Assessing Value of Biomedical Digital Repositories

Chun-Nan Hsu
Department of Medicine
Biomedical Informatics
University of California, San Diego

bioCADDIE

FORCE11
The Future of Research Communications and e-Scholarship

UNIVERSITY OF CALIFORNIA · SAN DIEGO ·

Impact vs. Influence

- Impact: Actual changes that the work brings to the filed.
  - outcomes, practices, and methodologies
- Influences: How widely the work has been disseminated and viewed.
- High impact not proportional to high influence, and vice versa

Measure vs. Mention

- Traditional citation metrics measure *mentions*
- Digital repositories allow access *measures*
- *Mentions* reflect influences better
- *Mentions* not always imply actual use and impacts
- Measures and mentions not always correlate

RCSB Protein Data Bank (PDB)

PDB assigns each protein structure a PDB ID and their corresponding primary citations

Highly cited PDB entries differ from highly mentioned entries

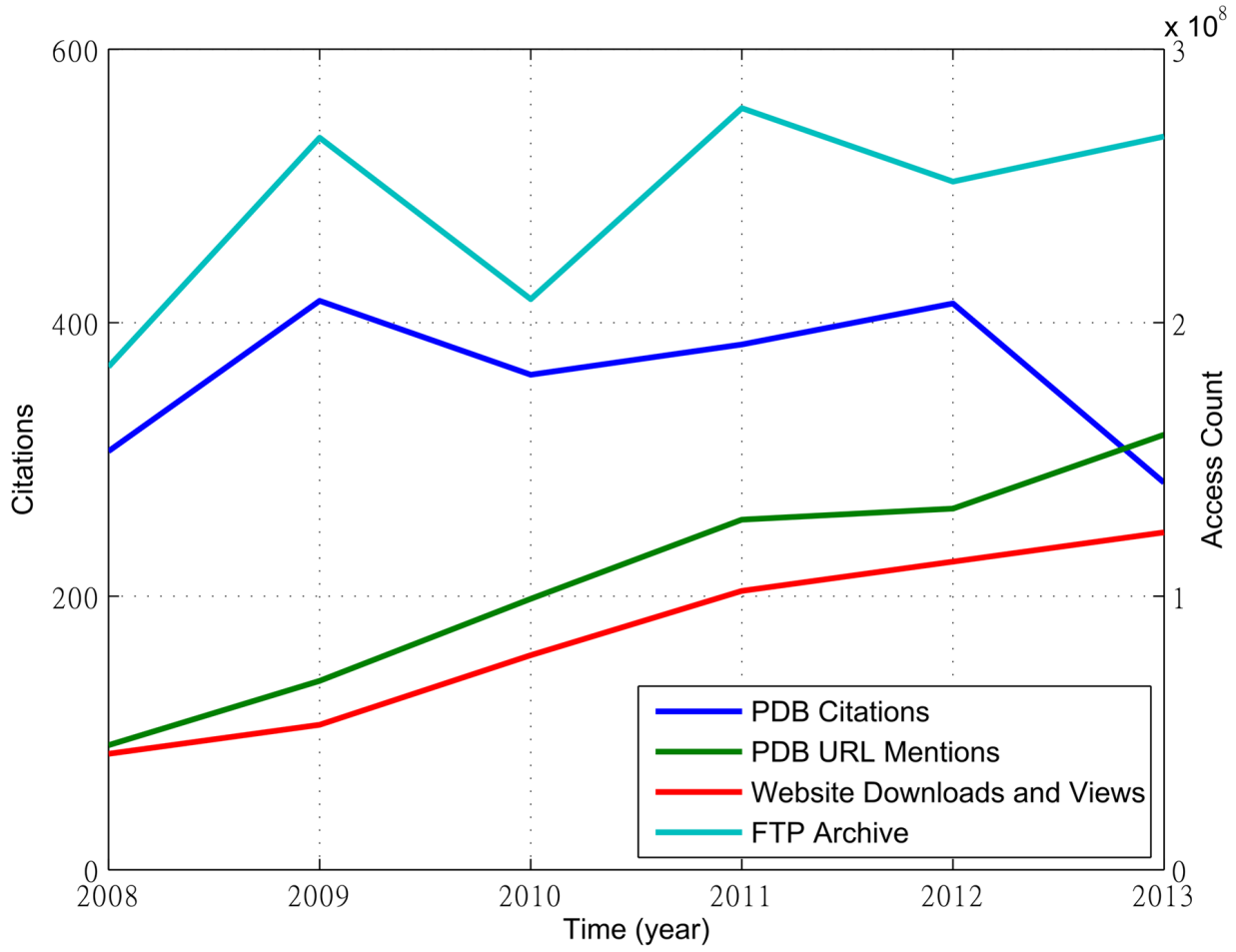
Paper citations	PDB ID mentions	Web access (http views)	FTP access
1BL8	2RH1|62	3GCB|619704	1SMW|73305
1F88	1UBQ|45	**1ATP|529687**	2HQT|65162
1GC1	3EML|35	3RAT|475538	2BDI|56950
1FFK	2R9R|34	3PIC|414549	100D|55744
1RV1	1K4C|33	1CRN|341635	3REC|51584
2A79	**2A79|32**	1A00|281788	2VXJ|51375
2RH1	2B4C|32	1I0Z|255816	2E7S|50580
1AIK	1KX5|30	1HSG|193644	2BX5|49845
1YSG	2VT4|30	1JGN|187364	1JUS|48526
1ENV	1U19|27	3QB5|186577	2FJF|47238
1GIX, 1GIY	1JJ2|27	2RH1|186374	3DTD|46606
1SFC	1J5E|24	1CBS|184458	2GX2|46346
1J5E	3DNA|24	1I10|179304	12E8|46262
2MYS	1YCR|23	3KUS|170478	1ILU|45114
1ORQ	1JFF|23	1TUP|166056	2DG0|44850
1K4C	2AW4|23	3KUT|158233	1JT6|44503
3CSS	2J00|22	1DWF|157638	2FK3|44143
2BG9	**1F88|22**	4HHB|156361	1KLF|43570
1HTM	**1ATP|21**	1EMA|152395	2ZHX|43502
1TSR, 1TUP	2AVY|21	1HEW|150732	2VUT|42880

