Trends in the development of Bioinformatics Resources

Jason Hennessey[1], Constantin Georgescu[2], Jonathan D. Wren[2*]

[1]Computer Science Department, Boston University, 111 Cummington Street, Boston, MA 02215 [2]Arthritis & Clinical Immunology Department, Oklahoma Medical Research Foundation, Oklahoma City, Oklahoma 73104-5005; Adjunct Assistant Professor, Department of Biochemistry & Molecular Biology, OU Health Sciences Center. *Corresponding author: jdwren@gmail.com

Summary/Abstract:

Motivation: As the amount of scientific data grows, peer-reviewed Scientific Data Analysis Resources (**SDARs**) such as published **downloadable software programs, databases and online web servers** have had a strong impact on the productivity of scientific research. SDARs are typically linked to using an Internet URL, which have been shown to decay in a time-dependent fashion (Figure 1). What is less clear is whether or not SDAR-producing group size or prior experience in SDAR production correlates with SDAR persistence or whether certain institutions or regions account for a disproportionate number of peer-reviewed resources.

Results: We identified 23,820 non-archival URLs produced between 1996 and 2013, out of which 11,977 were classified as SDARs. Production of SDARs as measured with the Gini coefficient is far more widely distributed among institutions (.62) and ZIP codes (.65) than scientific research in general, which tends to be disproportionately clustered within elite institutions (.91) and ZIPs (.96). The top 1% of institutions produced 68% of all published research whereas for SDARs they only produced 16%. Some labs produced many SDARs (maximum detected=64), but 74% of SDAR-producing authors have only published one SDAR. Interestingly, decayed SDARs have significantly ($p< 8.32 \times 10^{-4}$) fewer average authors (4.33 +/- 3.06), than available SDARs (4.88 +/- 3.59). Approximately 3.4% of URLs, as published, are errors of entry or formatting, including DOIs and links to clinical trials registry numbers.

Conclusions: SDAR production is less dependent upon institutional location and resources, and SDAR online persistence does not seem to be a function of infrastructure or expertise. Yet, SDAR team size correlates positively with SDAR accessibility. While a detectable URL entry error rate of 3.4% is relatively low, automated URL checking would be an inexpensive solution. We also observe a 2.4% error rate in entry of simple ratio percentage calculations across MEDLINE, suggesting a base error rate may exist literature-wide and might correlate with the complexity of the entry task.

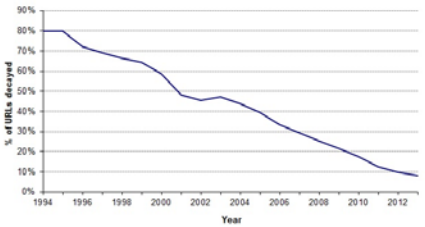

Figure 1: URLs decay in a time-dependent manner. This has been well documented across fields, and has led to concerns about reproducibility and permanence of SDARs. As of Feb 2015, in MEDLINE there were approximately 27,000 unique SDARs.

Introduction

As the amount of biological data grows rapidly in both its abundance and its public availability, Scientific Data Analysis Resources (SDARs) have become indispensable in biology for analysis. In fact, Table 1 shows that an SDAR has been the most cited scientific paper in the biomedical sciences 8 of the last 18 years and have placed in the top 3 most cited papers 13 of the last 18 (72%). So far, at least one SDAR has never failed to place in the top 10 most cited papers annually.

Disappointingly, top SDARs tend to be published in journals with a significantly lower average journal impact factor (JIF) (5.7) relative to the top non-SDAR paper (41.5, $p<0.007$). This suggests the impact of software development is by-and-large either underappreciated or underestimated.

Most highly cited paper	Year publish	Cites	Top SDAR rank	Top SDAR JIF	Top non-SDAR JIF
MEGA5: Molecular Evolutionary Genetics Analysis	**2011**	**4,999**	**1**	**--**	**--**
Cancer statistics, 2010	2010	4,970	3	--	--
Cancer statistics, 2009	2009	5,413	8	8	153*
A short history of SHELX	**2008**	**32,099**	**1**	**2.2**	**153***
MEGA4: Molecular evolutionary genetics analysis	**2007**	**17,043**	**1**	**10**	**35.7***
Induction of pluripotent stem cells from fibroblast cultures by defined factors	2006	5,609	7	5.3	32*
The transcriptional landscape of the mammalian genome	2005	9,600	3	1.3	31*
Electric field effect in atomically thin carbon films	2004	11,182	2	5.3	31*
MrBayes 3: Bayesian phylogenetic inference under mixed models	**2003**	**10,635**	**1**	**5.3**	**30***
Risks & benefits of estrogen plus progestin in healthy postmenopausal women	2002	7,561	2	2.1	30*
Analysis of relative gene expression data using real-time quantitative PCR	2001	23,894	4	5.3	3.6
The Protein Databank	**2000**	**11,849**	**1**	**8.2**	**32***
Mechanisms of disease - Atherosclerosis - An inflammatory disease	1999	12,365	7	3.3	51.6*
A new software suite for macromolecular structure determination	**1998**	**14,499**	**1**	**14**	**31***
Gapped BLAST and PSI-BLAST	**1997**	**34,343**	**1**	**8.2**	**31***
Generalized gradient approximation made simple	1996	28,725	5	2.3	7.9
Controlling the False Discovery Rate (Benjamini & Hochberg)	1995	12,766	2	2	2
CLUSTAL-W - Multiple Sequence Alignment	**1994**	**39,011**	**1**	**8.2**	**3.7**

Table 1: Most cited papers in the Internet Age (from 1994-2011) according to the Institute for Scientific Information (ISI) Web of Knowledge as of August 15th, 2013. The rank of the most cited SDAR is shown (bolded when #1), as well as the 2012 journal impact factor (JIF) of the publishing journal (data for 2011 & 2010 not shown because these years are included in the 2012 IF and will be skewed). *Elite journals, defined as the top quarter of the top one percent (0.25%) of all journals by IF.

