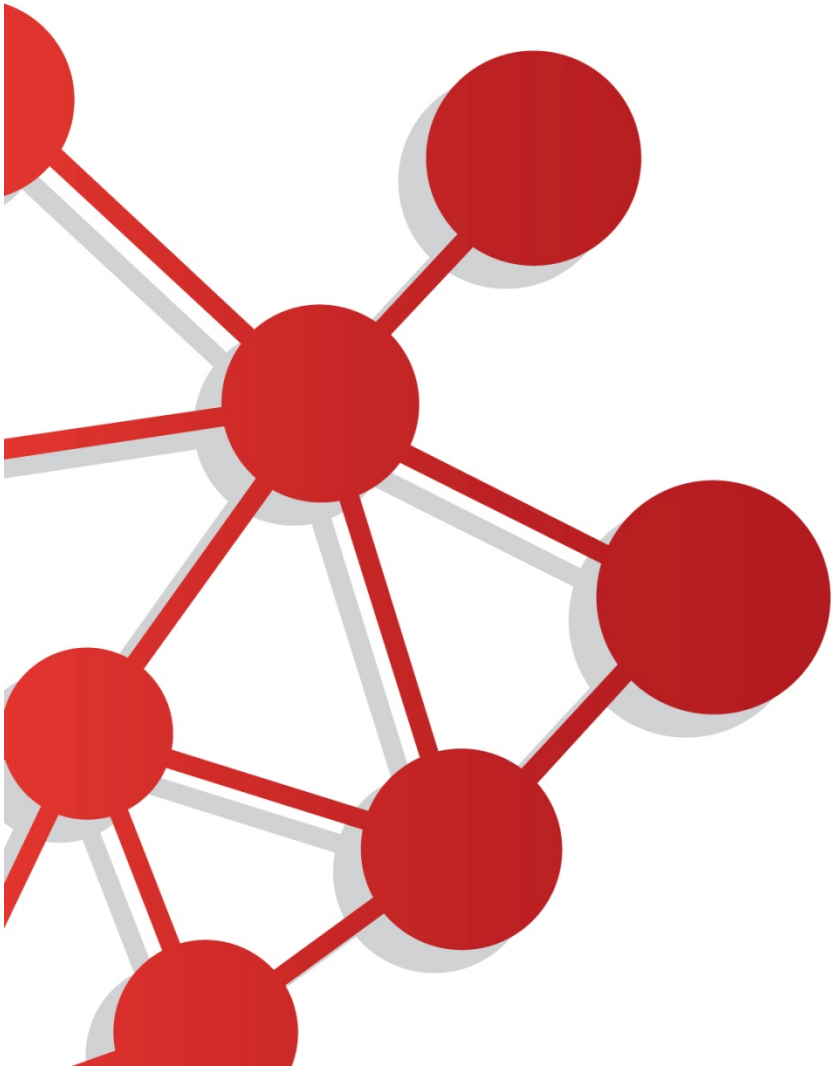
The *Nature* of research, publishing and data…



Jason Wilde

Nature Publishing Group

11th December 2012

nature publishing group npg

Article of the future…

- Evolution of research
- Text data metadata…
- The article of the future?
- Data, databases, data files
- The data problem
- Supplementary Information
- External data/information
- Accession codes/data citation and data publication
- Closing remarks

Evolution of research

No. 4356 April 25, 1953 NATURE 737

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

[1] Vinograd, J. R., Gierer, H., and Jerome, W., *Proc. Nat. Acad. Sci.*, 39, 149 (1953).

[2] Longuet-Higgins, M. S., *Nature, Lond.*, 171, 603 (1953).

[3] Van Arkel, W. G., *Wissen. Rede Papers in Phys. Genootsch. Natuur.*, 11 (1953).

[4] Ekman, C. W., *Arkiv. Mat. Astron. Fysik (Stockholm)*, 2 (1953).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three interwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

©1953 Nature Publishing Group

DNA Structure 1953

1 Page

2 Authors

1 Figure

no data

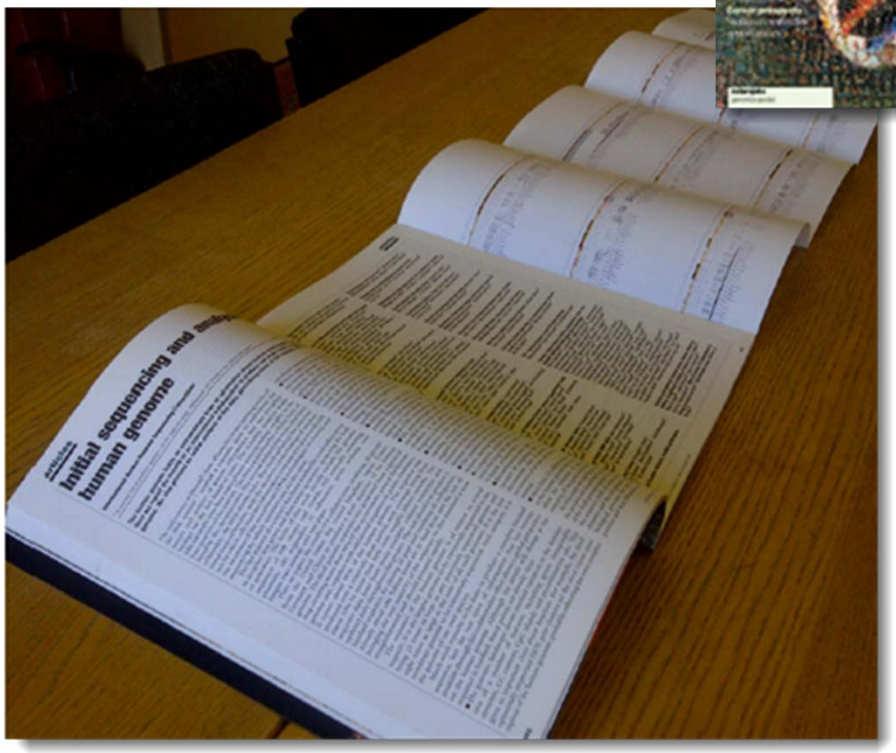
nature publishing group

Evolution of research

Human Genome 2001

62 Pages, 150 Authors,

49 Figure, 27 tables



Evolution of research

1000 Genome Project 2010

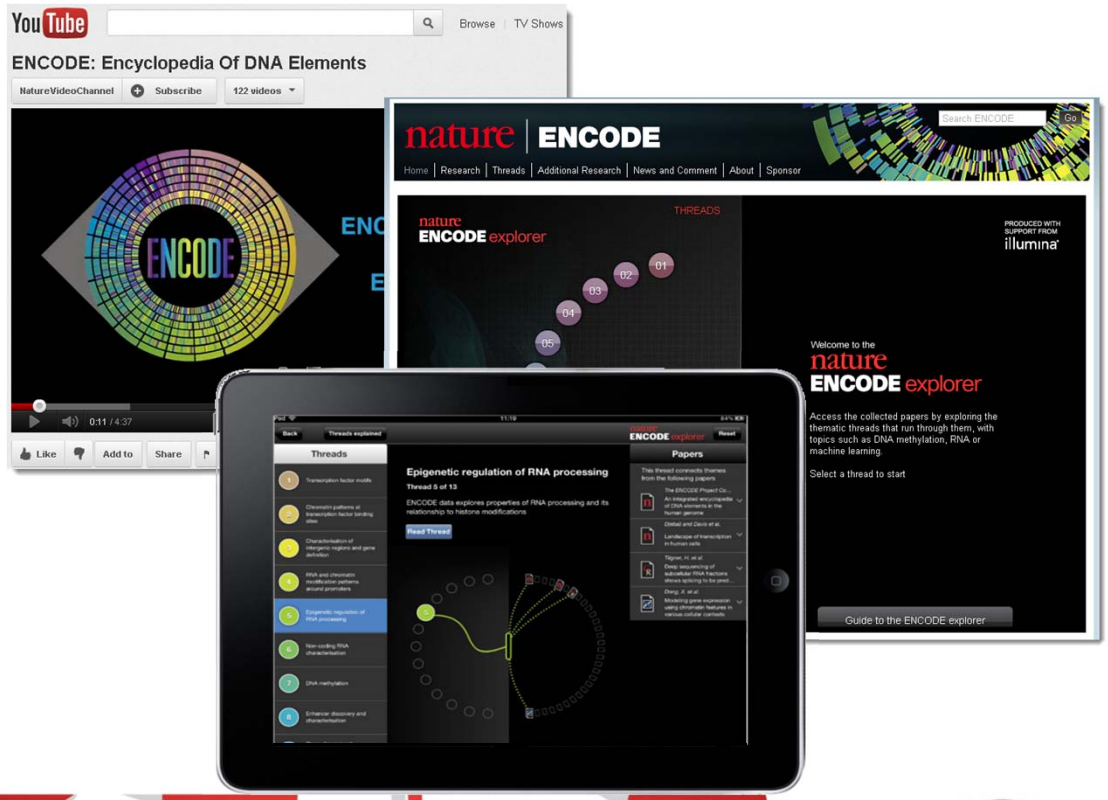
12 page, 76 Institutions, 12,145 SRA run ids

