Data Citation in Biomedicine: What, Why and How

Tim Clark
Harvard Medical School & Massachusetts General Hospital

Editorial Manager User Group Meeting
Taj Hotel, Boston - June 16 2016

Background

- NIH, NAS, other science policy makers very concerned about scientific reproducibility & robustness of results [1].

- Significant science policy studies recommend archiving & direct citation of primary data in research articles [2, 3, 4].

- NIH Big Data to Knowledge (BD2K) Program:
  "Facilitate broad use of biomedical digital assets by making them discoverable, accessible and citable." (NIH 2015) [5]

- Technology and many recommendations in place [6, 7].

→ NIH BioCADDIE / FORCE11 Data Citation Pilot in progress[8].

Some reasons to cite data

1

- Transparency & Validation
- Reproducibility & Robustness

=> better science

2

- Big Data meta-analyses
- Extract new knowledge

=> re-use & discovery

3

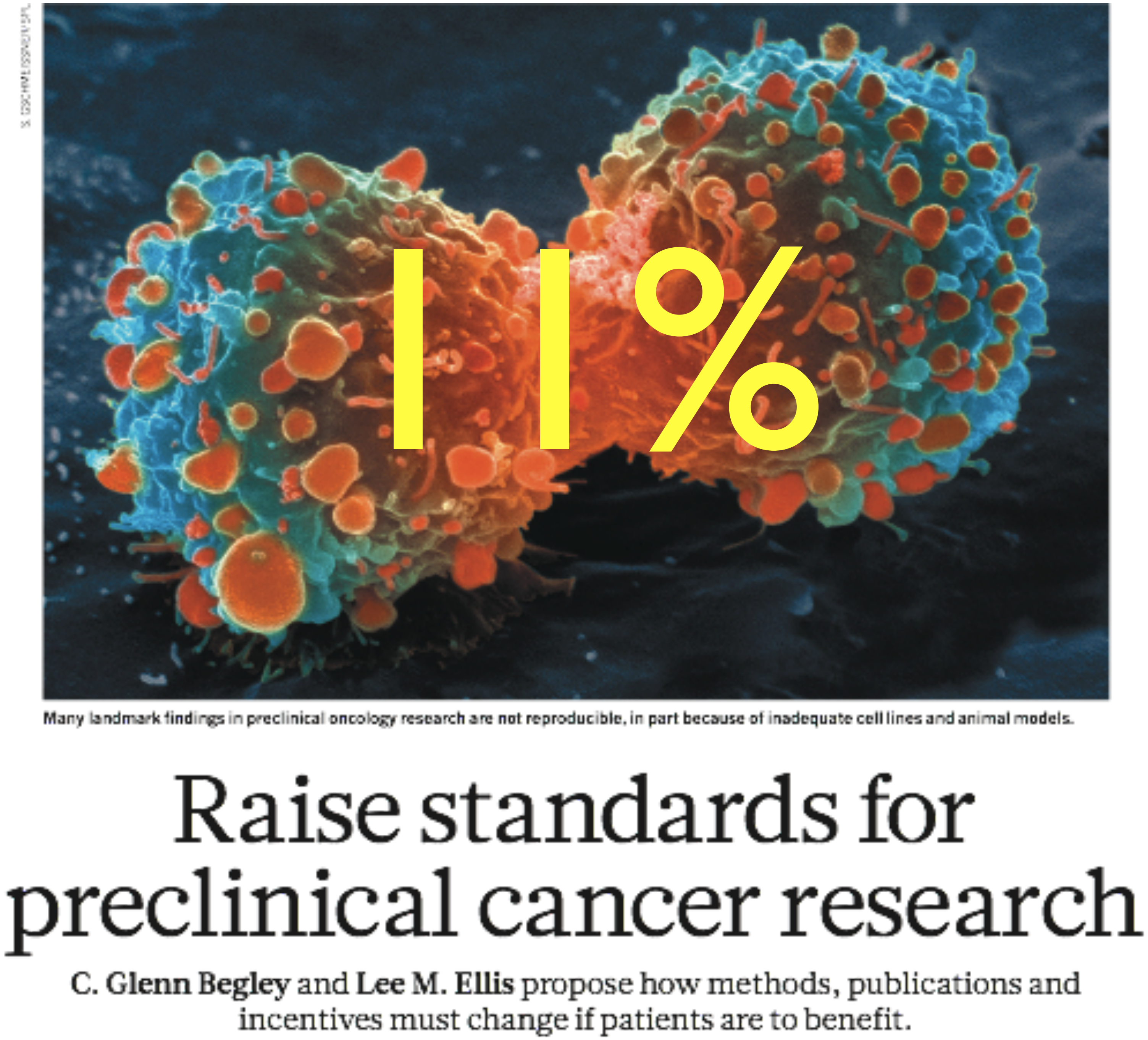
- Radically Improve Biomedical Translation

=> cure diseases

FAIR Data

- **Findable** – just as articles are findable
- **Accessible** – with appropriate permissions
- **Interoperable** – break down silos
- **Reusable** – across the life sciences ecosystem

Non-reproduciblity is a big issue in biomedicine



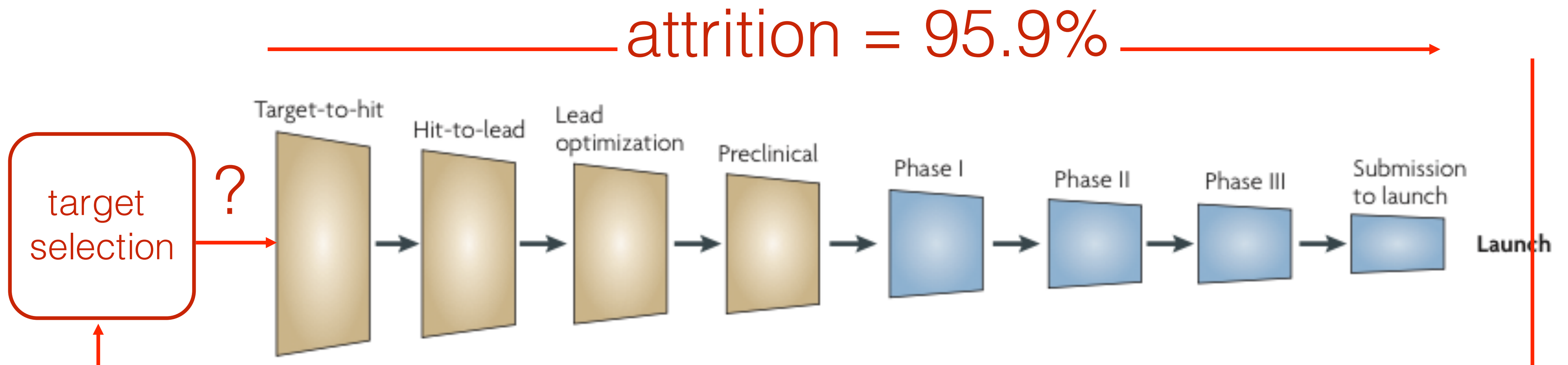
Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and **Lee M. Ellis** propose how methods, publications and incentives must change if patients are to benefit.

