To: Office of Science and Technology Policy (OSTP) [*digitaldata@ostp.gov*]

Fr: Samuel Helfaer, Bachelor of Science, Physics, Yale University [*samuel.helfaer@yale.edu*]

Re: Proceed with Public Access to Digital Data From Publicly Funded Scientific Publications

EXECUTIVE SUMMARY

- The United States Government spends in excess of $145 Billion a year to support research and development to produce new knowledge.
- This new knowledge fuels innovation in many diverse fields such as medicine, technology, climate science, defense and agriculture. This improves the American economy through both higher productivity and new sources of productivity.
- Open access to non-confidential data and publications produced using government funds would increase the government's return on investment for research, improve our economy by increasing productivity and decreasing unemployment, and, in turn, provide further incentive to invest more in science.
- An open access protocol will allow more collaboration and less unnecessarily repeated science as well as the opportunity to improve academic rigor and repurpose data.
- This effort will also enable common people to educate themselves and attempt to distill data, thus encouraging STEM education and creating an opportunity for research with low barriers to entry. This will catalyze an improvement in STEM research as well as STEM literacy in the American population, further improving our economy.
- Open data will make peer review more effective and promote a more collaborative model of discovery while maintaining the incentives provided by competition.
- While there are significant logistical challenges to opening access to all data produced through federally funded research, the benefits far outweigh the costs.
- The Human Genome Project as well as the NIH and NSF data sharing policies demonstrate the feasibility and promise of open data in federally funded research.
- Open access to data from federally funded research will spur innovation by other federal researchers as well as those in private industry.
- Open data will contribute to better policy and decision-making by individuals, corporations and government agencies and improve the American economy through job growth, productivity increases and happier, healthier lives for all American citizens.
- Opening data to the taxpayers that fund federal research will ensure American economic prosperity, strengthen national security and propel the pursuit of knowledge forward.

BACKGROUND: FEDERALLY FUNDED RESEARCH

Every year the United States government invests more than $145 billion in funding for research and development (R&D)[1]. Countless additional dollars are spent elsewhere (or not collected in taxation) in order to incentivize innovation and improve the welfare of the American people and people throughout the world. This funding has yielded incredible advances in virtually every field of science and technology from defense and agriculture to artificial intelligence and medicine. These research dollars come from taxpayers, including students, other researchers and scientists, as well as many hard-working men and women. While all of these constituencies serve to benefit, they generally do not have access to the resultant data.

Remarkable advances have been made in both basic and applied research and spurred inventions that have improved every facet of our lives. Still, we, the American people, are not realizing the full potential benefits of this research. Much of the data collected, programs and algorithms created and insights gleaned remain in the hands of a precious few. Even when an academic paper is published with open access, the clean, peer-reviewed, edited academic paper often belies reams of data and years of work that don't make it to the final draft. The academic paper has become a vital means of communicating research and serves a valuable purpose, however, there is much to be gained from sharing all that went into its creation.

Allowing others to utilize all of the data produced from federally funded researchers will promote innovation through both collaboration and competition. It will enable other scientists to avoid conducting the same experiments as well as confirm the rigor and validity of studies

[1] Sargent Jr., John F. "Federal Research and Development Funding: FY2011."*Congressional Research Service Report for Congress* R41098 7.5700 (2011). <http://www.fas.org/sgp/crs/misc/R41098.pdf>.

conducted on the taxpayer's dime. Through open data, research dollars will be used more effectively and efficiently. This will yield a higher return-on-investment and, in turn, higher impact research will trigger more research dollars and more innovation. Similarly, through open access to data, STEM students will have an unprecedented ability to engage with real research data and make a contribution from an early stage in their education. As the world economy becomes increasingly globalized, strong domestic research and development efforts coupled with strong STEM education will keep the American economy competitive in the global economy.

