ILLINOIS

Dynamic Data Citations: The Current State

Ruth Duerr

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois

Ronin Institute

Overview

- A brief history of data citation in general
- Dynamic data citations

Note: Many of these slides or their content have been borrowed from various folks (e.g., Andreas Rauber, etc.)

A brief history of data citation

Data citation used to be common practice

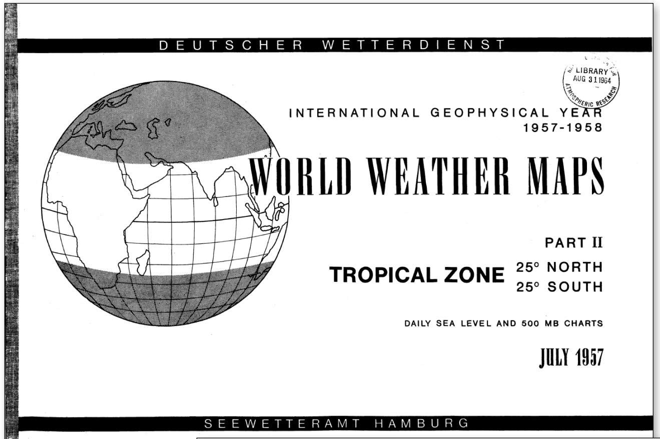
What!!

?

A brief history of data citation

Data was in the literature!

In Books and Technical Reports



DEUTSCHER WETTERDIENST

INTERNATIONAL GEOPHYSICAL YEAR
1957-1958

WORLD WEATHER MAPS

PART II

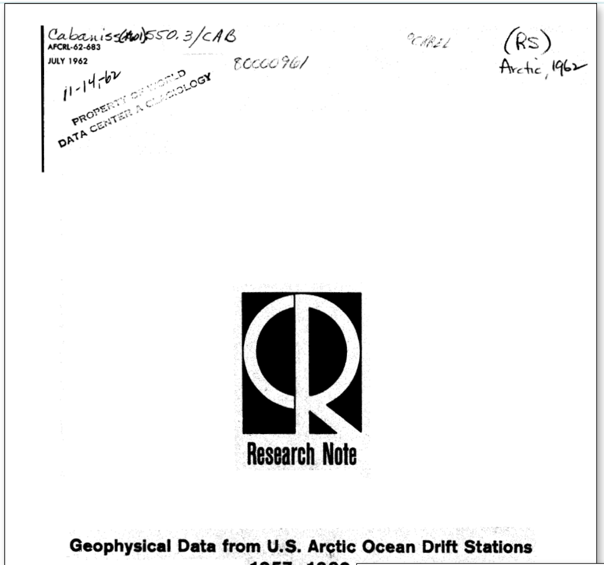
TROPICAL ZONE 25° NORTH 25° SOUTH

DAILY SEA LEVEL AND 500 MB CHARTS

JULY 1957

SEEWETTERAMT HAMBURG

DATA
90022



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

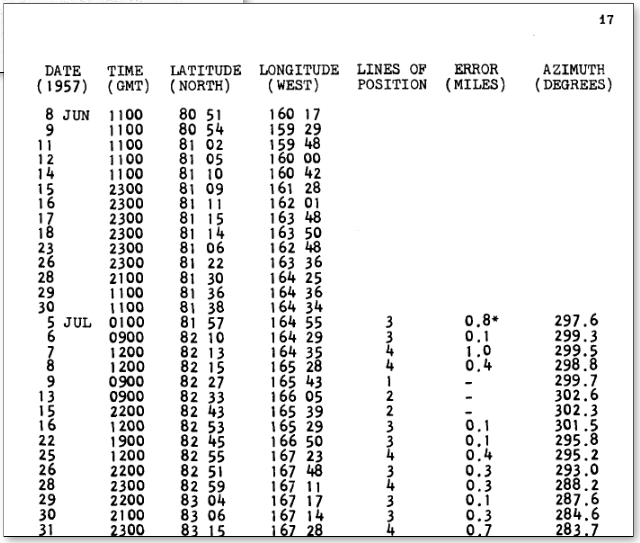
80000961

(RS)
Arctic, 1962

Research Note

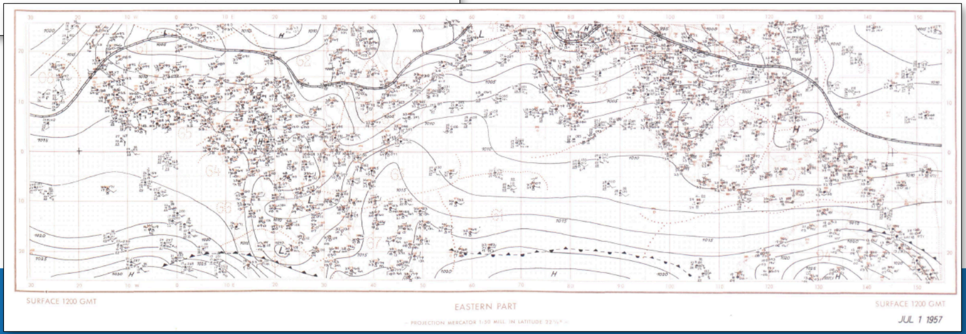
Geophysical Data from U.S. Arctic Ocean Drift Stations
1957—1960