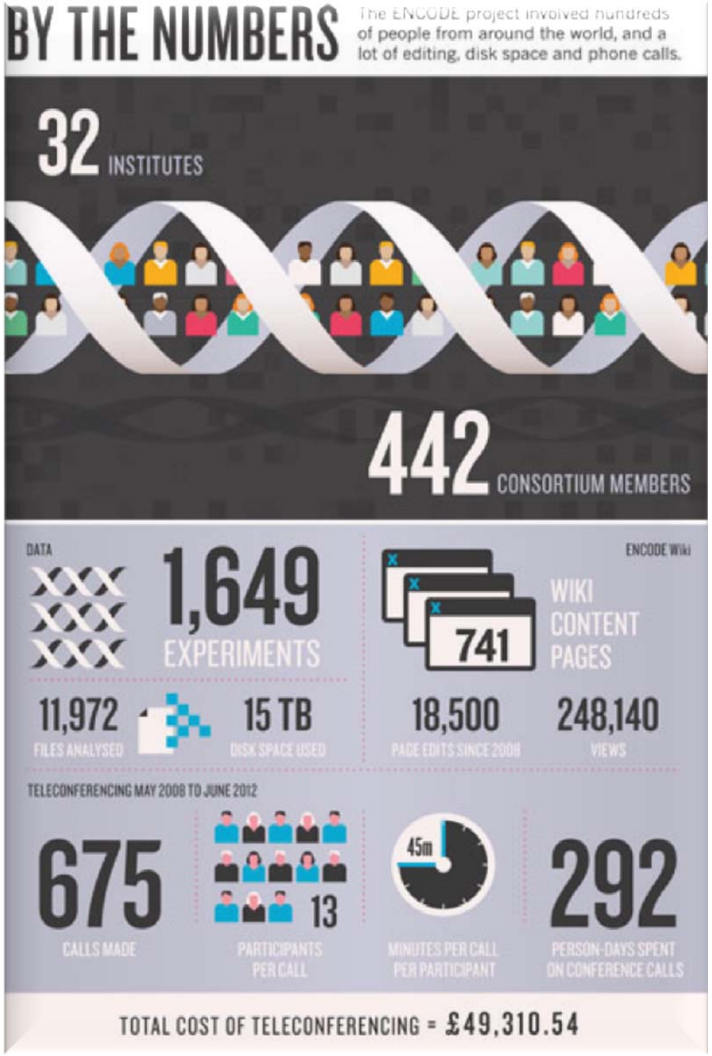
nature publishing group npg

Evolution of research

Encode Project 2012

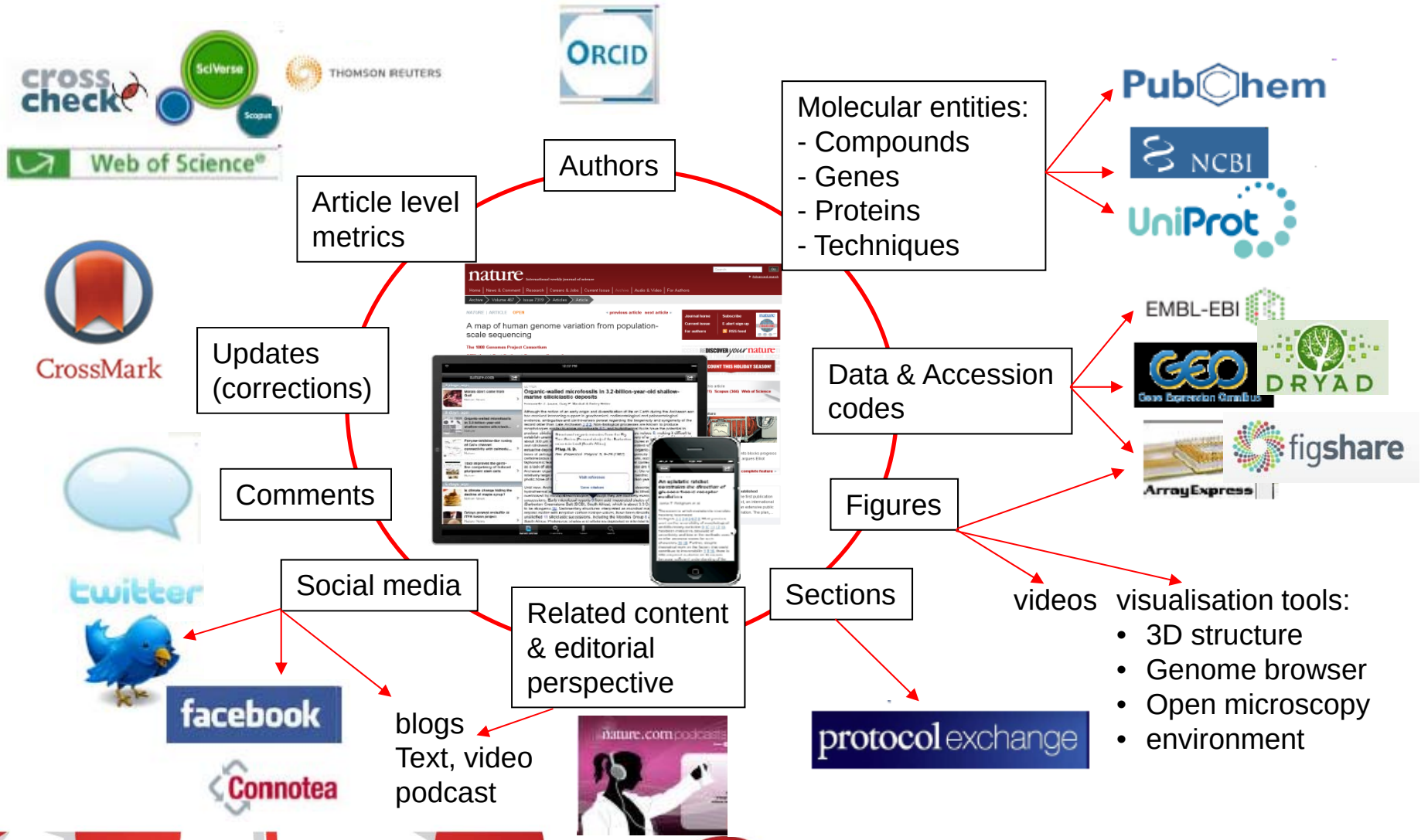
30 papers, 3 Journals





nature publishing group npg

Text, data, metadata…



nature publishing group npg

The article of the future?

Contextual advertising

ases

Patents

Articles about the same entity

Related data files

New sources of relevant information

Faceted sear

endors & Suppliers

NATURE

A WEEKLY ILLUSTRATED JOURNAL OF SCIENCE

"To the solid ground
Of Nature trusts the mind which builds for aye."—WORDSWORTH

THURSDAY, NOVEMBER 4, 1869

NATURE: APHORISMS BY GOETHE

NATURE! We are surrounded and embraced by her: powerless to separate ourselves from her, and powerless to penetrate beyond her.

Without asking, or warning, she snatches us up into her circling dance, and whirls us on until we are tired, and drop from her arms.

She is ever shaping new forms: what is, has never yet been; what has been, comes not again. Everything is new, and yet nought but the old.

We live in her midst and know her not. She is incessantly speaking to us, but betrays not her secret. We constantly act upon her, and yet have no power over her.

The one thing she seems to aim at is Individuality; yet she cares nothing for individuals. She is always building up and destroying; but her workshop is inaccessible.

Her life is in her children; but where is the mother? She is the only artist; working-up the most uniform material into utter opposites; arriving, without a trace of effort, at perfection, at the most exact precision, though always veiled under a certain softness.

Each of her works has an essence of its own; each of her phenomena a special characterisation: and yet their diversity is in unity.

She performs a play; we know not whether she sees it herself, and yet she acts for us, the lookers-on.

Incessant life, development, and movement are in her, but she advances not. She changes for ever and ever, and rests not a moment. Quietude is inconceivable to her, and she has laid her curse upon rest. She is firm. Her steps are measured, her exceptions rare, her laws unchangeable.

She has always thought and always thinks; though not as a man, but as Nature. She broods over an all-comprehending idea, which no searching can find out.

Mankind dwell in her and she in them. With all men she plays a game for love, and rejoices the more they win. With many, her moves are so hidden, that the game is over before they know it.

That which is most unnatural is still Nature; the stupidest philistinism has a touch of her genius. Whoso cannot see her everywhere, sees her nowhere rightly.