Begley CG and Ellis LM, *Nature* 2012, **483**(7391):531-533

Lack of robustness in pharmaceutical R & D



attrition = 95.9%

target selection

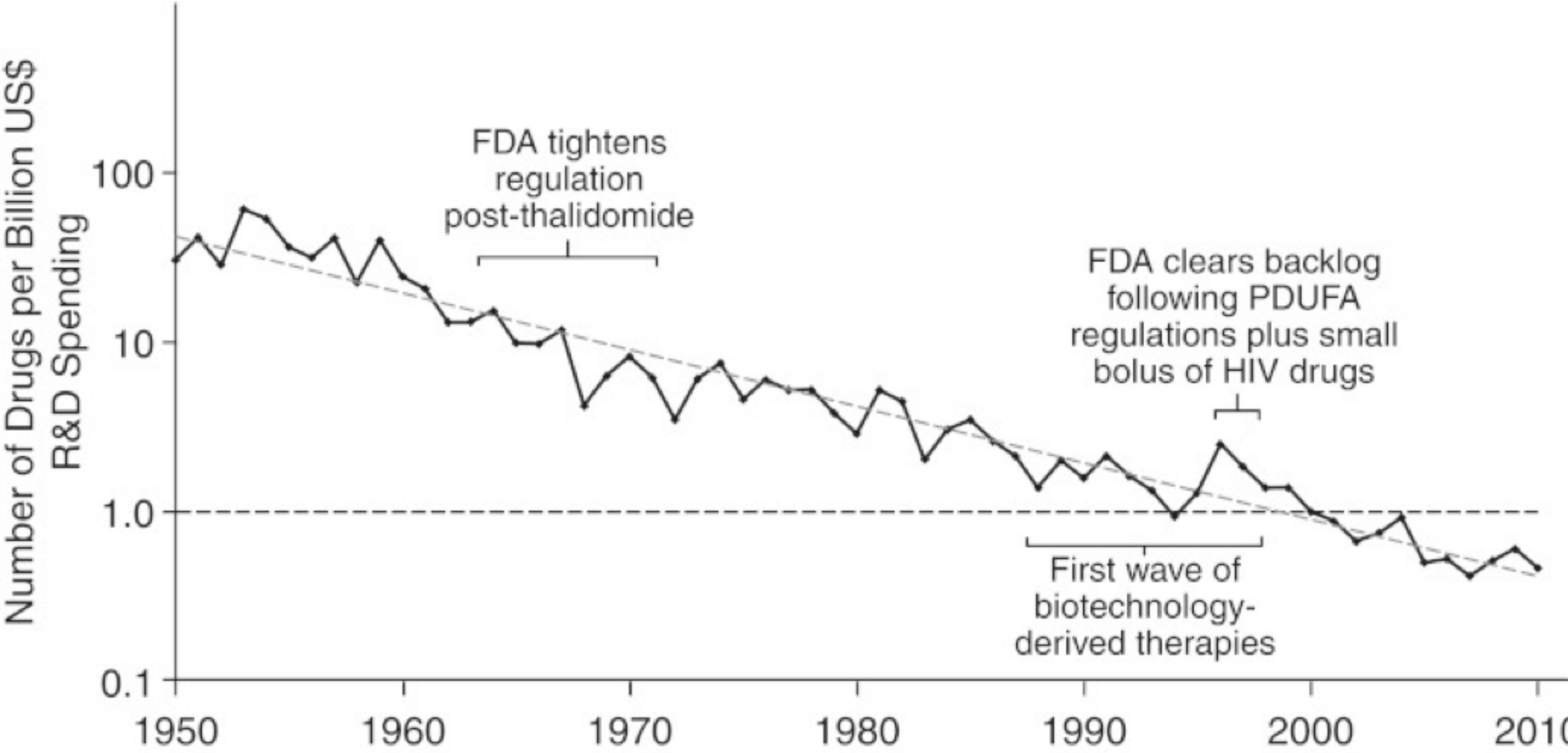
?

	Target-to-hit	Hit-to-lead	Lead optimization	Preclinical	Phase I	Phase II	Phase III	Submission to launch	Launch
p(TS)	80%	75%	85%	69%	54%	34%	70%	91%	
***WIP* needed for 1 launch**	**24.3**	**19.4**	**14.6**	**12.4**	**8.6**	**4.6**	**1.6**	**1.1**	1
Cost per WIP per Phase	$1	$2.5	$10	$5	$15	$40	$150	$40	
Cycle time (years)	1.0	1.5	2.0	1.0	1.5	2.5	2.5	1.5	
Cost per launch (out of pocket)	**$24**	**$49**	**$146**	**$62**	**$128**	**$185**	**$235**	**$44**	**$873**
% Total cost per NME	3%	6%	17%	7%	15%	21%	27%	5%	
Cost of capital	11%								
Cost per launch (capitalized)	**$94**	**$166**	**$414**	**$150**	**$273**	**$319**	**$314**	**$48**	**$1,778**

□ Discovery □ Development

$1.78 billion per new drug

Paul, S.M., *et al.* (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge, *Nat Rev Drug Discov*, 9, 203-214.

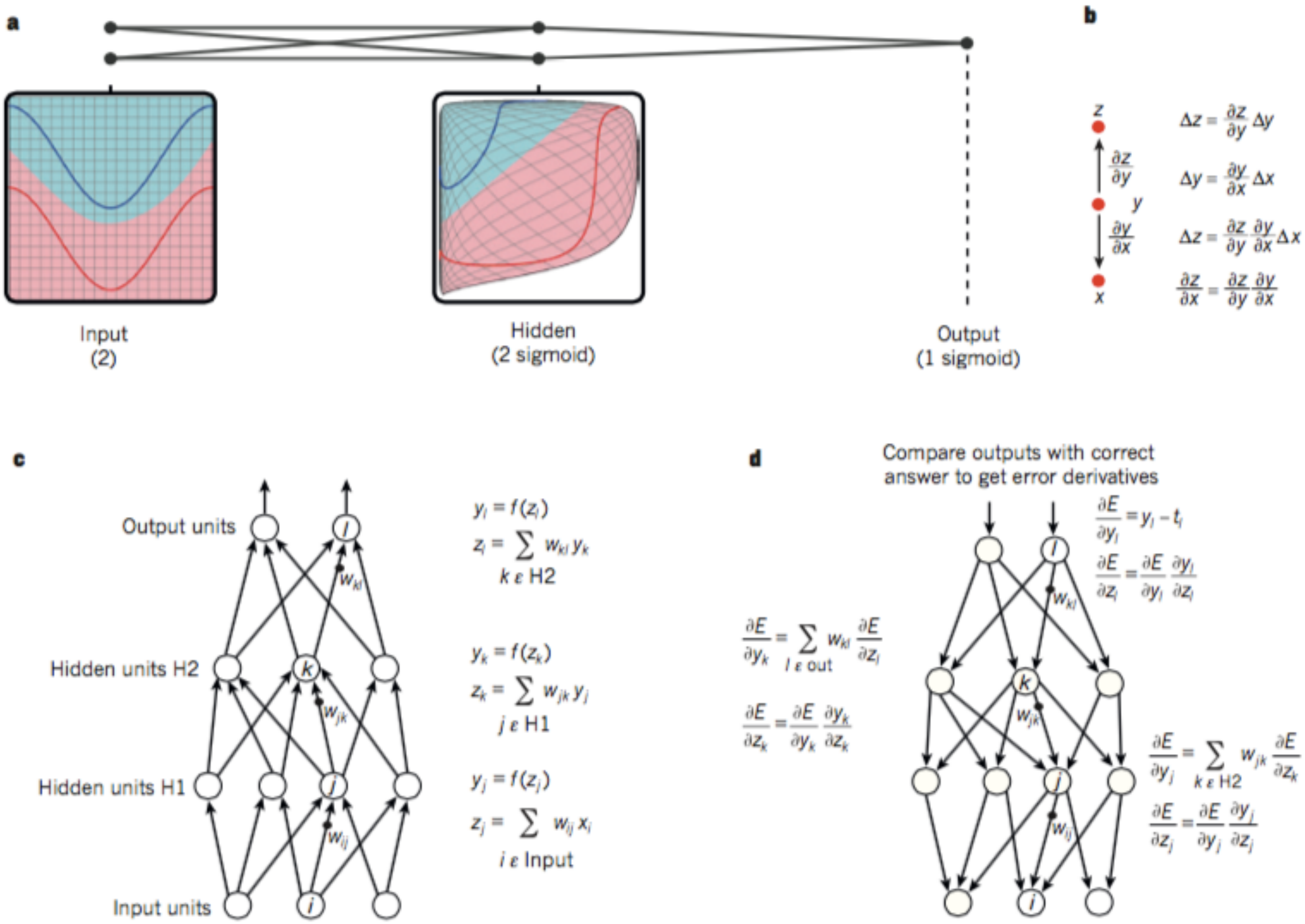


Scannell et al. 2012. Nat Rev Drug Discov, 2012;11(3):191–200 [9].