G. H. CABANISS
EDITOR



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



SURFACE 1200 GMT

EASTERN PART

SURFACE 1200 GMT

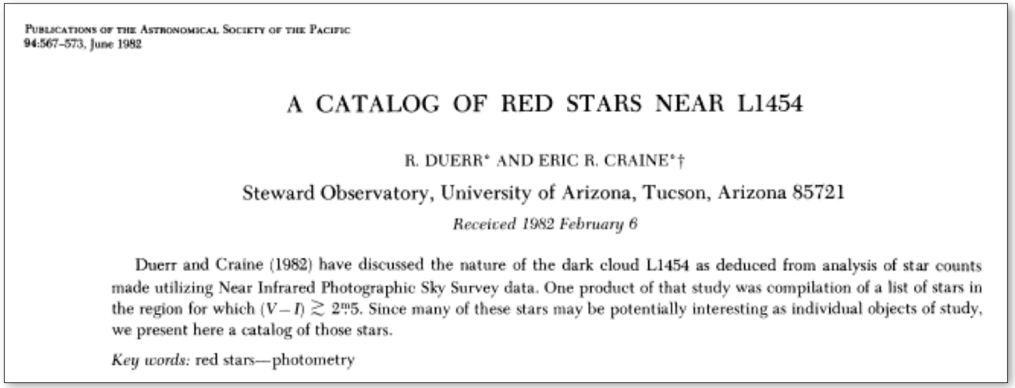
— PROJECTION MERCATOR 1:50 MILL. IN LATITUDE 22½° —

JUL 1 1957

A brief history of data citation

Data was in the literature!

and Journals



Publications of the Astronomical Society of the Pacific
94:567–573, June 1982

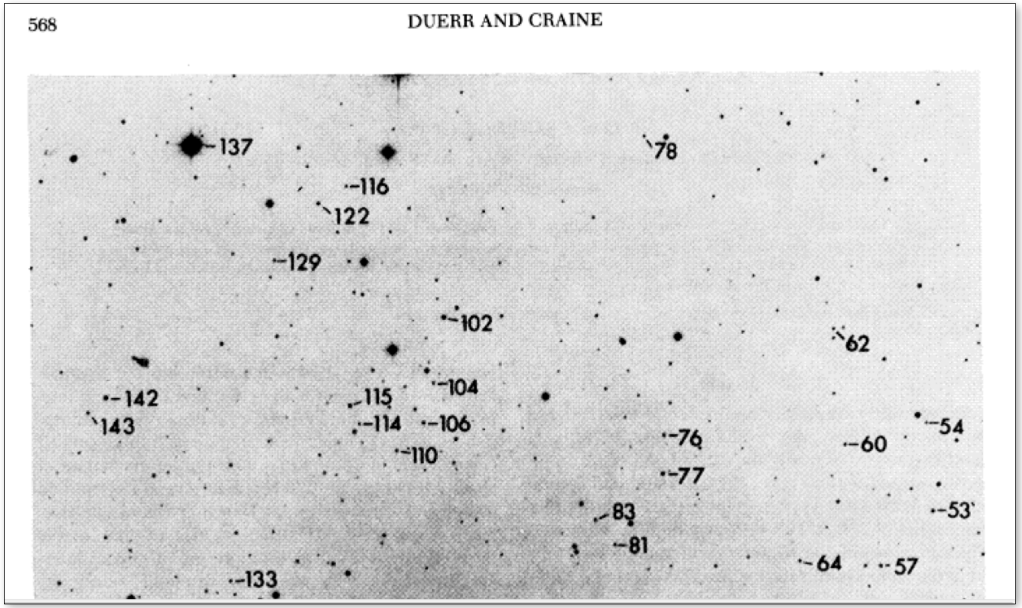
A CATALOG OF RED STARS NEAR L1454

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 $(V-I) \gtrsim 2\overset{\mathrm{m}}{.}5$. 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

A brief history of data citation

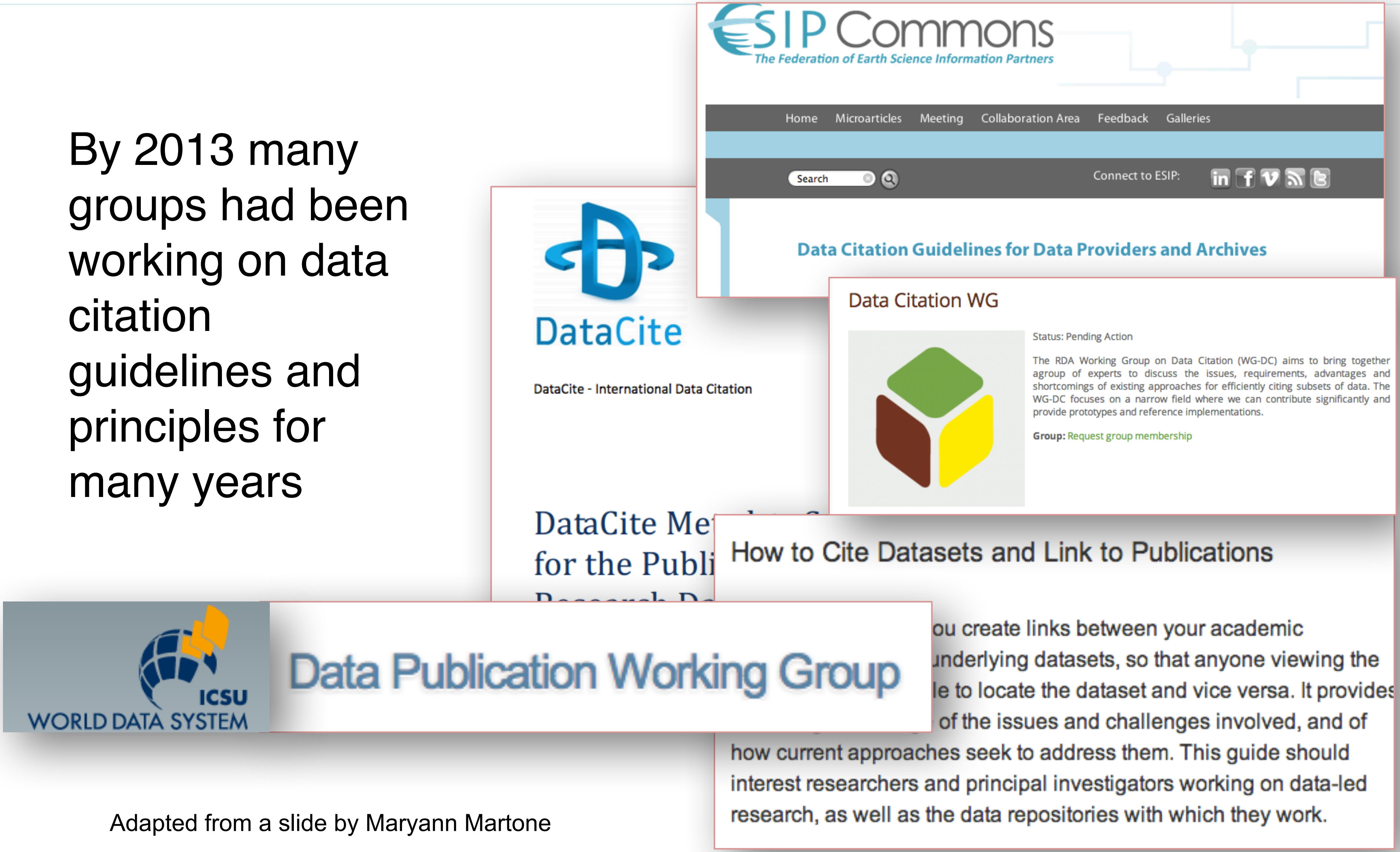
- That started changing with the advent of digital data
    - At first because the publications were still paper
        - Why would you want to make your data less accessible to the computers needed to analyze it?
        - Now how do you represent a multi-dimensional data set in a two-dimensional medium?
        -
        -
        -
    - Later because often the data was voluminous

