

# Open (Research) Data

## Why do we care and what can we do?

Sünje Dallmeier-Tiessen

CERN

Scientific Information Service



Temporal and Dominion as enbled at Westminster  
lawfully, fully and freely representing all the Estates  
of the People of this Realm did upon the thirteenth  
day of February in the year of our Lord one  
thousand six hundred eighty eight present unto  
their Majesties then called and known by the  
Names and Title of William and Mary Prince  
and Princess of Orange being present in their  
proper persons or in their Declaration in a  
Writing made by the said Lords and Commons  
in the words following vizt Whereas the  
late King James the second by the Assistance  
of diverse evil Counsellors Judges and Ministers  
employed by him did endeavour to subvert and  
extirpate the Protestant Religion and the Lawes  
and Liberties of this Kingdome By assuming  
and exercising a power of Dispensing with  
and suspending of Lawes and the Execution of  
Lawes without Consent of Parliament By  
committing and prosecuting diverse wrong  
Offences for humbly petitioning to be excused from  
concurring to the said assumed Power By issuing  
and causing to be executed a Commission under  
the Great Seale for erecting a Court called the  
Court of Commissioners for Ecclesiasticall Causes  
By levying money for and to the use of the Crowne  
by pretence of Prerogative for other times and in  
other manner then the same was granted by  
Parliament By raising and keeping a standing  
Army within this Kingdome in time of Peace  
without Consent of Parliament and Quartering  
Soldiers contrary to Law By raising severall  
good Subjects being Protestants to be disarmed  
at the same time when Papists were both armed  
and employed contrary to Law By violating the  
freedome of Election of Members to serve in  
Parliament By prosecutions in the Court of  
King's Bench for matters and causes requirable

Zoom in

Hide details

Close viewer

## Description

Full title:

The Bill of Rights

Created:

1689

Formats:

Manuscript

Held by:

© Parliamentary Archives, London HL/PO/JO  
/10/1/1430, memrs. 2-3

Shelfmark:

HL/PO/JO/10/1/1430

A determined attempt by King James II (r. 1685–88) to reinstate Catholic worship in England, coupled with his increasingly authoritarian responses to resistance, resulted in a wave of unrest in 1688. In November, a Dutch force led by Prince William of Orange (the future King William III, r. 1689–1702) invaded England in support of the king's opponents. After James's army had crumbled and he had fled to France, William (husband of James's elder daughter, Mary) summoned a new Parliament, the 'Convention'.

The Convention assembled on 22 January 1689, and within two weeks had voted that King James II had 'abdicated the government' by 'breaking the original contract between King and people'. Having 'withdrawn himself out of the government', James had left the throne vacant. In parallel with debates on whether James should be formally removed from the throne, both Houses of Parliament agreed a statement to assert and confirm what were seen as ancient laws and liberties, and to underline the arbitrary and illegal

# What is this “research data” thing?

“the evidence used to inform or support research conclusions. Some key facets include:

- Method of creation or collection: from observations, experiments or simulations, derived from existing data or obtained from reference/canonical sets
- Readiness for use: raw, cleaned and calibrated, summarised or visualised
- Format: text; tables; time series; images; video and audio recordings
- Type of content: interviews;
- Size: large or small files; individual files or large file sets; sometimes just a single number
- Storage location: local drives; institutional filestores/repositories; national/international data centres; cloud services”

# Who cares?



# Swiss National Science Foundation

The SNSF therefore expects all its funded researchers

- to store the research data they have worked on and produced during the course of their research work,
- to share these data with other researchers, unless they are bound by legal, ethical, copyright, confidentiality or other clauses, and
- to deposit their data and metadata onto existing public repositories in formats that anyone can find, access and reuse without restriction.

Research data is collected, observed or generated factual material that is commonly accepted in the scientific community as necessary to document and validate research findings.

