figshare

Introducing a more user-friendly and aesthetic approach to managing peer-reviewed data

SCIENTIFIC DATA

About

- **Andrew Hufton** is Managing Editor of *Scientific Data*, a journal from Nature Publishing Group

Goals

- To provide a service that enables authors to submit and store data-rich articles, so that the data can be easily visualized within a published document

Approach

- Integrate the figshare Datastore and Viewer into the *Scientific Data* editorial-submission and data-article platforms

Results

- The embedded figshare data uploader has become, with Dryad, the most widely used uploader by authors submitting articles among the 80-plus data repositories used by *Scientific Data*
- Articles with figshare-hosted and visualized data are heavily used. The most cited article from Google Scholar using figshare data has 45 citations in other sources, and the third most cited article, also with data at figshare, has been visited 4,720 times in the last 18 months
- Although it took more than 12 months after launch, the *Scientific Data* editorial team are starting to see data re-use, and new analysis and publications arising out of this re-use

Andrew Hufton

The data journal *Scientific Data* has been using the figshare Datastore and Viewer solution for the past 18 months to help manage its data article submission, review and publication process. We interviewed its Managing Editor Andrew Hufton to find out how the project is progressing.

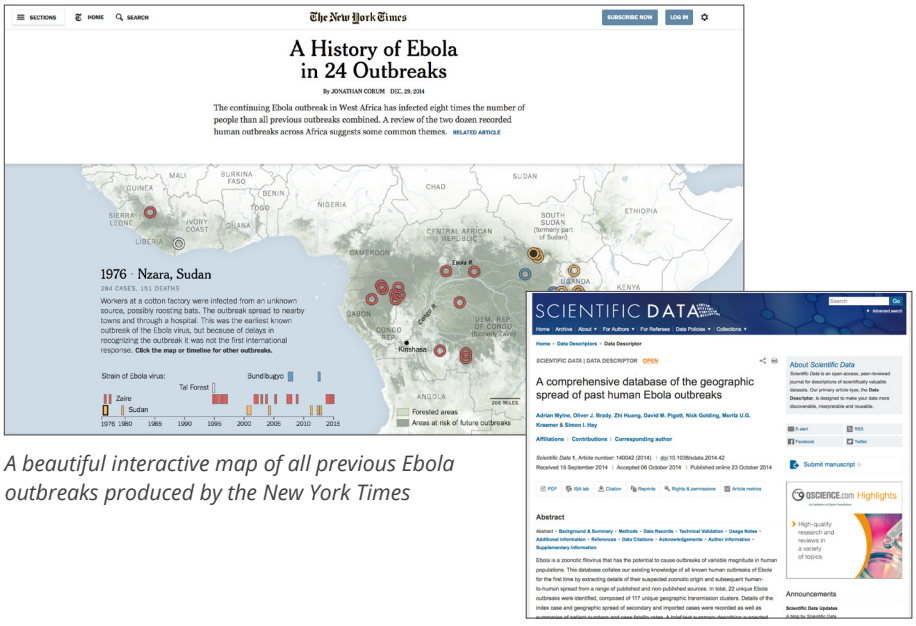
Andrew Hufton trained in genetics and worked on topics in developmental biology and genome evolution. He is a former editor of Molecular Systems Biology, an open-access journal for a community of scientists who are particularly interested in publishing and studying data-rich articles.

"Systems biologists, in general, have a particular obsession with data, and a progressive, even aggressive, attitude to data sharing," he remembers. "In molecular biology, you have two of the most compelling cases of data sharing: the Protein Data Bank and the INSDC repositories, including Genbank. By creating a very integrated, structured data-sharing environment, they collectively allowed the field of bioinformatics to be created. In biology, data sharing completely transformed the scientific process."

With this track record, Andrew was later appointed as Managing Editor of *Scientific Data*, a journal whose mission is to provide a platform for the peer-reviewed publication of scientific data. Each of the articles goes through peer review and typically describes a valuable set of data that can be challenging to present because of its unique nature.

"Between 70 and 80% of what we publish comprises new and unique data sets," Andrew says. "We focus on getting the standards right around the way the data is presented. Another 20 to 30% expand on data reported in the

cont.