She loves herself, and her innumerable eyes and affections are fixed upon herself. She has divided herself that she may be her own delight. She causes an endless succession of new capacities for enjoyment to spring up, that her insatiable sympathy may be assuaged.

She rejoices in illusion. Whoso destroys it in himself and others, him she punishes with the sternest tyranny. Whoso follows her in faith, him she takes as a child to her bosom.

Her children are numberless. To none is she altogether miserly; but she has her favourites, on whom she squanders much, and for whom she makes great sacrifices. Over greatness she spreads her shield.

She tosses her creatures out of nothingness, and tells them not whence they came, nor whither they go. It is their business to run, she knows the road.

Her mechanism has few springs—but they never wear out, are always active and manifold.

The spectacle of Nature is always new, for she is always renewing the spectators. Life is her most exquisite invention; and death is her expert contrivance to get plenty of life.

She wraps man in darkness, and makes him for ever long for light. She creates him dependent upon the earth, dull and heavy; and yet is always shaking him until he attempts to soar above it.

Articles need to r nto a 3D entity in the information space.

Data, databases, data files

Two important factors are driving to make research data more available and reusable:

- To ensure the scientific process is transparent and can be scrutinised and research results reproduced
- To speed the scientific process, lead to new insights and reduce duplicated and repeated work

To achieve this research data needs to be

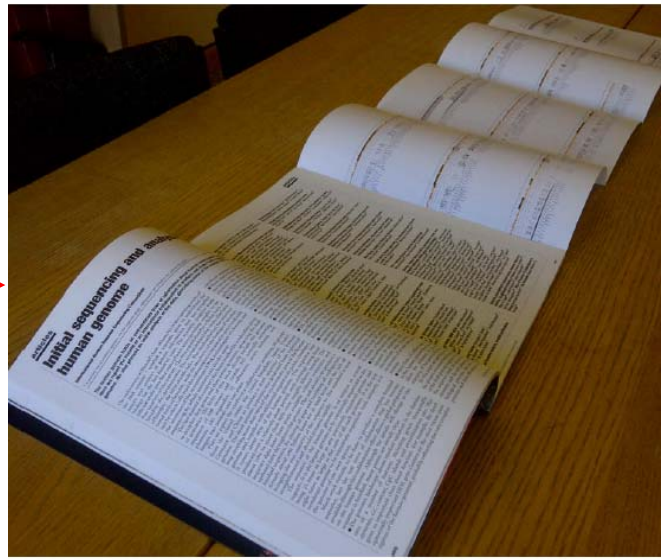
- Available
- Findable
- Interpretable
- Re-usable
- Citable

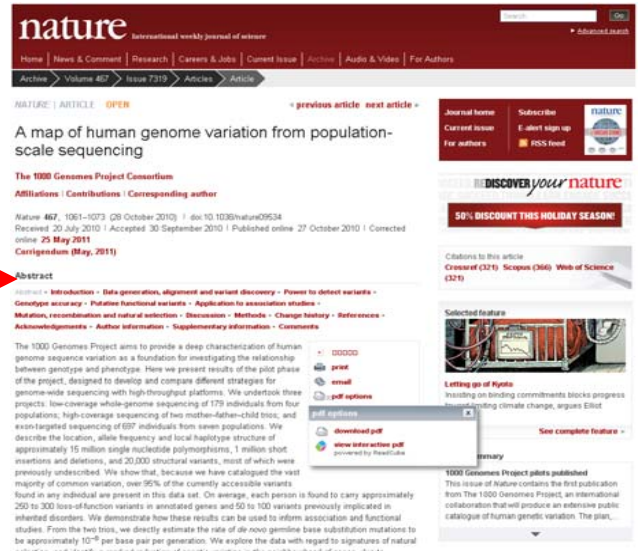
Data, databases, data files

No. 4356 April 25, 1953 NATURE 737

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid



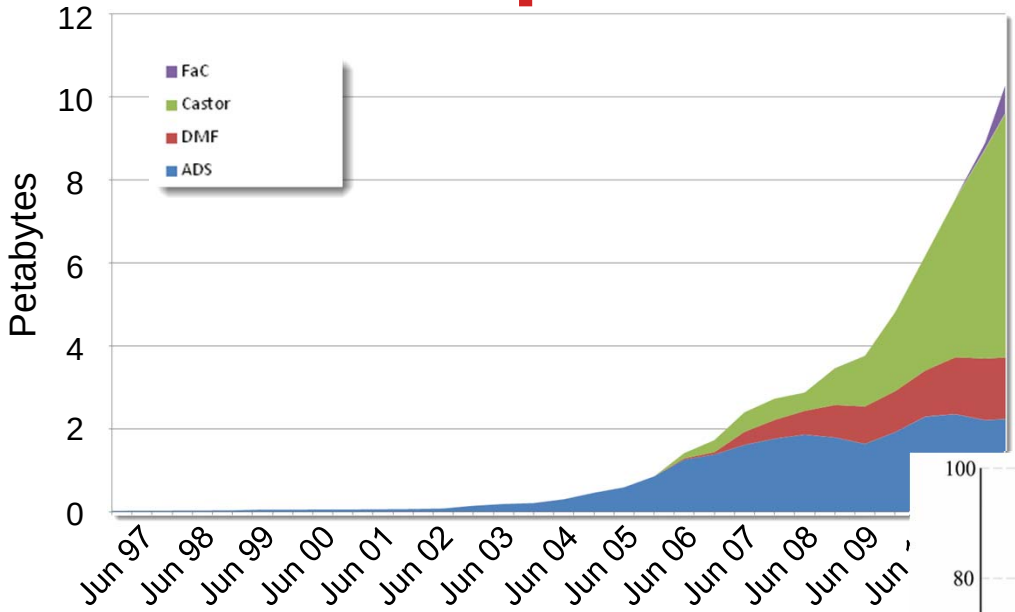


nature

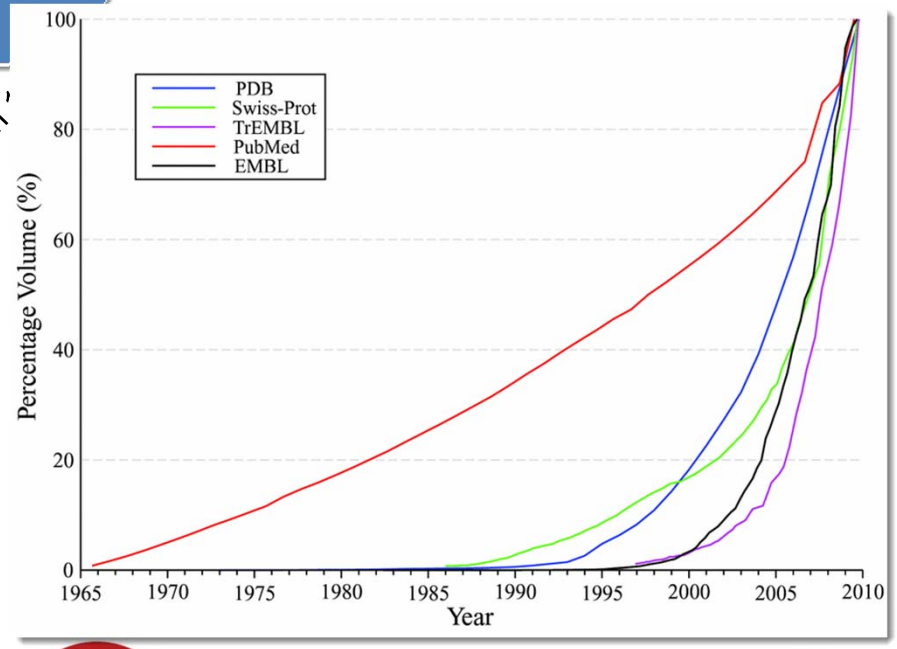
A map of human genome variation from population-scale sequencing

- Even in NPG titles the amount of data, over the last 60 years, has grown phenomenally.
- As with all publishers, very little of the research we publish is linked to the data underpinning it

The data problem



- **CERN** high energy physics data from the Atlas, LHCb, CMS and Alice. These data are stored using the CERN designed **Castor** infrastructure
- **Castor** is also used to store data from STFC's UK facilities (**FaC**).
- **ADS** is the old archive with lots of small backups.
- **DMF** includes the preserved documents from the CEDA Repository at the British Atmospheric Data Centre, and Tesella Safety Deposit Box (SDB) services to preserve data collected on the ISIS, Diamond and CLF facilities.