# US and UK



**National Science Foundation**  
WHERE DISCOVERIES BEGIN

QUICK LINKS

SEARCH

HOME FUNDING AWARDS DISCOVERIES NEWS PUBLICATIONS STATISTICS ABOUT NSF FASTLANE

**Office of Budget, Finance and Award Management (BFA)**



[DIAS Home](#)  
[CAAR Branch](#)  
**Policy Office**  
[Systems Office](#)  
[View DIAS Staff](#)

Search DIAS Staff

[BFA Organization](#)  
**Office of Budget, Finance, & Award Management**  
[Budget Division](#)  
[Division of Acquisition and Cooperative Support](#)

## Dissemination and Sharing of Research Results

### NSF Data Sharing Policy

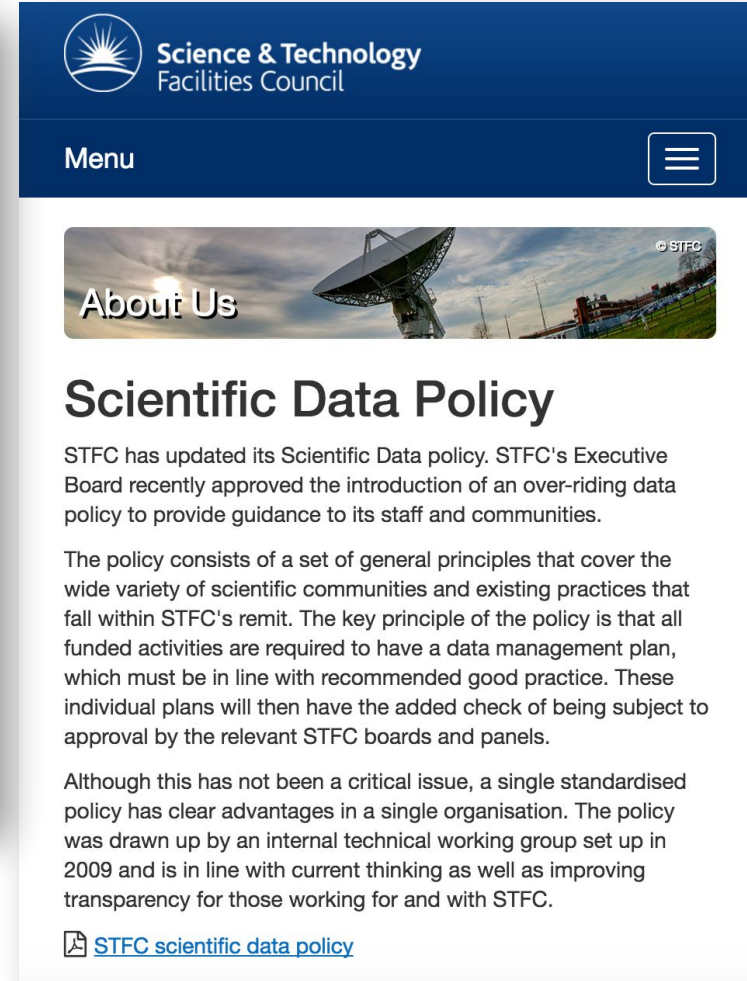
Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. See [Award & Administration Guide \(AAG\) Chapter VI.D.4](#).

### NSF Data Management Plan Requirements

Proposals submitted or due on or after January 18, 2011, must include a supplementary document of no more than two pages labeled "Data Management Plan". This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results. See [Grant Proposal Guide \(GPG\) Chapter II.C.2.i](#) for full policy implementation.


### Requirements by Directorate, Office, Division, Program, or other NSF Unit

Links to data management requirements and plans relevant to specific Directorates, Offices, Divisions, Programs, or other NSF units, are provided below. If guidance specific to the program is not provided, then the requirements established in [Grant Proposal Guide, Chapter II.C.2.i](#) apply.



**Science & Technology Facilities Council**

Menu



About Us

## Scientific Data Policy

STFC has updated its Scientific Data policy. STFC's Executive Board recently approved the introduction of an over-riding data policy to provide guidance to its staff and communities.