Creation of Scientific Data Analysis Resources

As shown in **Figure 2**, the average number of authors per paper indexed in MEDLINE has been steadily increasing over time. Figure 3 shows how the authorship byline has changed by decade. Taking this into account, we analyzed whether or not continued SDAR accessibility was affected by team size (# of authors per SDAR paper), team experience (# of SDARs published per senior author), and institution of origin.

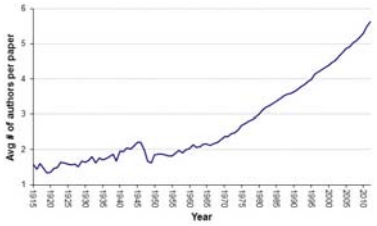

Figure 2: Growth in the number of authors per paper over time in MEDLINE

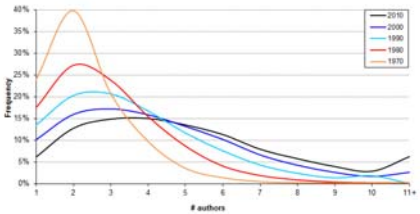

Figure 3: Smoothed histogram of how the # of authors per paper is changing. Recent years have seen the rise of "mega papers" with hundreds, even thousands of authors.

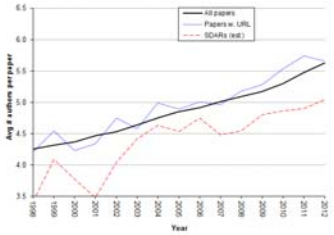

Figure 4: Average team size to create an SDAR has risen with time, but remained below the average.

of authors correlates with SDAR stability

The number of authors per team publishing a new SDAR tends to be smaller than the average team size in MEDLINE (**Figure 4**). There were 9,708 unique SDARs whose corresponding URL was accessible during this study (avg # of authors = 4.88 +/- 3.59) and 3,332 where the SDAR was not accessible (avg # of authors = 4.33 +/- 3.06) – data by year summarized in **Figure 5**. Because the number of authors is increasing with time along with the number of published SDARs, and the we furthermore know that availability is a function of time, we used logistic regression to model the probability of SDAR decay as a function of the number of authors, and year of publication, and the difference was significant ($p<= 0.0009$)

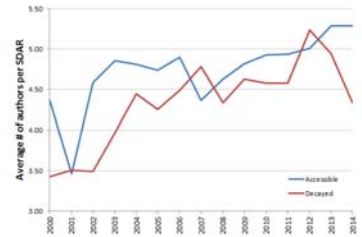

Figure 5: # of authors per created SDAR that are still accessible (blue) and decayed (red)

Experience in creating SDARs does not correlate with odds of future accessibility

We identified 6,600 unique senior author names associated with one or more SDARs. A total of 2,279 of these senior authors had published multiple SDARs. The average fraction of SDARs still available for multi-SDAR authors was 74% (+/- 33%) vs. 74% (+/- 44%) for single-SDAR authors, but the difference was not statistically significant ($p = 0.95$, 2-tailed t-test, unequal variance). Raising the threshold to authors that had published five or more SDARs (n=498) did not change the results. They had a slightly higher average of 76.2% (+/- 25%) accessible, but the difference was not statistically significant ($p = 0.089$).

SDAR production is widely distributed

It has been long known that a relatively small number of institutions generate a disproportionately large number of the scientific papers published annually, which is generally attributed to disproportionate amounts of infrastructure among institutions. We wanted to know if SDAR production followed a similar pattern.

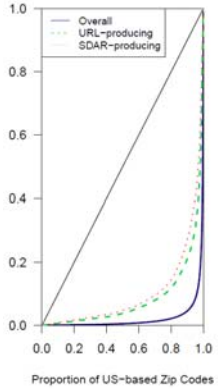

Figure 6: Lorenz curve for production of papers (dark blue), URLs (green) and SDARs (red). In a system where each institution publishes the same # of papers, the curve would be a straight line (shown for comparison). The graph indicates that SDAR production is more widely distributed among institutions than papers in general.

To examine distribution in SDAR production, we used the Gini coefficient, which is widely used to estimate income inequalities, but here can be used to estimate production inequalities. **Figure 6** shows the results.

Summary:

SDARs have had a dramatic impact on science and are increasingly becoming a team effort, much like science in general. But unlike traditional publications, SDARs can decay and their contribution disappears. This hurts the scientific record as well as future chances for publication of people who develop similar SDARs with increased robustness, because the new SDAR may be considered less novel even if the original SDAR has decayed.

Future Directions:

SDAR decay may merely be a byproduct of an evolving system whereby unused SDARs are neglected but widely used ones receive much attention from a user community and are maintained due to peer support and recognition. We plan to examine the evolutionary "niche" SDARs arise in – how many competitors are already in the area and how many citations serve as "fuel" for their continued maintenance and development. We want to know what fraction of SDAR production is driven by the prevailing skills/knowledge of SDAR producers and what fraction is demand-driven.

Acknowledgements

The PI would like to thank the National Science Foundation for their generous support of the DRIVE (Digital Resource Impact, Validation and Evolution) project (EAGER grant # ACI-1345426).

References:

1. Hennessey et al. *BMC Bioinformatics*. 2014 Oct 21;15 Suppl 11:S7
2. Wren JD. *Bioinformatics*. 2008 Jun 1;24(11):1381-5
3. Wren JD. *Bioinformatics*. 2004 Mar 22;20(5):668-72