

Publishing of Scientific Data

Jodi Schneider

jodi.schneider@deri.org

Twitter @jschneider

SFI Summit
2010-11-16
Athlone, Ireland



Data deposit may be required



Digital Enterprise Research Institute

www.deri.ie

■ Community norms

- Crystallography, astronomy, genomics, ...

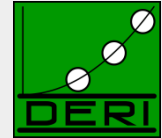
■ Peer-review and publication

- *Nature*: “Supporting data must be made available to editors and peer-reviewers **at the time of submission...**”

■ Funders

- NSF **proposals** must include a 2-page Data Management Plan

Data citation



- “Cite this paper if you use my dataset”
- DOI, handle, Repository ID
 - Tracking reuse is hard! <http://bit.ly/doi-fail>
- Universal Numerical Fingerprint (UNF)
 - Changes when the data does
 - Cryptographic hash of the data content
 - Micah Altman, Gary King (2007). “A Proposed Standard for the Scholarly Citation of Quantitative Data”. D-Lib 13(3/4) <http://www.dlib.org/dlib/march07/altman/03altman.html>
 - UNF:3:DaYIT6QSX9r0D50ye+tXpA==

Data *itself* as publication?



Digital Enterprise Research Institute

www.deri.ie

- Data-only journals
 - *Earth System Science Data*
- Databases as a research product
 - Ph.D. curators extracting information from papers
- Machine recording of experiments
 - Open Notebook Science
- Integration of data into publications
 - Phil Bourne (2005) **Will a Biological Database Be Different from a Biological Journal?** PLoS Comput Biol 1(3): e34. doi:[10.1371/journal.pcbi.0010034](https://doi.org/10.1371/journal.pcbi.0010034)

Interactive Data inside the PDF



Digital Ent

QuickTime Player File Edit View Share Window Help Stop Recording Utopia Documents - Calling International Rescue: knowledge lost in literature and data landslide! Sat 11:09

www.deri.ie

Rescuing knowledge lost in literature and data 325

Table 2 - Apparent permeability (P_{app}) with our artificial membrane (A.M.), % of drug recovery (R%), human fraction absorbed (F_a), apparent permeability with Caco-2 and PAMPA, octanol/water partition coefficient ($\log K_{ow}$) and distribution coefficient ($\log D$) for the compounds

Compound	A.M. P_{app} ($\times 10^{-3} \text{ cm s}^{-1}$) \pm S.D.	R%	F_a % ^a	Caco-2 ^b P_{app} ($\times 10^{-3} \text{ cm s}^{-1}$)	PAMPA ^c P_{app} ($\times 10^{-3} \text{ cm s}^{-1}$)	$\log K_{ow}$ ^d	$\log D^e$
1. Chlorothiazide	0.86 \pm 0.04	99.1	53	0.015	0.13	-0.24	-0.05
2. Aciclovir	0.91 \pm 0.02	99.9	21	0.025	0.00	-1.74	-1.86
3. Naloxol	1.37 \pm 0.03	95.5	32	0.388	0.00	0.71	0.68
4. α -Methyl-dopa	0.32 \pm 0.01	97.1	41	0.015	0.00	-1.80	-1.80
5. Atisocid	2.09 \pm 0.10	98.6	52	0.020	0.00	0.16	-1.29
6. Ranitidine	2.15 \pm 0.03	99.9	55	0.049	0.05	0.27	-0.29
7. Metformin	2.27 \pm 0.20	97.3	55	0.950	0.00	-1.43	-1.22
8. Furosemide	2.75 \pm 0.02	99.0	60	0.012	0.06	2.29	-0.69
9. Hydrochlorothiazide	3.10 \pm 0.05	98.3	70	0.051	0.00	-0.07	-0.12
10. Chloramphenicol	3.97 \pm 0.01	99.3	90	2.05	0.17	1.14	1.14
11. Hydrocortisone	4.28 \pm 0.07	99.8	91	1.40	0.34	1.61	1.55
12. Pindolol	3.74 \pm 0.07	99.2	92	1.67	0.49	1.75	0.19
13. Propranolol	3.97 \pm 0.08	99.8	93	4.19	2.35	1.25	1.25
14. Metoprolol	4.81 \pm 0.08	99.6	95	2.37	0.35	1.88	-0.16
15. Thiopropine	4.05 \pm 0.06	99.1	97	2.52	0.48	-0.29	-0.05
16. Thimethoprim	4.55 \pm 0.09	99.8	97	8.30	0.50	0.91	0.74
17. Naproxen	6.88 \pm 0.02	98.9	98	3.95	1.06	3.18	0.23
18. Verapamil	4.16 \pm 0.03	97.5	98	1.38	0.74	3.79	2.66
19. Acetylsalicylic acid	4.91 \pm 0.03	97.8	100	2.82	1.32	0.88	0.34
20. Ketoprofen	4.27 \pm 0.08	99.1	100	2.01	1.67	3.12	-1.51
21. Caffeine	4.11 \pm 0.08	99.3	100	3.08	1.08	-0.07	0.02