Big Data

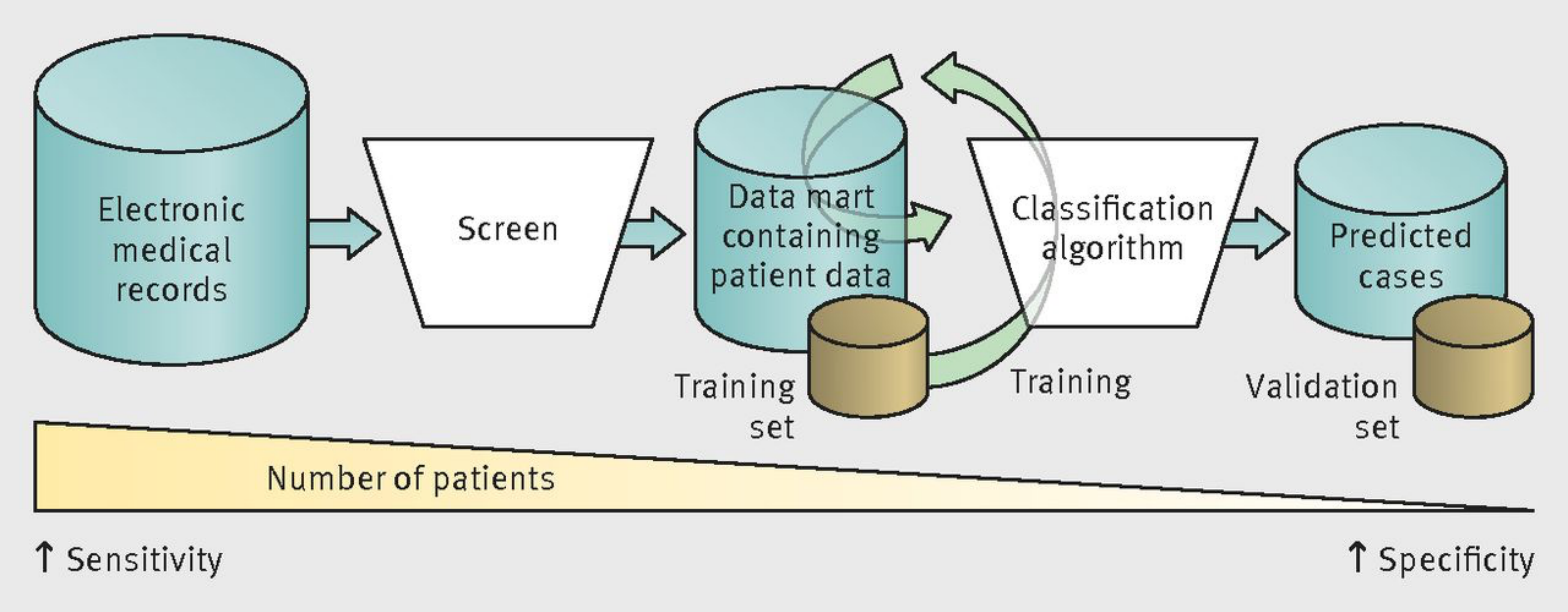
The Open Science Prize
111010110
10110101
010110110
01100101
0101101
010110111

Deep Learning Methods


a
Input
(2)
Hidden
(2 sigmoid)
Output
(1 sigmoid)
b
z
y
x
$\frac{\partial z}{\partial y}$
$\frac{\partial y}{\partial x}$
$\Delta z = \frac{\partial z}{\partial y} \Delta y$
$\Delta y = \frac{\partial y}{\partial x} \Delta x$
$\Delta z = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x} \Delta x$
$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x}$
c
Output units
Hidden units H2
Hidden units H1
Input units
l
k
j
i
w_{kl}
w_{jk}
w_{ij}
$y_l = f(z_l)$
$z_l = \sum_{k \varepsilon \mathrm{H2}} w_{kl} y_k$
$y_k = f(z_k)$
$z_k = \sum_{j \varepsilon \mathrm{H1}} w_{jk} y_j$
$y_j = f(z_j)$
$z_j = \sum_{i \varepsilon \mathrm{Input}} w_{ij} x_i$
d
Compare outputs with correct answer to get error derivatives
l
k
j
i
w_{kl}
w_{jk}
w_{ij}
$\frac{\partial E}{\partial y_l} = y_l - t_l$
$\frac{\partial E}{\partial z_l} = \frac{\partial E}{\partial y_l} \frac{\partial y_l}{\partial z_l}$
$\frac{\partial E}{\partial y_k} = \sum_{l \varepsilon \mathrm{out}} w_{kl} \frac{\partial E}{\partial z_l}$
$\frac{\partial E}{\partial z_k} = \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial z_k}$
$\frac{\partial E}{\partial y_j} = \sum_{k \varepsilon \mathrm{H2}} w_{jk} \frac{\partial E}{\partial z_k}$
$\frac{\partial E}{\partial z_j} = \frac{\partial E}{\partial y_j} \frac{\partial y_j}{\partial z_j}$

Adapted from LeCun et al. Deep Learning, Nature 521, 436–444. doi:10.1038/nature14539

EHR Predictive Data Mining

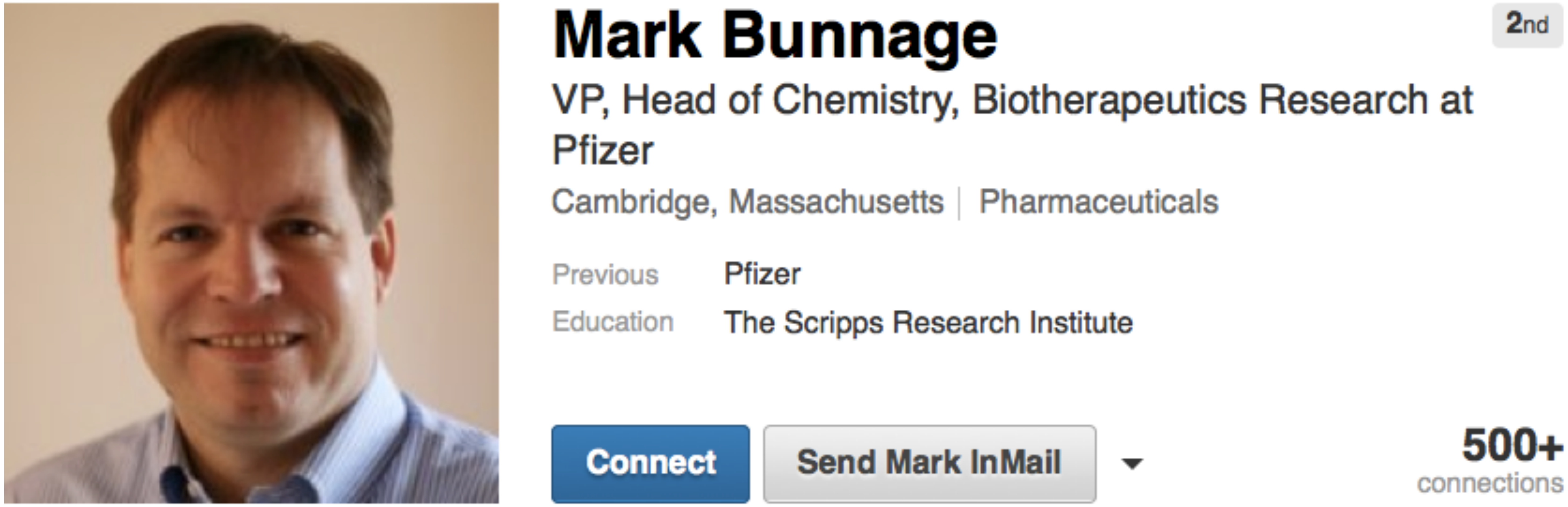


Adapted from Liao et al. BMJ 2015; 350 doi: 10.1136/bmj.h1885

A little historical perspective…

“Improving the quality of target selection is the single most important factor to transform industry productivity and bring innovative new medicines to patients.”

Bunnage, M.E. (2011) Getting pharmaceutical R&D back on target, *Nat Chem Biol*, **7**, 335-339.