The policy consists of a set of general principles that cover the wide variety of scientific communities and existing practices that fall within STFC's remit. The key principle of the policy is that all funded activities are required to have a data management plan, which must be in line with recommended good practice. These individual plans will then have the added check of being subject to approval by the relevant STFC boards and panels.

Although this has not been a critical issue, a single standardised policy has clear advantages in a single organisation. The policy was drawn up by an internal technical working group set up in 2009 and is in line with current thinking as well as improving transparency for those working for and with STFC.

[STFC scientific data policy](#)

# What about the other parties?

“A condition of publication in a Nature journal is that authors are required to make **materials, data, code, and associated protocols promptly available to readers without undue qualifications**. Any restrictions on the availability of materials or information must be disclosed to the editors at the time of submission. Any restrictions must also be disclosed in the submitted manuscript.”

## Over 600 Springer Nature journals commit to new data sharing policies

*All policies available under a Creative Commons license*

London, 6 December 2016

More than 600 journals across Nature Research, Springer, BioMed Central and Palgrave Macmillan have committed to encouraging good practice in the sharing and archiving and citation of research data by adopting new Springer Nature [research data policies](#). The text of the policies has today been made available under a Creative Commons Attribution (CC BY 4.0) license so that they can be re-used by the wider research community.

These easy-to-understand policies encourage the publication of more open and reproducible research, and aim to increase growth and innovation in research and data



# Open Science

Open Science is an umbrella term that refers to the opening of scholarly knowledge creation and dissemination towards a multitude of stakeholders. It comprises, for instance, forms collaboration among researchers through online tools, emerging publication formats, the involvement of non-experts in the research or the alternative assessment of impact.

Friesike, Sascha & Schildhauer, Thomas (2014)

Transparency

Freedom

Efficiency

# The “R”s

Reproducibility

Reusability

Repurposability

Replicability

Re...

Getting the most out of investment: time, funding, dedication, passion

Sharing unique results

# Reproducibility

**nature** International weekly journal of science

Search   [Advanced search](#)

[Home](#) | [News & Comment](#) | [Research](#) | [Careers & Jobs](#) | [Current Issue](#) | [Archive](#) | [Audio & Video](#) | [For Authors](#)

[Archive](#) > [Volume 533](#) > [Issue 7604](#) > [Editorial](#) > [Article](#)

NATURE | EDITORIAL

[Share](#) [Email](#) [Print](#) [E-alert](#) [RSS](#) [Facebook](#) [Twitter](#)

## Reality check on reproducibility

A survey of *Nature* readers revealed a high level of concern about the problem of irreproducible results. Researchers, funders and journals need to work together to make research more reliable.

25 May 2016

[PDF](#) [Rights & Permissions](#)

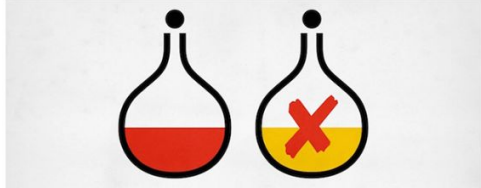
Is there a reproducibility crisis in science? Yes, according to the readers of *Nature*. [Two-thirds of researchers who responded to a survey by this journal](#) said that current levels of reproducibility are a major problem.

The ability to reproduce experiments is at the heart of science, yet failure to do so is a routine part of research. Some amount of irreproducibility is inevitable: profound insights can start as fragile signals, and sources of variability are infinite. But, the survey suggests, there is a bigger issue — and something that needs to be fixed. One-third of the survey respondents said that they think

### Related stories


- [The pressure to publish pushes down quality](#)
- [Research data: Silver lining to irreproducibility](#)

### Crisis talks




**1,500 scientists lift the lid on reproducibility**  
Survey sheds light on the 'crisis' rocking research.

[Like](#) [Share](#) Eamonn Maguire and 258,314 others like this.



