Publishing challenges for researchers, funders and publishers

Philip Campbell

FAPESP Sao Paolo

28 February 2013

FAPESP

Science as an open enterprise: open data for open science

28/02/2013 - FAPESP - São Paulo

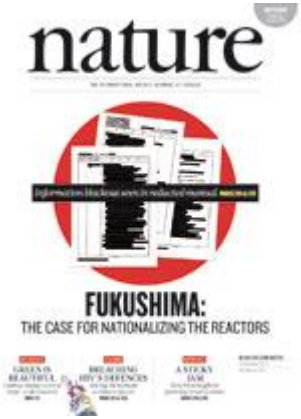
nature publishing group npg

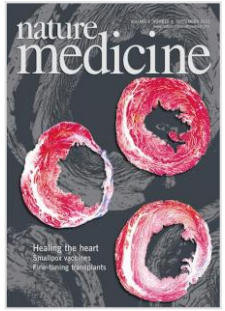
Content

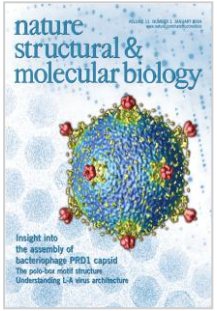
- Nature's selection process
- Issues surrounding the future literature
- Enhancing credit for researchers
- What happens after a Nature publication?
- Issues in replication
- A future for open publishing of papers and data?

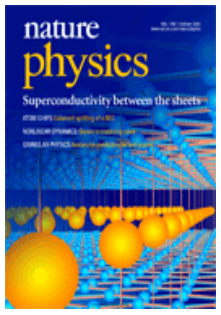
Nature & Nature Research Journals

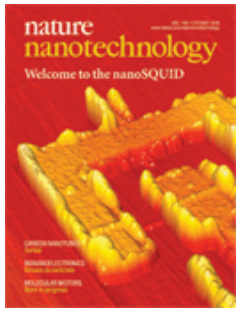
(The print versions…)

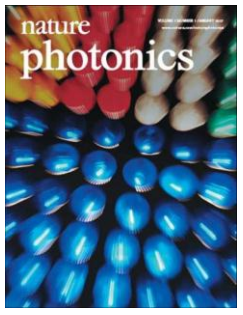




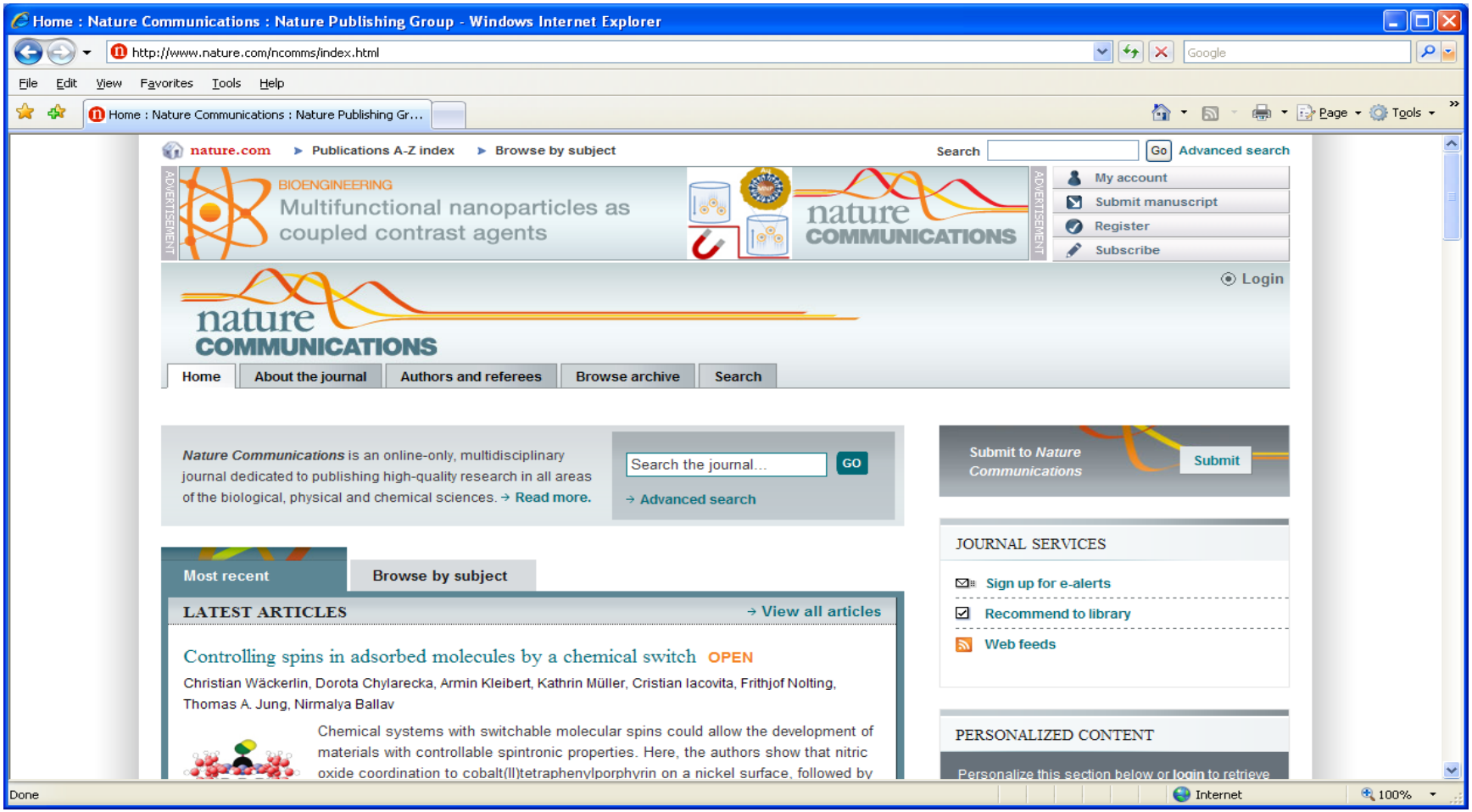




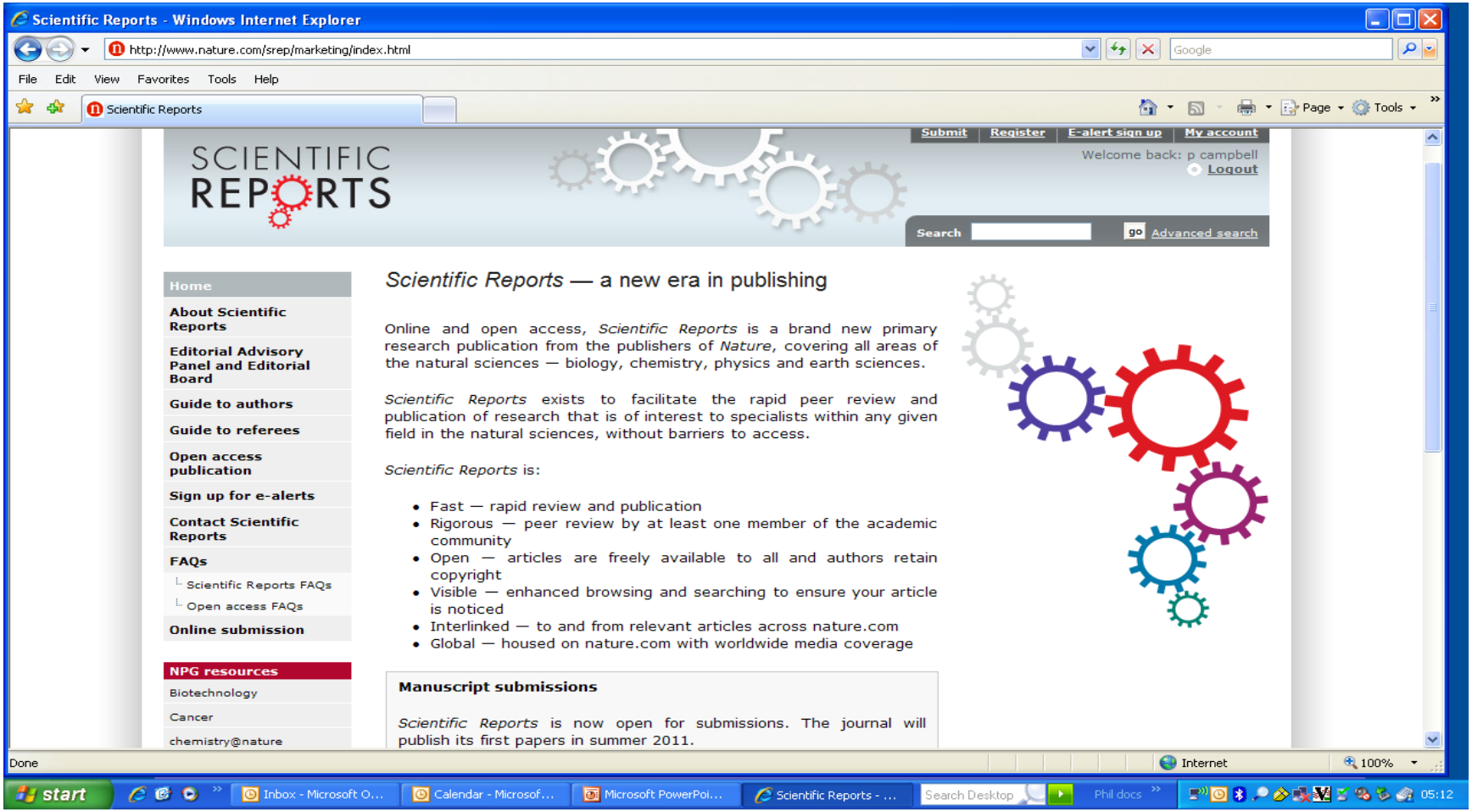




Evolving business model: Online-only 'hybrid' *Nature Communications*



Minimum-threshold publishing



The editorial resource

- Nature + research journals: about 100 chief eds and research editors.
- Review journals: about 50 commissioning editors
- Nature + NBT + N Med: about 40 magazine staff editors/reporters
- Copy editors, admin, production
- All of these are sources of added value (and are widely recognised as such ‘out there’)
- But how do we add value, exactly?