Graphical illustration of the 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 500000) and protein structures (**blue**; current total 60000)

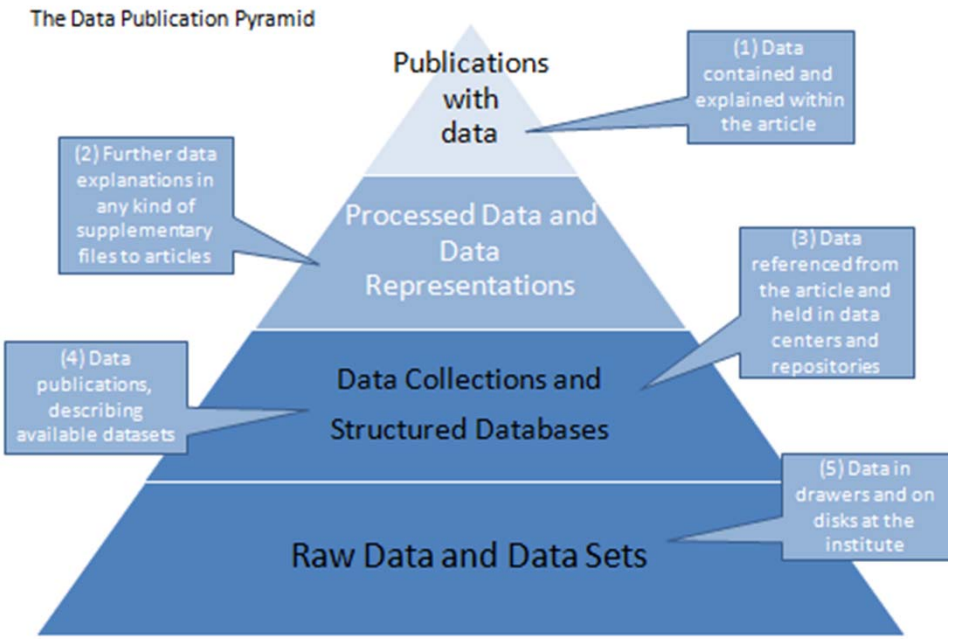
The data problem

- Fourth paradigm: science has moved into a data-intensive era[1]
- Data deluge: the quantity of data produced continues to increase, with exponential growth estimated in the near future, raising concerns about long term preservation:
  - 30,000 human genome sequences will have been completed by end 2011[2]
  - the amount of data stored per research project will grow from the megabyte-gigabyte range to the terabyte-petabyte range by 2015[3]
  - there will be 35 zettabytes[4] of digital data created per annum by 2020: 44 times the amount of digital data produced in 2009[5]

1 Jim Gray on eScience: A Transformed Scientific Method. Based on the transcript of a talk given by Jim Gray to the NRC-CSTB1 in Mountain View, CA, on January 11, 2007
2 A map of human genome variation from population-scale sequencing, Nature 467, 28 October 2010
3 Researcher estimates, in PARSE.Insight survey: Insight into digital preservation of research output in Europe, 2010
4 1 zettabyte = 1 trillion gigabytes
5 According to the International Data Corporation's Digital Universe Study (2010), in Ad -hoc Strategic Coordinating Committee on Information and Data, Interim Report to the ICSU Committee on Scientific Planning and Review, April 2011

The data problem

The linking between publications and research data is well illustrated by the publication pyramid

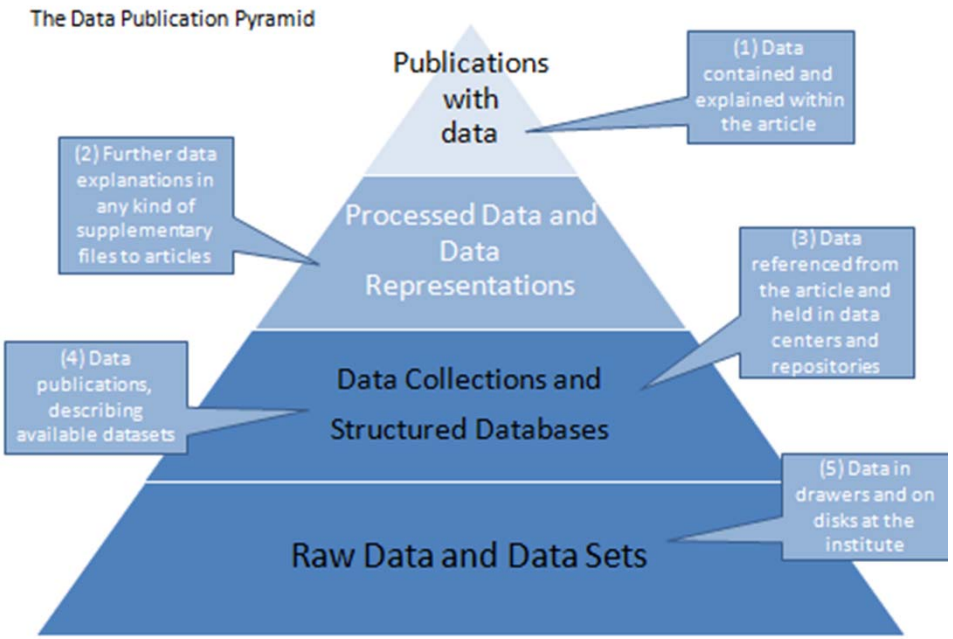


- As we move down the pyramid the exclusive relationship between publication and data is diminished
- At the bottom layer most datasets remain unpublished, hence unfindable and not reusable
- The data published today, within or with publications, is only the tip of the data iceberg

Source: Eefke Smit, Opportunities for Data Exchange, Data Enters Scholarly Communication, October 2011
http://www.stm-assoc.org/2011_12_5_ODE_Report_On_Integration_of_Data_and_Publications.pdf

The data problem

A variety of players have responded to the changing data landscape with offerings positioned at different levels of the publication pyramid, with high interaction between players



Journal /society publishers

1. Enhanced journal articles offering links to datasets (Elsevier, Springer, NPG)
2. Data publications akin to journal publications but focused on describing datasets (Giga Science, Earth Systems, F1000)

Data archives
Community-specific open repositories (Dryad, FigShare)
Institutional repositories

Other service providers
Assigning DOIs to datasets (DataCite)

Information providers
Enhancing discoverability and citability of datasets (Scopus, Web of Science)

Source: Eefke Smit, Opportunities for Data Exchange, Data Enters Scholarly Communication, October 2011
http://www.stm-assoc.org/2011_12_5_ODE_Report_On_Integration_of_Data_and_Publications.pdf

2. Supplementary information

“ It has become a limitless bag of stuff[1] ”

Emilie Marcus Editor-in-chief of Cell Press journals.

- Supplementary information (SI) is growing and contains all kinds of stuff (multimedia, data sets, figures, tables, experimental details…….)
- Both contains information that is only tenuously linked to the article and information that is ‘integral’ to the article.
- Puts additional pressures on authors, reviewers and readers
- Locked in PDF’s data is not easy to find or extract
- Different publishers/communities are approaching in different ways
  - Some no longer accept supplementary information (Journal of Neuroscience)
  - Some are restricting the amount of SI (Cell and NPG Nature branded life science journals)
  - Others are more critically looking at what should be in the article and what really is supplemental (NPG).