Mark Bunnage

VP, Head of Chemistry, Biotherapeutics Research at Pfizer

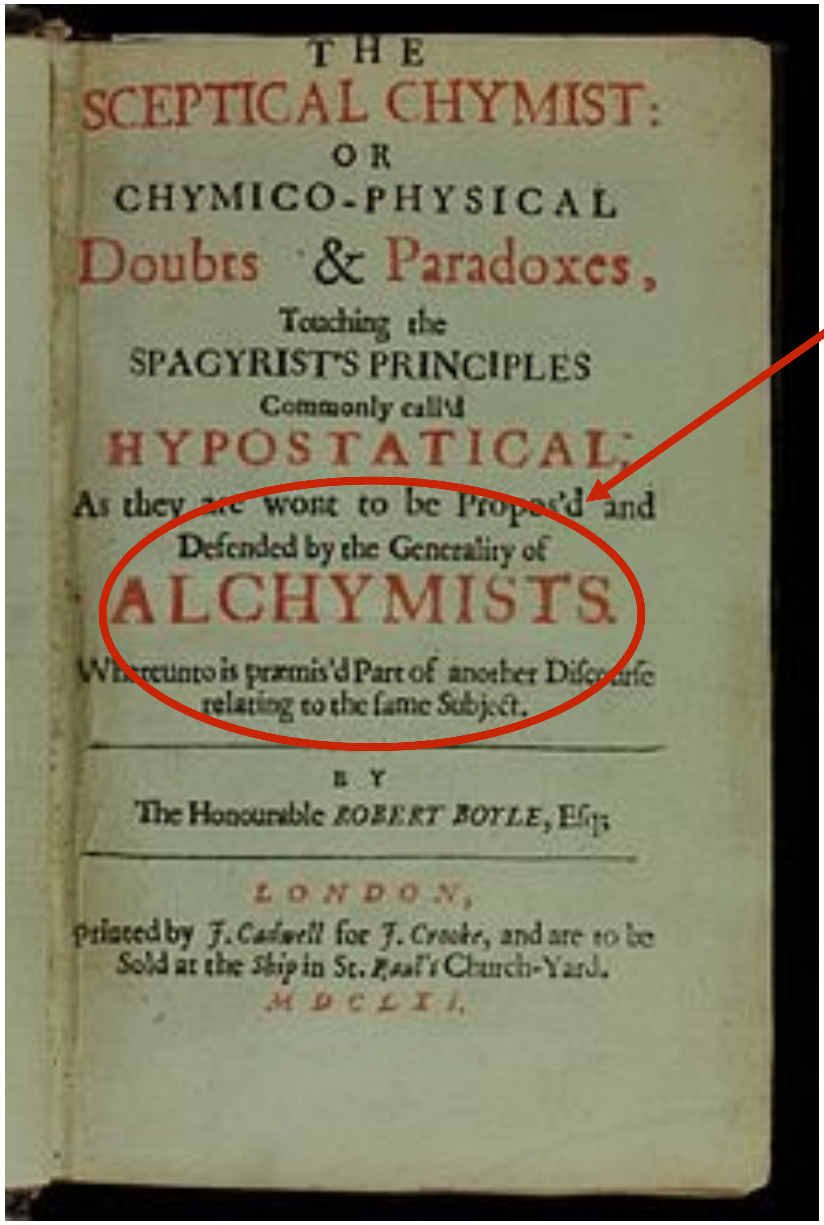
Cambridge, Massachusetts | Pharmaceuticals

Previous Pfizer

Education The Scripps Research Institute

500+ connections

Transparency c. 1660



THE
SCEPTICAL CHYMIST:
OR
CHYMICO-PHYSICAL
Doubts & Paradoxes,
Touching the
SPAGYRIST'S PRINCIPLES
Commonly call'd
HYPOSTATICAL;
As they are wont to be Propos'd and
Defended by the Generality of
ALCHYMISTS.
Whereunto is præmis'd Part of another Diſcourſe
relating to the ſame Subject.

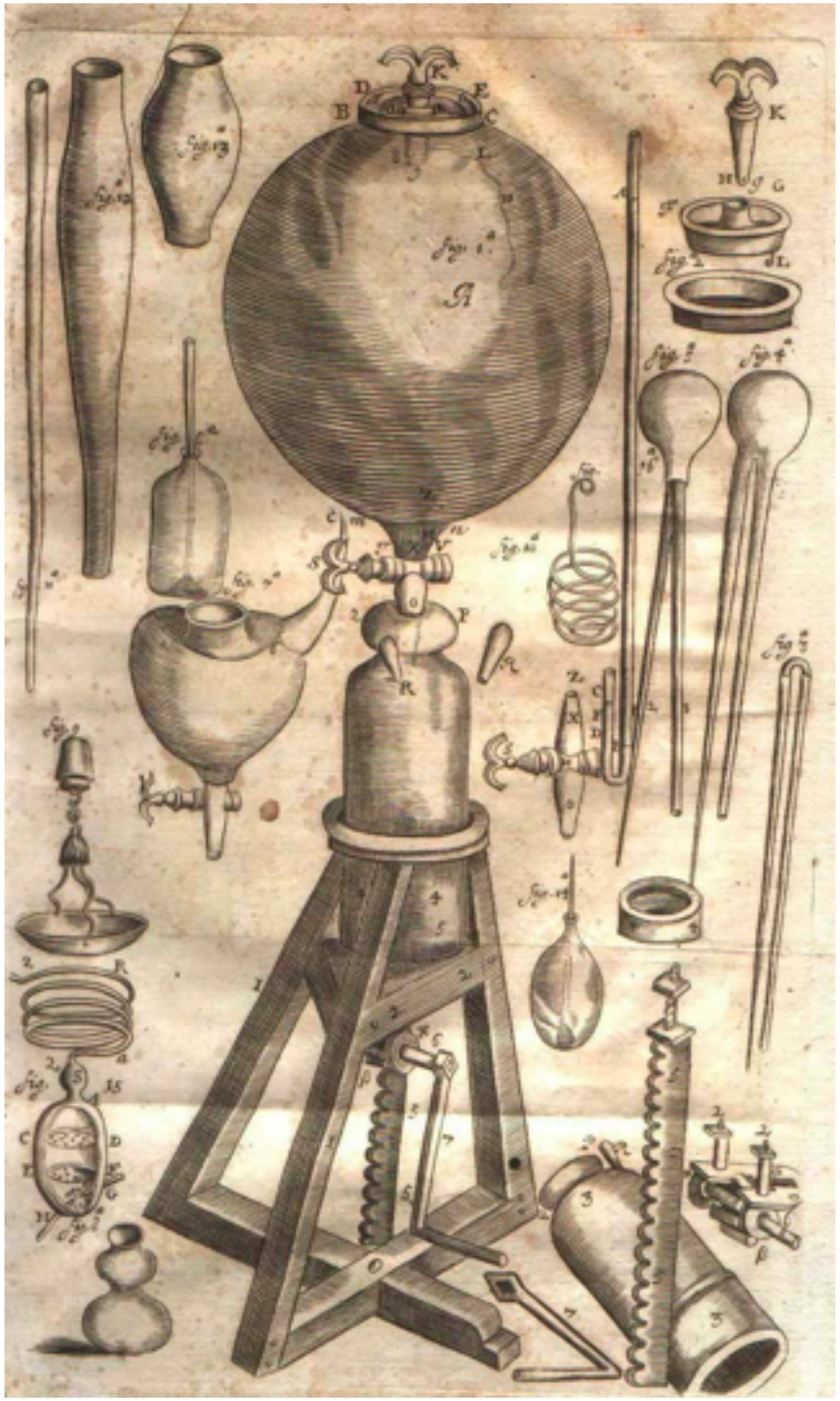
BY
The Honourable *ROBERT BOYLE*, Eſq;

LONDON,
Printed by *J. Cadwell* for *J. Crooke*, and are to be
Sold at the *Ship* in St. *Paul's* Church-Yard.
MDCLXI.

c. 1660: Robert Boyle and colleagues concerned *inter alia* with scientific fraud, e.g. “transformation of lead into gold”…

Boyle et al. promoted a “new natural philosophy” based on interrogating nature through **open experiment**…

Scientific facts will now be established by **public**, **reproducible** demonstration before a “jury of one’s peers”.



BOYLE: “We took a large and lusty frog and having included him in a small receiver we drew out the air not very much and left him very much swelled and able to move his throat from time to time - though not so fast as when he freely breathed before the exsuction (extraction) of the air. He continued alive about two hours that we took notice of, sometimes removing from one side of the receiver to the other, but he swelled more than before, and did not appear by any motion of his throat or thorax (chest) to exercise respiration. But his head was not very much swelled, nor his mouth forced open. After he had remained there somewhat above 3 hours, for it was not 3 hours and an half, perceiving noe signe of life in him, we let in the air upon him, at which the formerly tumid (swelled) body shrunk very much, but seemed not to have any other change wrought in it and though we took him out of the receiver yet in the free air it self, he continued to appear stark dead nevertheless to see the utmost of the experiment having caused him to be carried into a garden and layd upon the grass all night, the next morning we found him perfectly alive again.” (BP 18, fol. 127r)

adapted from Carusi 2015, “Virtual Witnessing”, in *Future of Research Communications & eScholarship,* Oxford UK*,* 11-12 January 2015.