A brief history of data citation

- Digital data repositories came into being in the final decades of the 20th century
  - Many collocated with existing data centers (e.g., World Data Centers set up during the International Geophysical Year 1957/8)
  - Many have been promoting data citation for decades

A brief history of data citation

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

A brief history of data citation

Photo: Flickr

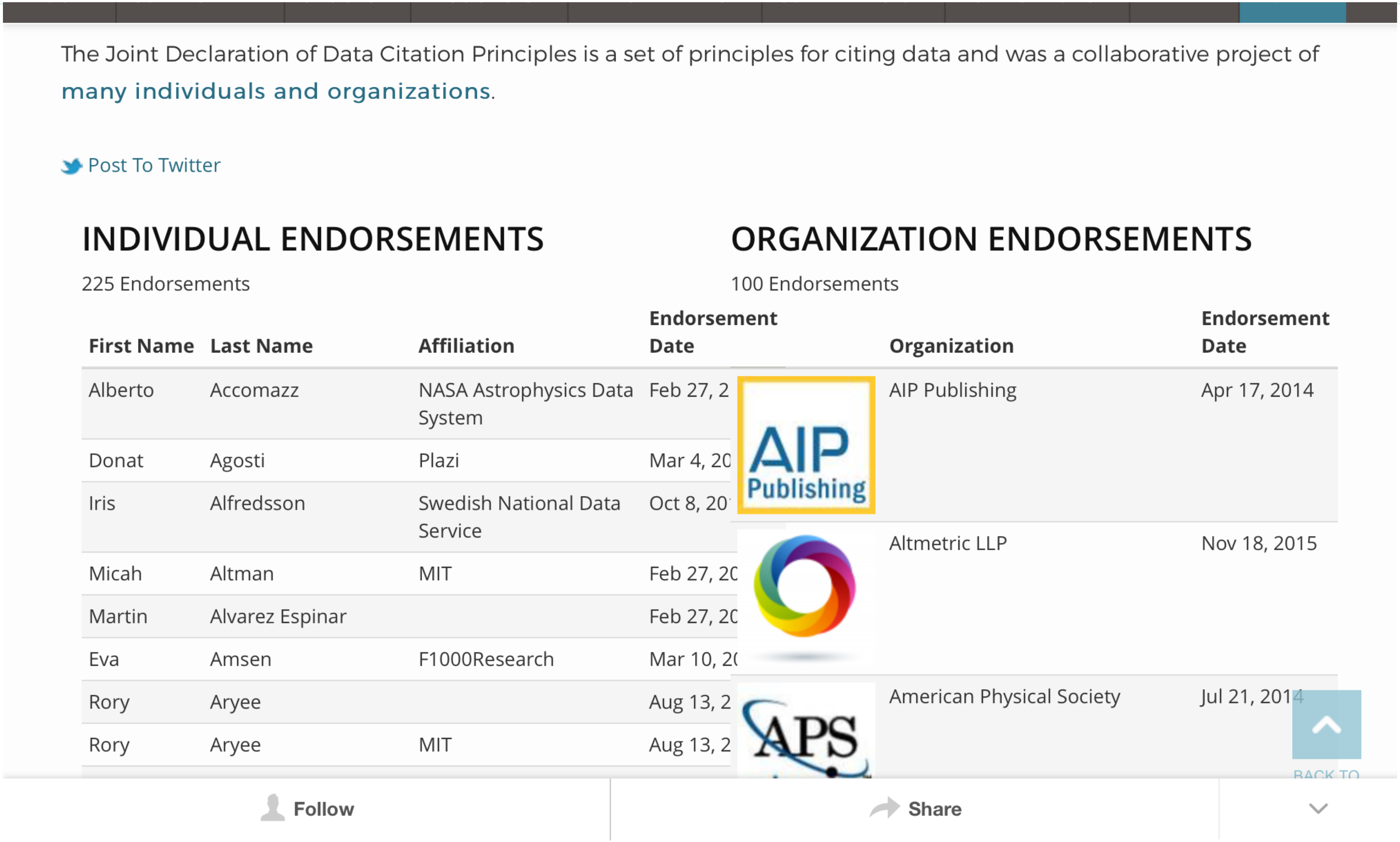
Paul Uhlir "...a plea to come together"

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.
- ***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

Joint Declaration of Data Citation Principles



The Joint Declaration of Data Citation Principles is a set of principles for citing data and was a collaborative project of many individuals and organizations.