The *Nature* editors who select papers
(*Nature* has never had an editorial board.)

- 25 editors, full-time professionals, age 30-50+
- Recruited from successful post-docs or faculty
- Selected for ability to comprehend and assess across a discipline and beyond
- UK, Indian, French, German, US, Dutch, Italian…
- Biological, chemical and physical sciences
- Five+ extended meetings/visits each per year, plus shorter trips
- Over 11,000 submissions per year, each editor considers ~10 per week
- Reject ~65% immediately
- Referee 35%,
- accept 8% of the total

Strong contender for review

- Addresses an interesting question
- Strong, well-controlled data
- Rules out some alternative explanations
- Speculation doesn’t “stretch the data”
- Discussion puts paper in perspective
- Provides strong insight or other scientific value

Referee Assessment

Manuscript#:

Referee:

Corresponding Author:

Due Date:

Title:

1. Comments for Author

These remarks will be sent to the authors and possibly other peer reviewers.
Guidelines for Referees can be found here:
http://www.nature.com/authors/editorial_policies/peer_review.html

Please incorporate the points below into your comments to Authors.

A. Summary of the key results
B. Originality and significance: if not novel, please include reference
C. Data & methodology: validity of approach, quality of data, quality of presentation
D. Appropriate use of statistics and treatment of uncertainties
E. Conclusions: robustness, validity, reliability
F. Suggested improvements: experiments, data for possible revision
G. References: appropriate credit to previous work?
H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

A good reviewer will:

- Provide a good review and also…
- Sometimes or often involve younger colleagues (in confidence)
- Nurture high standards in the next generation of peer reviewers

Multi-disciplinarity challenges: research assessment

Editorial and peer-review assessment challenges that affect funding agencies too:

- *finding appropriate referees across all relevant disciplines and integrating their assessments*
- *thinking imaginatively and holistically about potential interest in a submission or proposal*
- *understanding alien concepts and language*
- *independent overview and strong judgement*

Even if there are no technical flaws, editors often face contradictory reviewers' recommendations….

* Diverse technical expertise
* Diverse conceptual backgrounds
* Judged on their own terms

….. and always make their own decisions

- which sometimes requires them to overrule all the referees, and publish!

What will we add value to?

Evolving research literature: from this….

LETTERS

Common variants conferring risk of schizophrenia

A list of authors and their affiliations appears at the end of the paper

Schizophrenia is a complex disorder, caused by both genetic and environmental factors and their interactions. Research on pathogenesis has traditionally focused on neurotransmitter systems in the brain, particularly those involving dopamine. Schizophrenia has been considered a separate disease for over a century, but in the absence of clear biological markers, diagnosis has historically been based on signs and symptoms. A fundamental message emerging from genome-wide association studies of copy number variations (CNVs) associated with the disease is that its genetic basis does not necessarily conform to classical nosological disease boundaries. Certain CNVs confer not only high relative risk of schizophrenia but also of other psychiatric disorders[1–3]. The structural variations associated with schizophrenia can involve several genes and the phenotypic syndromes, or the 'genomic disorders', have not yet been characterized[4]. Single nucleotide polymorphism (SNP)-based genome-wide association studies with the potential to implicate individual genes in complex diseases may reveal underlying biological pathways. Here we combined SNP data from several large genome-wide scans and followed up the most significant association signals. We found significant association with several markers spanning the major histocompatibility complex (MHC) region on chromosome 6p21.3-22.1, a marker located upstream of the neurogranin gene (*NRGN*) on 11q24.2 and a marker in intron four of transcription factor 4 (*TCF4*) on 18q21.2. Our findings implicating the MHC region are consistent with an immune component to schizophrenia risk, whereas the association with *NRGN* and *TCF4* points to perturbation of pathways involved in brain development, memory and cognition.

To begin our search for sequence variants associated with schizophrenia, we performed a genome-wide scan of 2,663 schizophrenia cases and 13,498 controls from eight European locations (England, Finland (Helsinki), Finland (Kuusamo), Germany (Bonn), Germany (Munich), Iceland, Italy and Scotland; collectively called SGENE-plus) using the Illumina HumanHap300 and HumanHap550 BeadChips. In total, 314,868 SNPs meeting our quality control criteria were included in an allelic association analysis. To adjust for relatedness and potential population stratification, genomic control was applied to each study group.

None of the markers gave *P* values smaller than our genome-wide significance threshold of 0.05/314,868, or approximately 1.6×10^{-7} (see Supplementary Fig. 1 for a quantile-quantile plot and Supplementary Table 1 for markers with the smallest *P* values). Next, we combined findings from our top 1,500 markers with results for the same markers (or surrogates for them) from both the International Schizophrenia Consortium[5] (excluding the Scottish samples overlapping with samples in our study, 2,602 cases and 2,885 controls) and the European–American portion of the Molecular Genetics of Schizophrenia[6] (2,681 cases and 2,653 controls) study. Twenty-five of our top 1,500 markers (or eighteen counting very strongly correlated ($r^2 > 0.8$) markers only once) had *P* values less than 1×10^{-5} in the combined results (Supplementary Table 2). These top markers were followed up in as many as 4,999 cases and 15,555 controls from four sets of additional samples from Europe (set 1, 715 cases and 3,634 controls from the Netherlands; set 2, 3,330 cases and 6,892 controls from Denmark (Aarhus), Denmark (Copenhagen), Germany (Bonn), Germany (Munich), Hungary, the Netherlands, Norway, Russia and Sweden; set 3, 287 cases and 3,987 controls from Finland; set 4, 667 cases and 1,042 controls from Spain (Santiago) and Spain (Valencia)) (Supplementary Table 3).

Three markers, all in the extended MHC region on the short arm of chromosome 6, showed genome-wide significance in the combination of SGENE-plus and the follow-up samples described above (Table 1). In addition, four other markers—two in the MHC region, one at 11q24.2 and one at 18q21.2—showed genome-wide significance when results from the International Schizophrenia Consortium and the Molecular Genetics of Schizophrenia study were included (Table 1).

In the MHC region on chromosome 6p21.3-22.1, the five genome-wide significant markers (*P* ranging from 1.1×10^{-9} to 1.4×10^{-12} in all samples combined) have risk alleles with average control frequencies between 78% and 92% (Table 1). Combined odds ratios (ORs) for the markers range from 1.15 to 1.24 (Table 1) with no significant heterogeneity between the study groups ($P > 0.25$, Supplementary Table 4). For all of the markers, the multiplicative model for risk provides an adequate fit ($P > 0.62$).

Despite spanning about five megabases (Mb), the five chromosome 6p markers cover only about 1.4 centimorgans (cM) and substantial linkage disequilibrium exists between them (Supplementary Table 5). The association of rs6932590 (the most significant marker), however, cannot account for all of the association of the four remaining markers (Supplementary Table 6). Most notably, conditional on rs6932590, rs3131296 has an association *P* value of 3.4×10^{-6}, indicating that rs3131296 may be capturing a second susceptibility variant or that both rs6932590 and rs3131296 are correlated with a third, higher risk, variant not examined here.

To examine association of the genome-wide significant SNPs in the 5-Mb region on 6p21.3-22.1 with classical human leukocyte antigen (HLA) alleles, long-range phasing haplotypes[7] tagging the major alleles at the *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQA1* and *HLA-DQB1* loci in Icelanders were used. Only rs3131296 shows substantial ($r^2 > 0.5$) correlation with any of the classical HLA alleles tested; this marker has an r^2 of 0.86 with *DRB1*03* and an r^2 of 0.81 with *HLA-B*08*. Simplified tags for these two classical alleles, appropriate for the European samples of SGENE-plus, had effects that were not statistically distinguishable from the effect of rs3131296. In the case of both *DRB1*03* and *HLA-B*08*, the classical HLA allele is paired with the protective allele of rs3131296, making the results described here consistent with the under-transmission of *DRB1*03* to schizophrenic offspring reported previously[8].