AMERICA COMPETES REAUTHORIZATION ACT OF 2010: DIGITAL DATA

The America COMPETES Reauthorization Act of 2010 established a working group under the National Science and Technology Council (NSTC) to "coordinate federal science agency research and long-term stewardship of the results of unclassified research, including digital data and peer-reviewed scholarly publications, supported wholly, or in part, by funding from the federal science agencies." It also stipulates that "[OSTP] shall develop policies for the management and use of Federal scientific collections to improve the quality, organization, access, including online access, and long-term preservation of such collections for the benefit of the scientific enterprise."[2]

This brief intends to outline the benefits and implementation challenges of open data specifically related to the Request for Information (RFI) pertaining to "Public Access to Digital Data Resulting From Federally Funded Scientific Research"[3] released in accordance with

[2] 1861, 111 Cong., America COMPETES Reauthorization Act of 2011 3985-3987 (2011) (enacted). <http://www.gpo.gov/fdsys/pkg/PLAW-111publ358/pdf/PLAW-111publ358.pdf>.
[3] Wackler, Ted. "Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research." *Federal Register* 76.214 (2011): 68517-8518. <http://www.gpo.gov/fdsys/pkg/FR-2011-11-04/pdf/2011-28621.pdf>.

Section 103(b)(6) of the America COMPETES Reauthorization Act of 2010 (ACRA; Pub. L. 111-358).

HIGHER IMPACT, HIGHER RETURN ON INVESTMENT

Measuring the impact of scientific research across diverse fields can be difficult and highly subjective. Many metrics have been developed to assess the impact or value of a given study. These metrics often include the number of times other papers cite a given paper as well as the impact of those papers. Consistent with this, open access articles are found to be cited more frequently even when controlling for article quality, indicating that opening access to articles contributes to their impact.[4] A similar effect is seen with open data as other researchers can build on the research already conducted. By releasing more data, the opportunities for usage increase proportionally.

When more researchers use data gained through federally funded research, the research dollars go further. The opportunities for progress are larger and more numerous and represent a significant increase in return-on-investment (ROI) for taxpayers and government agencies. Higher impact, higher ROI scientific research means more funding which will directly benefit researchers who compete for research dollars as well as the American people who benefit from the new knowledge and technologies.

CULTURAL CHALLENGES: THE REPUTATION CONUNDRUM

At present, primary external motivators for researchers include money, reputation, and time. Money and time, along with space and other resources allow researchers to pay themselves and

[4] Gargouri Y, Hajjem C, Larivière V, Gingras Y, Carr L, et al. 2010 Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. PLoS ONE 5(10): e13636. doi:10.1371 <http://www.plosone.org/article/fetchObjectAttachment.action;jsessionid=55B7DC69A50002ADB4E7B4F02FF2FBA0?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0013636&representation=PDF>.

others in their group, provide the necessary equipment for experimentation, and have the time to think and do research. Reputation, attained through successful research, ensures continued access to time, money and other resources. Given the importance of reputation, much of what drives scientists is derived from how reputation is measured. In our current construct, reputation is often measured by citations, h-indices, number of papers published, and download counts.[5] This incentive structure places the primary focus on the publishing phase of research and paints this as an ultimate step, downplaying continued engagement with published research. In this "publish or perish" environment, any time spent away from publishing and grant-writing serves as essentially wasted time.

CULTURAL CHALLENGES: REALIGNING INCENTIVES

While academic papers play a valuable role in the spread of information, they lack much of the information gleaned during the research project. Time spent by researchers cleaning up and publishing data, meta-data and code and engaging with other researchers on a blog or other communication interface is time spent furthering human knowledge and science. By not incentivizing or requiring this information to be published, we are sacrificing millions of dollars in potential innovation and unnecessarily repeated research and creating a lack of transparency that can allow academic dishonesty and poorly conducted research.

This incentive structure has become deeply ingrained and institutionalized in many academic disciplines. Changing this culture requires a realignment of incentives to favor publishing of, and engagement with, data. This can be achieved through open data stipulations accompanying federal research funds as well as consideration of the open access contributions of researchers (both previous and in the context of a proposed project) when evaluating federal grant

[5] Siemens, George. "What, Exactly, Is Open Science? | The OpenScience Project." *The OpenScience Project | Open Source Scientific Software*. Web. 05 Dec. 2011. <http://www.openscience.org/blog/?p=269>.

applications. Through enhanced options and requirements of open data, tenure committees and hiring decisions as well as other reputational measures will increasingly hinge on a researcher's participation and impact in open science.