**Sign up for FREE today**



Recent **Read** Commented

http://www.nature.com/news/reality-check-on-reproducibility-1.19961

**Tales**

# E coli outbreak: German organic farm officially identified

Eat cucumbers, tomatoes and lettuce again, say German health authorities, but avoid bean sprouts



**i** The cause of the E coli outbreak has been officially linked in Germany to the consumption of bean sprouts. Photograph: Christian Charisius/EPA

Bean sprouts from an organic farm in northern [Germany](#) caused the *E coli* outbreak that has killed 31 people and infected thousands more, German officials said on Friday.

Health inspectors have identified the source of the infections after linking patients who fell ill with the bug to 26 restaurants and cafes known to have received produce from the farm in Lower Saxony.

## E Coli outbreak in Germany

Previous post

[Report plots strategy for UK taxonomy](#)

Next post

[A Picasso fetches £13.5 million for obesity research](#)

NEWS BLOG

# The German E. coli outbreak: 40 lives and hours of crowdsourced sequence analysis later

20 Jun 2011 | 17:03 BST | Posted by Brian

Posted on behalf of Marian Turner.

The outbreak of *E. coli* infections in Germany still reported over the weekend. Since the first case, 40 people have been infected, of which 849 contracted the severe form.

On 10 June, the Robert Koch Institute announced that *E. coli* O104:H4, had been found on an organic strawberry. The most burning question of where the bacteria came from remains unanswered.

Information is pouring out of collaborative efforts. The Beijing Genomics Institute released the full genome of the strain. An international group of scientists jumped on the news and sequenced using Ion Torrent technology, generating

<https://github.com/ehec-outbreak-crowdsourced/BGI-data-analysis/wiki>

# Crowdsourced approach

This screenshot shows a GitHub repository page for 'ehc-outbreak-crowdsourced / BGI-data-analysis'. The repository has 16 watchers, 64 stars, and 0 pull requests. The active tab is 'Wiki', showing a page titled 'Home' edited by Peter Cock on 22 Aug 2013. The main heading of the wiki page is 'E. coli O104:H4 Genome Analysis Crowdsourcing'. The text on the page states: 'In this wiki we aim to gather all the results of the analysis of the E. coli O104:H4 strain responsible for the May/June 2011 outbreak in Germany and Europe. TEN isolates from the outbreak have been sequenced so far:'. On the right side, there is a 'Pages' sidebar with 25 pages, including 'Home' and 'Alignment of all three assemblies with 55989 and plasmids'.

# Open Street Map

Open

Collaborative, crowdsourced

Flexible for users and developers

Free, full access and control

Trust - Conflicts

The screenshot displays the OpenStreetMap web interface. At the top left is the OpenStreetMap logo. To its right are navigation buttons: 'Bearbeiten' (with a dropdown arrow), 'Chronik', 'Export', 'Mehr' (with a dropdown arrow), 'Anmelden', and 'Registrieren'. Below the logo is a search bar containing the text 'epfl|', with 'Los' and a search icon button to its right. Under the search bar, the heading 'Suchergebnisse' is followed by a close button (X). Below this, it says 'Ergebnisse von OpenStreetMap Nominatim'. Two search results are listed:

- Universität École Polytechnique Fédérale de Lausanne (EPFL), Place Cosandey, Ecublens, District de l'Ouest lausannois, Waadt, Schweiz
- Öffentliches Gebäude EPFL, Rue Robert Blum, Moselbrück, Nancy, Mörthe und Mosel, Elsass-Champagne-Ardennen-Lothringen, Metropolitanes Frankreich, 54700, Frankreich

The right side of the screenshot shows a map of the EPFL campus area, including labels for 'Bassenges', 'EPFL', 'UNIL-Sorge', and 'Rofex Learning Center'. The map includes standard navigation controls on the right edge: a vertical zoom slider, a compass, a home button, a full-screen button, and a help button (question mark).

