

What is



DDBJ

DNA Data Bank of Japan

<http://www.ddbj.nig.ac.jp/>

A Platform for International Data Sharing

Collaboration of **3**
3 Inlets for One content in 3 formats



The screenshot shows the International Nucleotide Sequence Database (INSDC) homepage. The header features the 'INSDC' logo with a circular emblem and the text 'International Nucleotide Sequence Database'. Below the header, there are three main navigation links: 'ABOUT INSDC', 'POLICY', and 'ADVISORS'. The 'ABOUT INSDC' section contains logos for EMBL, NCBI, and DDBJ, each with their respective names and brief descriptions. The 'POLICY' section includes a bulleted list of collaborative details between the databases. The 'ADVISORS' section is partially visible.

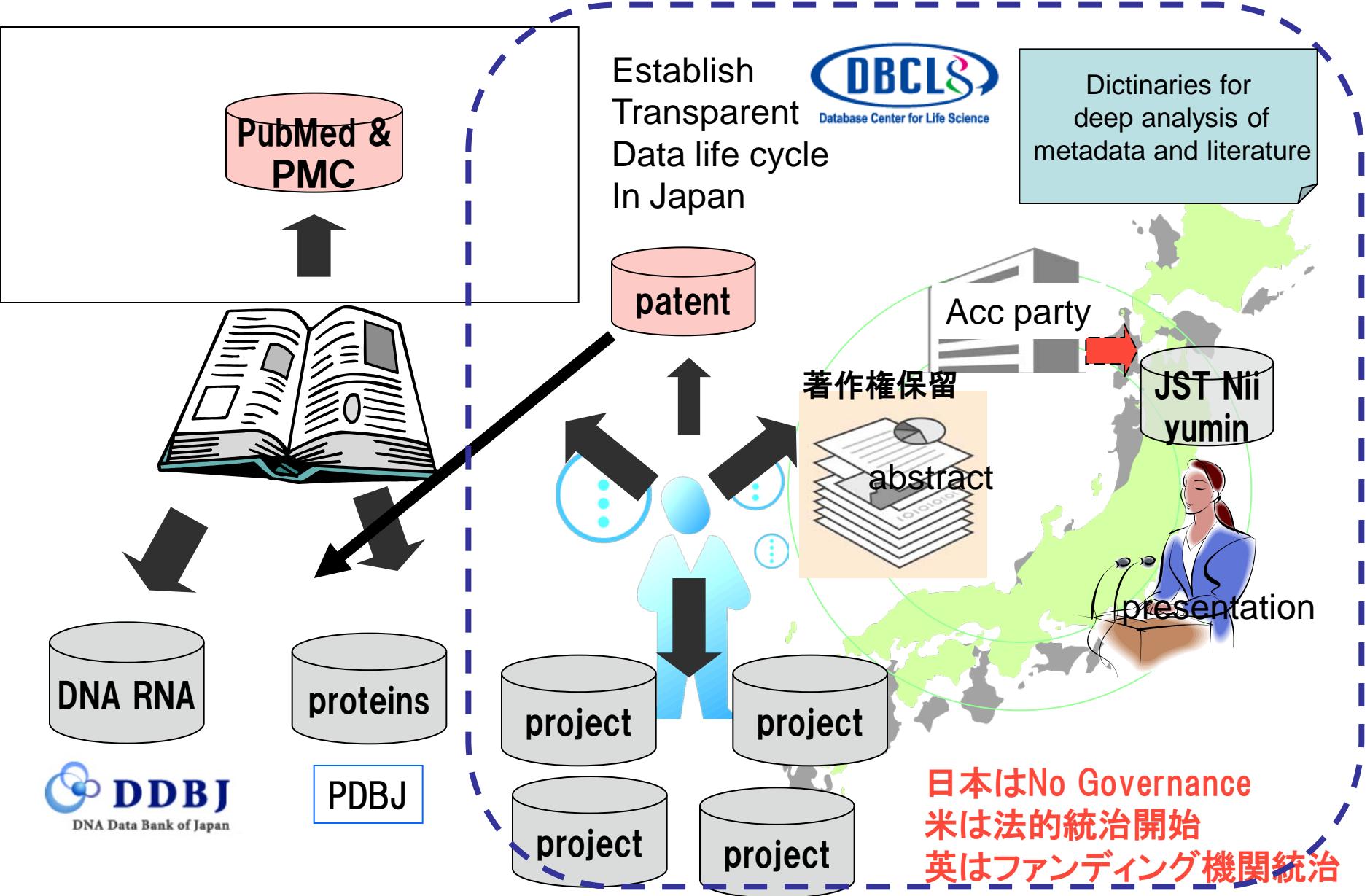
I
International Nucleotide Sequence Database

- The International Nucleotide Sequence Database is maintained collaboratively between [EMBL](#), [NCBI](#), and [DDBJ](#).
- The INSDC advisory board, the [International Nucleotide Sequence Database Advisory Committee](#), consists of members of each of the databases' advisory committees. Members of this committee unanimously adopted the data-sharing policy of the three databases below.
- Individuals submitting data to the international database should follow the [INSDC policy](#).