Many models are currently being used to catalyze this shift toward open access. ORCHID (Open linguistic Resources CHannelled toward InterDisciplinary research), for example, explicitly incentivizes blog post and other less formal peer review while the NIH and NSF both have data sharing stipulations in their research grants. This is not enough. Without requiring data to be published across all federal funding sources, we are missing out on a substantial portion of the potential research impact and wasting money on redundant research. We are not optimizing the performance of our research and development minds and resources if they do not have access to the most current and complete datasets possible.

CONFIDENTIALITY CONCERNS

As with all experiments, it is vital to ensure that all privacy and confidentiality concerns are considered with the utmost care. Any information that could compromise personal privacy, national security or intellectual property rights should not be released. These concerns are especially relevant in the medical[6] and defense spaces and, as such, all grants that could include proprietary or confidential data should be granted in accordance with an explanation of what data will likely need to be withheld from open access publishing.

COMPETITION CONCERNS: INTELLECTUAL PROPERTY

Both collaboration and competition are essential to successful innovation and groundbreaking research. With open access to data and experiment code, many other researchers will have the

[6] See HIPAA, the Health Insurance Portability and Accountability Act.

potential to benefit from the hard work of those who have already published their data. While this will allow more minds to collaborate, it also has the potential to diminish the incentives of developing proprietary algorithms and methodologies as well as collecting proprietary data. By allowing researchers to maintain control of their data and code, we provide them an opportunity to conduct as much analysis and algorithmic optimization as they desire with little acute time pressure. In order to preserve the incentive that proprietary data and code provides, it is essential to allow scientists to request more time with their proprietary data and code before release (similar to a patent). This should be able to be included in their initial grant application or in a subsequent application before they have published their papers or data. Allowing a patent-like system for researchers will maintain the incentive to make large research advances.

This extended proprietary-data period should not apply to sectors with well-defined databases and methodologies in which the researcher likely received substantial benefit from the already established database and has merely made an incremental advance. While an extension-of-proprietary-usage provision incentivizes large innovation, it has the potential for abuse. This policy should be used sparingly and continually monitored to ensure it is fulfilling its intended role and not being overused.

CONCERN: PEER-REVIEW AND THE COST OF DATA-SCRUBBING

One of the tenets of modern research is peer-review. This review process is meant to keep scientists conducting honest, rigorous and important research through the oversight of a community of knowledgeable colleagues. While much peer-review is currently undertaken in conjunction with journals, with a realignment of incentives toward valuing community review and collaboration, open review can become a major component of research. By opening up

data to the greater research community and beyond, the opportunities for peer-review become significantly larger and more powerful.

Many arguments against requiring data and in-depth methodology publication hinge on the challenges of cleaning up data and study specifications. This process of data (and methodology) scrubbing can be time-consuming, however, it is worthwhile. This process provides this data for many other potential researchers. More importantly, any study that is being peer-reviewed should have the data and meta-data available to be audited and reviewed by a reviewer. A peer-reviewer is not merely supposed to read over the academic paper and check for grammatical mistakes and guarantee stylistic writing. It is their responsibility and goal to ensure the academic rigor and honesty of the study. While data is often made available in peer-review processes, it should be made generally available with low barriers to entry so that all stakeholders can evaluate the methodologies, data and conclusions without merely looking through the small window into what the researchers felt to be important or conclusive.

ACADEMIC RIGOR THROUGH OPEN SCIENCE

The dangers of closed science have come to light in a variety of cases recently.[7] In one highly publicized case, a famous social psychologist, Diederik Stapel, fabricated immense amounts of data over 15-20 years and 100 publications are now under investigation.[8] Similarly, a recent study entitled "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant" demonstrates how easy it is to

[7] Fanelli, Daniele. "How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data." Ed. Tom Tregenza. *PLoS ONE* 4.5 (2009): E5738. <http://www.plosone.org/article/fetchObjectAttachment.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0005738&representation=PDF>.