^a Literature F_a values (Chiu et al., 2000; Zhu et al., 2002).
^b Literature Caco-2 P_{app} values (Absenz and Haerel, 2003; Nicklin et al., 1996; Yamashita et al., 2000; Yazdani et al., 2004; Zhu et al., 2002).
^c Literature PAMPA P_{app} values (Gugano et al., 2001; Zhu et al., 2002).
^d Literature $\log K_{ow}$ values (Cheng et al., 2004; Moffat et al., 2003; Ziu et al., 2002).
^e Literature $\log D$ values (Moffat et al., 2003; Nicklin et al., 1996; Zhu et al., 2002).

Figure 9 Lynch imagines being able to toggle between a published table of numerical values and their graphical representation

For readers viewing this article using UD, from this typical table of data from the *European Journal of Pharmaceutical Sciences* [62], explore the result of clicking on the UD logo. Reproduced from Corti, G., Maestrelli, F., Cirri, M., Zerrouk, N. and Mura, P. (2006) Development and evaluation of an *in vitro* method for prediction of human drug absorption II. Demonstration of the method suitability. *European Journal of Pharmaceutical Science* 27, 354-362, Copyright (2006) with permission from Elsevier.

there are also parallels here with successful social/collaborative annotation models such as Wikipedia.

This project aims to exploit emerging Web technologies to spur a transition away from traditional 'solid' scientific papers (which crystallize fragments of scientific knowledge at a point in time) to Liquid Publications, which may adopt multiple shapes, evolve continuously and are enriched by multiple sources. The idea is to promote early circulation of innovative ideas, to optimize the processes by which researchers create, assess and disseminate knowledge, and to stimulate publishers to offer more advanced services (including the maintenance of scientific social networks, automatic notification of new contributions in certain areas, social

Lynch, for example, imagines a future in which there exists a wide range of specialized visualization tools for various forms of structured data [37]. It would be useful, he suggests, to be able to toggle between a rendered image and its underlying data-set, or between a published table of numerical values and their graphical representation, perhaps like the scenario shown in Figure 9?

In a similar state of reverie, Bourne has a vision in which journals provide software for visualizing and interpreting their published content, obviating the need for specialized knowledge in handling esoteric tools; he envisages such software ultimately allowing various forms of basic analysis (simple statistical tests,

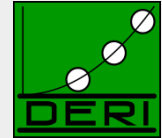
Teresa K. Attwood et al. (2009) Calling international rescue: knowledge lost in literature and data landslide! *Biochemical Journal*. doi:10.1042/BJ20091474

New jobs and roles



- Ph.D. scientists: Extract facts, populate databases, ...
- Computer scientists: Semantic tech, data mining, ...
- Embedded librarians: Metadata, provenance, ...
- Data scientists: Data capture, visualization, stats, ...
- Engineers: Self-documenting apparatus, sensors, ...

Research Assoc./Sci Data Curator



Digital Enterprise Research Institute

www.deri.ie

- Develop the biomedical ontology in OWL
- Annotate biomed resource metadata w/ the ontology
- Help with iterative design of annotation tools
- Participate in working groups to define requirements
- Determine database content
- Implement the data model
- Help with data load processes, data reconciliation, quality assurance, and OWL ontology software integration.

Scientific Data Curator



- Curate morphological data from the literature
- Populate a database
- Contribute new terms, definitions, and relationships to the ontologies where needed
- Work with the community to ensure consistency
- Review the data submitted by experts
- Work closely with software developers to develop the database, curatorial interface, web interface