A beautiful interactive map of all previous Ebola outbreaks produced by the New York Times

The underlying data article with data sets hosted by Figshare

other research papers, when the underlying data are sufficiently complex that they deserve this. We also try to break out the supplemental data from *Nature* articles where they can be presented as raw experimental data and made to be machine readable, reusable and citable."

"Researchers choose to share their data for many reasons - a commitment to open science, to help build collaborations, or funders or journal requirements. But if you are going to share your data, services like figshare help you get your data out quickly, and figshare and *Scientific Data* ensures that you get real credit for doing so."

Relatively young but maturing fast

Scientific Data is two years old in May and Andrew expects to obtain an Impact Factor for the data journal in 2017. "We have been accepted in Scopus and we are indexed in Medline and Pubmed," he says. The journal charges £890 per article under a gold open-access model and it takes on average 30-40 days for the peer-review process. The lead time until full publication of a data article is between three and six months. Typically, 10% of articles are rejected following peer review.

"There are a lot of people who care about data quality and who support us," says Andrew. "Quality does matter, and bad data is poisonous, so we need the peer-review process to verify the data." Andrew also believes that the way in which figshare assists the presentation of data represents "an important cultural shift for science".

New and emerging areas

Scientific Data focuses on new and emerging areas of research. "We thought the majority of our contributions would be in the field of genomics (and other "omics" fields like metabolomics and proteomics), but actually our biggest value is in areas where data sharing is difficult," Andrew says. "We get a lot of contributions on the subjects of neuroscience and climate science. These are fields that are very data heavy at the moment, and where standards are less clear, but everyone wants to get at all the data available as everyone is modelling the same system." The journal has also had important submissions from computational materials science. "The US has a funding initiative called the "Materials Genome" which has been putting money into computational materials," says Andrew. "This has led to a proliferation of data which explores a large chemical space on supercomputers." "We also see contributions from social sciences," he adds. "Statistical physicists are becoming attracted to this area, and we are seeing a proliferation of research looking at social networks and urban patterning. There is a real expectation for validation and quality, which is where we come in."

cont.

"It took us a year after launch to really start to see this re-use. We are definitely seeing people taking the data, doing analysis and publishing new findings on new methods, and other reference data sets which are cited in the way that major reference works would be cited."