How to submit data

- For full details of how to submit data to your partner, see the [partner's page](#).
- [DDBJ](#), [EMBL](#), [GenBank](#)

Collaboration among DBCLS, PDBj, DDBJ



科学データの3側面

- Syntax
- Semantics
- Pragmatics
 - コントロール権問題: IP, Privacy, Credit
 - サイズの問題: Disk space, search time

DBCLS

DDBJ

Pragmatic problem #1

Data Sharing



**WHY SHARING
DATA IS
IMPORTANT**

科学機構とは 知識・データの累積と共有の保障機構

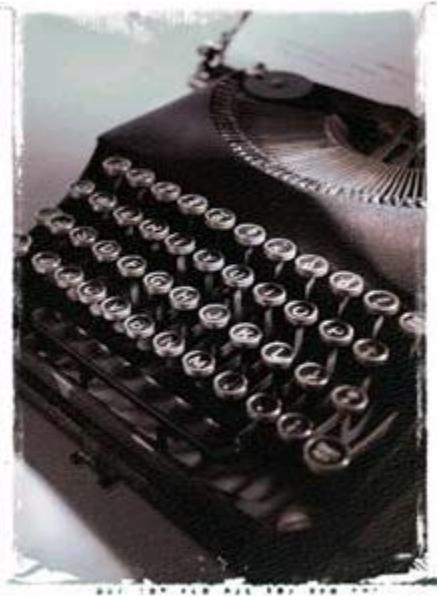
「もしも私が遠くまで見渡せるとすれば、それは
私が巨人の肩の上に載っているからだ」

アイザック・ニュートン卿

皆が知識やデータを分断所有してると
科学が進歩しない



1940's

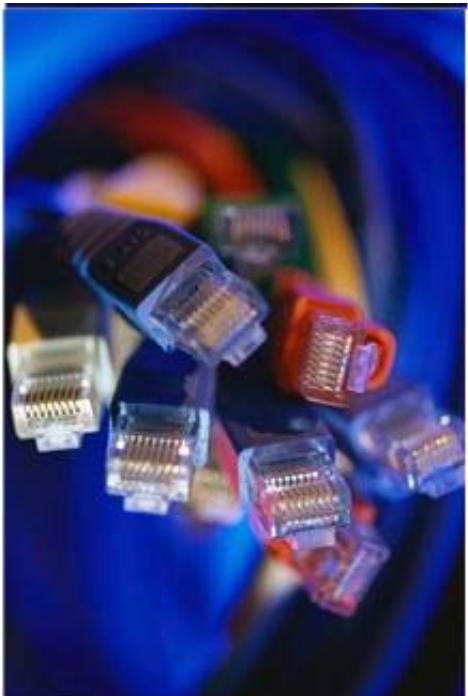


The key to success
is to find a man of
genius, give him
money, and leave
him alone.”

-----James Conant, *president, Harvard University*

論文があればいい時代

2000's



“ Not all the smart people in the world is working for you.”

--Billy Joy, *founder, Sum Microsystems*

論文よりデータが使いたい時代



**WHY SHARING
DATA IS
DIFFICULT**

TALKS | IN LESS THAN 6 MINUTES

Tim Berners-Lee: The year open data went worldwide

TED2010, Filmed Feb 2010; Posted Mar 2010



00:48 | 05:34

Share

Rate

Subtitles Available in:

19 languages [Off]

About this talk[Open interactive transcript](#)

At TED2009, Tim Berners-Lee called for "raw data now" -- for governments, scientists and institutions to make their data openly available on the web. At TED University in 2010, he shows a few of the interesting results when the data gets linked up.

About Tim Berners-Lee

Tim Berners-Lee invented the World Wide Web. He leads the World Wide Web Consortium, overseeing the Web's standards and development. [Full bio and more links](#)

About our sponsor

On the human network, video changes everything.
[See how »](#)



DATABASE HUGGING

- “They are very tempted to keep it. Hans called it “Database Hugging”.
- You hug your data until you’ve make a beautiful website for it.....
.....You have no idea the number of excuses people come up with to hang on to their data.

Tim-Bernars Lee @ TED (2009)

http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html



HOW CAN WE
MAKE THEM
SHARE



Greed

From a screenplay “ Greed” (1926)

Instructions to Authors (AJHG)

- Nucleic acid and protein sequences, single-nucleotide polymorphisms (SNPs), copy number variants
- (CNVs), microarray data, and macromolecular structures determined by X-ray crystallography (along
- with structure factors) must be deposited in the appropriate public database and must be accessible
- without restriction from the date of publication. The URL of the databases used must be included in the
- Web Resources section of the manuscript. All entry names and/or accession numbers must be included
- in the Material and Methods section. Microarray data should be MIAME compliant (for guidelines see
- <http://www.mged.org/Workgroups/MIAME/miame.html>).
- Newly described SNPs should be submitted to an appropriate database such as dbSNP
- (<http://www.ncbi.nlm.nih.gov/SNP/>) prior to submission of revised manuscripts. The identification
- numbers should be used to describe the SNPs in the manuscript.

平成22年度
科学研究費補助金公募要領
(研究成果公開促進費)
— 学術定期刊行物、学術図書、データベース —

A Grat in aid for
Dissemination of Research Results
--Periodicals Textbooks Databases--

NIH Data Sharing Policy

"Data should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data."