Post To Twitter

INDIVIDUAL ENDORSEMENTS

225 Endorsements

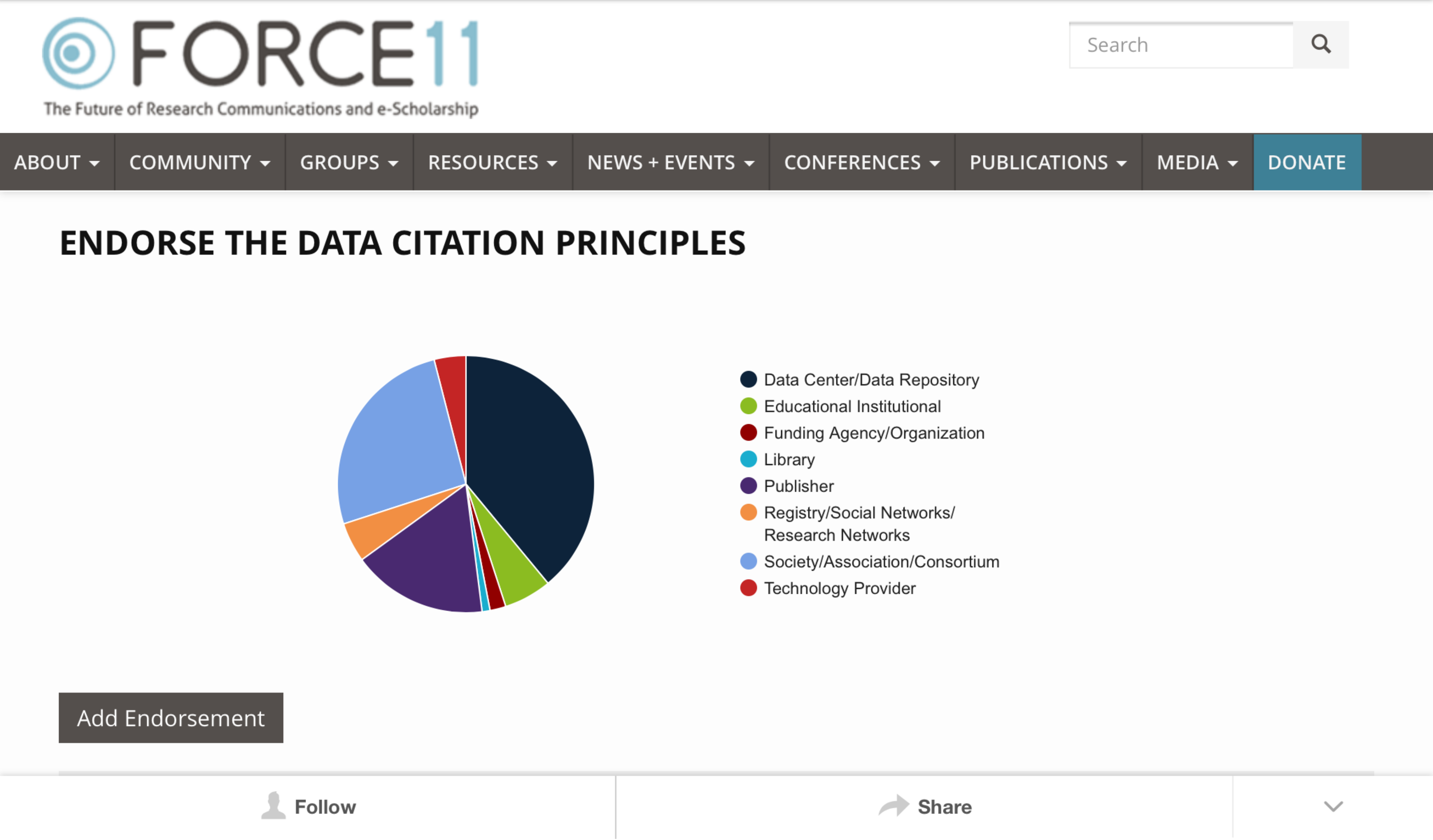
First Name	Last Name	Affiliation	Endorsement Date
Alberto	Accomazz	NASA Astrophysics Data System	Feb 27, 2
Donat	Agosti	Plazi	Mar 4, 20
Iris	Alfredsson	Swedish National Data Service	Oct 8, 20
Micah	Altman	MIT	Feb 27, 20
Martin	Alvarez Espinar		Feb 27, 20
Eva	Amsen	F1000Research	Mar 10, 2(
Rory	Aryee		Aug 13, 2
Rory	Aryee	MIT	Aug 13, 2

ORGANIZATION ENDORSEMENTS

100 Endorsements

	Organization	Endorsement Date
AIP Publishing	AIP Publishing	Apr 17, 2014
	Altmetric LLP	Nov 18, 2015
APS	American Physical Society	Jul 21, 2014

Joint Declaration of Data Citation Principles



FORCE11
The Future of Research Communications and e-Scholarship
Search
ABOUT
COMMUNITY
GROUPS
RESOURCES
NEWS + EVENTS
CONFERENCES
PUBLICATIONS
MEDIA
DONATE

ENDORSE THE DATA CITATION PRINCIPLES

● Data Center/Data Repository
● Educational Institutional
● Funding Agency/Organization
● Library
● Publisher
● Registry/Social Networks/ Research Networks
● Society/Association/Consortium
● Technology Provider
Add Endorsement
Follow
Share

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.

Data Citation in the NISO-JATS DTD

NISO-JATS is an open standard for representing full text articles in XML
Used widely, but not limited to, in life sciences.