[8] Crocker, Jennifer, and M. Lynn Cooper. "Addressing Scientific Fraud." *Science Magazine* 334.6060 (2011): 1182. <http://www.sciencemag.org/content/334/6060/1182.full>.

accumulate and report statistically significant evidence for a false hypothesis.[9] In their article, they point out that leaving out certain data can be as important as keeping certain data and that current practices encourage reporting exclusively "what worked": that is, those results that were statistically significant. Decisions about data manipulation are not made in advance of the data collection. Simmons et al. propose guidelines that would require researchers to, for example, disclose every question asked on a survey not merely the questions with statistically significant responses. This same goal can be achieved through requiring federally funded researchers to release all of their collected data. Similarly, releasing all collected data is a much more effective and powerful mechanism for preventing undue corporate (or other conflict of interest) influence than merely requiring authors' disclosure in small print.

As they explain "Our goal as scientists is not to publish as many articles as we can, but to discover and disseminate truth…We should embrace these [proposed rules about disclosing research methods] as if the credibility of our profession depended on them. Because it does." By requiring the release of data for federally funded research, the process of peer-review will become more effective and academic honesty and rigor will improve. This cultural shift will also catalyze the release of data in research not funded by federal dollars yielding a research multiplier effect. This will increase both the quality and value of research done in the United States and propel our pursuit of knowledge forward.

THE CURSE OF EXPERTISE: AMATEUR SCIENTISTS

Wikipedia provides a case study of the amazing advances that can be made through open access. Another oft-cited example of an open access success is Tim Gowers' Polymath

[9] Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22.11 (2011): 1359-366. <http://pss.sagepub.com/content/22/11/1359.full.pdf>.

project[10], which allowed 27 mathematicians to prove a theorem that had stumped many famous mathematicians using an online cooperative model. Still, academic research should not become Wikipedia. There is a valuable place for expertise. Data that is released should not need to meet an arbitrary standard of understandability to a layman (as Wikipedia does). Likewise, while the data should be accessible to everyone, there will certainly be means of stratifying the research conducted with the data. A tenured professor should not be required to directly reply to a high school student who thinks the researcher has made a mistake or if the high-schooler does not understand their methodology – this is the role for an online community forum. A confused young person or amateur scientist with their own theory can initiate a conversation in an online forum. This conversation could be seen by a colleague of the researcher who, in turn, may validate its legitimacy and ask the author to explain that portion of the study.

By allowing authors' colleagues to repeat their experiment or build on their experiment without having to recreate and back-engineer every method, open access to data and methodology will accelerate the pace of research. In this process, researchers may have to explain their data or process to some of the others in their field but those colleagues will ultimately reciprocate and a dynamic, collaborative model of research will flourish.

KEY COMPONENTS OF DATABASE IMPLEMENTATION

Given the current databases in place and rapidly evolving database landscape, an attempt to explicitly design and implement a single, all-inclusive database for federally funded research would be both logistically challenging and of questionable long-term value. Still, there are guiding principles that should steer our vision of successful database creation. These databases

[10] Parker, Matt. "Welcome to WikiMaths: Home of Hard Sums." *Shortcuts Blog*. The Guardian, 8 May 2011. Web. 05 Dec. 2011. <http://www.guardian.co.uk/science/2011/may/08/welcome-to-wikimaths>.

should allow anyone to access and search through the data and meta-data readily and without buying or downloading any proprietary software. In short, databases should be easy enough to use that an average taxpayer could use them (even if they do not necessarily understand all of the data, methodologies, etc.). In this way, we can democratize the data that taxpayers fund.

Databases should:

1) Be machine searchable and indexable by other databases (such as google).
2) Maintain data in non-proprietary formats (such as csv, xls, txt, etc.).
3) Aggregate and organize data based on usage, researchers and other criteria to ensure proper secondary accreditation and citation as well as ease-of-use.
4) Attempt to include all available information within its purview, as it is hard to anticipate the needs and desires of other researchers (and other users).
5) Automatically update when new data is released/published or be easily updateable to adapt to the new model of dynamic scientific research.