Citing entries in PDB

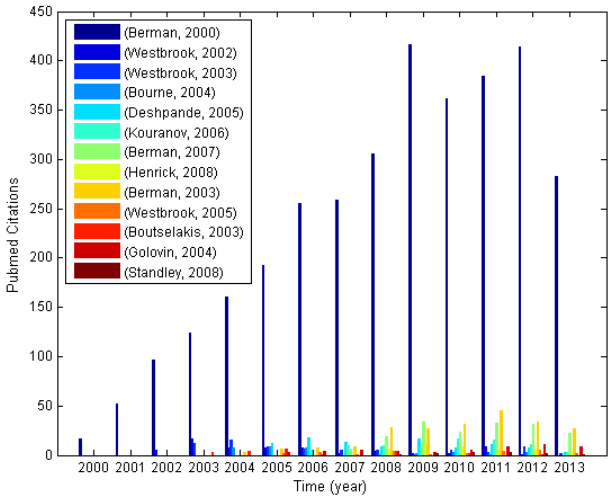
Identifier	Example	Machine Readable	Mentions (*)	%
PDB ID	PDB ID: **1STP**	Y	14,888	4.8
PDB DOI	http://dx.doi.org/10.2210/pdb**1stp**/pdb	y	155	0.05
External Link Tag	<ext-link .. ext-link-type="pdb" xlink:href="**1STP**">	y	32,108	10
PDB File Name	1stp.pdb	y	895	0.03
PDB URL	http://www.rcsb.org/../structureId=1stp	y, but URL may change	657	0.2
Non-standard PDB ID	PDB code: **1STP**, PDB reference **1STP**, PDB accession number **1STP**, Many variations …	y/n	22,081	7.1
PDB in Context	We employed the following **PDB** coordinates: glycogen phosphorylase, **1gpy** …	y/n with NLP or ML	16,726	5.4
Free Text	We first placed S2 bound to human PI3KC; (**3ene**) into the reference coordinates …	y/n with NLP or ML	221,287	72 (incl. many false positives!)

