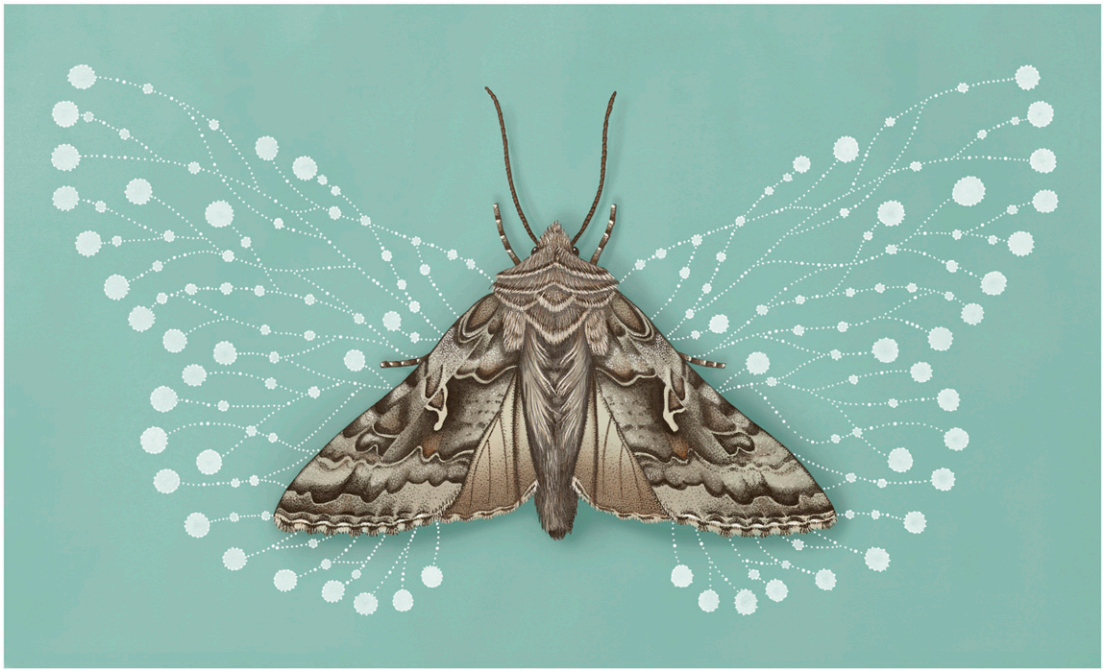
On Mr. Hyslop's prediction, content archives, and preprint servers

May R. Berenbaum, Editor-in-Chief, PNAS

The alfalfa looper moth, *Autographa californica*, is a nondescript grayish brown moth first described by German lepidopterist Adolph Speyer in 1875 (1) and then reclassified by American entomologist Rodrigues Ottolengui (2). Neither the moth's scientific name nor its common name is particularly descriptive; its range includes not just California but also much of the rest of western North America, and its diet includes not just alfalfa but also dozens of other species in multiple plant families. Among the first papers published about alfalfa loopers after the species' description was a US Department of Agriculture Bureau of Entomology bulletin article, "The alfalfa looper" (3), whose author acknowledged that there was no demand for a review of information about this insect; it wasn't considered a pest "of even the slightest importance." However, after finding a note from 1886 in the Bureau files reporting that 2,005 alfalfa looper moths were collected from an electric light globe in a single morning in Los Angeles, Hyslop (3) thought that populations of this moth might periodically explode, and, "in anticipation of such an outbreak the biological notes and other data at hand in this office" should be published. Two years later, alfalfa loopers appeared suddenly by the millions in sweet clover and alfalfa fields in Montana (4). The infestation was devastating, but, to the relief of farmers, the alfalfa loopers disappeared almost as quickly as they had appeared. Describing the outbreak, Parker (4) remarked that "Mr. Hyslop's prediction has certainly come true and the wisdom of collecting and publishing biological data that are not of immediate economic importance is again emphasized" (p. 291).