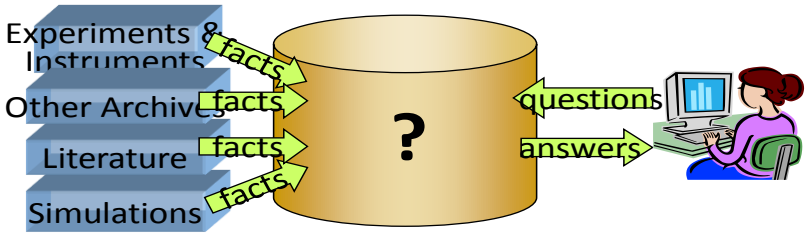
Many autoimmune and infectious diseases have been associated with *DRB1*03* and, indeed, inspection of top MHC region SNPs from recent genome-wide association scans of three of these—type I diabetes[9], coeliac disease[10] and systemic lupus erythematosus[11]—reveals, for each disease, SNPs having a HapMap CEU (Utah residents with ancestry from northern and western Europe) r^2 of at least 0.73

..to this: the scientific edifice

(image by the late Jim Gray, Microsoft)

X-Info

- The evolution of X-Info and Comp-X
  for each discipline X
- How to codify and represent our knowledge



The Generic Problems

- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to *re*organize it
- How to share with others

- Query and Vis tools
- Building and executing models
- Integrating data and Literature
- Documenting experiments
- Curation and long-term preservation

..which is a vision of..

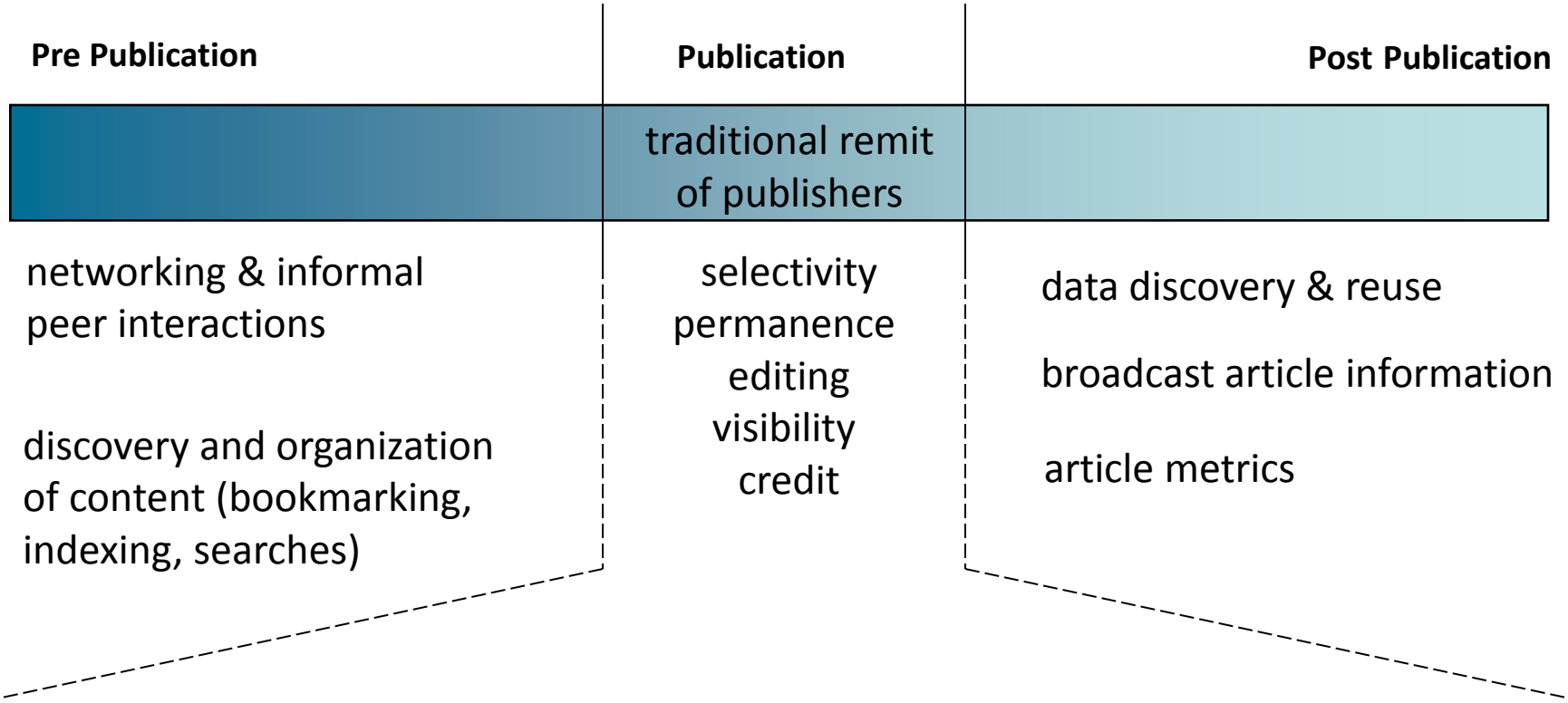
.. an edifice of text, numbers, equations, data, software, images, graphics, videos, tables, annotations and metadata that is:

- Seamless and hyperdimensional
- Readable, searchable and computable by humans and machines
- Checked and ranked by editors and crowds
- Sourced from journals, labs, community databases, institutional databases, meta-analyses, grey literatures,...
- Underpinned by community standards of nomenclature, annotation, sharing and integrity (technical and ethical)
- Free to all users
- Financially sustainable

How does/will a publisher of primary research add value?

- Selectivity
  - for validity, quality, impact and long-term scientific significance
- Editing
  - language, sense, technical accuracy, community standards
- Distribution and visibility
  - via multiple channels: print, online, mobile, index
  - enhance discoverability: metadata, navigation
  - Context and comments
  - Amplification to society eg via media
- Permanence and credit
  - an accurate permanent record in durable online formats
  - credit for scientific contributions
- Assisting the scientific community
  - Credit and accountabilities (funding, scientific, ethical)
  - Contribution to strategic thinking in the community
  - Providing tools that assist researchers in their publishing and lab activities

An expanding remit



credit: author contributions author disambiguation referees	*Editing*: Community standards Protocols Annotations	*Permanence*: formats corrections context	*Visibility*: via multiple platform content , data, metadata formats and standards

Who else needs to add value?

- Researchers are not much interested in post-publication peer review, but some individual and collective blogs are influential.
- Funding agencies need to provide funds for e-infrastructure and author-fees and may also to provide tools and facilities – eg NIH (US), ESRC (UK) are rare examples
- Universities and research institutions: libraries and IT departments to provide repositories and information-tools
- Researchers themselves will add value by also exploiting generic services – Google, Facebook, websites – and building tools that serve their communities, …
- But these roles will not be adopted easily and are not specifically funded.
- Timescale of change: several years

..and tools usable by researchers (examples from a workshop).

- **Platforms: Project collaboration software, "smart" laboratory software, and provenance systems**
  David de Roure, Professor of e-Research, Osford e-Research Centre (myExperiment)
- Tim Clark, Director of Bioinformatics, MassGeneral Institute for Neurodegenerative Disease, Harvard Medical School (SWAN)
- Rafael Sidi, Vice President, Product Management, Elsevier (SciVerse)

- **Media: Production, distribution, archiving (for example, video, 3-D modeling, and databases)**
  Phil Bourne, Professor of Pharmacology, University of California, San Diego (SciVee and BioLit)
- Moshe Pritsker, CEO, Editor-in-Chief, Co-founder, Journal of Visualized Experiements (JoVE)
- Chris Lintott, Director of Citizen Science at the Adler Planetarium (Zooniverse)
- Taliesin Beynon, Research Programmer Wolfram|Alpha (Computable Document Format—CDF)

- **Literature: Publications based on text and still images (creation, reviewing, dissemination, archiving, and reproducibility)**
  Kaitlin Thaney, Manager of External Partnerships, Digital Science (Read-Cube)
- Alex Wade, Director or Scholarly Communication, Microsoft Research (Chemistry Add-in for Word and Article Authoring Add-In)
- Charles Parnot, Member, Team Mekentosj (Demo of Papers for iPad) | video

- **Review: Standard publication-based systems, alternative rating systems, etc.**
- Peter Binfield, Publisher, Public Library of Science (Article Level Metrics)
- Sarah Greene, Editor-in-Chief, Faculty of 1000 (F1000)
- Jason Priem, University of North Carolina at Chapel Hill (Total-Impact)