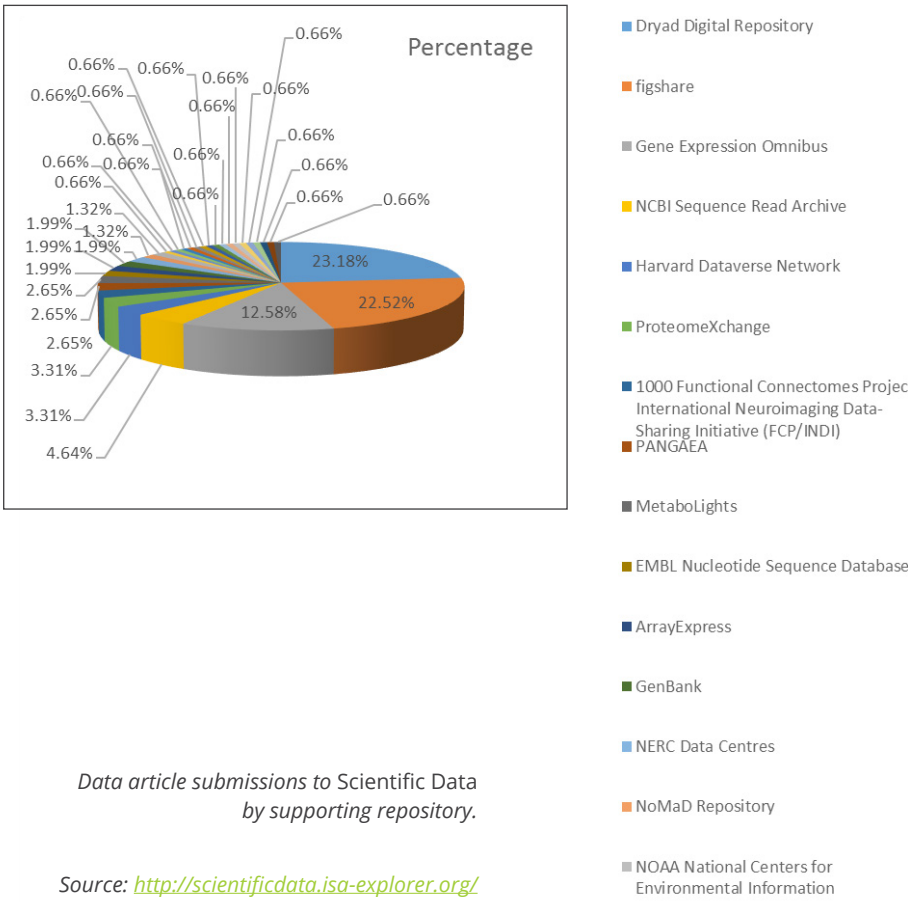
Broadest scope

Scientific Data is the data journal with the broadest scope and publishes between five and ten data descriptors every month. The data journal also has a 90% publication rate from peer-reviewed submissions. Traditional results articles are turned away, although the journal is now opening its doors to articles on data reuse or systems that support data sharing. "We have a delightful early adopter phenomenon delivering us very high quality data," says Andrew. "Some of our contributors are data managers, but most are PIs or postdoctoral researchers. Our approach gives the group an opportunity to have a new lead author, perhaps the person who conducted the experiment."

User-friendly

Andrew regards figshare as a very flexible and user-friendly service for the journal's authors and peer-reviewers. "It is an easy place for people to deposit the data," he says. "There is a submitting interface that integrates with the EJP submission system through an iframe. There is a logical journey with questions on data deposit, and there is a link to figshare." Submissions are neck and neck between figshare and Dryad. Andrew comments "Submitting data to figshare is easier, and the data presentation is better with figshare. Dryad has a high profile in ecology and environmental sciences and has active curators who help people with their files." *Scientific Data* has relationships with 80 other repositories but only figshare and Dryad are integrated directly. There is a long tail of deposits across the repositories but figshare and Dryad are at the top.

cont.



Data article submissions to Scientific Data *by supporting repository.*

Source: http://scientificdata.isa-explorer.org/

Beautiful presentation of the data

The figshare light box can be used to embed playable videos in articles. For example, Comparative, transcriptome analysis of self-organizing optic tissues (http://www.nature.com/articles/sdata201530). "We have put videos in, and we have used figshare often for supplementary figures where it presents in a really nice slide show."

Usage and citation

Although popular data articles in *Scientific Data* are downloaded thousands of times, data files are usually downloaded at lower rates. "We are seeing real cases of re-use of data through Google Scholar," says Andrew. "It took us a year after launch to really start to see this re-use. We are definitely seeing people taking the data, doing analysis and publishing new findings on new methods, and other reference data sets which are cited in the way that major reference works would be cited." The top cited *Scientific Data*-published article (http://www.ncbi. nlm.nih.gov/pmc/articles/PMC4322588/) on Google Scholar has a figshare data citation, with 45 citations for the article in other sources. Quantum chemistry structures and properties of 134 kilo molecules (http://www.nature.com/articles/sdata201422), our third most cited article, also with data at figshare, has been visited 4,720 times over about 1.5 years.

Enabling accurate data journalism

Making the data available in a highly organized and structured way is also creating new opportunities for powerful science communication. Andrew gives an example: "We have seen data journalists pick up the data on popular topics which we have released data on such as Ebola. Here's the Ebola data publication we published (http://www.nature.com/articles/sdata201442), and it helped create a beautiful interactive map of all previous Ebola outbreaks produced by the New York Times (http://www.nytimes.com/interactive/2014/12/30/science/history-of-ebola-in-24-outbreaks.html?_r=0). The actual data files are hosted in figshare".

figshare strengths

The advantages of figshare are its relative ease of use and the ability to visualize a wide range of files in an attractive Google-style experience.

Future directions

Andrew would like to see the *Scientific Data* relationship with figshare evolve, with figshare helping to capture standardized metadata for every dataset described in *Scientific Data*'s articles, to further aid data discovery and re-use. "We're pleased that figshare are open to developing custom solutions for our and our authors' needs," he says.