# CERN Open Data (opendata.cern.ch)

## Education

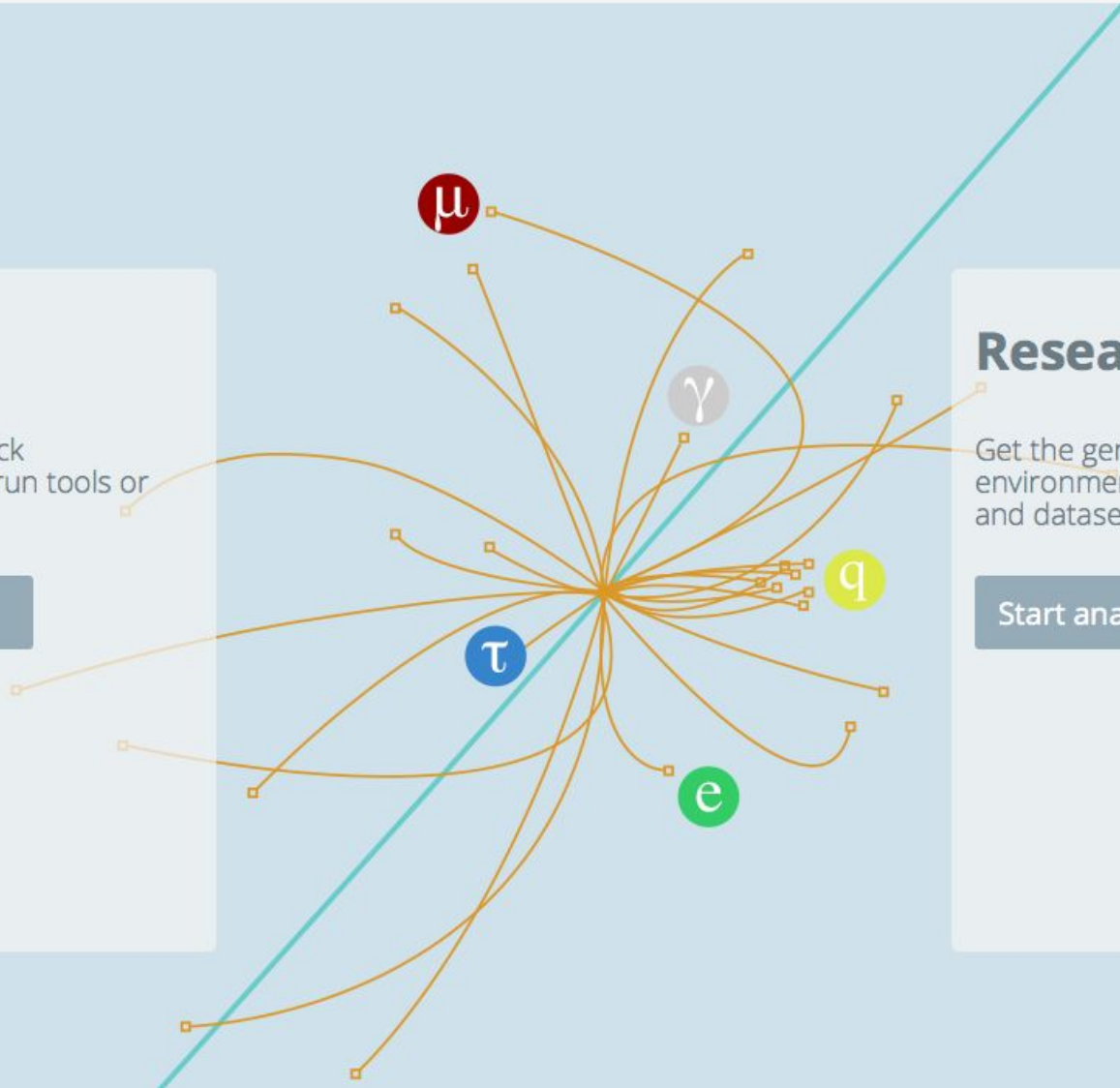
Visualise events, check reconstructed data, run tools or build your own!