Biologist Excuses for Not Sharing Data

- "If I publish my data I will get scooped."
- "I did all the work why should anyone else have any benefits?
- "It is ***my precious***…"

- "Eh…my postdoc has it somewhere…"

Data Science Extremism

- “All data must be published in RDF format.”
- “Column headers must be normalized to a formal ontology specified in W3C Web Ontology Language.”
- “So I can use all my cool semantic web tools on it.”

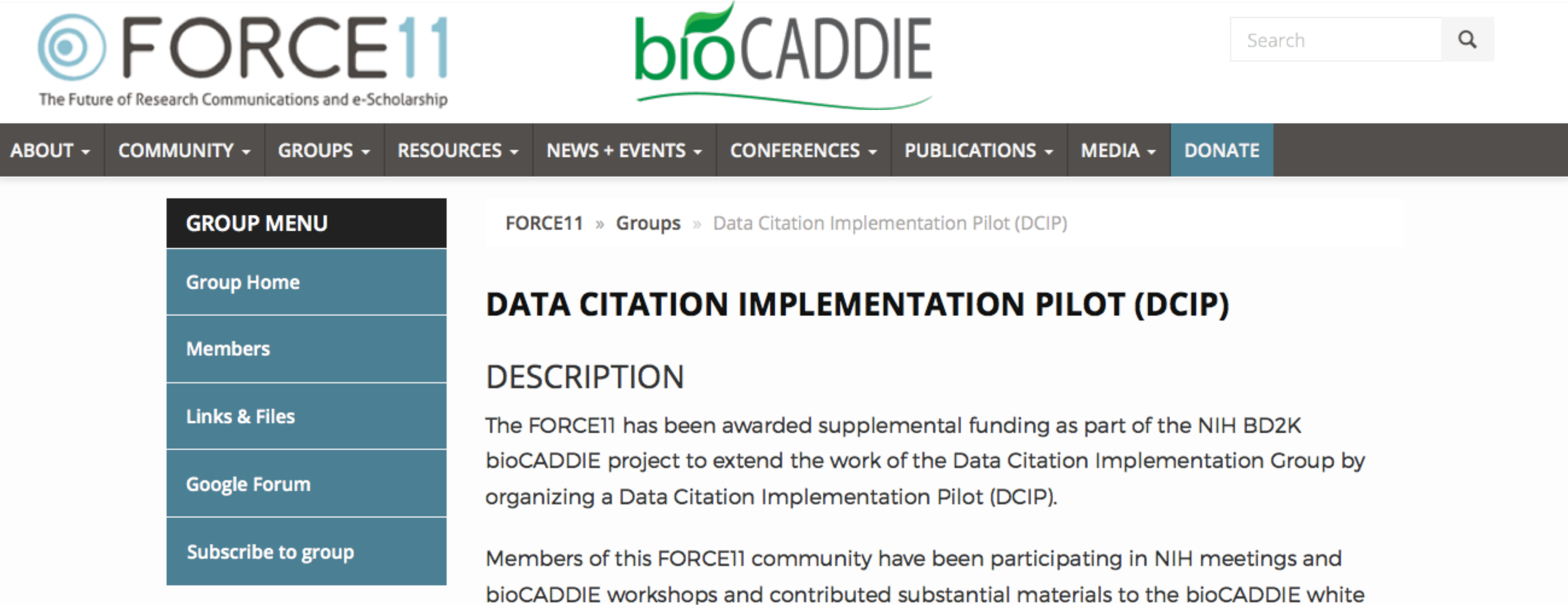
468 ontologies
6,435,788 classes

ENZOTYX.COM
JACO

Publishers

- Publishers are incentivized towards **open data**.
- Because:
  - You need **the article** to understand the data.
- Some are working toward very large **Big Data infrastructures** which they hope to **monetize**.

Incentives



FORCE11 » Groups » Data Citation Implementation Pilot (DCIP)

DATA CITATION IMPLEMENTATION PILOT (DCIP)

DESCRIPTION

The FORCE11 has been awarded supplemental funding as part of the NIH BD2K bioCADDIE project to extend the work of the Data Citation Implementation Group by organizing a Data Citation Implementation Pilot (DCIP).

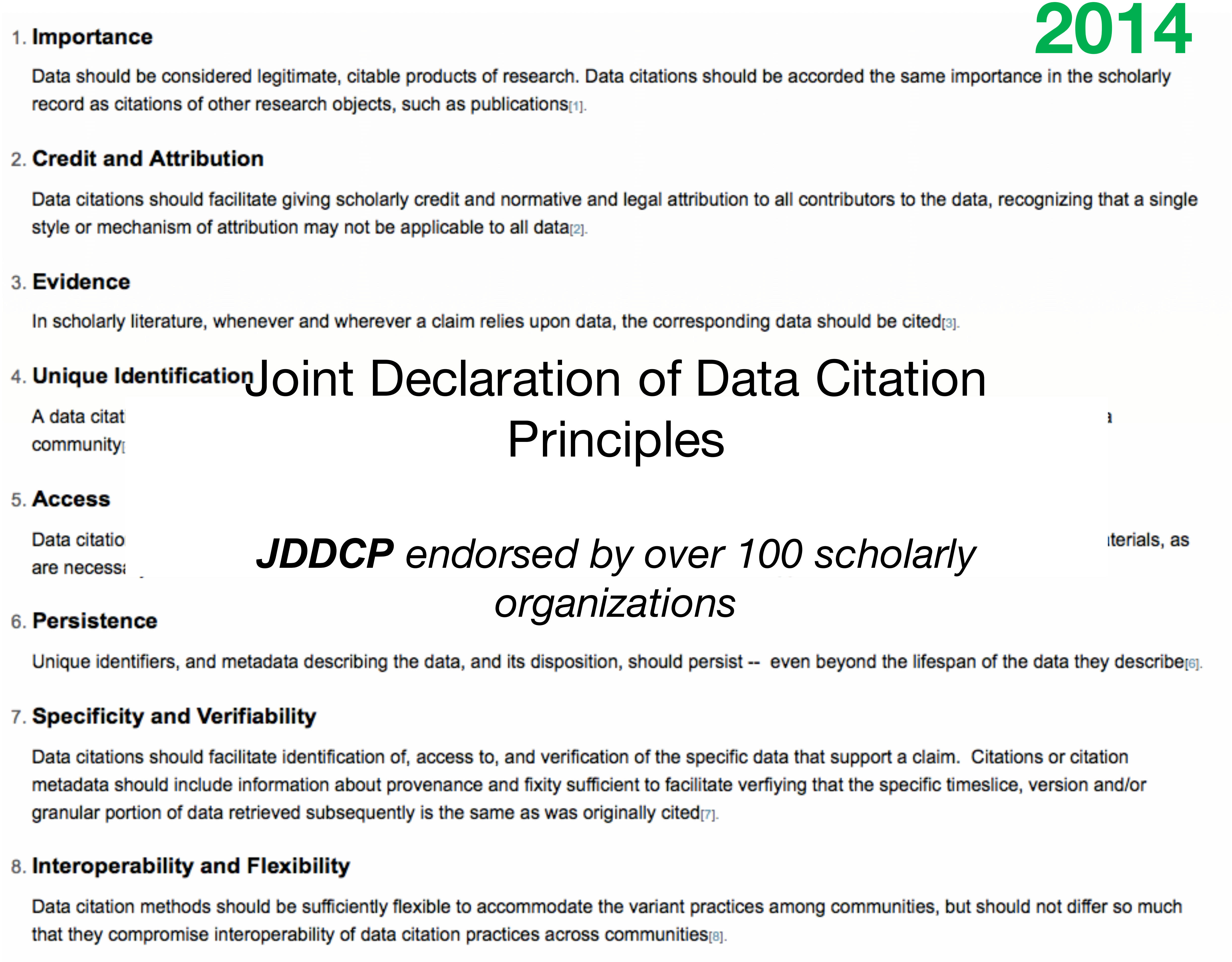
Members of this FORCE11 community have been participating in NIH meetings and bioCADDIE workshops and contributed substantial materials to the bioCADDIE white

https://www.force11.org/group/data-citation-implementation-pilot-dcip

2014

Joint Declaration of Data Citation Principles

***JDDCP** endorsed by over 100 scholarly organizations*



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[1].