The Metamorphosis of the Alfalfa Looper Literature

Since 1915, more than 2,200 papers have been published with "*Autographa californica*" as a search topic



***Autographa californica*. Image credit: Alex Boersma (artist).**

First published April 13, 2020.

www.pnas.org/cgi/doi/10.1073/pnas.2005725117

in the Web of Science Core Collection; these have collectively been cited more than 67,000 times. Sixteen of those papers have been in PNAS, and these have, together, been cited more than 1,700 times. That might seem surprising, given that PNAS aims to publish papers of broad interest across multiple scientific disciplines. For the record, three papers on *A. californica* were published in *Science*, two in *Nature*, and one in *Cell* during this period, as well.

As it happens, what has been most interesting about *A. californica* is not the caterpillar itself but rather a baculovirus, *A. californica* multiple nucleopolyhedrovirus (*Ac*MNPV), isolated from a sick alfalfa looper in 1971 (5). Initially, because the virus could infect caterpillar species in at least three families, it was of interest as a possible biological pesticide; among its virtues beyond its host breadth, the virus also could be produced efficiently and economically in large quantities in a cell line derived from another looper species (6).

But it wasn't even the use of *Ac*MNPV as a biopesticide that led to a decade of papers published outside of entomology journals. *Ac*MNPV gained enormous importance because it could be engineered to overexpress foreign genes, including human genes, using its polyhedrin gene promoter. The virus produces a massive matrix of polyhedrin for protecting virus particles in the environment; the key engineering insight was to harness the power of the polyhedron promoter for recombinant gene expression. Cheap and easy to engineer, the baculovirus expression system has proved to be vastly superior to alternative systems based on mammalian cells, handling individual or multiple proteins of almost any size and expressing them at high levels, properly folded, for the duration of the infection. AcMNPV remains the major virus species, along with its polyhedrin promoter, in use today as the expression vector in baculovirus expression vector systems for recombinant protein expression in a variety of contexts, including mass production of egg-free influenza vaccines such as FluBlok (7).

The proliferation of open access preprint servers is creating a revolution in data sharing, dramatically increasing the pace of scientific communication.

In the first PNAS paper about *Ac*MNPV, Hoopes and Rohrmann (8) described their in vitro transcription system for investigating host and viral transcription factors regulating baculovirus gene expression. The most widely cited among the 16 papers, Hofmann et al. (9), reported the successful uptake of recombinant AcMNPV by human hepatocytes, raising the possibility of its use for organ-specific directed gene therapy. The value of such a system was immediately recognized across multiple disciplines, and Hofmann et al. (9) has been cited more than 360 times, in journals ranging from *Journal of Virology*, *Biomedicine & Pharmacotherapy*, and *Nature Communications* to *Fish and Shellfish Immunology*, *Investigative Ophthalmology & Visual Science*, and even *Pesticide Science*.

Access—Past, Present, and Future

It's highly improbable that any of the authors of the papers citing Hofmann et al. (9) had ever heard of Adolph Speyer, much less read his description of *A. californica*, but it may be that, had Speyer never described the species (and identified key features to distinguish it from related species), versatile, efficient, and inexpensive baculovirus expression systems might never have been developed. Or maybe they would have been. One of the most difficult challenges in science publishing is predicting the fate of any particular scientific finding. There are a few conspicuous exceptions—Watson and Crick (10) famously ended their classic paper proposing "a structure for deoxyribose nucleic acid" with the statement, "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." In general, however, predicting the future utility of research is an imprecise enterprise at best. That's all right, as long as research findings are published so that they can be retrieved when circumstances change to make them relevant.

Herein lies a potential problem in the fast-changing world of science publishing. In recent years, there has been a growing awareness of the need to preserve historical literature, not just in dusty locked rooms in inaccessible libraries vulnerable to fire, flood, and administration budget cutters but in the online environment. One valuable initiative, the Biodiversity Heritage Library (BHL), was launched in 2005 by a consortium of university, museum, botanic garden, and science academy and government libraries to cooperate in a massive effort to digitize and make available to the public the "legacy literature of biodiversity." To date, more than 100,000 volumes, amounting to millions of pages, have been digitized and are freely available for use. Despite having access to the University of Illinois library, the third-largest university library in the United States, I have used BHL to obtain critical "legacy" references that I couldn't locate otherwise. Although, while writing this editorial, I could find Hyslop (3) in a few places without much trouble, I found Speyer (1) only with the help of a taxonomist colleague, in Zobodat, the Zoological-Botanical Database, created and maintained by the Biology Center of the Upper Austrian State Museum.