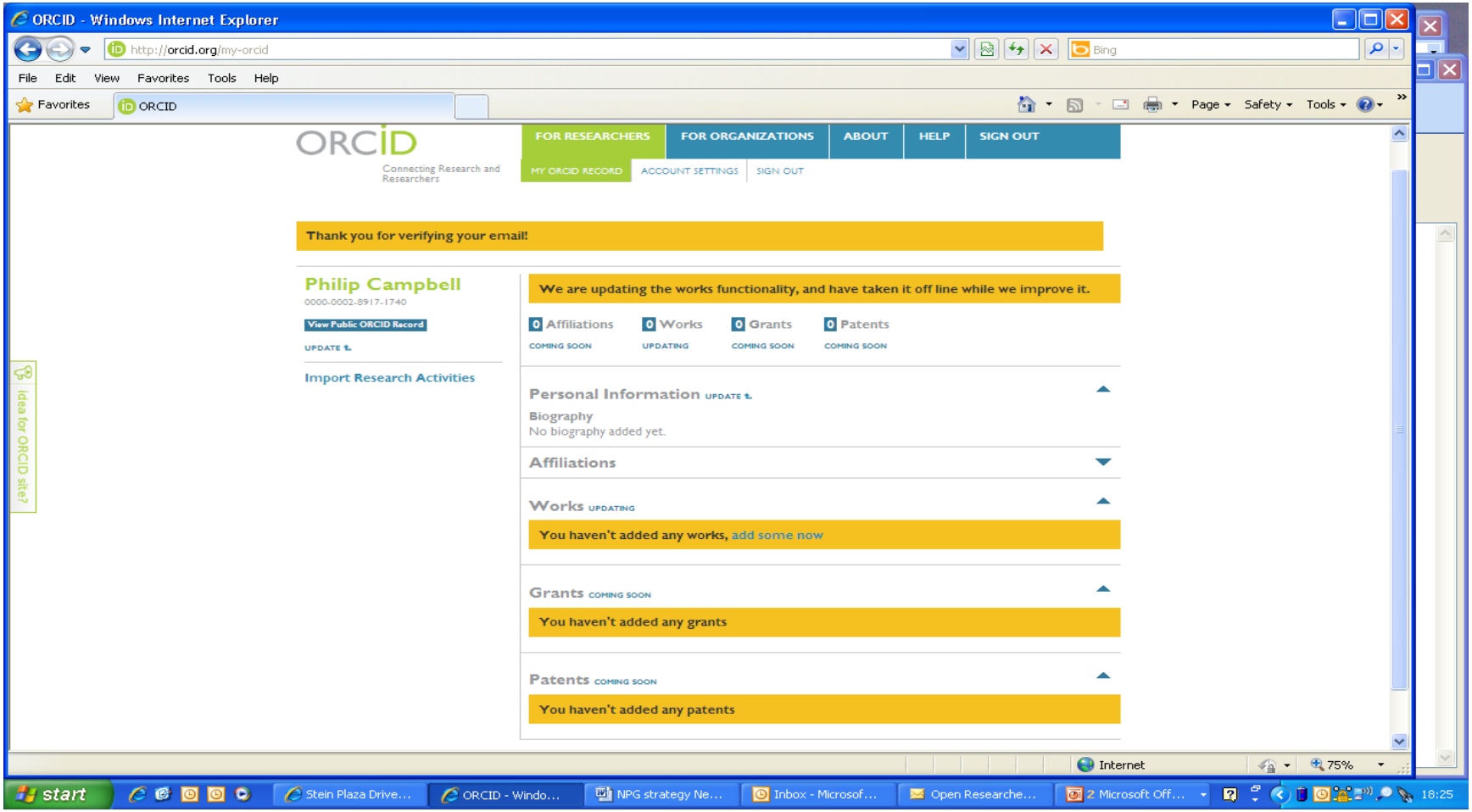
- **Resources: Seamless technologies for literature and data (literature/data search engines; cloud-based, group sharing, adjustable permissions, and integration with search)**
- Alberto Accomazzi, Project Manager, NASA Astrophysics Data System, Harvard-Smithsonian Center for Astrophysics (ADS and ADS Labs)
- Mercé Crosas, Director of Product Development, IQSS, Harvard University (Dataverse)
- Ashfaq Munshi, Founder and CEO, Terabitz (Terabitz)

- **Recognition: How can we best enable cooperation and adoption?**
  Jevin West, University of Washington (Eigenfactor)
- Jessica Mezei, Mendeley Community Liaison (Mendeley)
- Lee Dirks, Director of Education and Scholarly Communication, Microsoft Research Connections (Microsoft Academic Search)

Researchers' credit as….

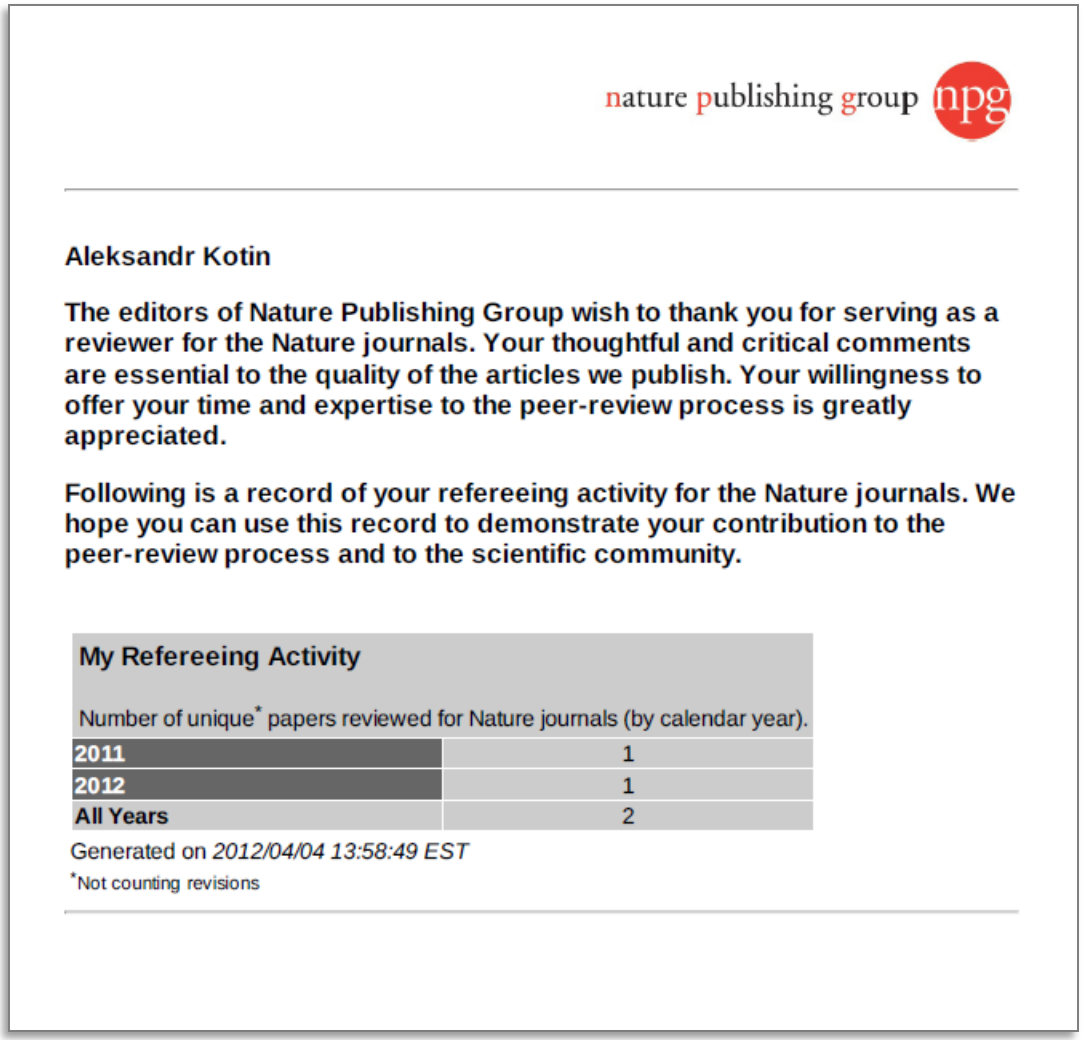
- Authors of papers and other research texts
- Scientific contributors – data, tools, …
- Referees
- Educators, communicators

Open Research and Contributor ID



Referee credit

Anyone having reviewed for a Nature journal may obtain an acknowledgement of their refereeing activity for the Nature journals



nature publishing group

Aleksandr Kotin

The editors of Nature Publishing Group wish to thank you for serving as a reviewer for the Nature journals. Your thoughtful and critical comments are essential to the quality of the articles we publish. Your willingness to offer your time and expertise to the peer-review process is greatly appreciated.

Following is a record of your refereeing activity for the Nature journals. We hope you can use this record to demonstrate your contribution to the peer-review process and to the scientific community.

My Refereeing Activity

Number of unique* papers reviewed for Nature journals (by calendar year).