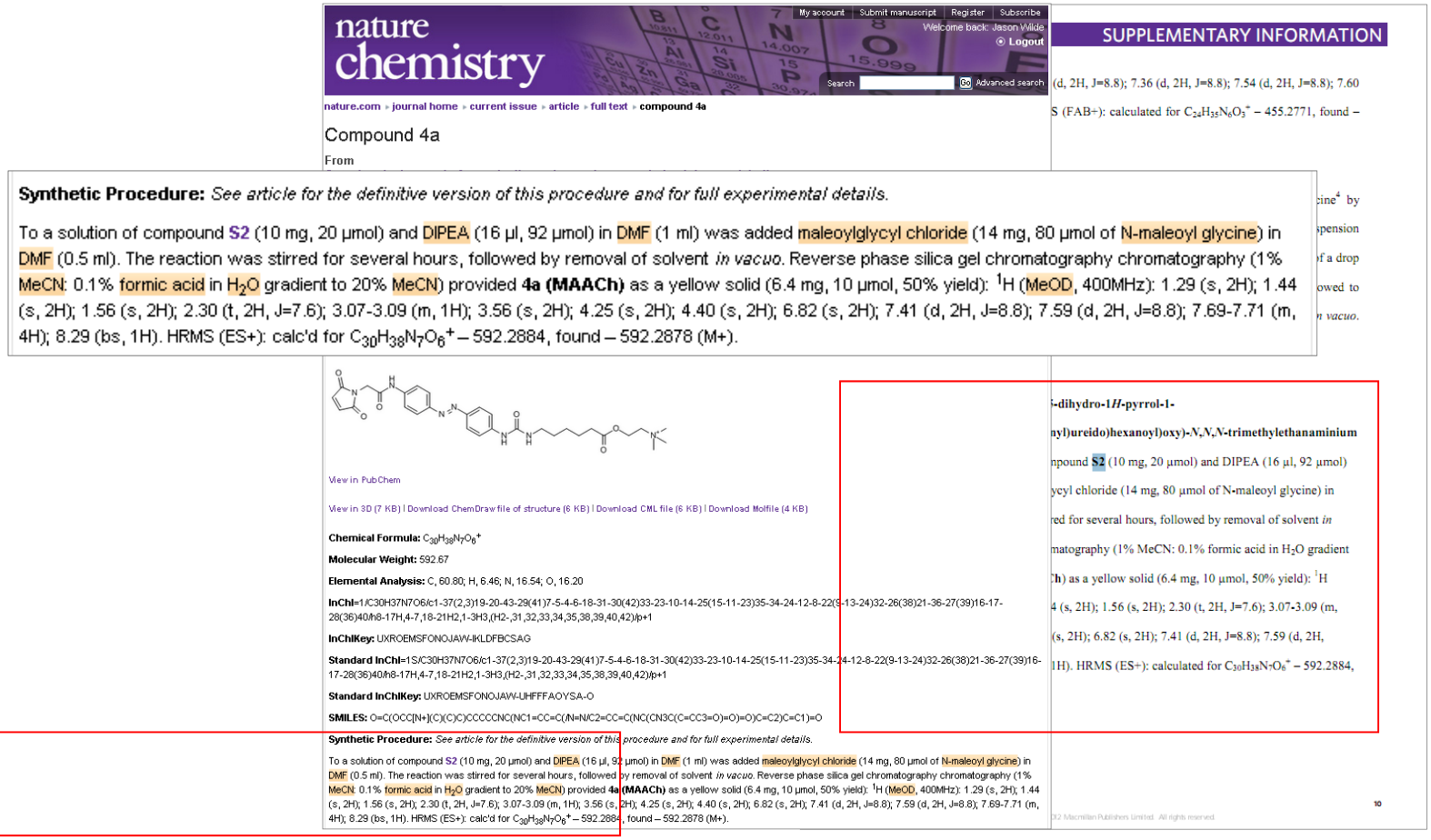
[1]The Scientist http://classic.the-scientist.com/news/display/58027/

2. Supplementary information

Supplementary Figure 2. Probabilistic tracts overlaid on the subject's anatomical image that was registered to the diffusion images. The sagittal sections on the right hemisphere show possible multisynaptic pathways that the probabilistic algorithm takes from extrastriate cortex, through the pons, and to the cerebellar pathways.

- Authors encouraged to submit one seamless paper, incorporating all of the essential information
- The editors will decide (with the referees and authors) elements that should be included in the paper and elements that should reside in the SI file.
- This is an experiment whose aim is to
  - make authors/reviewers/editors think more about data and what is integral to the paper
  - Ensure all information is peer reviewed.
- *Nature* is investigating an extended online version of the article (html and PDF) which will take data (figures) that would have been in the SI and elevate them to the article. Up to 14 figures and tables will be able to be included in the extended data version.

Supplementary information



nature chemistry

My account | Submit manuscript | Register | Subscribe

Welcome back: Jason Wilde

Logout

Search | Advanced search

nature.com ▸ journal home ▸ current issue ▸ article ▸ full text ▸ compound 4a

Compound 4a

From

Synthetic Procedure: *See article for the definitive version of this procedure and for full experimental details.*

To a solution of compound **S2** (10 mg, 20 µmol) and DIPEA (16 µl, 92 µmol) in DMF (1 ml) was added maleoylglycyl chloride (14 mg, 80 µmol of N-maleoyl glycine) in DMF (0.5 ml). The reaction was stirred for several hours, followed by removal of solvent *in vacuo*. Reverse phase silica gel chromatography chromatography (1% MeCN: 0.1% formic acid in H_2O gradient to 20% MeCN) provided **4a (MAACh)** as a yellow solid (6.4 mg, 10 µmol, 50% yield): ^{1}H (MeOD, 400MHz): 1.29 (s, 2H); 1.44 (s, 2H); 1.56 (s, 2H); 2.30 (t, 2H, J=7.6); 3.07-3.09 (m, 1H); 3.56 (s, 2H); 4.25 (s, 2H); 4.40 (s, 2H); 6.82 (s, 2H); 7.41 (d, 2H, J=8.8); 7.59 (d, 2H, J=8.8); 7.69-7.71 (m, 4H); 8.29 (bs, 1H). HRMS (ES+): calc'd for $C_{30}H_{38}N_7O_6^+$ – 592.2884, found – 592.2878 (M+).

View in PubChem

View in 3D (7 KB) | Download ChemDraw file of structure (6 KB) | Download CML file (6 KB) | Download Molfile (4 KB)

Chemical Formula: $C_{30}H_{38}N_7O_6^+$

Molecular Weight: 592.67

Elemental Analysis: C, 60.80; H, 6.46; N, 16.54; O, 16.20

InChI=1/C30H37N7O6/c1-37(2,3)19-20-43-29(41)7-5-4-6-18-31-30(42)33-23-10-14-25(15-11-23)35-34-24-12-8-22(9-13-24)32-26(38)21-36-27(39)16-17-28(36)40/h8-17H,4-7,18-21H2,1-3H3,(H2-,31,32,33,34,35,38,39,40,42)/p+1

InChIKey: UXROEMSFONOJAVV-IKLDFBCSAG

Standard InChI=1S/C30H37N7O6/c1-37(2,3)19-20-43-29(41)7-5-4-6-18-31-30(42)33-23-10-14-25(15-11-23)35-34-24-12-8-22(9-13-24)32-26(38)21-36-27(39)16-17-28(36)40/h8-17H,4-7,18-21H2,1-3H3,(H2-,31,32,33,34,35,38,39,40,42)/p+1

Standard InChIKey: UXROEMSFONOJAVV-UHFFFAOYSA-O

SMILES: O=C(OCC[N+](C)(C)C)CCCCCNC(NC1=CC=C(/N=N/C2=CC=C(NC(CN3C(C=CC3=O)=O)=O)C=C2)C=C1)=O

Synthetic Procedure: *See article for the definitive version of this procedure and for full experimental details.*

To a solution of compound **S2** (10 mg, 20 µmol) and DIPEA (16 µl, 92 µmol) in DMF (1 ml) was added maleoylglycyl chloride (14 mg, 80 µmol of N-maleoyl glycine) in DMF (0.5 ml). The reaction was stirred for several hours, followed by removal of solvent *in vacuo*. Reverse phase silica gel chromatography chromatography (1% MeCN: 0.1% formic acid in H_2O gradient to 20% MeCN) provided **4a (MAACh)** as a yellow solid (6.4 mg, 10 µmol, 50% yield): ^{1}H (MeOD, 400MHz): 1.29 (s, 2H); 1.44 (s, 2H); 1.56 (s, 2H); 2.30 (t, 2H, J=7.6); 3.07-3.09 (m, 1H); 3.56 (s, 2H); 4.25 (s, 2H); 4.40 (s, 2H); 6.82 (s, 2H); 7.41 (d, 2H, J=8.8); 7.59 (d, 2H, J=8.8); 7.69-7.71 (m, 4H); 8.29 (bs, 1H). HRMS (ES+): calc'd for $C_{30}H_{38}N_7O_6^+$ – 592.2884, found – 592.2878 (M+).