Technical Workshop: June 2014, London
18 (publishers, JATS users, and JATS committee reps)

Workshop Goals

- JATS recommendations to support structured data citations according to the F11 Data Citation Principles
- Decide adoption and implementation strategy by publishers

FORCE11
The Future of Research Communications and e-Scholarship

OpenAIRE

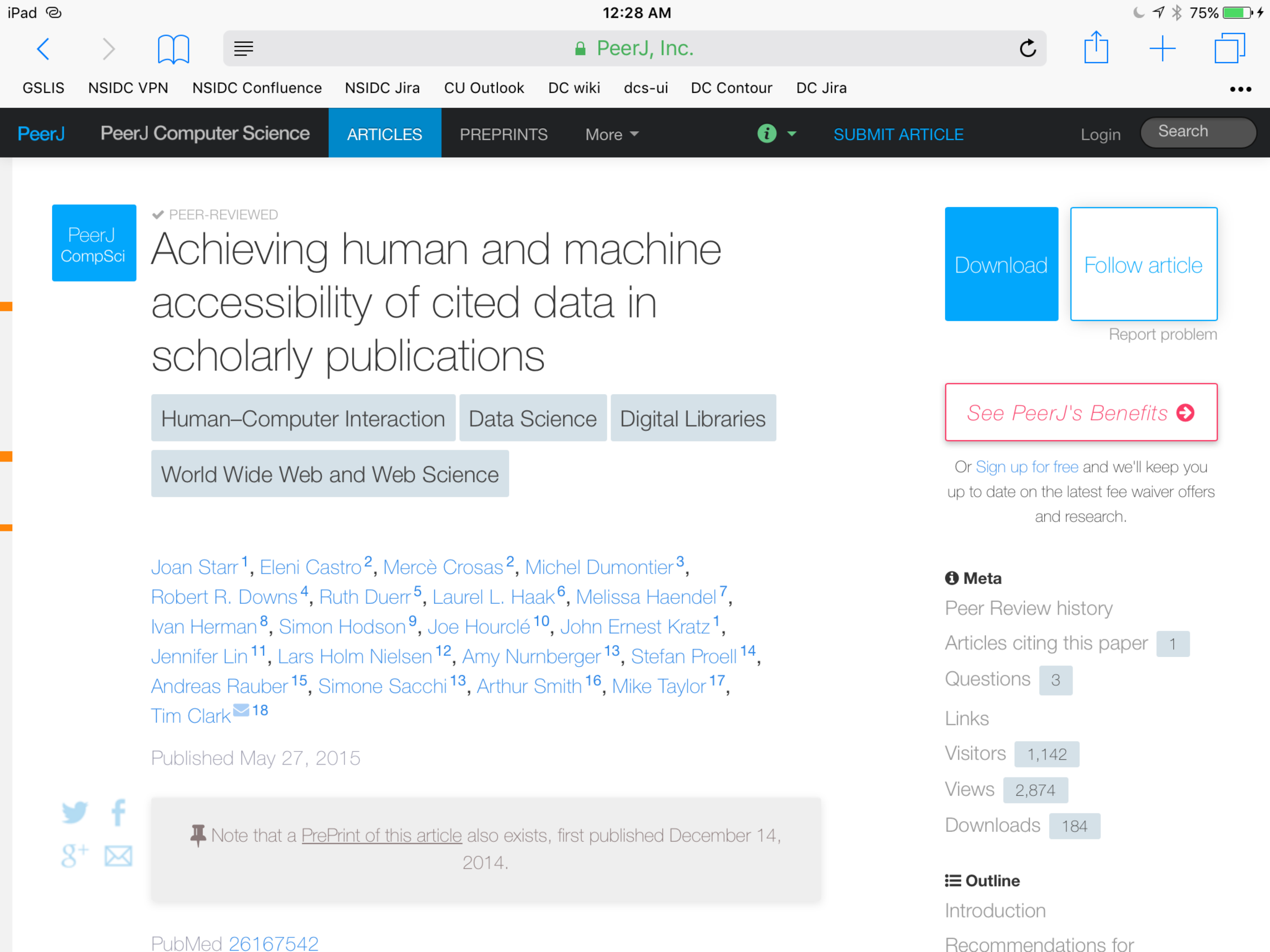
Implications of NISO-JATS support for data citation