2011	1
2012	1
All Years	2

Generated on *2012/04/04 13:58:49 EST*

*Not counting revisions

Available since April 2012 for download from 'My Account' page in manuscript submission system of all Nature journals – reflects activity across all journals

Transmission of electrical signals by spin-wave interconversion in a magnetic insulator

Y. Kajiwara[1,2], K. Harii[1], S. Takahashi[1,3], J. Ohe[1,3], K. Uchida[1], M. Mizuguchi[1], H. Umezawa[5], H. Kawai[5], K. Ando[1,2], K. Takanashi[1], S. Maekawa[1,3] & E. Saitoh[1,2,4]

1. Institute for Materials Research, Tohoku University, Sendai 980-8577, Japan
2. Department of Applied Physics and Physico-Informatics, Keio University, Yokohama 223-8522, Japan
3. CREST,
4. PRESTO, Japan Science and Technology Agency, Sanbancho, Tokyo 102-0075, Japan
5. FDK Corporation, Shizuoka 431-0495, Japan

Correspondence to: E. Saitoh[1,2,4] Correspondence and requests for materials should be addressed to E.S. (Email: saitoheiji@imr.tohoku.ac.jp).

The energy bandgap of an insulator is large enough to prevent electron excitation and electrical conduction[1]. But in addition to charge, an electron also has spin[2], and the collective motion of spin can propagate—and so transfer a signal—in some insulators[3]. This motion is called a spin wave and is usually excited using magnetic fields. Here we show that a spin wave in an insulator can be generated and detected using spin-Hall effects, which enable the direct conversion of an electric signal into a spin wave, and its subsequent transmission through (and recovery from) an insulator over macroscopic distances. First, we show evidence for the transfer of spin angular momentum between an insulator magnet $Y_3Fe_5O_{12}$ and a platinum film. This transfer allows direct conversion of an electric current in the platinum film to a spin wave in the $Y_3Fe_5O_{12}$ via spin-Hall effects[4, 5, 6, 7, 8, 9, 10, 11]. Second, making use of the transfer in a $Pt/Y_3Fe_5O_{12}/Pt$ system, we demonstrate that an electric current in one metal film induces voltage in the other, far distant, metal film. Specifically, the applied electric current is converted into spin angular momentum owing to the spin-Hall effect[7, 8, 10, 11] in the first platinum film; the angular momentum is then carried

Author contributions statement

- **Author Contributions** Y.K., K.H., K.U. and K.A. performed the measurements and analysed the data; J.O. carried out the numerical analysis; S.T., S.M. and E.S. provided the theoretical analysis; H.U. and H.K. contributed to the sample fabrication; Y.K., K.H, K.U., M.M. and K.T. contributed to the experimental set-up; Y.K., S.T., J.O., K.U., M.M., H.U., K.T., S.M. and E.S. wrote the manuscript; all authors discussed the results and commented on the manuscript; and E.S. planned and supervised the project.

Question about developing author-contribution transparency

- Author contribution statements in Nature journals are informal, unstructured, non-templated.
- Should this change? How? (Possible goals: increased credit, increased accountability for potential flaws.)
- How granular should this information become?

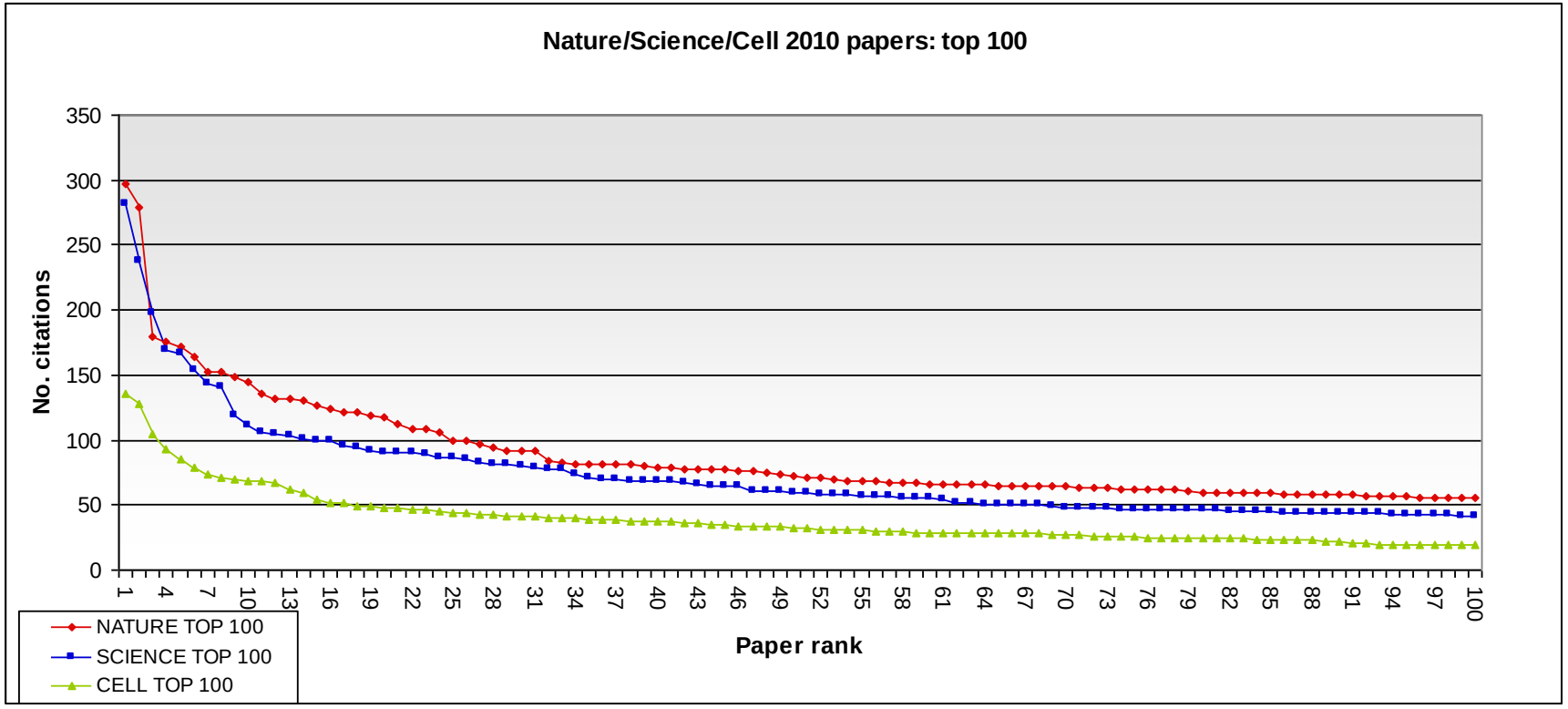
Impacts of research in *Nature*

The impacts of research papers can be measured in many ways, including:

- ***Citations in the scientific literature***
- Discussion in society
- Fame, fortune – or notoriety
- Technological applications

Citation tracking

measured Nov 2011



Other post-publication histories: fame and fortune

- Media coverage
- Chair at Stanford
- Substantial pay bonuses eg in Chinese labs
- Grants: “one Nature paper equivalent to 1 million Euros”
- Investments in companies.

Notoriety: frauds

- Libel laws may prevent open discussion of reasons for retraction
- Hwang and Schoen mega-misconducts both highlight the duty of co-authors to validate spectacular results, even if they are in other labs and/or countries.

Irreproducibility: manifestations

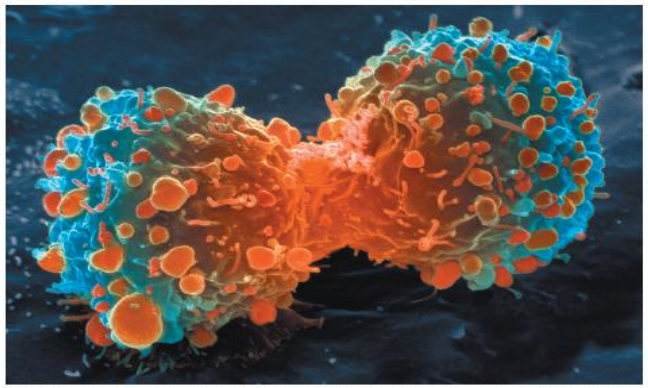
- Growth in formal corrections
- Failures to replicate by biotech and pharma
- Public discussions