—Final NIH Statement on Sharing Research Data February 26, 2003

**What You
Need to Know
for Successful
Funding**

Data Sharing and Scoring

Reviewers will not use the data sharing plan in determining scientific merit or priority score for applications. Program staff are responsible for overseeing the data sharing policy and for assessing the appropriateness and adequacy of the data sharing plan. Program concerns must be resolved prior to making an award.

Cost of Data Sharing

Applicants may request funds for data sharing and archiving. The financial issues should be addressed in the budget section of the application. If the data have been collected already, a competitive or administrative supplement may be available.

We placed our National Survey of Adolescent Males data in the Data Archive on Adolescent Pregnancy and Pregnancy Prevention where it has been accessed for analysis and classroom projects. The advantage of this approach is that the Archive provides both on-line assistance and publicity on data availability to the research and education communities.

Fraya Sonenstein, The Urban Institute

NIH Staff Can Help

Instructions related to data sharing can be found in the specific Request for Proposal (RFP), Request for Application (RFA), or Program Announcement (PA). However, NIH encourages investigators to consult with NIH program staff prior to submitting an application to determine the appropriateness of data sharing and a suitable mechanism for disseminating data.

For more information on the NIH Data Sharing Policy, including a link to the Final NIH Statement on Sharing Research Data, visit the NIH Office of Extramural Research Website on Data Sharing Policy:

http://grants.nih.gov/grants/policy/data_sharing



DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health

Office of Extramural Research

9000 Rockville Pike

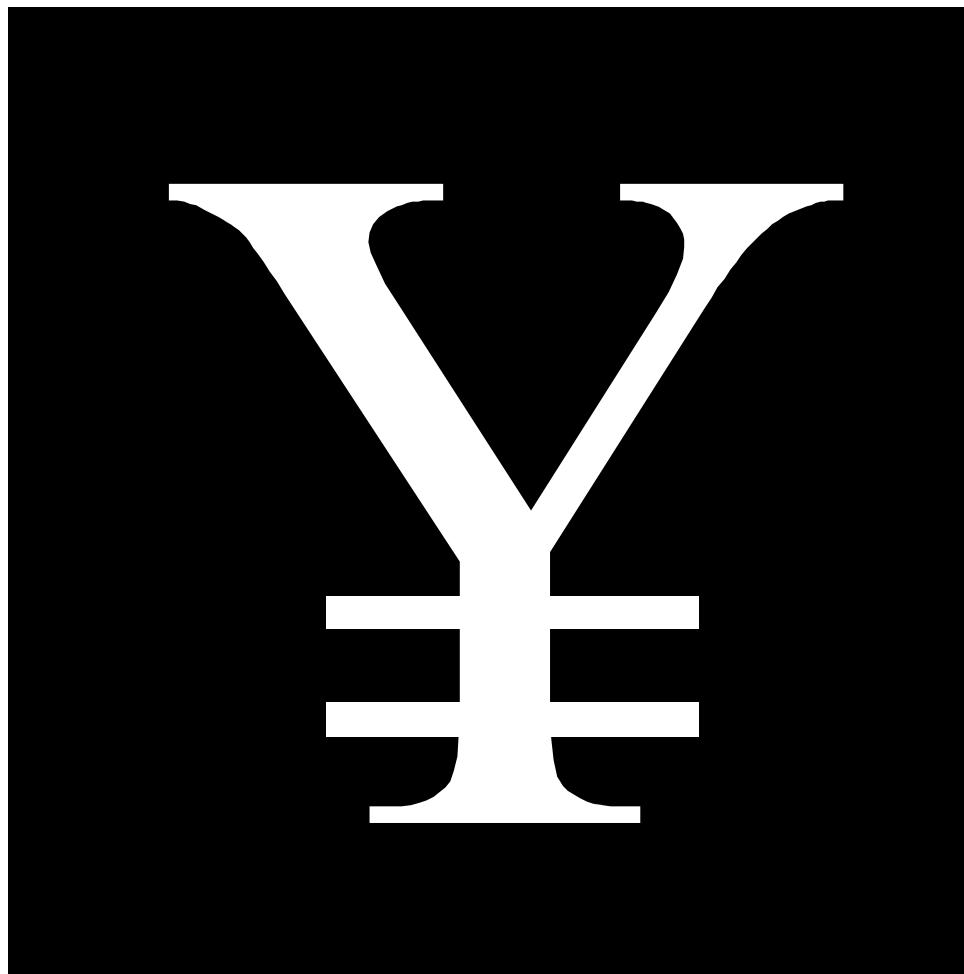
Rockville, MD 20850

Tel: 301-496-1096

<http://grants.nih.gov/grants/oer.htm>

The National Institutes of Health (NIH), part of the Department of Health and Human Services (DHHS), is the principal health research agency of the Federal Government. The Office of Extramural Research (OER) provides policies and guidelines for extramural research grants administration. OER has primary responsibility for developing and implementing NIH Grants Policy, including policies related to data and safety monitoring and protection of human subjects; monitoring compliance with humane use and care of laboratory animals policy; coordinating program guidelines; and developing and maintaining the information systems for grants administration.