- 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

For more info: mcentyre@ebi.ac.uk

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

Meta

Peer Review history

Articles citing this paper 1

Questions 3

Links

Visitors 1,142

Views 2,874

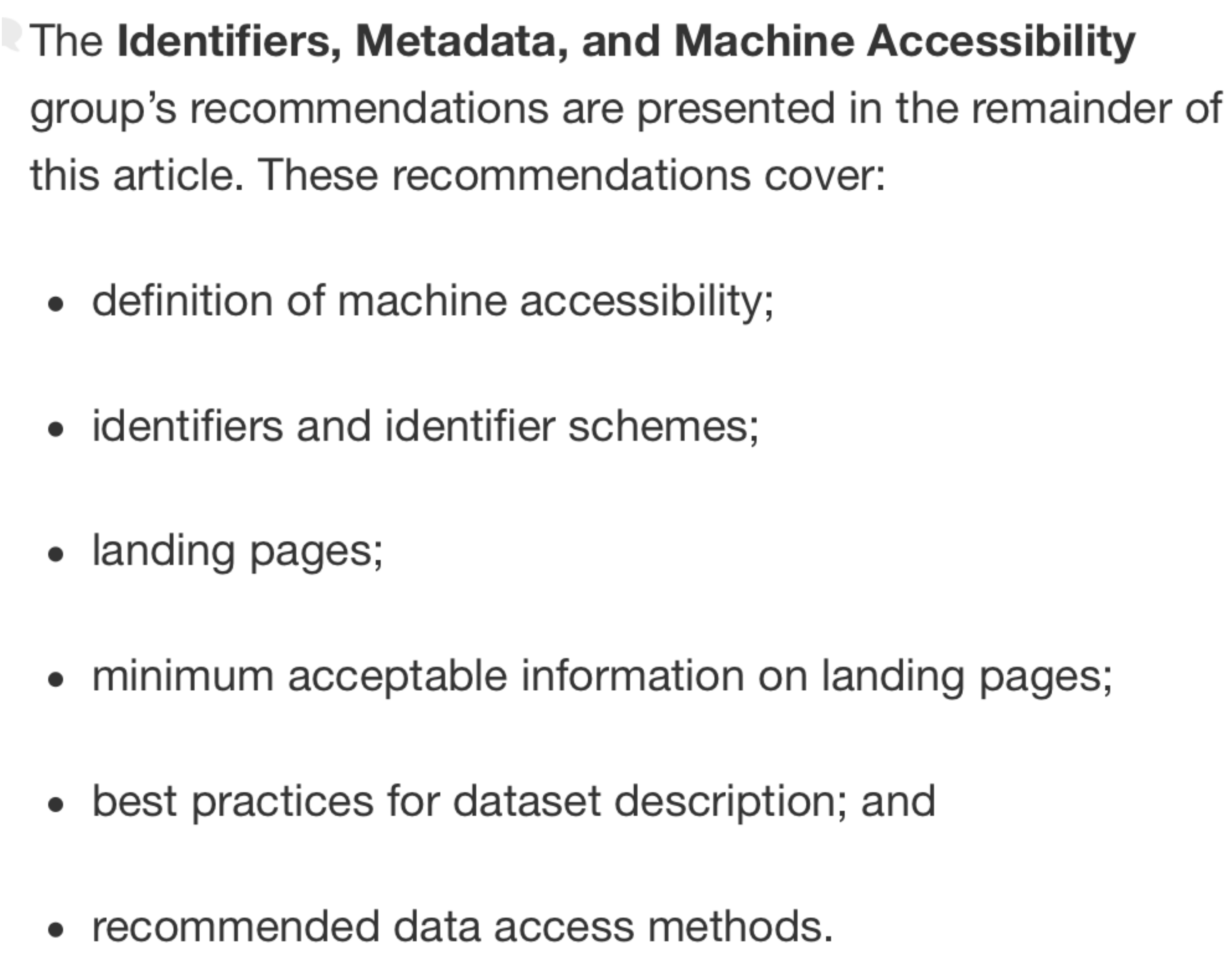
Downloads 184

Outline

Introduction

Recommendations for

Data Citation Implementer's Group



The **Identifiers, Metadata, and Machine Accessibility** group's recommendations are presented in the remainder of this article. These recommendations cover:

- 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.

Moving Forward

- Research Data Alliance has several working groups working on data citation also
  - Data bibliometrics
  - Data services
  - Data Workflows in conjunction with Force 11 group
  - Cost recovery for data centers
  - Dynamic data citation

Moving Forward in Earth Sciences

- Brooks Hanson (AGU) and Kerstin Lehnert (IEDA) held two publisher's round table meetings
    - Statement of Commitment from Earth and Space Science Publishers and Data Facilities
    - Data management policies
    - Training/Ethics - E.g., for NSF program managers
    - Ongoing collaboration between publishers and data centers
    - Index of data facilities (now an extension to re3data.org schema)
- AGU has endorsed both the Joint Principles and the COPDESS statement.

RDA Dynamic Data Citation – Recommendations

Preparing Data & Query Store

- R1 – Data Versioning
- R2 – Timestamping
- R3 – Query Store

When Resolving a PID

- R11 – Landing Page
- R12 – Machine Actionability

When Data should be persisted

- R4 – Query Uniqueness
- R5 – Stable Sorting
- R6 – Result Set Verification
- R7 – Query Timestamping
- R8 – Query PID
- R9 – Store Query
- R10 – Citation Text

Upon Modifications to the Data Infrastructure

- R13 – Technology Migration
- R14 – Migration Verification

Data Citation of Evolving Data

Data Citation – Recommendations

A) Preparing the Data and the Query Store

- R1 – Data Versioning: Apply versioning to ensure earlier states of data sets the data can be retrieved
- R2 – Timestamping: Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp
- R3 – Query Store: Provide means to store the queries and metadata to re-execute them in the future

Data Citation – Recommendations

B) Persistently Identify Specific Data sets (1/2)

When a data set should be persisted:

- R4 – Query Uniqueness: Re-write the query to a normalised form so that identical queries can be detected. Compute a checksum of the normalized query to efficiently detect identical queries
- R5 – Stable Sorting: Ensure an unambiguous sorting of the records in the data set
- R6 – Result Set Verification: Compute fixity information/checksum of the query result set to enable verification of the correctness of a result upon re-execution
- R7 – Query Timestamping: Assign a timestamp to the query based on the last update to the entire database (or the last update to the selection of data affected by the query or the query execution time). This allows retrieving the data as it existed at query time

Data Citation – Recommendations

B) Persistently Identify Specific Data sets (2/2)

When a data set should be persisted:

- R8 – Query PID: Assign a new PID to the query if either the query is new or if the result set returned from an earlier identical query is different due to changes in the data. Otherwise, return the existing PID

- R9 – Store Query: Store query and metadata (e.g. PID, original and normalized query, query & result set checksum, timestamp, superset PID, data set description and other) in the query store

- R10 – Citation Text: Provide citation text including the PID in the format prevalent in the designated community to lower barrier for citing data.

Data Citation – Recommendations

C) Resolving PIDs and Retrieving Data

- R11 – Landing Page: Make the PIDs resolve to a human readable landing page that provides the data (via query re-execution) and metadata, including a link to the superset (PID of the data source) and citation text snippet

- R12 – Machine Actionability: Provide an API / machine actionable landing page to access metadata and data via query re-execution

EU Pilot's

- Pilot workshops and implementations by
  - Various EU projects (TIMBUS, SCAPE,…)
    - Linguistics transcriptions - XML database
    - CSV data
    - SQL databases
  - DEXHELPP – Social Security Data
  - NERC (UK Natural Environment Research Council Data Centres)
  - Virtual Atomic and Molecular Data Centre

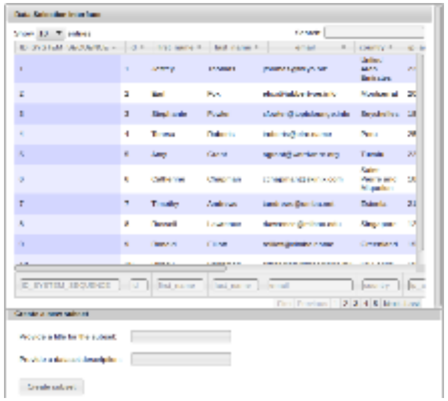
see reports at https://rd-alliance.org/groups/data-citation-wg.html