Start learning

## Research

Get the genuine working environments, virtual machines and datasets to start your research

Start analysing





Explore more than **1 petabyte**  
of open data from particle physics!

Start typing...

Search

search examples: [collision datasets](#), [keywords:education](#), [energy:7TeV](#)

## Explore

[datasets](#)  
[software](#)  
[environments](#)  
[documentation](#)

## Focus on

[ATLAS](#)  
[ALICE](#)  
[CMS](#)  
[LHCb](#)

▾ Get started ▾

# Impact

The Washington Post

Speaking of Science

## Open sourcing the secrets of the universe huge amount of data from Hadron Collider now online

By Sarah Kaplan April 26 



Science

## Cern makes 300TB of data available

By EMILY REYNOLDS

25 Apr 2016



## Teilchenbeschleuniger LHC: 300 Terabyte freigegeben

 heise online 26.04.2016 11:34 Uhr - Martin Holland





High Energy Physics – Phenomenology

# Jet Substructure Studies with CMS Open Data

Aashish Tripathee, Wei Xue, Andrew Larkoski, Simone Marzani, Jesse Thaler

(Submitted on 19 Apr 2017 (v1), last revised 8 May 2017 (this version, v2))

We use public data from the CMS experiment to study the 2-prong substructure of jets. The CMS Open Data is based on 31.8/pb of 7 TeV proton-proton collisions recorded at the Large Hadron Collider in 2010, yielding a sample of 768,687 events containing a high-quality central jet with transverse momentum larger than 85 GeV. Using CMS's particle flow reconstruction algorithm to obtain jet constituents, we extract the 2-prong substructure of the leading jet using soft drop declustering. We find good agreement between results obtained from the CMS Open Data and those obtained from parton shower generators, and we also compare to analytic jet substructure calculations performed to modified leading-logarithmic accuracy. Although the 2010 CMS Open Data does not include simulated data to help estimate systematic uncertainties, we use track-only observables to validate these substructure studies.

Comments: 35 pages, 19 figures, 6 tables, source contains sample event and additional plots; v2: references updated and figure formatting improved

Subjects: **High Energy Physics – Phenomenology (hep-ph)**; High Energy Physics – Experiment (hep-ex)

Report number: MIT-CTP 4890

Cite as: [arXiv:1704.05842 \[hep-ph\]](#)

(or [arXiv:1704.05842v2 \[hep-ph\]](#) for this version)

## Submission history

From: Jesse Thaler [[view email](#)]

[v1] Wed, 19 Apr 2017 18:00:00 GMT (28272kb,AD)

[v2] Mon, 8 May 2017 01:19:34 GMT (25903kb,AD)

## Download:

- PDF
- [Other formats](#)

(license)

## Ancillary files (details):

- [sample\\_mod\\_file.mod](#)

## Current browse context:

hep-ph

< [prev](#) | [next](#) >

[new](#) | [recent](#) | [1704](#)

## Change to browse by:

[hep-ex](#)

## References & Citations

- [INSPIRE HEP](#)  
([refers to](#) | [cited by](#))
- [NASA ADS](#)

## Bookmark (what is this?)



# Data as a **Community Resource**

In 1996, the International Human Genome Sequencing Consortium adopted the "Bermuda Principles"

*that expressly called for the automatic, rapid release of sequence assemblies of 1-2 kb or greater to the **public domain**.*

To implement the Bermuda Principles, in April 1997 the NHGRI adopted a data release policy

*that called upon those of its grantees engaged in large-scale genomic DNA sequencing to release DNA sequence assemblies of 2 kb or greater within 24 hours of their generation.*

"community resource project" was defined as a research project specifically devised and implemented to create a set of data, reagents or other material whose primary utility will be as a resource for the broad scientific community