SUPPLEMENTARY INFORMATION

(d, 2H, J=8.8); 7.36 (d, 2H, J=8.8); 7.54 (d, 2H, J=8.8); 7.60

S (FAB+): calculated for $C_{24}H_{35}N_6O_3^+$ – 455.2771, found –

10

02 Macmillan Publishers Limited. All rights reserved.

3. External data/information

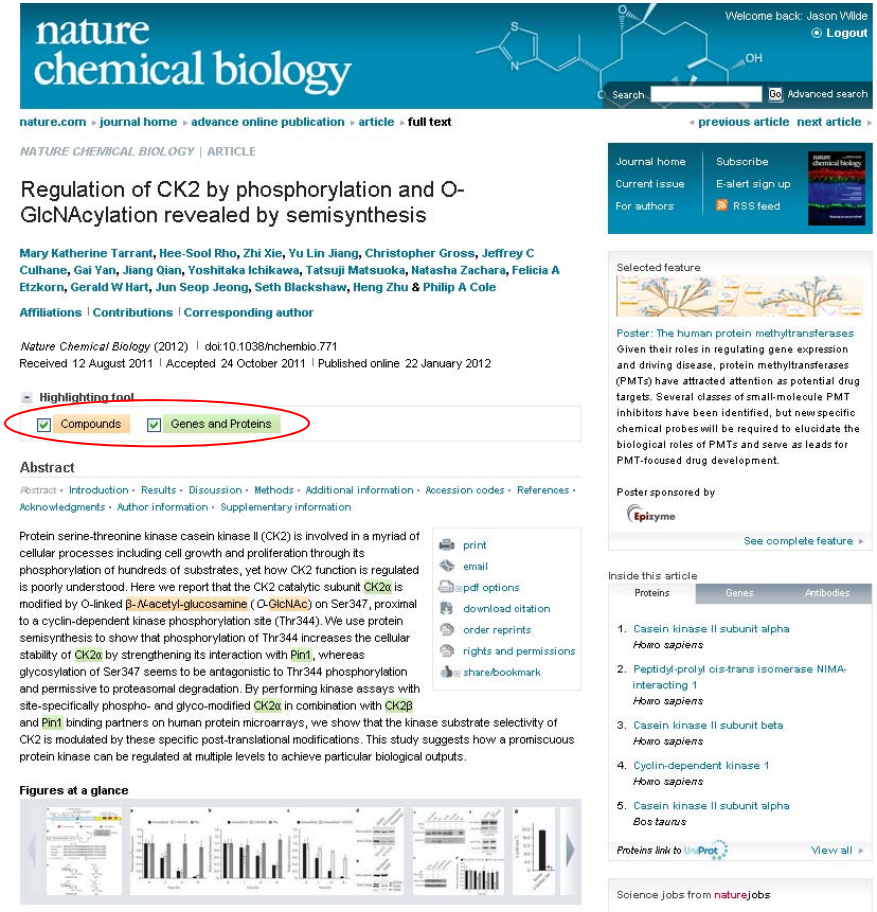
NPG materials and data policies

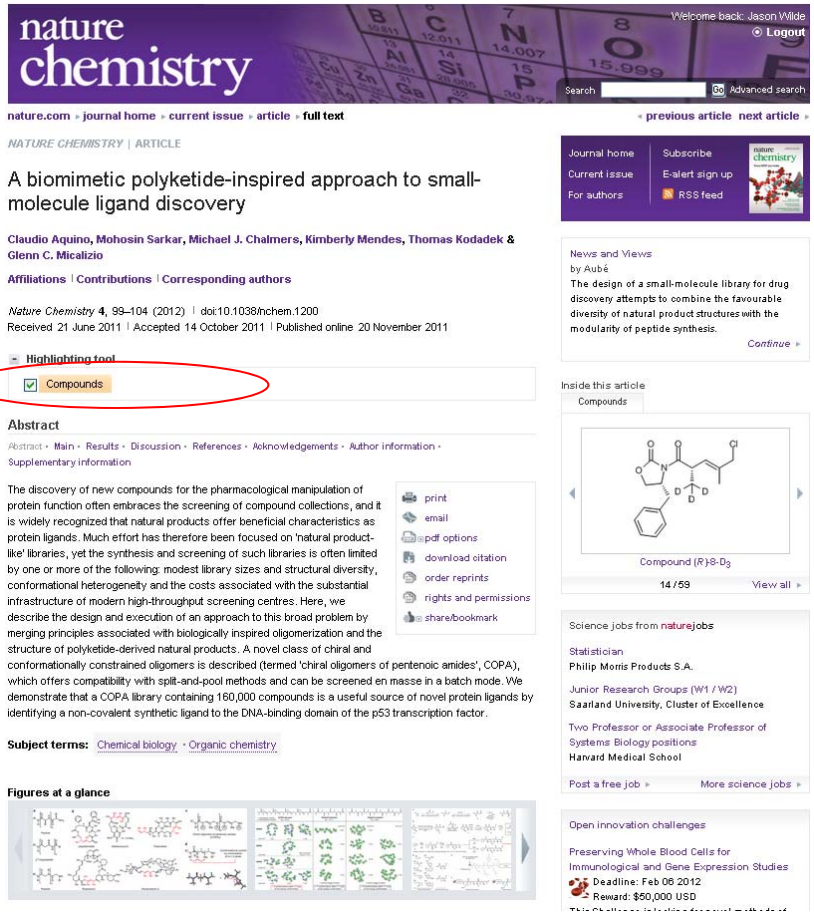
……A condition of publication in a Nature journal is that authors are required to make materials, data and associated protocols promptly available to others without undue qualifications…..

For the following types of data set, submission to a community-endorsed, public repository is mandatory.

Nature journals strongly recommend deposition of other types of data sets into appropriate public repositories that are at an earlier stage of development. ……