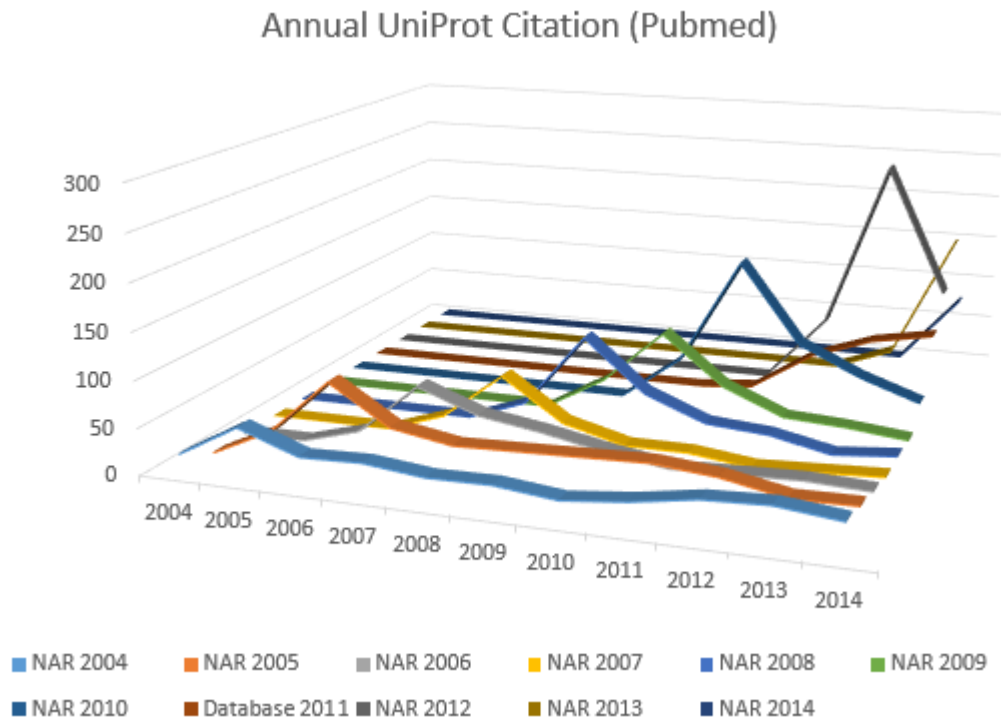
* Preliminary data, includes duplicate mentions within same article

How does data usage statistics correlate to paper citations and URL mentions?



Authors tend to follow the instructions of “how to cite”





PDB Identifier is not unique

- Currency: Each participant received Ksh 200 (**1USD** = +/-75Ksh) as
- Year: were abruptly interrupted in **1914** with the (example of an integer PDB ID!)
- Postal code: 385 Euston Road, London, NW1 **3AUT**, UK
- Room number: 110 Irving Street NW, Room **2A56**, Washington
- Floating point number: 1E10, 1D10
- Grant type: Parent study (NIH R01 NR04749; NIH **2R01** NR04749).
- Catalog number: selective detergent method kit (ultra HDL) cat no. **3K33** supplied by
- Chemical formulas: ellipsoid plot of Zn(H2O)2(C5H5N3O2)2 **2NO3** at the 50%
- Chemical name: Glycolysis under anaerobic condition produces **2ATP** per molecule
- Gene name: The polymorphisms of cytochrome P450 **2C19** (CYP2C19) gene
- Antibody: The primary detection antibody was unlabelled Mab **4B11**
- Technique: were subjected to 2D-gel electrophoresis (**2DGE**)
- Technique: when the recommendations of the NMR and **3DEM** VTFs are
- Instrument: using an Olympus Inverted Microscope (Olympus **1X71**, Tokyo, Japan)
- Instrument: were obtained with Hamamatsu C5810 color chilled **3CCD** camera
- Software: involved in base-pairing as computed by the **3DNA** program
- Software: domain definitions from SCOP, CATH, DALI, **3DEE**, and MMDB are

- Identifier needs a prefix to minimize ambiguities
- Tagging in text document will further disambiguate identifier