# Contents

16 FEBRUARY 2001  
VOL 291, ISSUE 5507

## Special Issue **The Human Genome**

### INTRODUCTION TO SPECIAL ISSUE

#### VIEWPOINT

### **Science Genome Map**


SCIENCE | 16 FEB 2001 : 1218 | 

Summary Full Text

#### VIEWPOINTS

### **The Human Genome and Our View of Ourselves**

BY SVANTE PÄÄBO

SCIENCE | 16 FEB 2001 : 1219-1220 | 

Summary Full Text

### **Proteomics in Genomeland**

BY STANLEY FIELDS

SCIENCE | 16 FEB 2001 : 1221-1224 | 

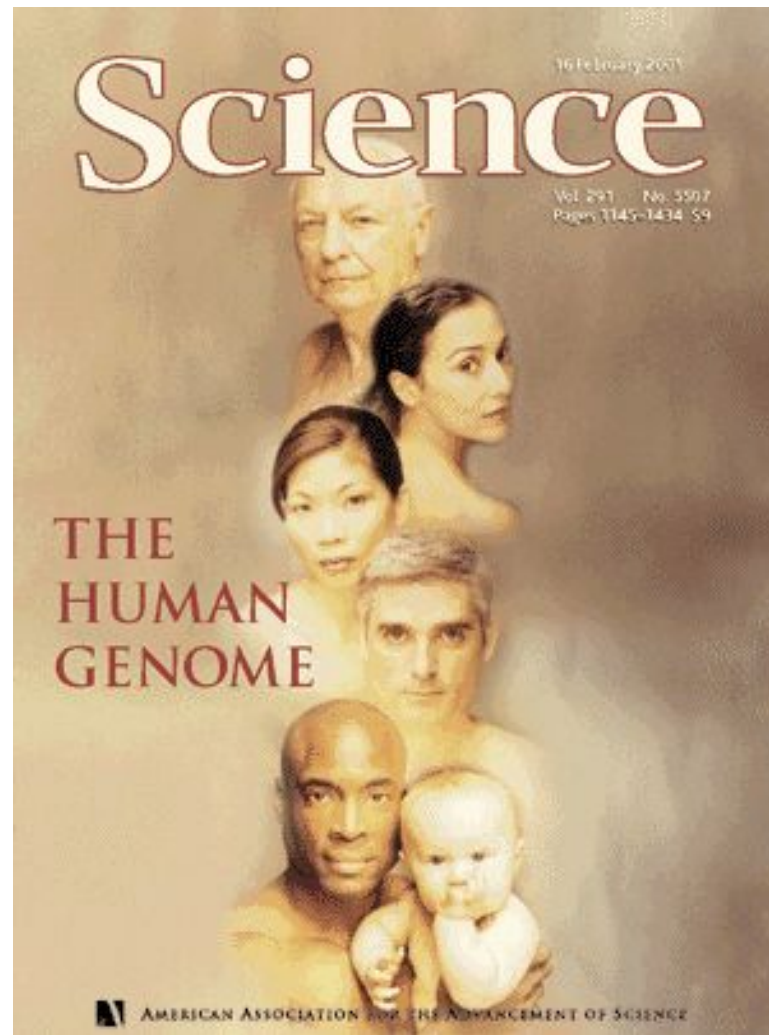
Summary Full Text

### **Dissecting Human Disease in the Postgenomic Era**

BY LEENA PELTONEN, VICTOR A. MCKUSICK

SCIENCE | 16 FEB 2001 : 1224-1229 | 

Summary Full Text



<http://science.sciencemag.org/content/291/5507>

What can we do

# Standard response: “Manage your data”

- Have an actual plan (DMP)
- Repositories
  - Discipline
  - Institution
  - General ones
- Other “hosts”
  - Collaborative platforms
  - The rest
- Journals

# Data Management Plan (DMP)

Helps understanding the “outputs” of your project

How do you want to manage these outputs?

How do you preserve them? (or who does it?)

How do you publish/share them?

Do you require funding to manage the resources?

Plan is often required by institutions, funders