*F*ear

CC-By-SA
Photo by MyName ([Bantosh](#))



OECD Principles and Guidelines for Access to Research Data from Public Funding

(2009)

「21世紀の科学技術とイノベーションの為の OECD 科学技術委員会」

第 11 回科学技術政策委員会閣僚級会合 2004 最終共同声明への 付録 1

公的資金由来の研究データへのアクセスについての宣言

DECLARATION ON ACCESS TO RESEARCH DATA FROM PUBLIC FUNDING

(2004 年 1 月 30 日 パリにて採択)

オーストラリア、オーストリア、ベルギー、カナダ、中国、チェコ共和国、デンマーク、フィンランド、フランス、ドイツ、ギリシャ、ハンガリー、アイスランド、アイルランド、イスラエル、イタリア、日本、韓国、ルクセンブルグ、メキシコ、オランダ、ニュージーランド、ノルウェイ、ポーランド、ポルトガル、ロシア連邦、スロバキア共和国、スイス、トルコ、英国、米国 の各国政府（欧州共同体を含む）は次のような認識を共有しております、

- ① データ、情報、知識の最適な国際交換は必ず科学研究の進展とイノベーションに貢献する。
- ② データへのオープンアクセスと制約のない利用は科学の進歩と研究者の育成を促進する。
- ③ オープンアクセスはデータ収集努力に対する公的な投資のもたらす価値を最大化する。
- ④ コンピュータの目覚しい進歩は公的資金由来の膨大な量の研究データを世界中の多数の研究施設の多様な目的の研究への利用転用を可能にし、それによって研究のスコープやスケールを増大させる。
- ⑤ 公的資金由来の研究データへの不等なアクセス制限は研究とイノベーションの質と効率を損ないうる。
- ⑥ 最適な研究データへのアクセスは発展途上国の地球規模の科学システムへの参加を促し、それによって途上国の社会的、経済的な発展に寄与する。
- ⑦ 公的資金由来の研究データの公表は安全保障、市民のプライバシー保護、知的財産権、保護を要する営業秘密などに関わる国内法により制約を受けるかもしれない。
- ⑧ 公的資金由来の研究データへのアクセスのいくつかの側面に関して OECD 加盟国毎にこれまでに方策が講じられており、また今後講じられると思われるが、バラバラな国内規制は国内外での公的資金研究の最適な利用の妨げになりうる。

<http://vimeo.com/1899536>

The image shows a screenshot of a Vimeo video player. At the top, the Vimeo logo is on the left, and 'Join vimeo' and 'Log In' buttons are on the right. Below the logo, the video title 'Sharon Terry, Patient Advocate' is displayed in large bold letters. To the left of the title is a circular orange icon containing a white letter 'a' and the text 'OPEN ACCESS'. Below the title, it says 'by Open Access Videos' and '2 years ago'. The main video frame shows a woman with long brown hair smiling. A cursor icon is visible on her forehead. On the right side of the video frame, there is overlaid text: 'Sharon Terry' on top and 'Patient Advocate' below it. To the right of the video frame are four black rectangular buttons with white icons: 'LIKE' (heart), 'SHARE' (link), 'EMBED' (code), and 'HD IS ON' (HD). At the bottom of the video frame, there is a progress bar showing '01:01' and a play button icon.

Sharon Terry, Patient Advocate

by Open Access Videos
2 years ago

Sharon Terry
Patient Advocate

01:01

Sharon Terry advocates for Public Access Act



National Institutes of Health Public Access

The Public Access Policy ensures that the public has access to the published results of NIH funded research to help advance science and improve human health.