DEXHELPP Pilot

- Routine / secondary data in the medical domain
- Accounting / reimbursement data from the social insurance providers for doctors and hospitals
- Collected for 99% of the Austrian population
- Full data for a 2-year span
  - For some provinces for a longer period
- Structured data (relational database)
- Around 2.5 billion records

Data Citation in DEXHELPP

- Data exchange format between institutions: CSV
- Subset creation process is based on the CSV Prototype
- Reproducibility
  - By tracing the creation process
  - Versioned data
  - Query based mechanism
- On demand subsets:
  - By re-executing the queries



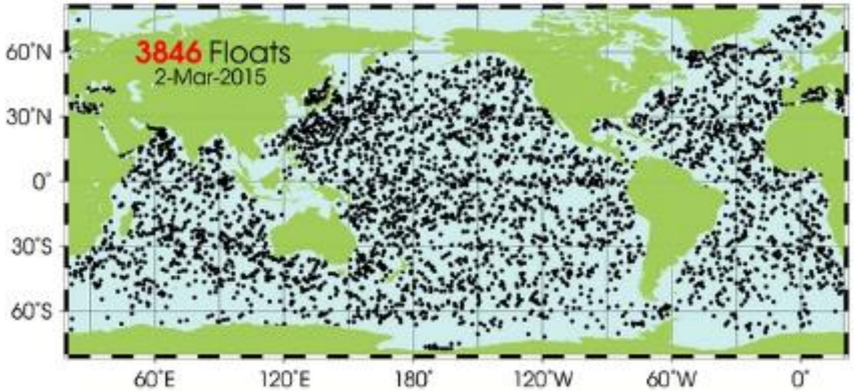
- Citation process preserves privacy and adds security
  - Different privacy levels per user group (k-anonymity)
  - Watermark data sets
  - Add fingerprints to identify to individual creator

Data Citation WG – UK NERC progress to date

- Reported to RDA Plenary 5 on progress especially in marine sector
- The ARGO buoy network approach to DataCite and publishing houses to establish dynamic data citation
- Student Fellow with the Earth Science Information Partners (ESIP) Data Stewardship Committee (Sophie Hou @ Uni Illinoise) using UK River Flow Archive as case study for gaining credit for dynamic research data
- NERC data centre experiencing increasing need for data DOIs leading to pressure to dynamic data citation mechanisms

What is the Argo global array?

- Argo is a global array of more than 3,000 free-drifting profiling floats
- Each measures the temperature and salinity of the upper 2000 m of the ocean

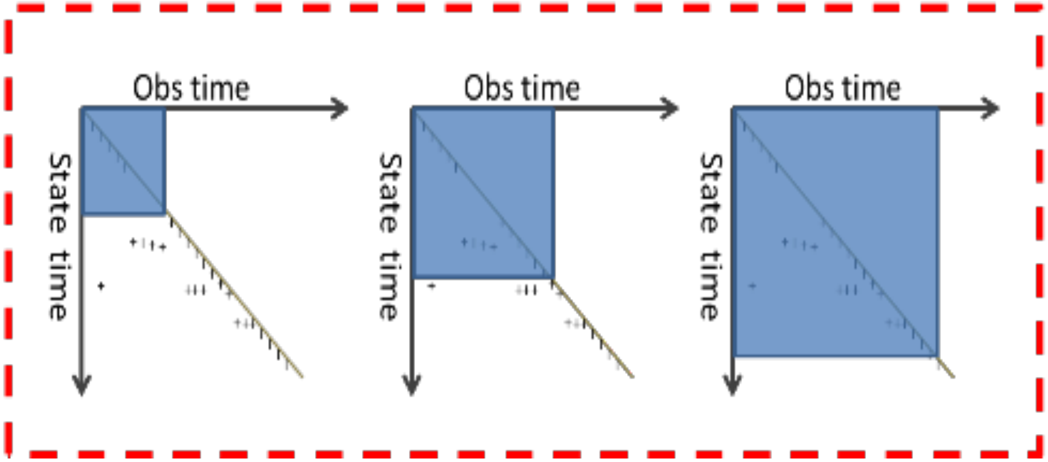


- This allows, for the first time, continuous monitoring of the temperature, salinity, and velocity of the upper ocean, with all data being relayed and made publicly available within hours after collection.

Progress on the Argo data archive

- The US NODC have proposed methods for snap-shotting of the NetCDF archives with DOIs minted at Ifremer, France
- The RDA conceptual model is being used to guide how the DOIs would be contracted and resolved

NODC Archive (collection of snapshots/granules)



Progress on the Argo data archive

- Argo data are cited by using the URI for the archive of Argo snapshots, followed by a "?" or a "#", followed by a query string identifier for the snapshot:
- e.g. http://dx.doi.org/10.7289/[Argo_accession_DOI]?[time_slice _information]
  - ? Client/browser side snapshot resolving service via a specific javascript for the accession
  - # Server side snapshot resolving service, preferred but not currently supported by DataCite.