LOGISTICAL CHALLENGES: DATA STORAGE AND EXPENSE

Federally funded research produces an enormous amount of data and meta-data each year. While much of this data may seem unimportant or excessive, especially as compared to the data presented within and alongside the final academic paper, it has the potential to benefit other researchers and citizens. As the cost of data storage and aggregation continue to fall, the cost to value ratio of storage and maintenance of data continues to improve.

Many databases are currently being developed both in connection with the government and through private endeavors. Given this current fragmented landscape, the key objectives going forward should be 1) increased access to data, 2) aggregation and organization of that data, and

3) interoperability between platforms. The government can play a direct role in goal 1 and a more indirect role in goals 2 and 3.

The Human Genome Project, the NIH Data Sharing Policy, and the NSF Data Management Policy all provide frameworks for successful implementation of government data sharing programs. Other smaller ventures outside of the government's direct purview include PLoS (Public Library of Science), SPIRES/INSPIRE (High-Energy Physics), and JASPAR (DNA transcription factors binding preferences) and VIVO (an academic-institution-only research network. All of these models are slightly different and reflect the realities in their given academic discipline(s), however, they would all benefit from more open data. By creating more open data, the value of aggregation, organization, and interoperability increases exponentially[11], incentivizing these goals by others. The government can also pursue goals 2 and 3 by providing grants to fund the creation of new databases and to improve those currently in operation.

STEM EDUCATION: ENABLING AMATEUR SCIENCE

The primary purpose of open data is not to allow amateur science or to improve STEM education. However, this will likely be one of its primary long-term positive effects. Open data will promote a collaborative model that allows amateurs and students access to much more data and resources than they do under the current model as well as opportunities to interact more readily with the greatest minds in the fields they are interested in through their research. This positive "collateral damage" will create future generations of scientists who are more interested in helping young people develop their analytical and experimental skills, further improving STEM education and literacy in America.

[11] Similar to Metcalfe's Law of Network Value, which says that network value is proportional to n^2, where n is the number of users in the network. In the research case, there is an additional positive feedback mechanism.

By opening access to data produced with federal dollars, we will promote a more scientifically inclined, academically rigorous society. This will help improve the productivity of the American economy and maintain our impressive track record as the leaders in global R&D and education.

SUCCESS WITHIN INDUSTRY

Another positive of open data is allowing private industry to benefit and innovate more directly from basic and applied research conducted through federal grants. This will, in turn, benefit both the United States economy as well as the American people who will have access to improved medical care, new technologies and less expensive manufactured products. Many companies already operate through a combined collaborative and competitive model in which data is shared even as various divisions compete for funding. This model has provided many corporations with record profits and rapid innovation. Adapting this model to the public sector will allow federal dollars to produce more knowledge, innovation and economic benefit.

INFORMED AND DYNAMIC POLICY-MAKING

One of the key benefits of open access to data outside the realm of academia is that it enables individuals and organizations to make informed decisions based on the best and most-recent data available. This allows the actors in our economy to perform more efficiently and improves the overall productivity of our economy. Remarkably, many of our government bodies do not have access to relevant research or data that could allow them to make better policy-related decisions. While economic data can be easily extracted from the bureau of labors statistics

online[12] (as one example), attempts to access scientific data are often met with pay-walls or simply academic papers containing a small percentage of the actual data collected.

The lack of accessible, organized, trustworthy data in our economy prevents individuals, corporations and government agencies from making the best fact-based decisions about vital issues such as agricultural policy, national defense, climate change and healthcare. This promotes inefficiencies in our economy and can have disastrous consequences as many of these actors (especially large corporations and large government agencies) have the ability to affect large changes.

OPEN DATA: FOR A BETTER AMERICAN FUTURE

"Improving the way that science is done means speeding us along in curing cancer, solving the problem of climate change and launching humanity permanently into space. It means fundamental insights into the human condition, into how the universe works and what it's made of. It means discoveries not yet dreamt of. In the years ahead, we have an astonishing opportunity to reinvent discovery itself. But to do so, we must first choose to create a scientific culture that embraces the open sharing of knowledge."[13] As Michael Nielsen, a pioneer in quantum computing, aptly describes, we stand at a turning point in human knowledge. Never before has the pace of innovation and acceleration of discovery been as rapid. We must seize this opportunity to use what we are learning to its fullest and propel discovery forward.