Home

1. Determine Applicability

2. Address Copyright

3. Submit Papers

4. Include PMCID in Citations

Policy Details

Info on Journals

Training/Communications

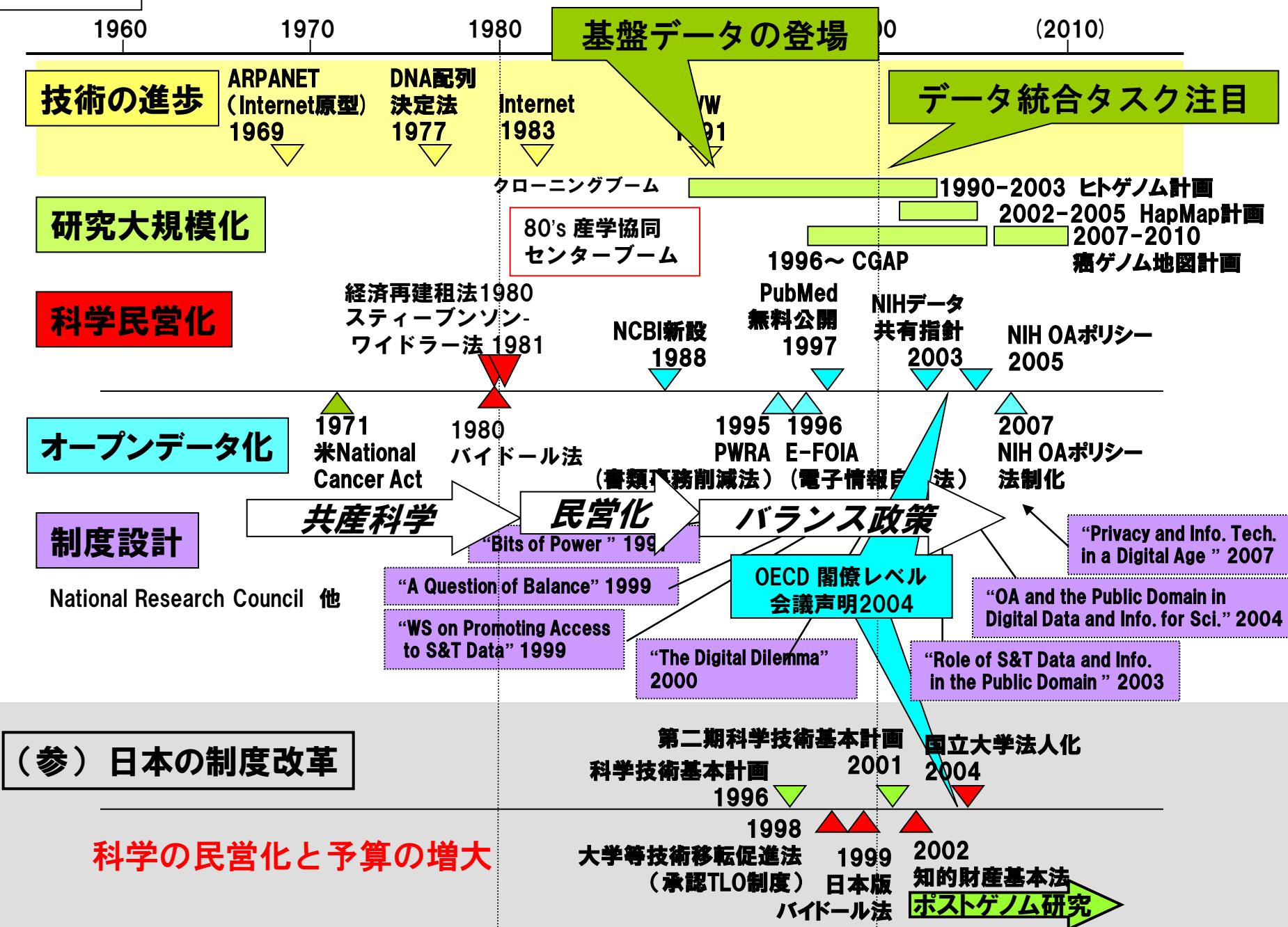
Overview

The [NIH Public Access Policy](#) ensures that the public has access to the published results of NIH funded research. It requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive [PubMed Central](#) *upon acceptance for publication*. To help advance science and improve human health, the Policy requires that these papers are accessible to the public on PubMed Central no later than 12 months after publication.