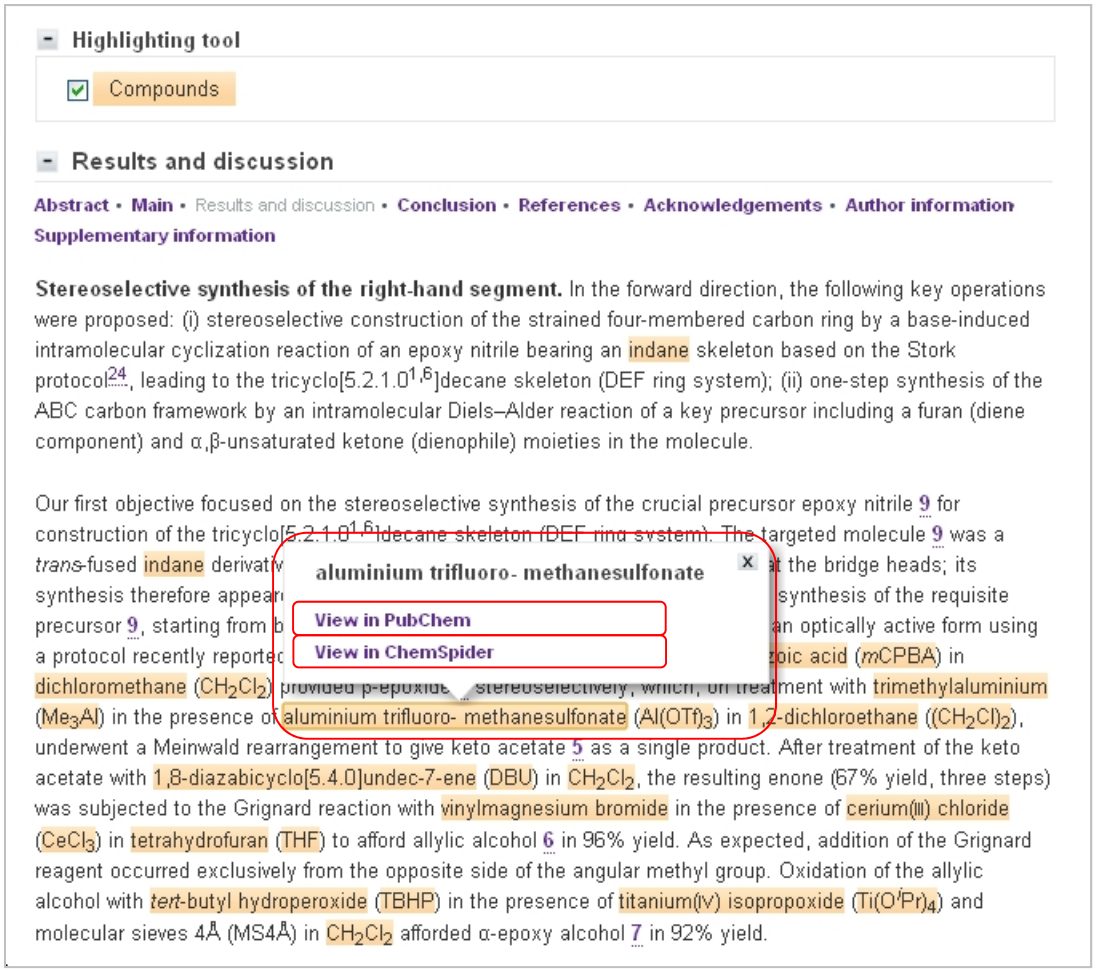
External data/information

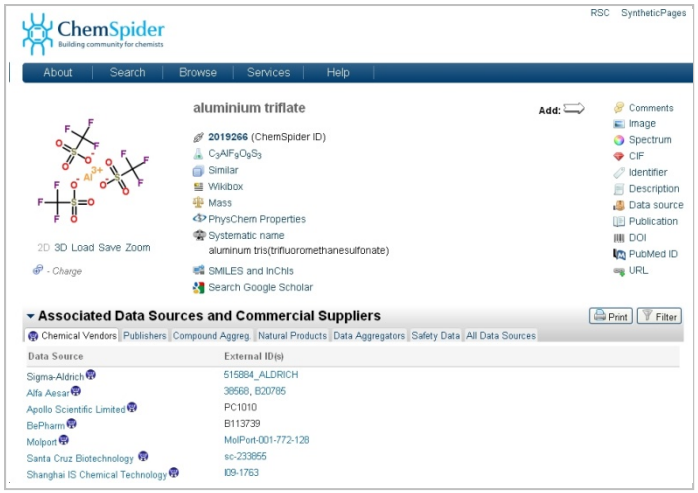




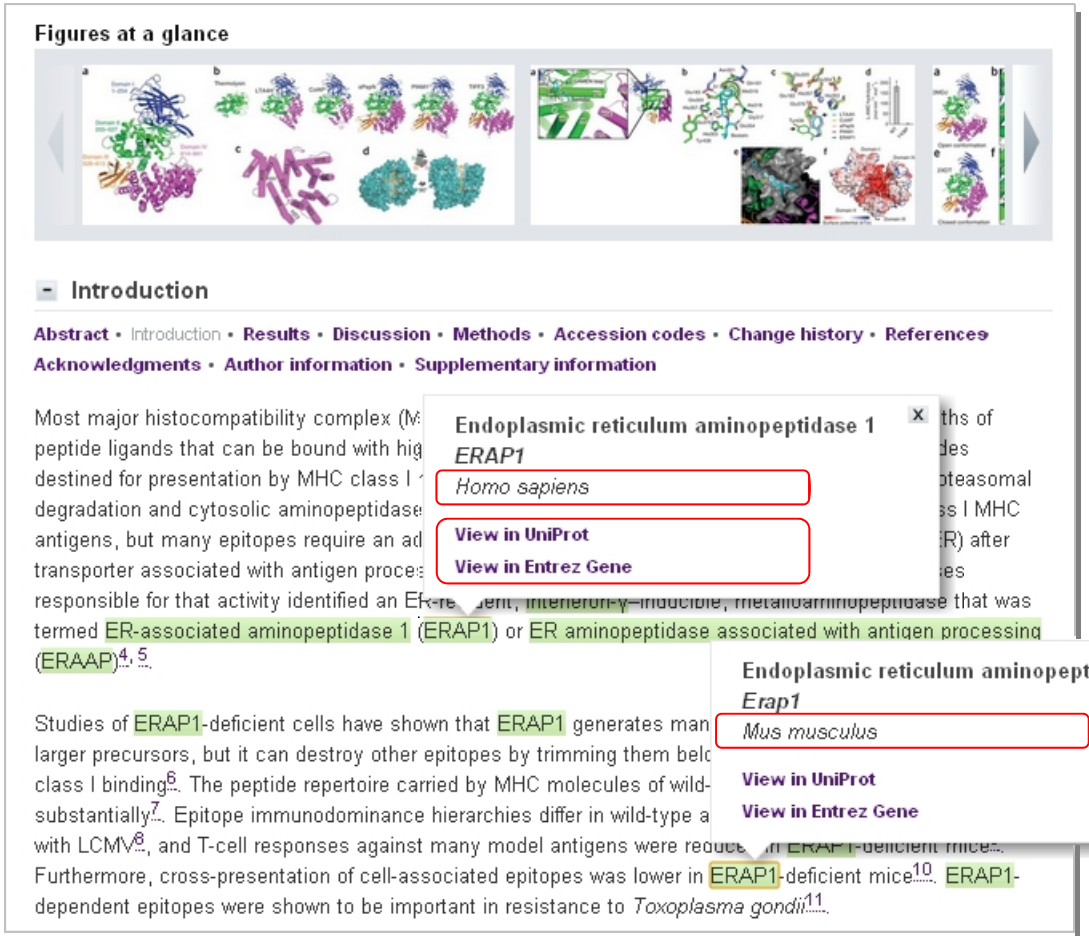
nature publishing group npg

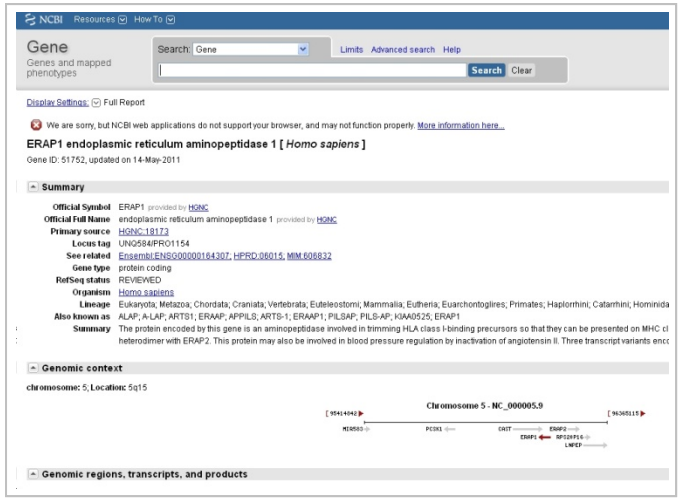
External data/information

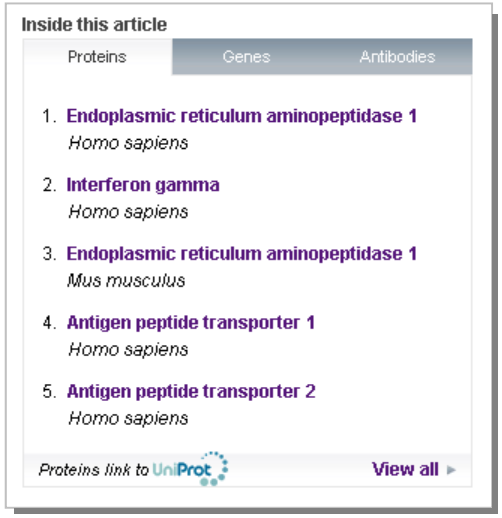




External data/information







Accession codes/data citation and data publication

Accession codes.

Coordinates of the models were deposited in the PDB under the following accession codes: 3TLS (E19'P mutant), 3TLT (H11'F mutant), 3UU3 (loop2-20' mutant oxidized form), 3UU5 (loop2-20' mutant reduced form), 3TLW (loop2-21' mutant oxidized form), 3UUB (loop2-21' m solution), 3UU4 (loop2-21' mutant reduced in crystal), 3TLV (loop2-22' mutant oxi (loop2-22' mutant reduced form), 3TLU (loop2-24' mutant oxidized form) and 3UU8 mutant reduced form).

Accession codes

Abstract • Introduction • Results • Discussion • Methods • Accession codes • References
Acknowledgments • Author information • Supplementary information

Primary accessions

Protein Data Bank

PDB 3TLS	3D 3TLS
PDB 3TLT	3D 3TLT
PDB 3UU3	3D 3UU3
PDB 3UU5	3D 3UU5
PDB 3TLW	3D 3TLW
PDB 3UUB	3D 3UUB
PDB 3UU4	3D 3UU4
PDB 3TLV	3D 3TLV
PDB 3UU6	3D 3UU6
PDB 3TLU	3D 3TLU
PDB 3UU8	3D 3UU8

Referenced accessions

Protein Data Bank

PDB 2VL0	3D 2VL0
PDB 3EAM	3D 3EAM
PDB 3EHZ	3D 3EHZ
PDB 3RHW	3D 3RHW

The GLIC pentameric Ligand-Gated Ion Channel E19'P mutant in a locally-closed conformation (LC2 subtype)

DOI: 10.2210/pdb3tls/pdb

Primary Citation

A locally closed conformation of a bacterial pentameric proton-gated ion channel.

Prevost, M.S., Sauguet, L., Nury, H., Van Renterghem, C., Huon, C., Poitevin, F., Baaden, M., Delarue, M., Corringer, P.J.

Journal: (2012) Nat.Struct.Mol.Biol. **19:** 642-649

PubMed: 22580559
DOI: 10.1038/nsmb.2307
Search Related Articles in PubMed

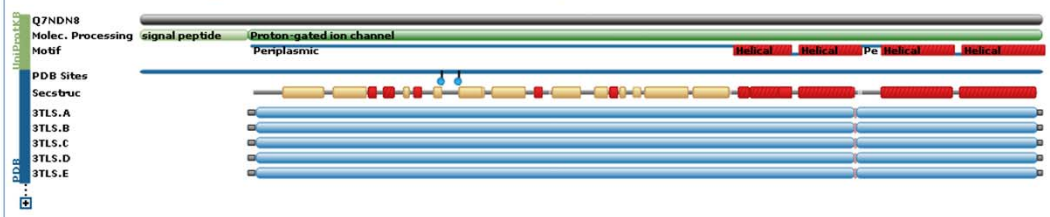
PubMed Abstract:

Pentameric ligand-gated ion channels mediate signal transduction through conformational transitions between closed-pore and open-pore states. To stabilize a closed conformation of GLIC, a bacterial proton-gated homolog from Gloeobacter violaceus whose open structure is known, we separately generated either four cross-links...
[Read More & Search PubMed Abstracts]

Molecular Description Hide

Classification: Membrane Protein, Transport Protein
Structure Weight: 183419.89

Molecule:	Glr4197 protein				
Polymer:	1	**Type:**	protein	**Length:**	321
Chains:	A, B, C, D, E				
Fragment:	UNP residues 44-359				
Mutation:	E285P				
Organism	Gloeobacter violaceus				
UniProtKB:	Protein Feature View | Search PDB | Q7NDN8				