Standardizing Mentions and Use

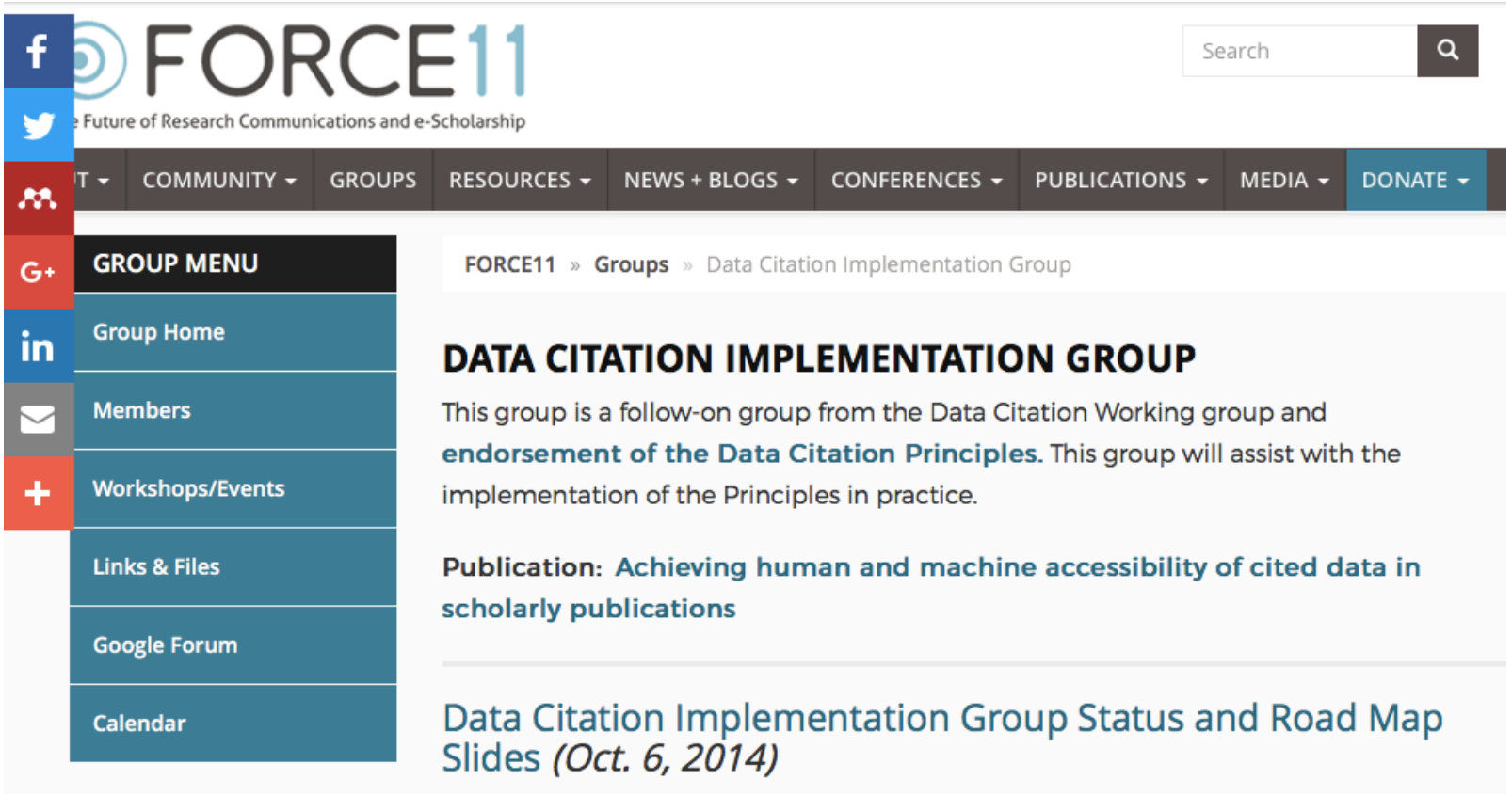
- Research Resource Identifier Initiative
- Data Citation Implementation Project

Research Resource Identifier Initiative (#RRID)

RRID dissemination

- The project has been running since 2014.
- RRID entries
  - 13K software tools/databases
  - 2M antibodies
  - ~500K model animals
- Over 1226 papers have appeared with RRID's from over 160 biomedical journals.
- Cell Press has just adopted the standard (http://www.cell.com/star-methods)
- eLife and the Endocrine Society just announced that they will be strongly encouraging authors to use RRID's in their journals.
- BMC, PLoS, Elservier, and more.

Data Citation Implementation Project (DCIP)