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[2].

3. **Evidence**

   In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited[3].

4. **Unique Identification**

   A data citat … a
   community[

5. **Access**

   Data citatio … terials, as
   are necess

6. **Persistence**

   Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe[6].

7. **Specificity and Verifiability**

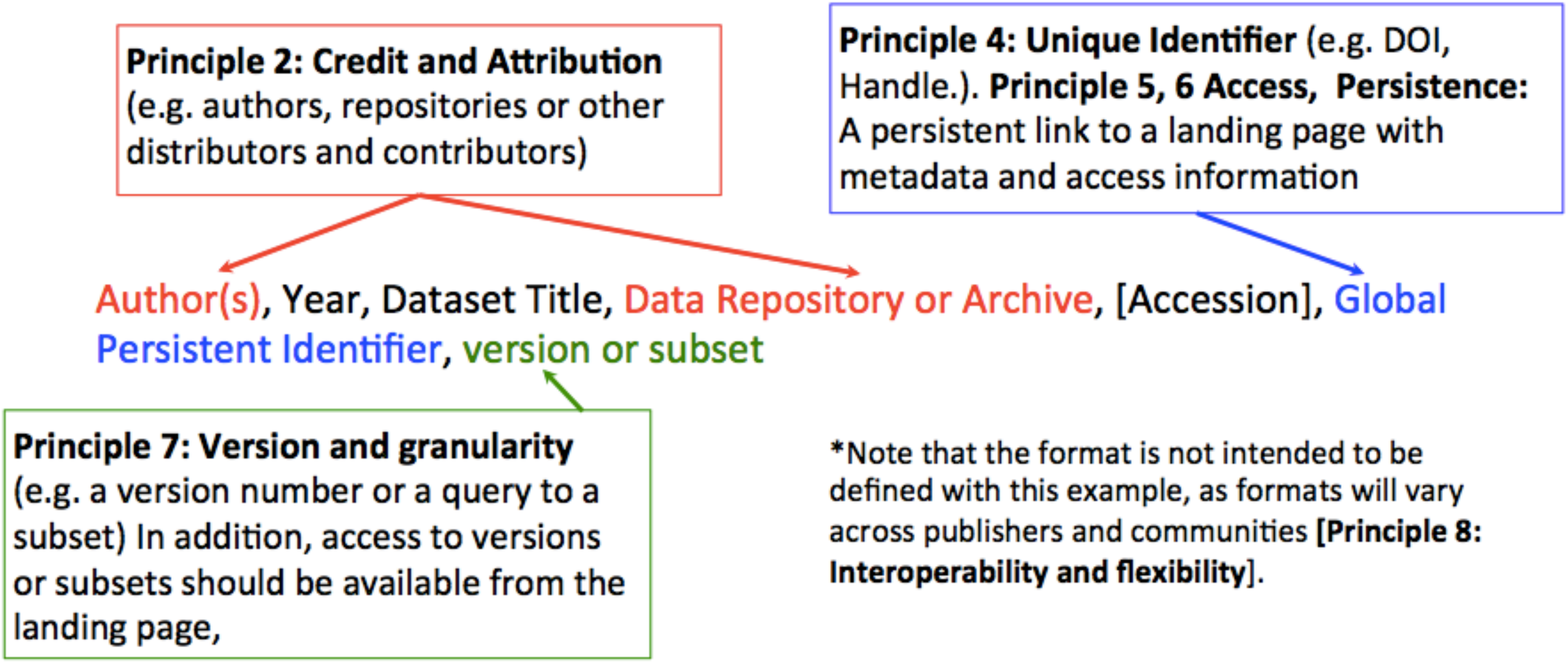
   Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verfiying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited[7].

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[8].

Data Citation Generic Example

example of a data citation as it would appear *in a reference list**



Principle 2: Credit and Attribution (e.g. authors, repositories or other distributors and contributors)

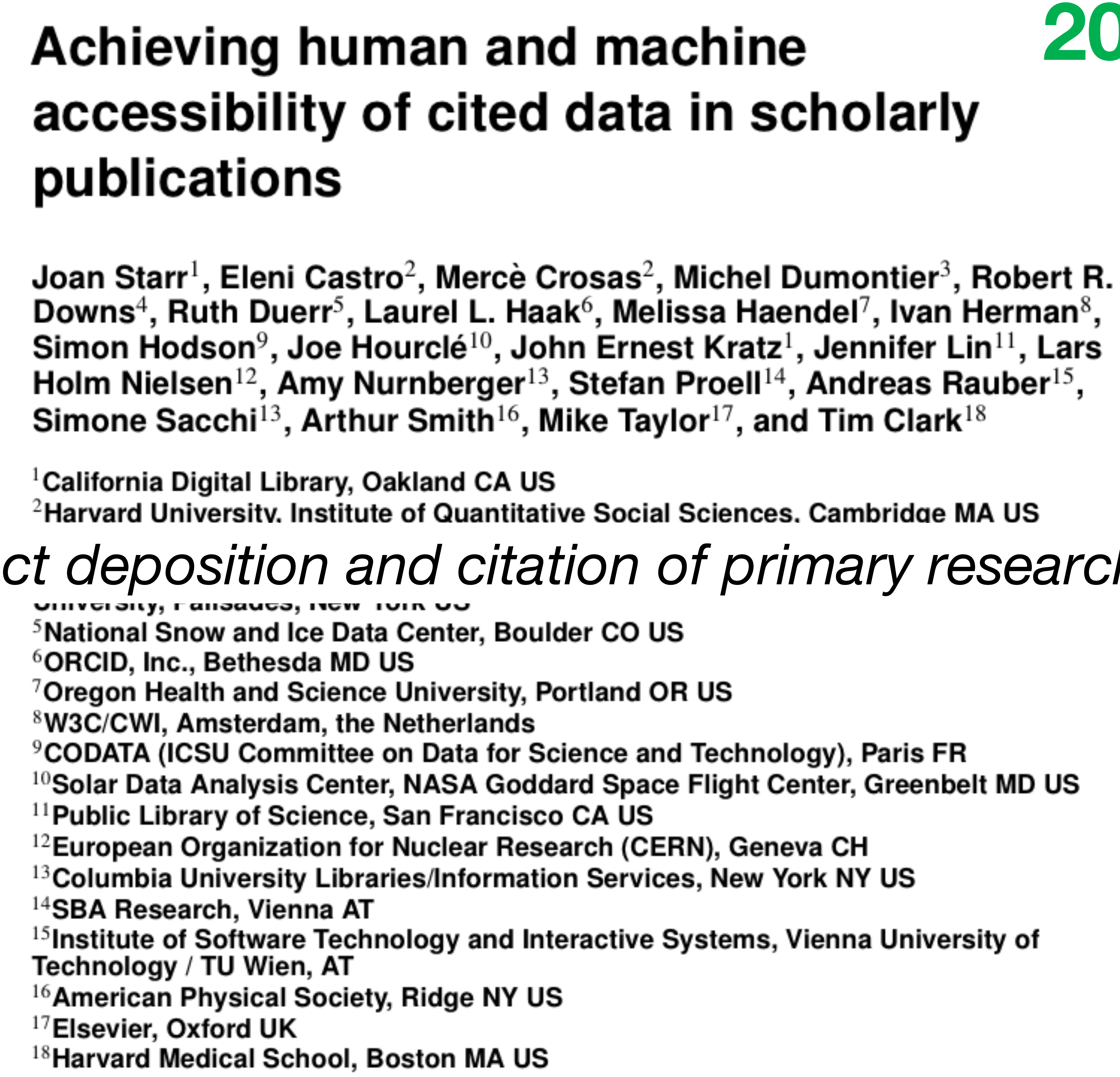
Principle 4: Unique Identifier (e.g. DOI, Handle.). **Principle 5, 6 Access, Persistence:** A persistent link to a landing page with metadata and access information

Author(s), Year, Dataset Title, Data Repository or Archive, [Accession], Global Persistent Identifier, version or subset

Principle 7: Version and granularity (e.g. a version number or a query to a subset) In addition, access to versions or subsets should be available from the landing page,

*Note that the format is not intended to be defined with this example, as formats will vary across publishers and communities **[Principle 8: Interoperability and flexibility]**.