Nature Structural & Molecular Biology **19**, 642–649 (2012)
doi:10.1038/nsmb.230

nature publishing group npg

Accession codes/data citation and data publication

Pew Hispanic Center. (2004). Changing channels and crisscrossing cultures: A survey of Latinos on the news media [Data file and code book]

Jackson, T. J., and M. H. Cosh, 2003: SMEX02 watershed soil moisture data, Walnut Creek, Iowa. National Snow and Ice Data Center, Boulder, CO, digital media. [Available online at http://

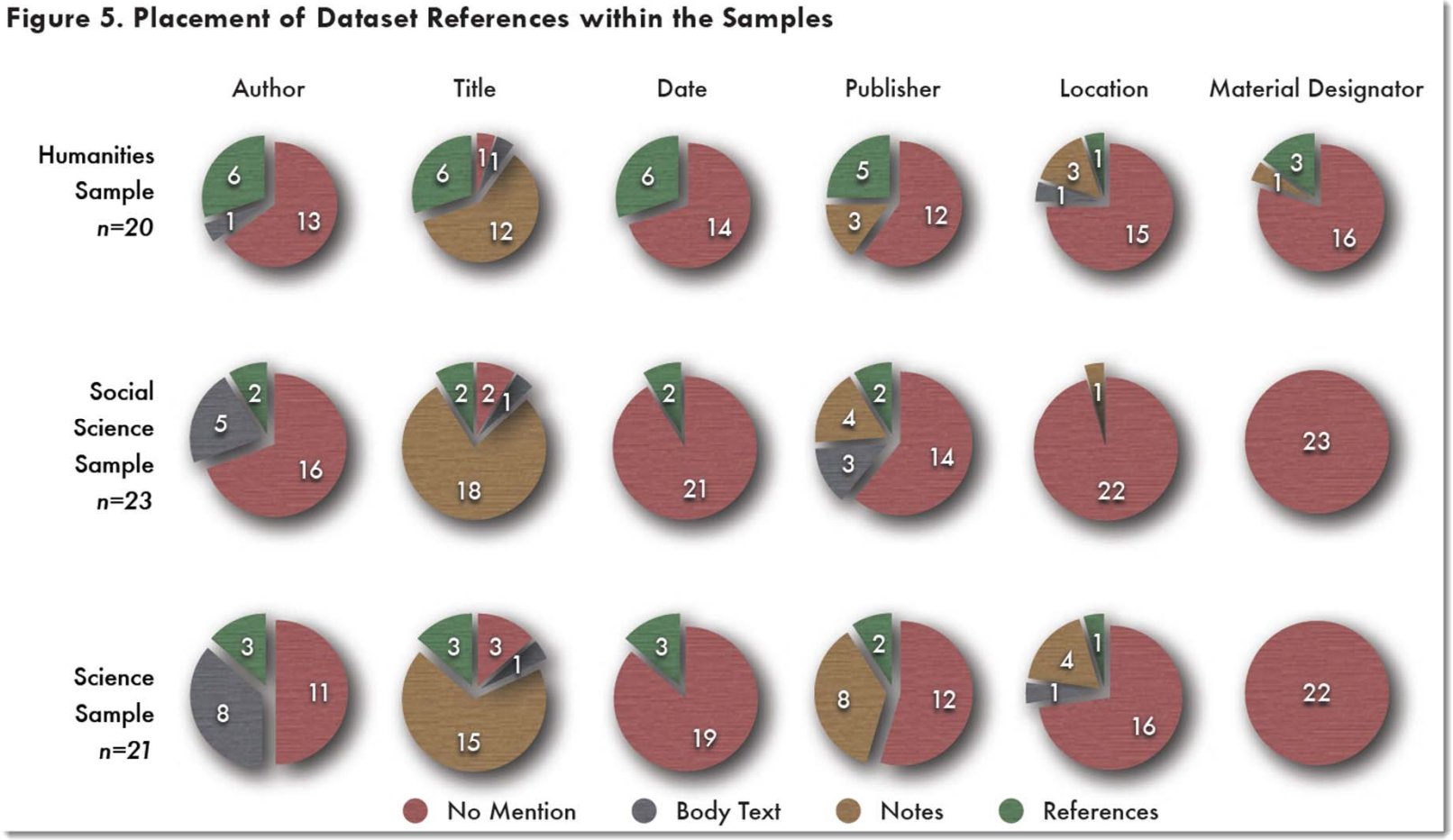
PDQ: NCI's Comprehensive Cancer Database. Bethesda, MD: National Cancer Institute. 1996. http://www.cancer.gov/cancerinfo/pdq/can-cerda

American Institute of Public Opinion. 1976. Gallup Public

GenBank (for RP11-322N14 BAC [accession number AC017046]; accessed on October 6, 2009), http://www.ncbi.nlm.nih.gov/Genbank/.

IMGT/HLA Sequence Database [Internet]. Release 2.9.0 Cambridge (England): European Bioinformatics Institute. 2003 - [update 2005 Jun 1; cited 2005 Jun 22]. Available from http://www.ebi.ac.uk/imgt/hla/

Accession codes/data citation and data publication



Mooney, H, Newton, MP. (2012). The Anatomy of a Data Citation: Discovery, Reuse, and Credit. *Journal of Librarianship and Scholarly Communication* 1(1):eP1035. http://dx.doi.org/10.7710/2162-3309.1035

Accession codes/data citation and data publication

- There are no agreed standards for data citation between a) publishers b) scientific disciplines or c) repositories
- For data to be truly discoverable and interoperable there needs to be both human and machine readable references to data

- There are a fledgling number of data publications and publishers are showing increased interest in this area.
- With the pressures to make research data available is this a way to credit those who do?

Earth System Science Data
The Data Publishing Journal

(GIGA)n SCIENCE

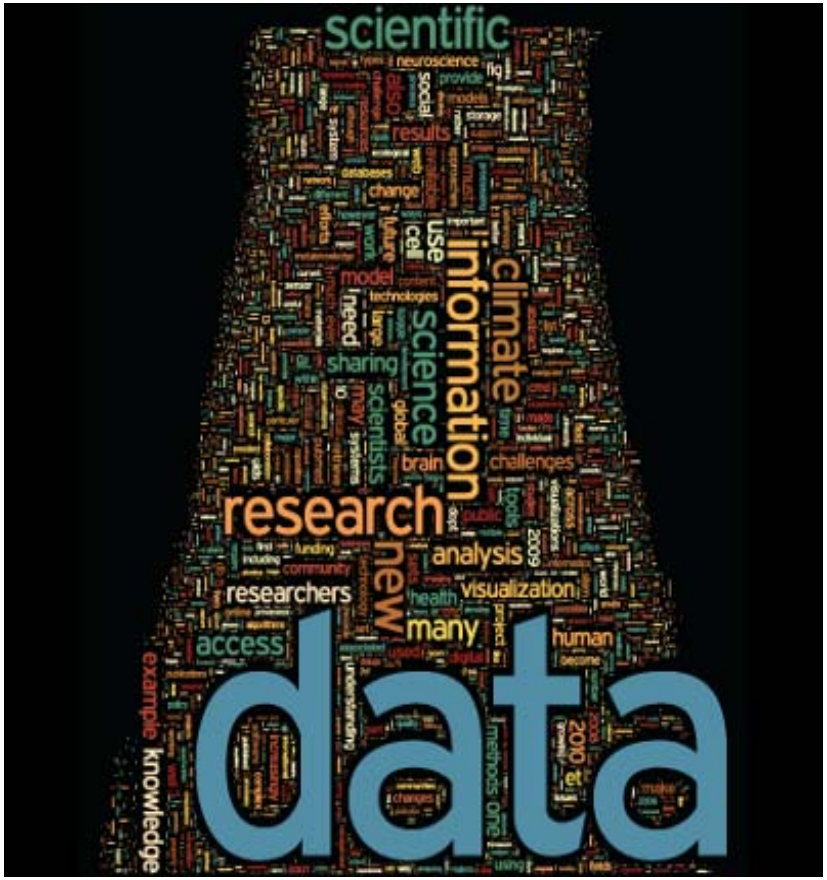
Ecological Archives

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

F1000Research
OPEN SCIENCE • OPEN DATA • OPEN PEER REVIEW

nature publishing group npg

Closing remarks

- Linking publication and underlying research data strengthen the scientific record and increases transparency.
- Funder policies are a key driver to make research data more available
- Journals can and do influence data deposition
- Not one size fits all…..
- Partnerships will be important (institutions, repositories, publishers, researchers, funders)

- However roles are not well established.
- Business models are to be determined.