483531a[2].pdf - Adobe Reader

File Edit View Document Tools Window Help

531 (1 of 3) 66.7% Find

COMMENT



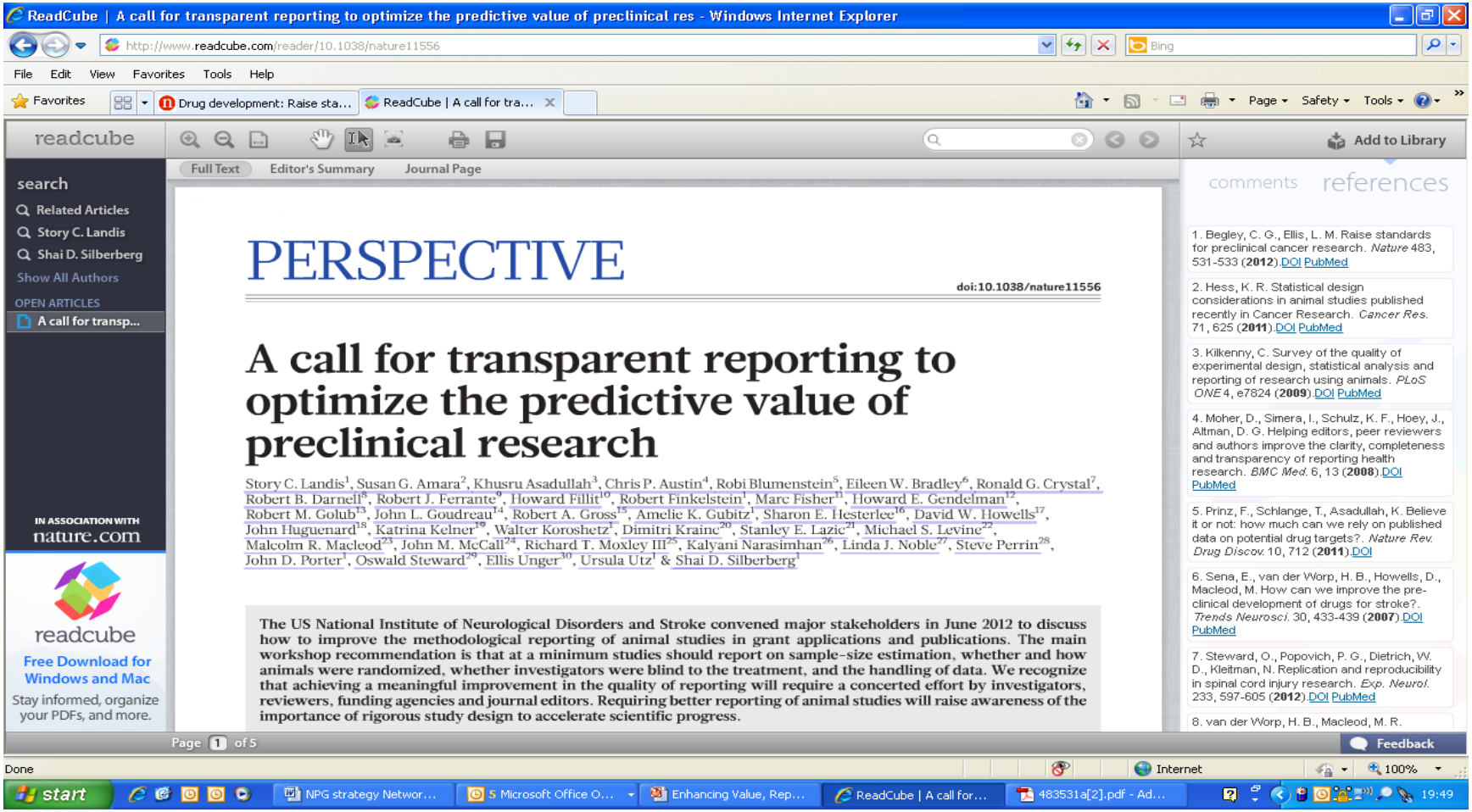
Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and **Lee M. Ellis** propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to | trials in oncology have the highest failure | investigators must reassess their approach to

start | NPG strategy Networ... | 5 Microsoft Office O... | Enhancing Value, Rep... | Drug development: R... | 483531a[2].pdf - Ad... | 19:44


ReadCube | A call for transparent reporting to optimize the predictive value of preclinical res - Windows Internet Explorer
http://www.readcube.com/reader/10.1038/nature11556
File Edit View Favorites Tools Help
Favorites
Drug development: Raise sta...
ReadCube | A call for tra...
Page Safety Tools
readcube
Add to Library
Full Text
Editor's Summary
Journal Page
search
Related Articles
Story C. Landis
Shai D. Silberberg
Show All Authors
OPEN ARTICLES
A call for transp...
IN ASSOCIATION WITH
nature.com
readcube
Free Download for Windows and Mac
Stay informed, organize your PDFs, and more.
comments
references
PERSPECTIVE
doi:10.1038/nature11556
A call for transparent reporting to optimize the predictive value of preclinical research
Story C. Landis[1], Susan G. Amara[2], Khusru Asadullah[3], Chris P. Austin[4], Robi Blumenstein[5], Eileen W. Bradley[6], Ronald G. Crystal[7], Robert B. Darnell[8], Robert J. Ferrante[9], Howard Fillit[10], Robert Finkelstein[1], Marc Fisher[11], Howard E. Gendelman[12], Robert M. Golub[13], John L. Goudreau[14], Robert A. Gross[15], Amelie K. Gubitz[1], Sharon E. Hesterlee[16], David W. Howells[17], John Huguenard[18], Katrina Kelner[19], Walter Koroshetz[1], Dimitri Krainc[20], Stanley E. Lazic[21], Michael S. Levine[22], Malcolm R. Macleod[23], John M. McCall[24], Richard T. Moxley III[25], Kalyani Narasimhan[26], Linda J. Noble[27], Steve Perrin[28], John D. Porter[1], Oswald Steward[29], Ellis Unger[30], Ursula Utz[1] & Shai D. Silberberg[1]
The US National Institute of Neurological Disorders and Stroke convened major stakeholders in June 2012 to discuss how to improve the methodological reporting of animal studies in grant applications and publications. The main workshop recommendation is that at a minimum studies should report on sample-size estimation, whether and how animals were randomized, whether investigators were blind to the treatment, and the handling of data. We recognize that achieving a meaningful improvement in the quality of reporting will require a concerted effort by investigators, reviewers, funding agencies and journal editors. Requiring better reporting of animal studies will raise awareness of the importance of rigorous study design to accelerate scientific progress.
1. Begley, C. G., Ellis, L. M. Raise standards for preclinical cancer research. Nature 483, 531-533 (2012).DOI PubMed
2. Hess, K. R. Statistical design considerations in animal studies published recently in Cancer Research. Cancer Res. 71, 625 (2011).DOI PubMed
3. Kilkenny, C. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PLoS ONE 4, e7824 (2009).DOI PubMed
4. Moher, D., Simera, I., Schulz, K. F., Hoey, J., Altman, D. G. Helping editors, peer reviewers and authors improve the clarity, completeness and transparency of reporting health research. BMC Med. 6, 13 (2008).DOI PubMed
5. Prinz, F., Schlange, T., Asadullah, K. Believe it or not: how much can we rely on published data on potential drug targets?. Nature Rev. Drug Discov. 10, 712 (2011).DOI
6. Sena, E., van der Worp, H. B., Howells, D., Macleod, M. How can we improve the pre-clinical development of drugs for stroke?. Trends Neurosci. 30, 433-439 (2007).DOI PubMed
7. Steward, O., Popovich, P. G., Dietrich, W. D., Kleitman, N. Replication and reproducibility in spinal cord injury research. Exp. Neurol. 233, 597-605 (2012).DOI PubMed
8. van der Worp, H. B., Macleod, M. R.
Page 1 of 5
Feedback
Done
Internet
100%
start
NPG strategy Networ...
5 Microsoft Office O...
Enhancing Value, Rep...
ReadCube | A call for...
483531a[2].pdf - Ad...
19:49

"A call for transparent reporting to optimize the predictive value of preclinical research"

- **Randomization**
- •Animals should be assigned randomly to the various experimental groups, and the method of randomization reported.
- •Data should be collected and processed randomly or appropriately blocked.

- **Blinding**
- •Allocation concealment: the investigator should be unaware of the group to which the next animal taken from a cage will be allocated.
- •Blinded conduct of the experiment: animal caretakers and investigators conducting the experiments should be blinded to the allocation sequence.
- •Blinded assessment of outcome: investigators assessing, measuring or quantifying experimental outcomes should be blinded to the intervention.

- **Sample-size estimation**
- •An appropriate sample size should be computed when the study is being designed and the statistical method of computation reported.
- •Statistical methods that take into account multiple evaluations of the data should be used when an interim evaluation is carried out.

- **Data handling**
- •Rules for stopping data collection should be defined in advance.
- •Criteria for inclusion and exclusion of data should be established prospectively.
- •How outliers will be defined and handled should be decided when the experiment is being designed, and any data removed before analysis should be reported.
- •The primary end point should be prospectively selected. If multiple end points are to be assessed, then appropriate statistical corrections should be applied.
- •Investigators should report on data missing because of attrition or exclusion.
- •Pseudo replicate issues need to be considered during study design and analysis.
- •Investigators should report how often a particular experiment was performed and whether results were substantiated by repetition under a range of conditions.