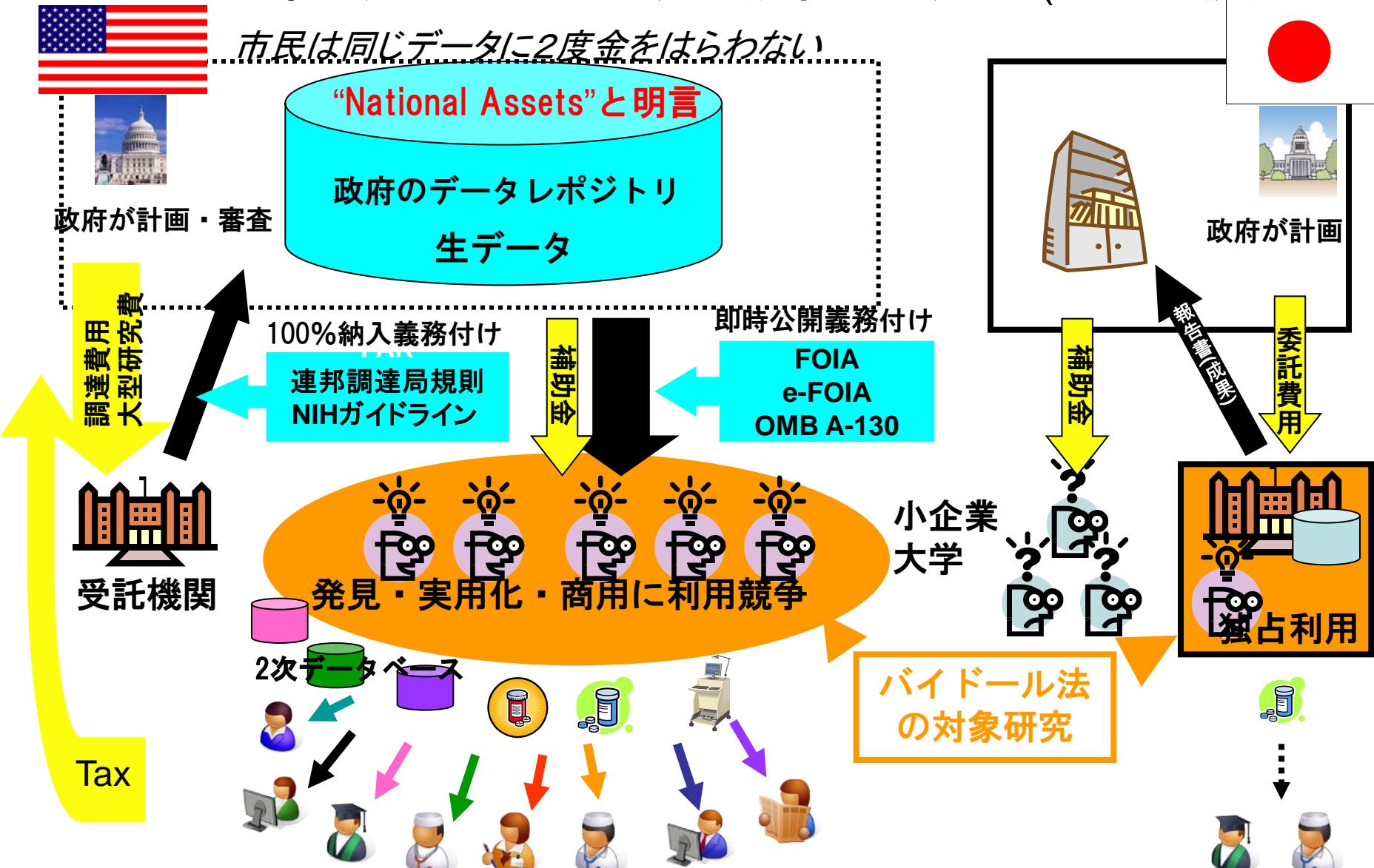
How to Comply

日米制度比較

米国公的科学制度の変遷（日米比較）



政策の違いによる下流の賑わいの違い(日米比較)



FAR : Federal Acquisition Regulation

FOIA : Freedom of Information Act

OMB : Office of Management and Budget

世界の動向

諸外国のオープン科学（論文アーカイブとデータアーカイブの義務付け）状況

	ファンディング機関	論文 データ
欧州連合	CERN (European Organization for Nuclear Research)	<input type="radio"/>
	European Commission	<input type="radio"/>
	European Research Council (ERC)	<input type="radio"/> <input type="radio"/>
オーストラリア	Australian Research Council (ARC)	
	National Health and Medical Research Council (NHMRC)	
オーストリア	Fonds zur Förderung der wissenschaftlichen Forschung (FWF)	<input type="radio"/>
ベルギー	Fonds Wetenschappelijk Onderzoek (Vlaanderen) (FWO)	<input type="radio"/>
カナダ	Canadian Breast Cancer Research Alliance (CBCRA)	
	Canadian Institutes of Health Research (CIHR)	<input type="radio"/> <input type="radio"/>
	Genome Canada	<input type="radio"/> <input type="radio"/>
	National Cancer Institute of Canada (NCIC)	<input type="radio"/>
フランス	Agence Nationale de la Recherche (ANR)	<input type="radio"/>
独逸	INSEERM (Institut national de la sante' et de la recherche médicale)	<input type="radio"/>
	Fraunhofer-Gesellschaft	<input type="radio"/>
アイルランド	Higher Education Authority (HEA)	<input type="radio"/> <input type="radio"/>
イタリア	IRCSET (Irish Research Council for Science, Engineering and Technology)	<input type="radio"/>
スイス	Instituto Superiore di Sanita (ISS)	<input type="radio"/>
	Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung	<input type="radio"/>
英国	Medical Research Council (MRC)	<input type="radio"/> <input type="radio"/>
	Biotechnology and Biological Sciences Research Council	<input type="radio"/> <input type="radio"/>
	Arts and Humanities Research Council (AHRC)	<input type="radio"/>
	Science and Technology Facilities Council (STFC)	<input type="radio"/>
	Economic and Social Research Council (ESRC)	<input type="radio"/> <input type="radio"/>
	Engineering and Physical Sciences Research Council (EPSRC)	
	Natural Environment Research Council (NERC)	<input type="radio"/> <input type="radio"/>
	Department of Health (DoH)	<input type="radio"/>
	Arthritis Research Campaign (arc)	<input type="radio"/>
	British Heart Foundation (BHF)	<input type="radio"/>
米国	Cancer Research UK	<input type="radio"/>
	Chief Scientist Office, Scottish Executive (CSO)	<input type="radio"/>
	JISC (Joint Information Systems Committee)	<input type="radio"/>
	Wellcome Trust	<input type="radio"/> <input type="radio"/>
	Howard Hughes Medical Institute (HHMI)	<input type="radio"/>
	National Institutes of Health (NIH)	<input type="radio"/> <input type="radio"/>



DDBJ as a public computer



14 annotators
20 engineers
Dr. Takagi
Dr. Nakamura
Dr. Kaminuma
Dr. Ogasawara
Dr. Takeuchi

DDBJ computation power

500 servers

1,000 +4,000 CPU

5 Terabytes Memory

0.75 Petabytes Disk

Pragmatic problem #2

Big Data

DNAデータベース (DDBJ/EMBL/GenBank=INSD) 総覧と検索

バージョン:DDBJ リ

DNAデータベースをプロジェクト単位で俯瞰、検索、分析、取得 詳細 >

トップ キーワード検索 (プロジェクト検索 ○) レコード検索 ○ 実行&二元分類 配列検索(bl)

全容 ダウンロード HELP

すべてのレコード

生物群区分
(研究プロジェクト数)

ヒト
(99666)

霊長
(7792)

齧歯
(65617)

哺乳
(25297)

脊椎
(36567)

無脊椎
(61239)

植物
(127050)

バクテリア
(92930)

ウイルス
(40723)

ファージ
(2035)

合成
(46654)

環境
(11618)

研究の型別分類
(研究プロジェクト数)

mRNA
(137095)

機能RNA・RNAゲノム
(23616)

免疫
(6028)

嗅覚
(146)

マーカー
(66412)

エクソン構造
(173670)

