifier number SID: 85098567 (3) and/or the unique chemical structure identifier CID: 5934766 (4)."

Excerpt of corresponding references from the article's bibliography:

(1) National Center for Biotechnology Information. PubChem BioAssay Database; AID=2299, Source=Scripps Research Institute Molecular Screening Center, http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=2299 (accessed Feb. 22, 2011).

(2) NIH Molecular Libraries. Probe Report for RBBP9 Inhibitors. Chapter ML114 IN *Probe Reports from the Molecular Libraries Program* [Internet], National Library of Medicine (US), National Center for Biotechnology Information, Bethesda, MD, 2010 (accessed 2011 Feb 22). Available from http://www.ncbi.nlm.nih.gov/books/NBK50690/ (or http://www.ncbi.nlm.nih.gov/books/n/mlprobe/ml114) in Entrez Books (http://www.ncbi.nlm.nih.gov/books).

(3) National Center for Biotechnology Information. PubChem Substance Database; SID=85098567, Source=Scripps Research Institute Molecular Screening Center, http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?sid=85098567 (accessed Feb. 22, 2011).

(4) National Center for Biotechnology Information. PubChem Compound Database; CID=5934766, http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=5934766 (accessed Feb. 22, 2011).