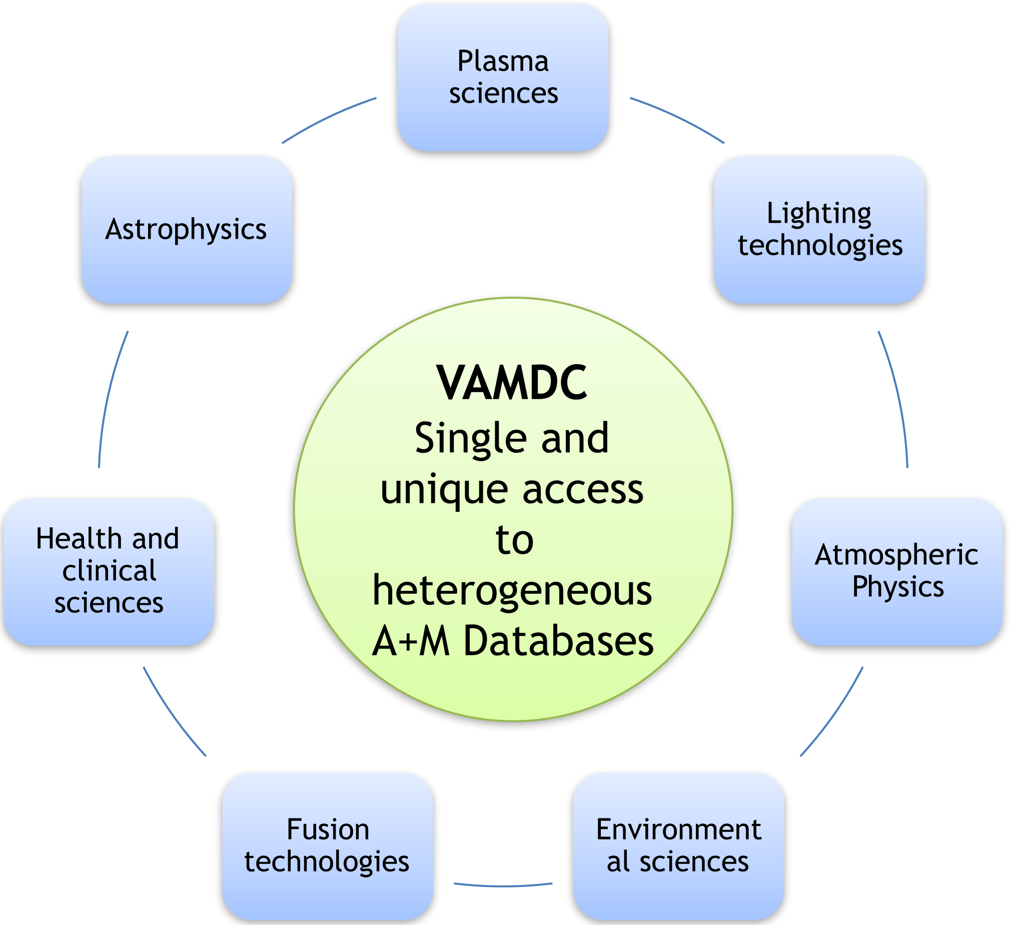
Where 7289 is the NOAA or Ifremer DOI prefix code

- http://dx.doi.org/10.7289/argo_doi_identifier?result_time=2005−01−11T16:22:25.00

Progress on the Argo data archive

- Current proposals are being discussed within Ifremer to determine approach, "?" may by necessary until # is supported by DataCite
- Discussions have started with publishing houses such as Royal Society, Elsevier, Springer, and Wiley as to tracking Argo data use in publications. The Thompson Reuters prototype hosted at ANDS looks promising.
- Issues for RDA discussion:
  - Increasing use of short DOIs by journals which impact on syntax
  - Metadata held by DataCite e.t.c. in dealing with versioning and 'access dates' for snapshot DOIs?
  - Using "#" or "?", is client side resolving an acceptable solution

The Virtual Atomic and Molecular Data Centre



- Federates 28 heterogeneous databases http://portal.vamdc.org/
- The "V" of VAMDC stands for Virtual in the sense that the e-infrastructure does not contain data. The infrastructure is a wrapping for exposing in a unified way a set of heterogeneous databases.
- The consortium is politically organized around a Memorandum of understanding (15 international members have signed the MoU, 1 November 2014)
- High quality scientific data come from different Physical/Chemical Communities
- Provides data producers with a large dissemination platform
- Remove bottleneck between data-producers and wide body of users

VAMDC consortium

Concluding remarks / open questions about query store

- How to deal with confidentiality of the information?
  - Should we need an authentication/authorization policy on the query store?
  - Is the sketched log service compliant with the EU law about confidentiality?

- We are providing to users the tools for efficiently cite our dynamic data, but
  - How can we be sure that they will use it for citing our data?
  - In other words, how to enforce the ‘citation instincts’ in our final users?

- We are thinking at proposing a ‘reverse approach’:
  - We may cite the users accessing to our data.
  - They will accept these terms, that will be explained in the condition of usage of the VAMDC services.

- How to prevent plagiarism?:
  - A user might extract data, modify and cite them as the original extracted ones.
  - Do we have tools for preventing such behaviors? MD5 of extracted data on query-store?

ESIP View of Dynamic 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.

Background

Two sessions held in the afternoon on the last day of the ESIP Winter Meeting

The first session focussed on

- Overview by Andreas on the concept of dynamic citations
- Several presentations of different data situations that might be of interest to examine in more detail

During the second session the group voted to examine 3 use cases in detail:

- MODIS Level 2 500 m snow product
- BCO-DMO ship and aerosol data
- LASP Interactive Solar Irradiance Datacenter (LISIRD)

Breakout groups were formed for each use case

Results (1 of 2)

- LISIRD system only needs minor tweaks to be able to do this
- BCO-DMO folks need to investigate costs for implementation
- MODIS case turned out to be the most difficult due to the number of different access services provided and the federated nature of some of them
  - FTP
  - Reverb/ECHO
  - subsetting by parameter, etc.

Results (2 of 2) - MODIS

- Identified a simple tool that would be helpful
  - Researcher would point to the directory tree containing the files they really used
  - The tool would record file names and checksums for each file in that directory tree
  - The tool would communicate this to the repository, which would provide a subset identifier in return
- NASA considering conducting a pilot of this through their ESDSWG

Other Work

- EGU 2016 session "20 years of persistent identifiers - where do we go next?"
  - South African Environmental Observation Network (SAEON) - Challenges in using PIDs for citation of dynamic data
  - Australian Resources Research Centre (ARRC)
- Hesburgh Libraries, U. Notre Dame
- Minnesota Population Center for Social Science Data
- IEEE, Portico, JHU work on repository, publisher interfaces

Questions?