The open access movement advocates not only for immediate free access to new papers but also for free access to back issues of journals; this access is exceedingly helpful in searching for historical findings that may have acquired new relevance years after their original publication date. Access to archived issues varies across publishers, but open access going back in time has considerable value, particularly in the context of, for example, carrying out systematic reviews or metaanalyses or documenting changing species distribution patterns. The two PNAS papers I

cited in this editorial, Hoopes and Rohrmann (8) and Hofmann et al. (9), are freely available, as are all papers published in PNAS older than six months, dating back to the first issue from January 1915.

Then there's the future to worry about. The proliferation of open access preprint servers is creating a revolution in data sharing, dramatically increasing the pace of scientific communication. The extraordinary value of this revolution has been evidenced of late in the midst of the ongoing coronavirus disease 2019 (COVID-19) pandemic, a moment in time during which rapid communication can save lives. The first cases of COVID-19 appeared in Wuhan, China, in December 2019; as of April 1, 1,069 COVID-19 SARS-CoV-2 preprints had been posted on medRxiv and bioRxiv (11). The existence of preprint servers may have an ancillary benefit by influencing how conventional science communication outlets respond to crises; dozens of academic journals, scientific societies, and other scientific enterprises have pledged to make available, without cost, relevant information on the virus during the course of the epidemic (12), PNAS among them (see the Editorial https://www.pnas.org/cgi/doi/10.1073/pnas.2006488117 and https://www.pnas.org/coronavirus).

On the downside, however, the same forces that allowed alfalfa looper literature to languish, hidden behind both paywalls and literal brick walls, may also adversely affect scientific disciplines whose pursuit involves research that might not appear immediately relevant to anyone beyond their relatively small community of specialists. The number of preprint servers is increasing steadily, but they have not yet caught up to the number of specialty journals, many of which are published by scientific societies. These journals have been the bulwark for disciplines characterized by profoundly basic or highly specialized research. Individually, such disciplines or subdisciplines are small, but, together, they represent a critical proportion of the global research enterprise. There are, for example, nearly 100 entomological journals, which are variously geography-restricted, taxon-restricted, or subdiscipline-restricted, including, for example, the Society for Invertebrate Pathology's *Journal of Invertebrate Pathology*, founded in 1959 as *Journal of Insect Pathology*, and source of Vail et al. (6), cited in this editorial. To date, BioRxiv, founded in 2013, lists only 27 disciplinary categories, and it's not immediately clear where manuscripts on insect diseases should be posted so that they're likely to be found by people searching for them. The preprint environment is becoming crowded, preprint servers lack interoperability, and finding certain types of relevant content is becoming more challenging. Moreover, whereas journals published by scientific societies adhere to the same ethical policies that are operative within the societies, it's unclear how preprint servers can offer a similar level of assurance about the legitimacy of their content. Most large preprint servers screen all articles to exclude offensive, nonscientific, or potentially dangerous content and, as well, check for plagiarism and encourage readers who detect possible ethics violations to report them, but it's not clear that small preprint servers have the resources or the inclination to provide this level of oversight.

Preprint servers typically allow posting of manuscripts irrespective of "perceived importance," which is basically a good thing, in view of the difficulties associated with anticipating the potential significance of any given research finding at any point in time. It is a policy that's well suited to large fields, where the number of qualified voluntary reviewers with interest, expertise, and time is substantial; the argument is made that authors benefit from peer reviews from a larger community of scholars than would be tapped for journal review. For some specialized disciplines, however, there may be only a handful of scholars in the world sufficiently qualified to provide an informed evaluation of certain manuscripts. Whereas journal editors actively search for reviewers, typically within a field they're familiar with, preprint reviewing, at least for the moment, depends on whether there are experts actively searching for papers. As well, several preprint servers are funded by philanthropic organizations that have missions and goals and thus have their own criteria for what is or isn't important or appropriate, which might influence what kind of content they are inclined to support.

Science progress depends on building on the knowledge acquired in the past, irrespective of why it was acquired.

If preprint servers continue to depend on support from philanthropic organizations, one consequence may be the homogenization of or elitism in science if specialized and emerging disciplines and research fail to find a niche. An additional critical question that has not yet been answered has plagued scientific communication since the Ebers Papyrus, a 110-page scroll documenting herbal remedies of ancient Egypt, was written (probably based on an even older text, which didn't survive). If not all manuscripts posted on preprint servers eventually get published in a scientific journal, or, for that matter, if scientific journals don't survive our current era, will preprint servers continue to be funded in perpetuity?