Landis *et al.*, *Nature* **490** 187–191 (11 October 2012) doi:10.1038/nature11556

Irreproducibility: actions

- Awareness raising - meetings: NINDS, NCI, Nature……
- Awareness raising: publications
- Editorial check lists
- Guidance for authors
- Guidance for referees

Irreproducibility: underlying issues

- Experimental design: randomization, blinding, sample size determinations, independent experiments vs technical replicates, replicability achieved plus reasons
- Statistics
- Big data: false positives, gut scepticism/tacit knowledge
- Gels, microscopy images,
- Reagents validity – antibodies, cell lines
- Animal studies description
- Methods description
- Data deposition
- IP confidentiality – replication failures unpublishable
- Publication of refutations – where?
- **Lab supervision**
- **Lab training**
- **Pressures to publish**

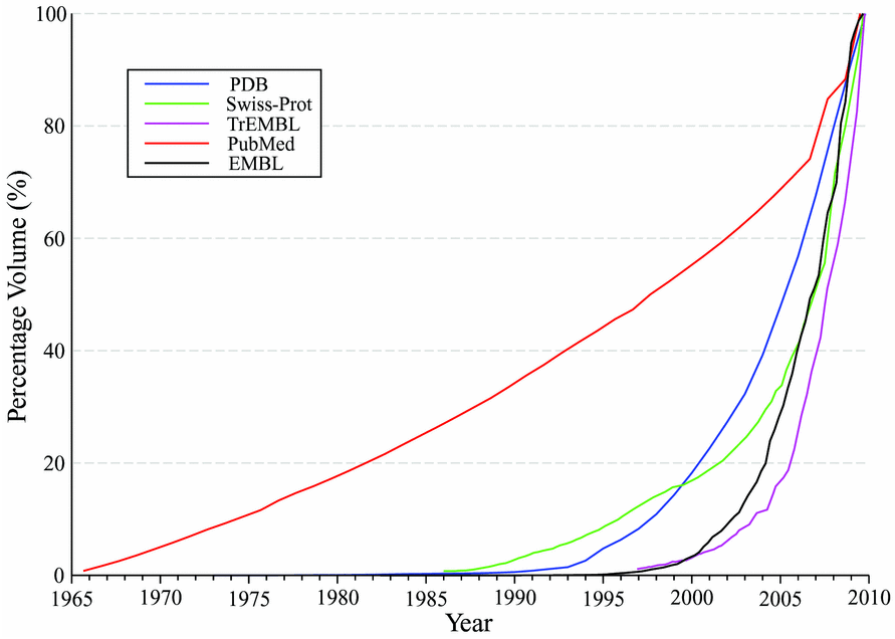
Science as an open enterprise

June 2012

THE ROYAL SOCIETY

Data, data, data….

Depositions of datasets in archives continue to grow, surpassing journal articles in biomedical research



Growth of biomedical research publications (red; current total >19 million), alongside the accumulation of research data, including nucleic acid sequences (black; current total ~163 million), computer-annotated protein sequences (magenta; current total 9 million), manually annotated protein sequences (green; current total 500,000) and protein structures (blue; current total 60,000)

Source: Biochemical Journal 2009 424, 317-333 - Teresa K. Attwood, Douglas B. Kell and others.

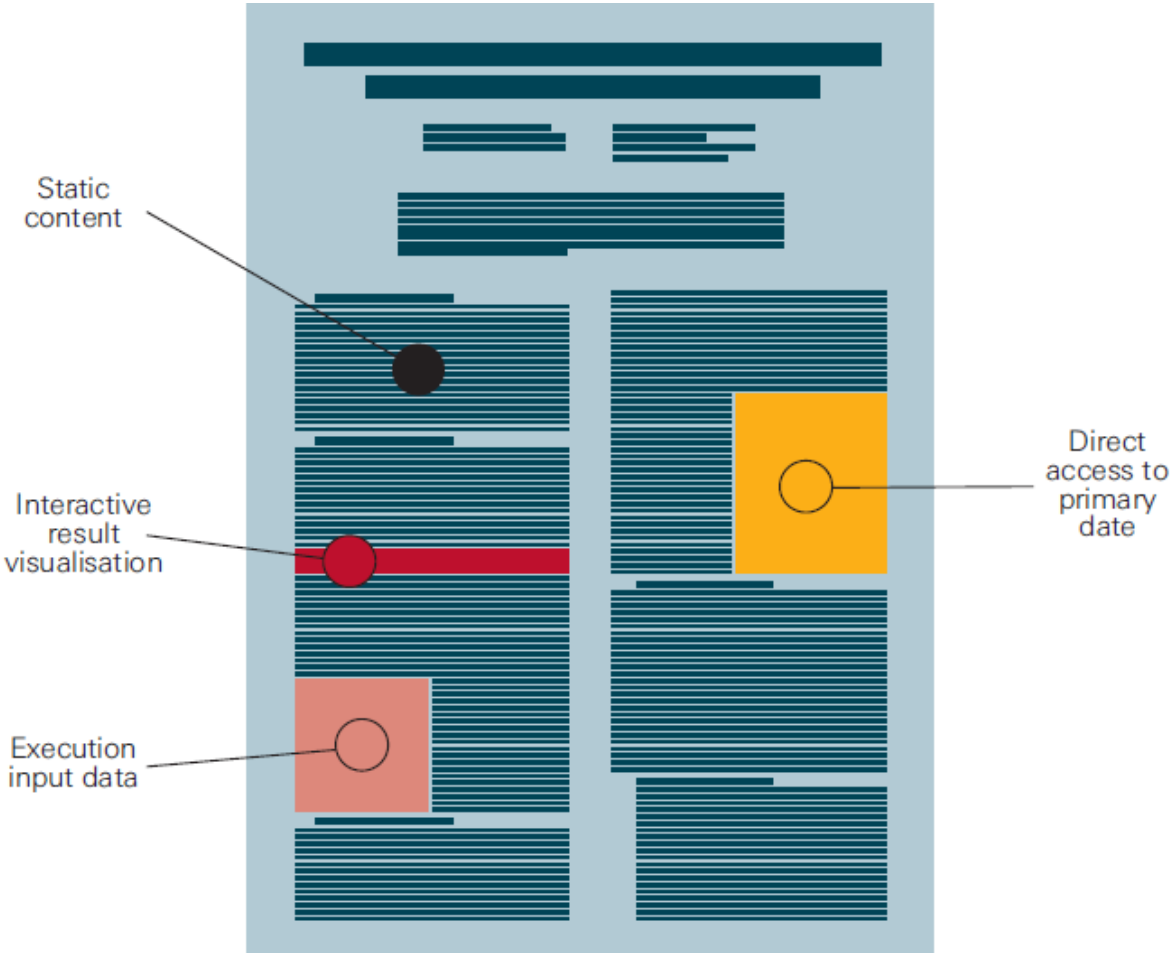
Why is open data an urgent issue?

- Maintaining and improving trust in science
- Sustaining replication and reproducibility
- Combating fraud
- Exploiting the data deluge & computational potential – eg looking for the unexpected, eg addressing planetary challenges
- Supporting citizen science
- Responding to citizens' demands for evidence

EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

Aspiration: all scientific literature online, all data online, and for them to interoperate



nature publishing group npg

Intelligent openness

Openness of data *per se* has no value. Open science is more than disclosure

Data must be:

- Accessible
- Intelligible
- Assessable
- Re-usable

METADATA

Only when these four criteria are fulfilled are data properly open

Boundaries of openness?

Legitimate commercial interests

Privacy (complete anonymisation is impossible)