The possibilities for open data extend beyond science. By requiring that all federally funded research and development release data and meta-data into the public domain, we will improve

[12] http://www.bls.gov/data/
[13] Nielsen, Michael. "The New Einsteins Will Be Scientists Who Share." *The Wall Street Journal*. 29 Oct. 2011. Web. 05 Dec. 2011. <http://online.wsj.com/article/SB10001424052970204644504576653573191370088.html>.

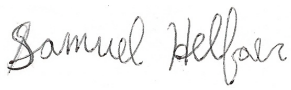
the American economy and our national security and allow us to make vital data-driven decisions about how we will face the seemingly insurmountable challenges that lie before us. Open data will allow us to take full advantage of our rapidly growing and evolving scientific knowledge to guide humanity forward on a successful trajectory.

No one can fully predict how open data will evolve. I am certain, however, that open access to our data will lead to a better tomorrow.

Thank you for your time and consideration.

If you have any questions for me, please do not hesitate to contact me at Samuel.Helfaer@yale.edu.

Sincerely,



Samuel Helfaer

Constituent, Pennsylvania's 6th Congressional District

Bachelor of Science Candidate, Physics, Yale University

References

1861, 111 Cong., America COMPETES Reauthorization Act of 2011 3985-3987 (2011) (enacted). <http://www.gpo.gov/fdsys/pkg/PLAW111publ358/pdf/PLAW111publ358.pdf>.

Crocker, Jennifer, and M. Lynn Cooper. "Addressing Scientific Fraud." *Science Magazine* 334.6060 (2011): 1182. <http://www.sciencemag.org/content/334/6060/1182.full>.

Fanelli, Daniele. "How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data." Ed. Tom Tregenza. *PLoS ONE* 4.5 (2009): E5738. <http://www.plosone.org/article/fetchObjectAttachment.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0005738&representation=PDF>.

Gargouri Y, Hajjem C, Larivière V, Gingras Y, Carr L, et al. 2010 Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. PLoS ONE 5(10): e13636. doi:10.1371 <http://www.plosone.org/article/fetchObjectAttachment.action;jsessionid=55B7DC6A50002ADB4E7B4F02FF2FBA0?uri=info%3Adoi%2F10.1371%2Fjournal.pone.001636&representation=PDF>.

Larson, Phil. "Request for Information on Public Access to Digital Data and Scientific Publications." *The White House*. 7 Nov. 2011. Web. 05 Dec. 2011. <http://www.whitehouse.gov/blog/2011/11/07/request-information-public-access digital-data-and-scientific-publications>.

Nielsen, Michael. "The New Einsteins Will Be Scientists Who Share." *The Wall Street Journal*. 29 Oct. 2011. Web. 05 Dec. 2011. <http://online.wsj.com/article/SB10001424052970204644504576653573191370088.html>.

Parker, Matt. "Welcome to WikiMaths: Home of Hard Sums." *Shortcuts Blog*. The Guardian, 8 May 2011. Web. 05 Dec. 2011. <http://www.guardian.co.uk/science/2011/may/08/welcome-to-wikimaths>.

Sargent Jr., John F. "Federal Research and Development Funding: FY2011."*Congressional Research Service Report for Congress* R41098 7.5700 (2011). <http://www.fas.org/sgp/crs/misc/R41098.pdf>.

Siemens, George. "What, Exactly, Is Open Science? | The OpenScience Project." *The OpenScience Project | Open Source Scientific Software*. Web. 05 Dec. 2011. <http://www.openscience.org/blog/?p=269>.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22.11 (2011): 1359-366. <http://pss.sagepub.com/content/22/11/1359.full.pdf>.

Wackler, Ted. "Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research." *Federal Register* 76.214 (2011): 68517-8518. <http://www.gpo.gov/fdsys/pkg/FR-2011-11-04/pdf/2011-28621.pdf>.