New Bottles for Old Wine?

It may seem peculiar, at best, and disingenuous, at worst, for the editor-in-chief of a multidisciplinary journal that publishes 3,200 papers (and rejects more than 15,000 papers) every year across biological, physical, and social sciences to express concern about tiny fields that aren't perceived by the world at large as contributing meaningfully to the progress of science and betterment of humanity. But this is the essence of multidisciplinarity—all new fields necessarily start out small. As well, history has provided innumerable examples of marginalized "old-fashioned" fields gaining new relevance with advances in other fields. Morphology may provide an example; if, as Vogel and

Wainwright (13) wrote, "structure without function is a corpse," then the science of genomics reanimated the corpse in a way that would make Dr. Frankenstein proud. And it seems a profound loss to humanity to discard hard-earned knowledge just because people today aren't sure what to do with it. Science progress depends on building on the knowledge acquired in the past, irrespective of why it was acquired.

There may even be selfish reasons for opening up past literature for public use. According to Mukherjee et al. (14), across virtually all branches of science and technology, manuscripts citing literature with low mean age and high age variance occupy a citation "hotspot" and double their odds of being within the top 5% of citations. So, if for no other reason, citing literature from a range of time periods (as, for example, in this editorial) might boost an h-index now and then.

Acknowledgments

I'm unspeakably grateful to Terry Harrison and Steven Passoa, for providing guidance through the multilingual labyrinth of historical taxonomic entomological literature, and to Bryony Bonning, for helping to illuminate the steps in the path leading from diagnosing sick caterpillars to creating indispensable tools for, among other things, preventing human disease.

1 A. Speyer, Europäisch-amerikanische Verwandtschaften. II. *Stett. Ent. Zeit* **36**, 97–175 (1875).
2 R. Ottolengui, *Plusia* and allied genera with descriptions of new species. *J. N.Y. Entomol. Soc.* **10**, 57–82 (1902).
3 J. A. Hyslop, The alfalfa looper. *USDA Bureau Entomology Bull.* **95**, 109–118 (1912).
4 J. R. Parker, An outbreak of the alfalfa looper (*Autographa gamma californica* Speyer). *J. Econ. Entomol.* **8**, 286–291 (1915).
5 P. V. Vail, T. Whitaker, H. Toba, A. N. Kishaba, Field and cage tests with polyhedrosis viruses for control of the cabbage looper. *J. Econ. Entomol.* **64**, 1132–1136 (1971).
6 P. V. Vail, D. L. Jay, W. F. Hink, Replication and infectivity of the nuclear polyhedrosis virus of the alfalfa looper, *Autographa californica*, produced in cells grown in vitro. *J. Invertebr. Pathol.* **22**, 231–237 (1973).
7 T. W. Kwang, X. Zeng, S. Wang, Manufacturing of AcMNPV baculovirus vectors to enable gene therapy trials. *Mol. Ther. Methods Clin. Dev.* **3**, 15050. (2016).
8 R. R. Hoopes Jr, G. F. Rohrmann, In vitro transcription of baculovirus immediate early genes: accurate mRNA initiation by nuclear extracts from both insect and human cells. *Proc. Natl. Acad. Sci. U.S.A.* **88**, 4513–4517 (1991).
9 C. Hofmann *et al.*, Efficient gene transfer into human hepatocytes by baculovirus vectors. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 10099–10103 (1995).
10 J. D. Watson, F. H. C. Crick, Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
11 https://connect.medrxiv.org/relate/content/181. Accessed 1 April 2020.
12 E. Yasinksi, Journals open access to coronavirus resources. *Scientist*, 13 February 2020. https://www.the-scientist.com/news-opinion/journals-open-access-to-coronavirus-resources–67105. Accessed 1 April 2020.
13 S. Vogel, S. A. Wainwright, *A Functional Bestiary* (Addison-Wesley, Reading, MA, 1969).
14 S. Mukherjee, D. M. Romero, B. Jones, B. Uzzi, The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot. *Sci. Adv.* **3**, e1601315 (2017).