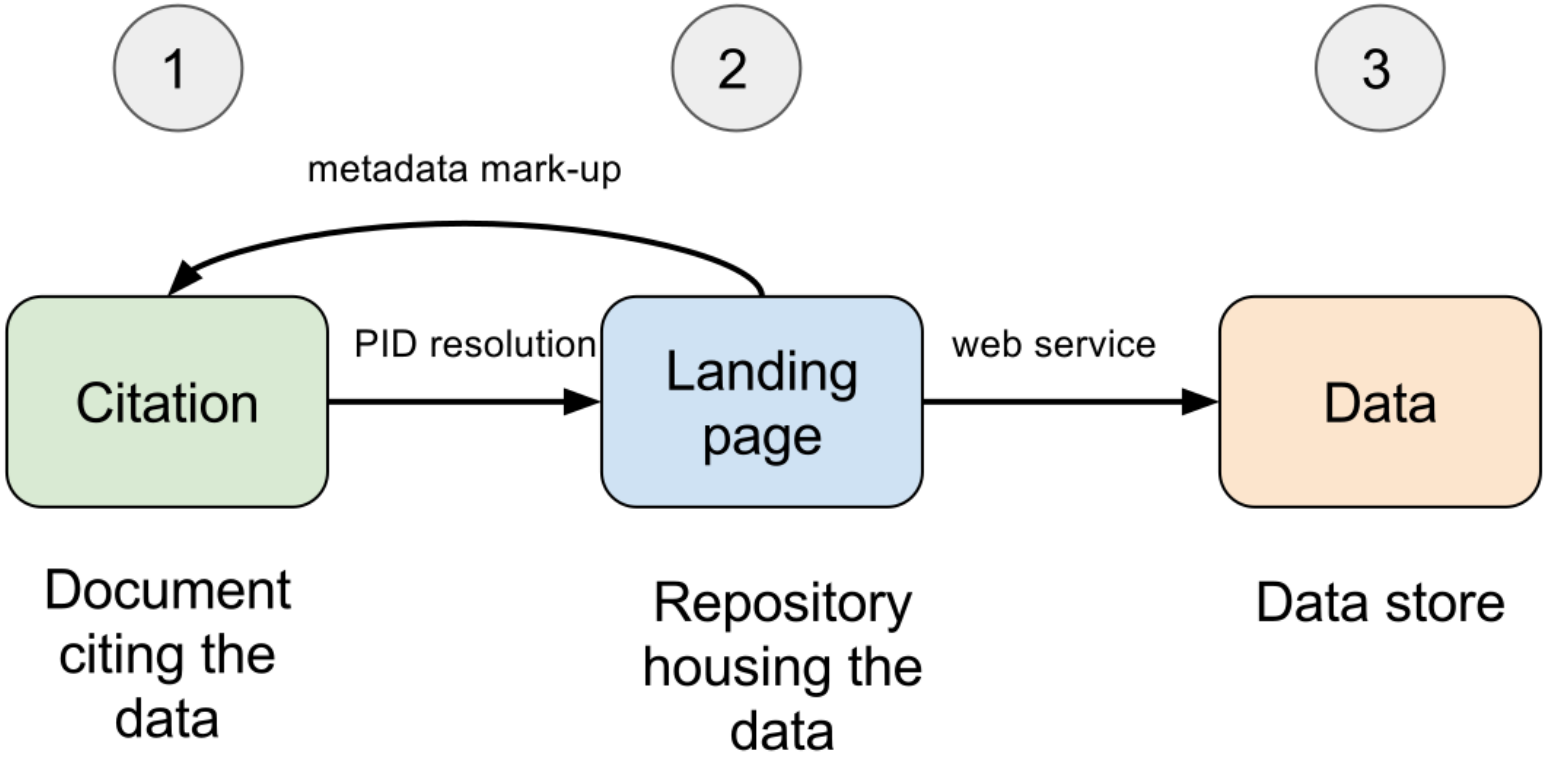
Joint Declaration of Data Citation Principles (JDDCP)

The Principles

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

https://www.force11.org/datacitation

Landing Page



Towards Reliable and Accurate Metrics

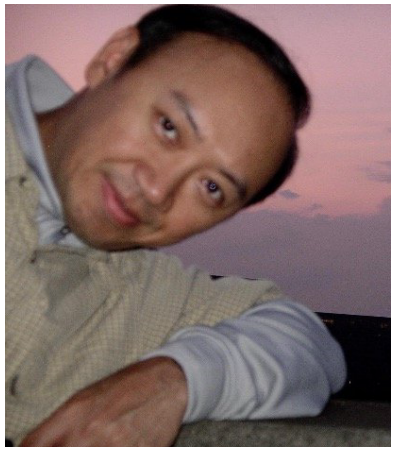
- Standardized RRID and Data Citation may not be the single perfect metric but wide adaptation of these standards will definitely lead to a more reliable and comparable metric than the status quo
- Also: Cite a data repository by
  - Distinguishing *actual use* and *merely related*
  - Distinguishing positive or negative sentiment wrt the cited resource
- Standard brings to unambiguous and persistent references to digital repositories

Funding statement

The research reported here is supported in part by

- Grant U24AI117966 National Institutes of Health Big Data to Knowledge (BD2K) Initiative,
- U24DK097771 National Institute of Diabetes and Digestive and Kidney Diseases, and
- U24DA039832 National Institute on Drug Abuse.

Contributors



Chunnan Hsu

Anita Brandowski

Jeff Grethe

Maryann Martone

Thank you for your attention!