2015



Achieving human and machine accessibility of cited data in scholarly publications

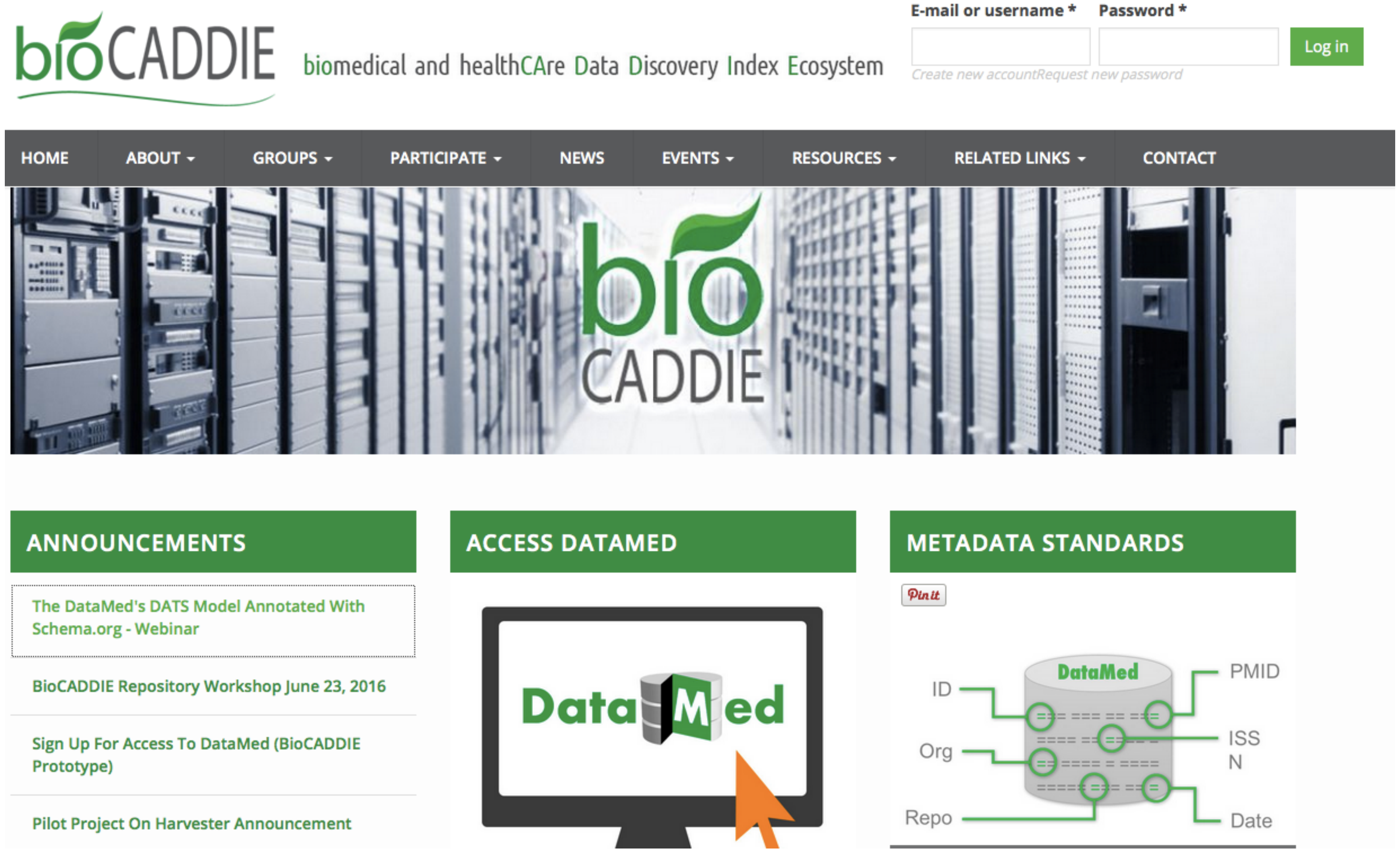
Joan Starr[1], Eleni Castro[2], Mercè Crosas[2], Michel Dumontier[3], Robert R. Downs[4], Ruth Duerr[5], Laurel L. Haak[6], Melissa Haendel[7], Ivan Herman[8], Simon Hodson[9], Joe Hourclé[10], John Ernest Kratz[1], Jennifer Lin[11], Lars Holm Nielsen[12], Amy Nurnberger[13], Stefan Proell[14], Andreas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

[1]California Digital Library, Oakland CA US
[2]Harvard University, Institute of Quantitative Social Sciences, Cambridge MA US
University, Palisades, New York US
[5]National Snow and Ice Data Center, Boulder CO US
[6]ORCID, Inc., Bethesda MD US
[7]Oregon Health and Science University, Portland OR US
[8]W3C/CWI, Amsterdam, the Netherlands
[9]CODATA (ICSU Committee on Data for Science and Technology), Paris FR
[10]Solar Data Analysis Center, NASA Goddard Space Flight Center, Greenbelt MD US
[11]Public Library of Science, San Francisco CA US
[12]European Organization for Nuclear Research (CERN), Geneva CH
[13]Columbia University Libraries/Information Services, New York NY US
[14]SBA Research, Vienna AT
[15]Institute of Software Technology and Interactive Systems, Vienna University of Technology / TU Wien, AT
[16]American Physical Society, Ridge NY US
[17]Elsevier, Oxford UK
[18]Harvard Medical School, Boston MA US

Direct deposition and citation of primary research data

2016

Data Citation Implementation Pilot



bioCADDIE biomedical and healthCAre Data Discovery Index Ecosystem

E-mail or username * Password * Log in

Create new account Request new password

HOME ABOUT ▾ GROUPS ▾ PARTICIPATE ▾ NEWS EVENTS ▾ RESOURCES ▾ RELATED LINKS ▾ CONTACT

ANNOUNCEMENTS

- The DataMed's DATS Model Annotated With Schema.org - Webinar
- BioCADDIE Repository Workshop June 23, 2016
- Sign Up For Access To DataMed (BioCADDIE Prototype)
- Pilot Project On Harvester Announcement

ACCESS DATAMED

METADATA STANDARDS

FORCE11

The Future of Research Communications and e-Scholarship

Pilot Strategic Objectives

a. Provide coordination & guidance for early adopters.
b. Help establish benchmark implementations.
c. Focus on archiving and citing primary research data.
d. Provide report on lessons learned to the community.
e. Make cited data discoverable.
f. Life sciences and biomedical domain.

Major Outputs

a. Identifiers: harmonization CDL / EBI.

b. Publishers: roadmap to data citation.

c. Repositories: implement landing page metadata for data citation.

d. FAQs: guidance for common implementations based on JDDCP.

Some Participants

- PLoS, Elsevier, Nature, BioMed Central, IOS Press, F1000 Research, GigaScience.
- European Bioinformatics Institute, National Library of Medicine, Dryad, FigShare, Dataverse.
- Harvard University, Columbia University, UCSD
- CrossRef, DataCite, California Digital Library

Participants

npg

University of California
CDL
California Digital Library

DataCite

NATIONAL LIBRARY OF MEDICINE

BioMed Central
The Open Access Publisher

Springer

(GIGA)^n SCIENCE

And you!