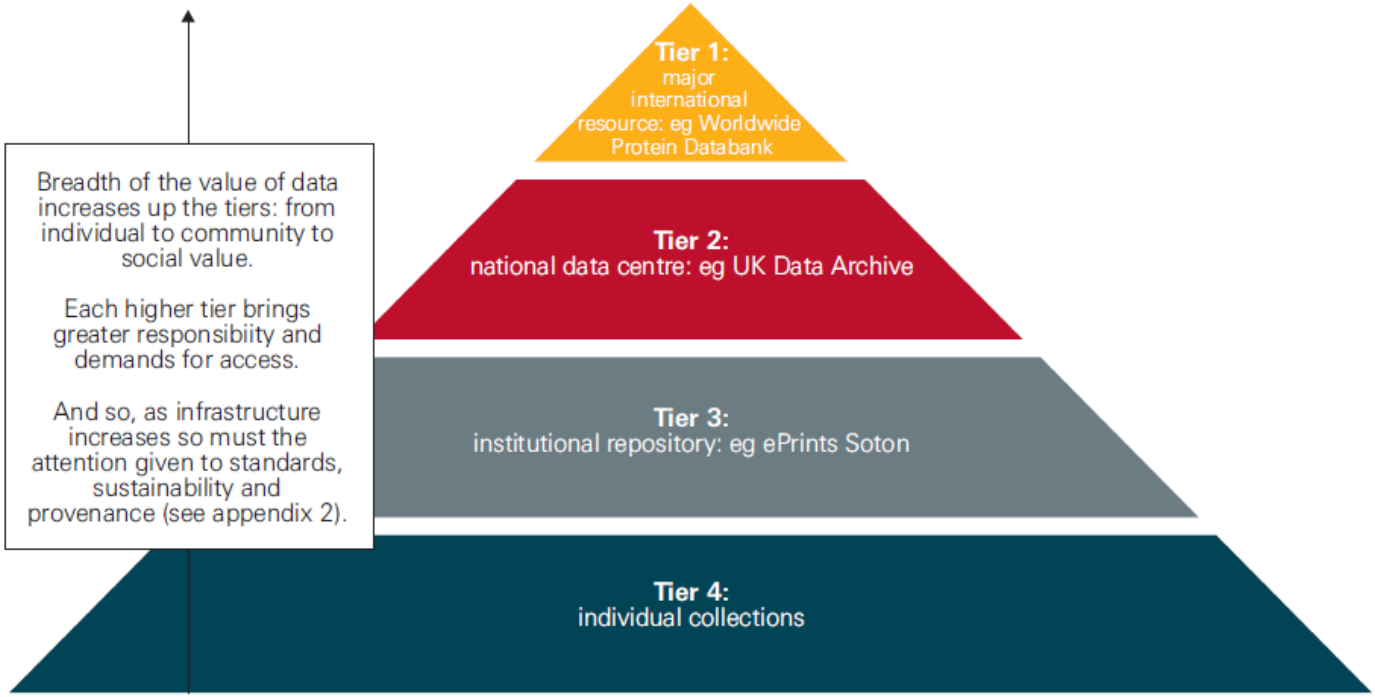
Safety

Security

But the boundaries are fuzzy & complex

The transition to open data

Pathfinder disciplines where benefit is recognised and habits are changing



Worldwide Protein Data Bank (wwPDB)

- The Worldwide Protein Data Bank (wwPDB) archive is the single worldwide repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids. It was founded in 1971, and is managed by the Worldwide PDB organisation (wwpdb.org).

- As of January 2012, it held 78477 structures. 8120 were added in 2011, at a rate of 677 per month. In 2011, an average of 31.6 million data files were downloaded per month. The total storage requirement for the repository was 135GB for the archive.

- **The total cost for the project is approximately $11-12 million per year (total costs, including overhead), spread out over the four member sites. It employs 69 FTE staff.** wwPDB estimate that $6-7 million is for “data in” expenses relating to the deposition and curation of data.

Databases as publications

- Hosts/suppliers of databases are publishers
- They have a responsibility to curate and provide reliable access to content.
- They may also deliver other services around their products
- They may provide the data as a public good or charge for access

Worldwide Protein Data Bank (wwPDB)

wwPDB – Services Provided	
Platform provision, maintenance and development?	Yes
Multiple formats or versions (eg PDF, html, postscript, latex; multiple revisions of datasets)?	Yes
'Front end' - web-access to pages?	Yes
Registration and access control?	No
Input quality control: consistency with format standards, adequate technical standards?	Yes
Input quantity control: ensure community coverage?	Yes
Add metadata and references to describe authorship, provenance and experimental or simulation context?	Yes
Provide accession number to log deposition?	Yes
Alert registrants to new additions?	Yes
Provide means by which the data can be cited and credited to originators?	Yes
Host or link to relevant analysis tools (eg visualisation, statistics)?	Yes
Measure and document impact: downloads, data citations?	Yes

UK Data Archive

- The UK Data Archive, founded 1967, is curator of the largest collection of digital data in the social sciences in the United Kingdom. It contains several thousand datasets relating to society, both historical and contemporary.
- The main storage ‘repository’ holds multiple versions of approx 1.26 million files (ie digital objects).
- On average around 2,600 (new or revised) files are uploaded to the repository monthly. The baseline size of the main storage repository is <1Tb, though with multiple versions and files outside this system, a total capacity of c.10Tb is required.
- **The UKDA currently (26/1/2012) employs 64.5 people. The total expenditure of the UK Data Archive (2010-11) was approx £3.43 million.**

arXiv.org

arXiv.org is internationally acknowledged as a pioneering and successful digital archive and open-access distribution service for preprints and e-prints of research articles. Funded and hosted by Cornell University Library and contributing institutions.

As of January 2012, it held over 750,000 articles. Around 7,300 are added per month. The size of the repository is currently 263GB.

arXiv.org employs just over six people. Its projected running costs for 2012 (including indirect costs) are in the region of $810,000 per year, of which roughly $670,000 are staff costs. Storage and computing infrastructure accounts for around $45,000 per year.

Dryad

Dryad (datadryad.org) is a repository of data underlying peer reviewed articles in the basic and applied biosciences.

Dryad closely coordinates with journals to integrate article and data submission. The repository is community driven, governed and sustained by a consortium of scientific societies, publishers, and other stakeholder organisations. Dryad currently hosts data from over 100 journals, from many different publishers, institutions, and countries of origin.

As of 24 January 2012, Dryad contained 1280 data packages and 3095 data files, associated with articles in 108 journals. 79 new data packages in December, 2011, with approximately 2.3 files per data package. Its current size is 0.05 TB.

Dryad has 4-6 FTE, with 50% devoted to operational core and 50% to R&D. Its total budget is around $350,000 per year, with staff costs of approximately $300,000, and $5,000-$10,000, of infrastructure costs including subscription services (eg DataCite, LOCKSS, etc.).

Business plan: revenues from payments for the submission of new data deposits cover the repository's operating costs (including curation, storage, and software maintenance).

The primary production server is maintained by the North Carolina State University Digital Library Program. The Dryad is currently applying to the State of North Carolina and the US IRS to be recognised as an independent not-for-profit organisation.

Institutional Repositories

The Repositories Support Project survey in 2011 received responses from 75 UK universities. It found that the average university repository employed a total 1.36 FTE – combined into Managerial, Administrative and Technical roles. 40% of these repositories accept research data. In the vast majority of cases (86%), the library has lead responsibility for the repository.

ePrints Soton

ePrints Soton, founded in 2003, is the institutional repository for the University of Southampton. It holds publications including journal articles, books and chapters, reports and working papers, higher theses, and some art and design items. It is looking to expand its holdings of datasets.

It currently has metadata on 65,653 items. The majority of these lead to an access request facility or point to open access material held elsewhere.

There are 46,758 downloads per month, and an average of 826 new uploads every month. The total size of the repository is 0.25TB.

It has a staff of 3.2 FTE (1FTE technical, 0.9 senior editor, 1.2 editors, 0.1 senior manager). Total costs of the repository are of £116, 318.

Priorities for action

1. a shift away from a research culture where data is viewed as a **private preserve**

2. expanded criteria used to **evaluate research** to give credit for useful data communication and novel ways of collaborating

3. **common standards** for communicating data

4. intelligent openness for **data relevant to published scientific papers**

5. strengthening the cohort of **data scientists**

6. new **software tools** to automate and simplify the creation and exploitation of datasets

Roles of traditional publishers

- Mandating open data
- Hosting the data behind some papers (‘small-science’)
- Citing data as well as papers
- Crediting data originators

Open access publishing of research papers and data

- ‘Green’ open access: authors’ final versions of papers should be freely available within a maximum period after journal publication. (Common maximum: 12 months. Nature journals: 6 months.)

- ‘Gold’ open access: final published version of a paper freely available to all from the moment of publication. No subscription barriers. Financed by Author Processing Charge. (eg Scientific Reports)

- ‘Hybrid’: a journal combining subscriptions with an option of gold open access for authors. (eg Nature Communications)

Nature journals

- Nature titles other than *Nature Communications* may go hybrid sometime – publishers are willing but subject to financial viability.
- Driven/enabled by national policies and funding agencies. The drive varies between disciplines, between countries.
- At current rate will take a few years to settle.

Roles of governments: UK

- Finch report 2012, Research Councils UK policy
- Support for gold open access by direct government funding to institutions. First block grants in April 2013, to enable 45% of papers funded by UK to be gold open access.
- After five years, intend to achieve 75% of papers funded by UK to be gold open access, 25% green
- This would be at expense of library budgets for subscriptions.

Roles of governments: US

- Policy announced last week about peer reviewed publications and digital data
- To Federal Agencies spending over $100 million per year, from White House: “Prepare plans to make your funded results free to read within 12 months”. Deadline late July.
- No mandate to fund equivalent of NIH’s PubMed Central for other agencies
- Require grant applications to include data management plans
- Allow grant applications to include provision for data management and curation
- Support training and development of workforce in data management
- Consult stakeholders and involve public and private sectors in implementation

Targets for data recommendations

Scientists – changing cultural assumptions Employers (universities/institutes) – data responsibilities; crediting researchers Funders of research - the cost of curation is a cost of research Learned societies – influencing their communities	Publishers of research – mandatory open data Business – exploiting the opportunity; awareness & skills Government – efficiency of the science base; exploiting its data Governance processes for privacy, safety, security - proportionality

Thanks for your attention