集団
(38675)

オルガネラ
(1457)

EST
(9275)

GS
(21)

全生物種 × すべての研究プロジェクト型 詳細 >

レコード数: 118,234,855

表示切替:

研究プロジェクト数: 635,410

研究プロジェクト単位

登録元の国一覧

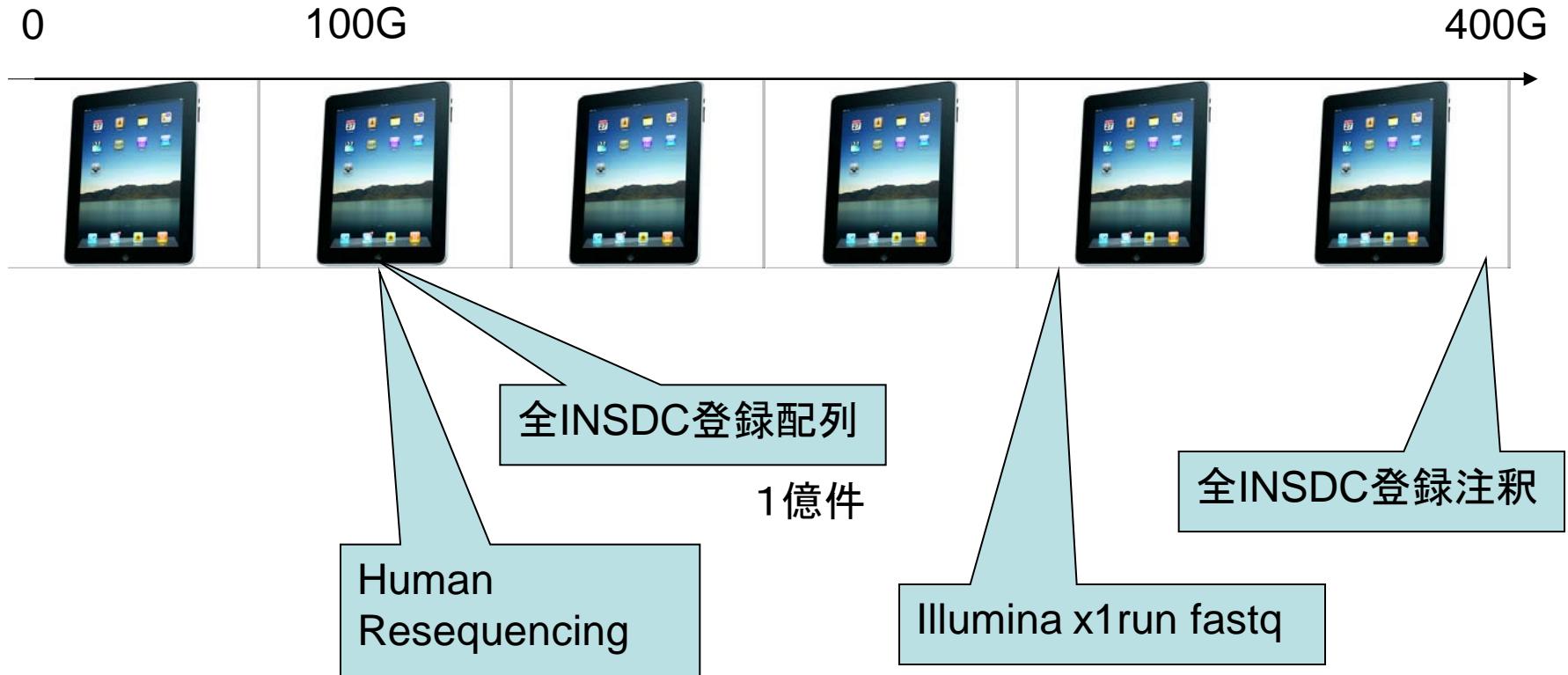
(選択:国名で絞り込む)

登録元の地理分布



	登録元の国	レコード数 研
1	アメリカ合衆国 地図	63,173,552
2	日本 地図	15,389,318
3	EPO 地図	5,320,939
4	カナダ 地図	5,101,864
5	イギリス 地図	4,375,024
6	USPTO 地図	3,697,797
7	フランス 地図	3,347,562
8	ドイツ 地図	2,778,359
9	JPO 地図	2,467,703
10	ブラジル 地図	2,120,249
11	中国 地図	2,049,857
12	ニュージーランド 地図	1,018,473