F1000Research
Open for Science

The Dataverse Project

ELSEVIER

THE EMBO JOURNAL

IOS Press

eLIFE

crossref

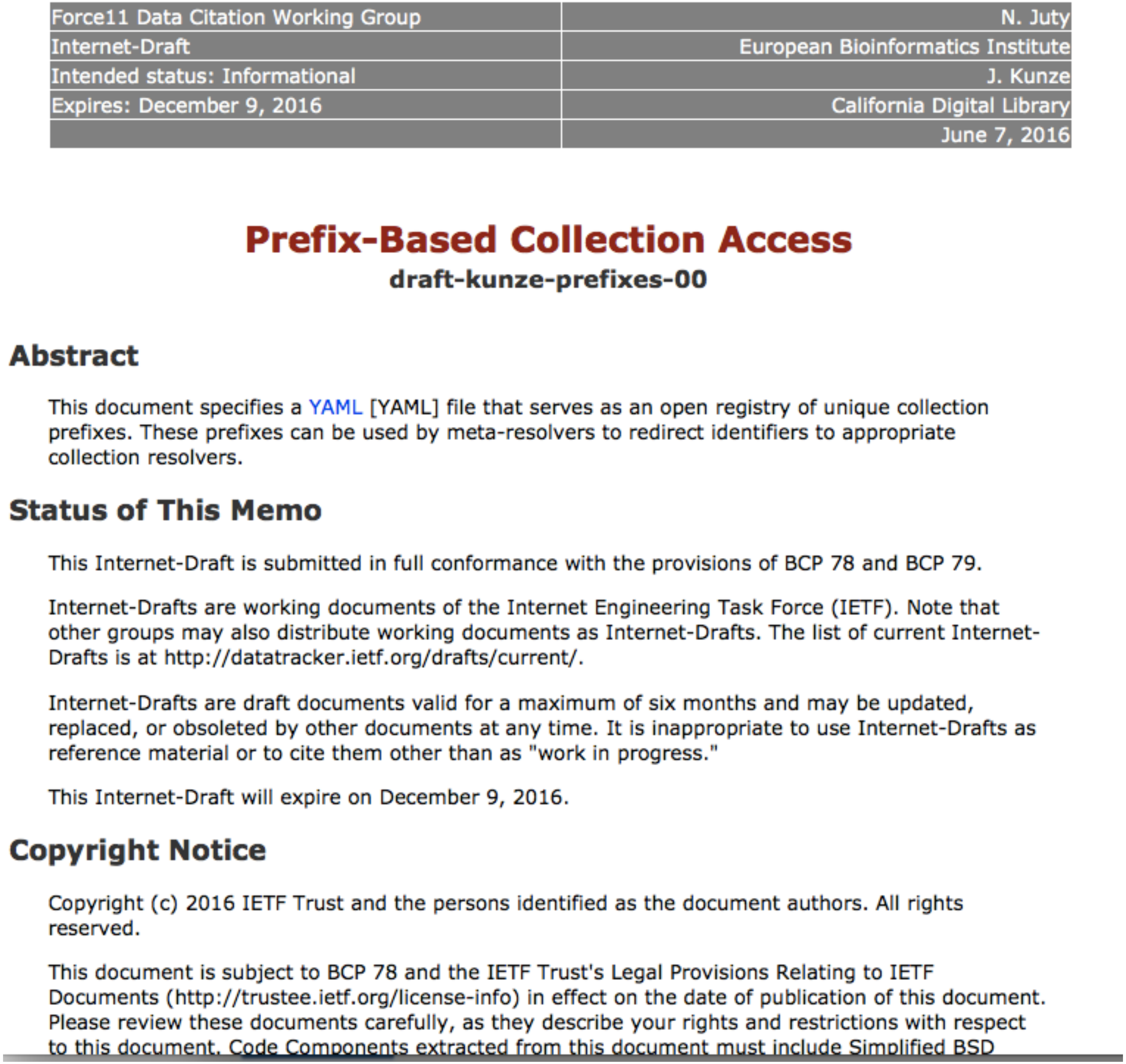
EMBL-EBI

Identifier Harmonization Group

- California Digital Library (EZID / Name2Thing)
- European Bioinformatics Institute (identifiers.org)
- co-representation from ELIXIR, BioCADDIE, NIH
- Harmonize identifier resolution for all standard bioinformatics databases across EU & US
- Workshop @ Harvard on **June 2**

DCIP Identifiers Workshop, June 2, 2016, Harvard University, Cambridge MA

John Kunze (CDL), Niall Beard (Manchester), Tim Clark (Harvard),Nick Juty (EBI), Ian Fore (NIH), Julie McMurry (UCSB), Jeff Grethe (UCSD), Rafa Jimenez (ELIXIR), Sarala Wimalaratne (EBI)



Force11 Data Citation Working Group	N. Juty
Internet-Draft	European Bioinformatics Institute
Intended status: Informational	J. Kunze
Expires: December 9, 2016	California Digital Library
	June 7, 2016

Prefix-Based Collection Access

draft-kunze-prefixes-00

Abstract

This document specifies a YAML [YAML] file that serves as an open registry of unique collection prefixes. These prefixes can be used by meta-resolvers to redirect identifiers to appropriate collection resolvers.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on December 9, 2016.

Copyright Notice

Early Adopter Repositories

- Leads: Martin Fenner & Mercè Crosas
- Workshop **June 22 @ UCSD** precedes BioCADDIE Repositories Outreach meeting.
- Goal: develop proposed landing page metadata and outreach plan for repository adoption.
- Also Discuss - extension of metadata work to schema.org.

Publishers

- Leads: Amye Kenall & Helena Cousijn
- Elsevier, SpringerNature, eLife, PLoS, et al.
- Outreach to other publishers in progress.
- Workshop **July 22 @ SpringerNature (London)** to develop Publishers Roadmap for data citation.

DCIP Executive

- Maryann Martone, Hypothesis and UCSD, co-Chair
- Tim Clark, Harvard Medical School, co-Chair
- Carole Goble, The University of Manchester & ELIXIR
- Jeffrey Grethe, UCSD and bioCADDIE
- Jo McEntyre, EMBL-EBI & ELIXIR
- Joan Starr, California Digital Library
- Martin Fenner, DataCite
- Simon Hodson, CODATA
- Chun-Nan Hsu, UCSD

Conclusions

- We need to systematically cite data for improved scientific transparency, reproducibility, robustness.
- Persistent discoverable data archives with cited data will enhance capability for validation & re-use.
- Goal: significantly improve biomedical translation.
- BioCADDIE / FORCE11 data citation pilot will promote implementing data citation in journals at scale.

References

1. Collins FS, Tabak LA: **Policy: NIH plans to enhance reproducibility**. *Nature* 2014, **505**(7485):612
2. Uhlir P: **For Attribution - Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop (2012)** In.: The National Academies Press; 2012: 220 [http://www.nap.edu/catalog.php?record_id=13564].
3. CODATA/ITSCI Task Force on Data Citation: **Out of cite, out of mind: The Current State of Practice, Policy and Technology for Data Citation**. *Data Science Journal* 2013, **12**:1-75
4. Data Citation Synthesis Group: **Joint Declaration of Data Citation Principles**. Edited by Martone M. San Diego CA: Future of Research Communication and e-Scholarship (FORCE11); 2014 [https://www.force11.org/datacitation].
5. NIH: **About BD2K**. Accessed October 28, 2015. [https://datascience.nih.gov/bd2k/about].
6. JATS Standing Committee. **NISO JATS Standing Committee Recommended Changes Between NISO/JATS 1.0 and JATS 1.1d3**. March 2015. http://www.niso.org/apps/group_public/download.php/14543/JATS-SC-Recommendations-1.1d3.pdf
7. Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, Haak LL, Haendel M, Herman I, Hodson S,´ JH, Kratz JE, Lin J, Nielsen LH, Nurnberger A, Proell S, Rauber A, Sacchi S, Smith A, Taylor M, Clark T: **Achieving human and machine accessibility of cited data in scholarly publications**. *PeerJ* 2015, **1:** e1.
8. **Data Citation Implementation Pilot,** FORCE11.org. Accessed April 11, 2016. http://bit.ly/1TNQxZH
9. Scannell et al. 2012. Nat Rev Drug Discov, 2012;11(3):191–200 [9].
10. Begley CG and Ellis LM*, Nature* 2012, 483(7391):531-533
11. Paul, S.M.*, et al.* (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge, *Nat Rev Drug Discov*, 9, 203-214.
12. Bunnage, M.E. (2011) Getting pharmaceutical R&D back on target, *Nat Chem Biol*, **7**, 335-339.
13. LeCun et al. **Deep Learning.** Nature 521, 436–444. doi:10.1038/nature14539
14. Liao et al. **Development of phenotype algorithms using electronic medical records and incorporating natural language processing.** BMJ 2015; 350 doi: 10.1136/bmj.h1885