データサイズ



10億文字=1 G byte

1000人ゲノム: 100T byte

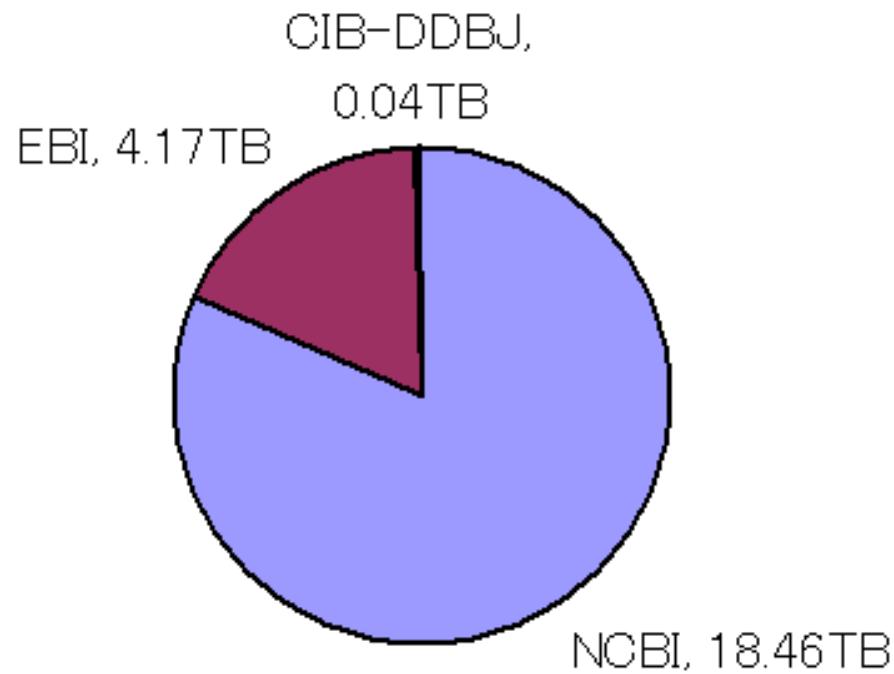
High-throughput "Next-Generation" Sequencing Facilities Map

There are **1166** total machines listed in the database situated in **397** centre

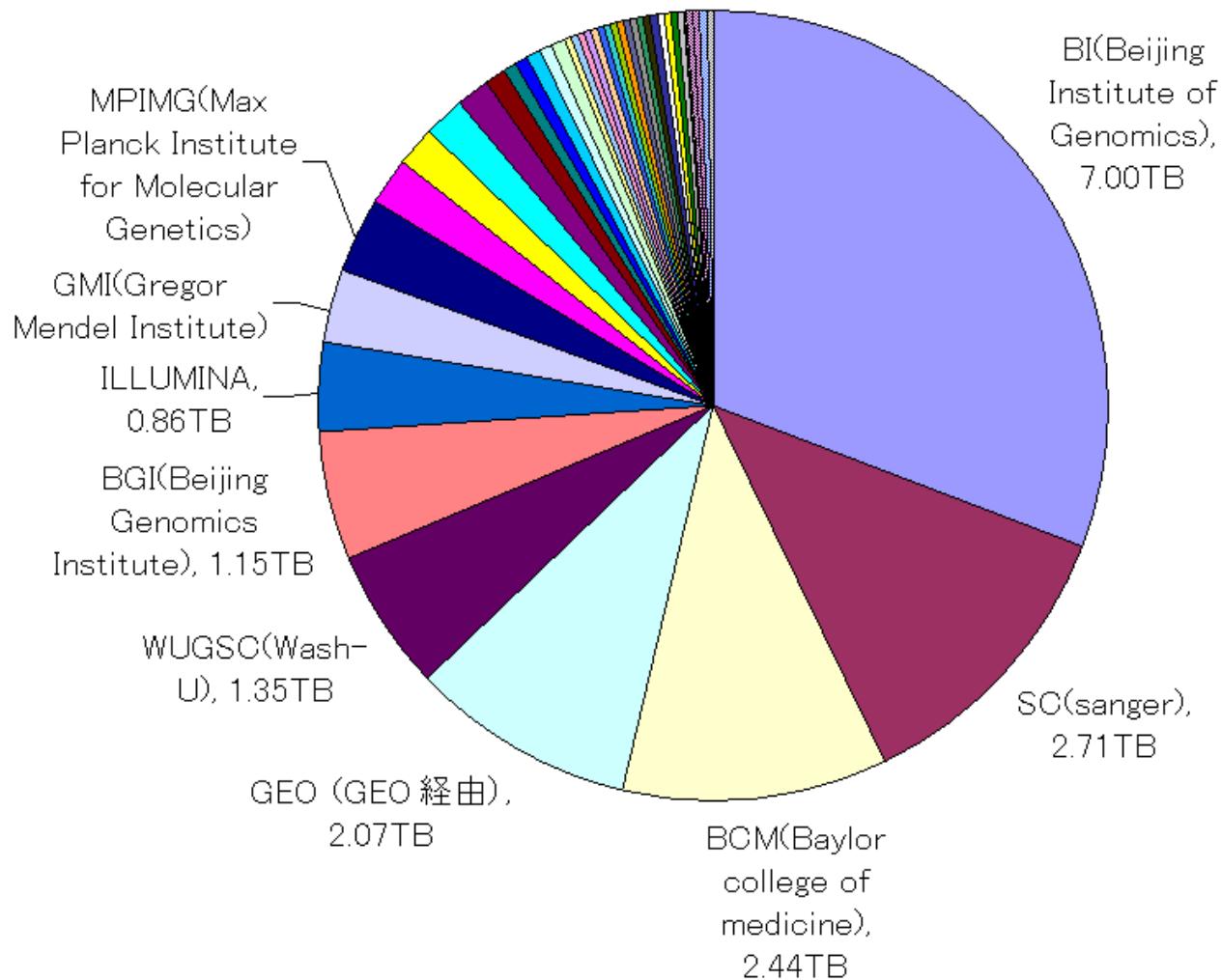


<http://pathogenomics.bham.ac.uk/hts/hts/stats>

Breakdown of SRA-INSI Open Access division

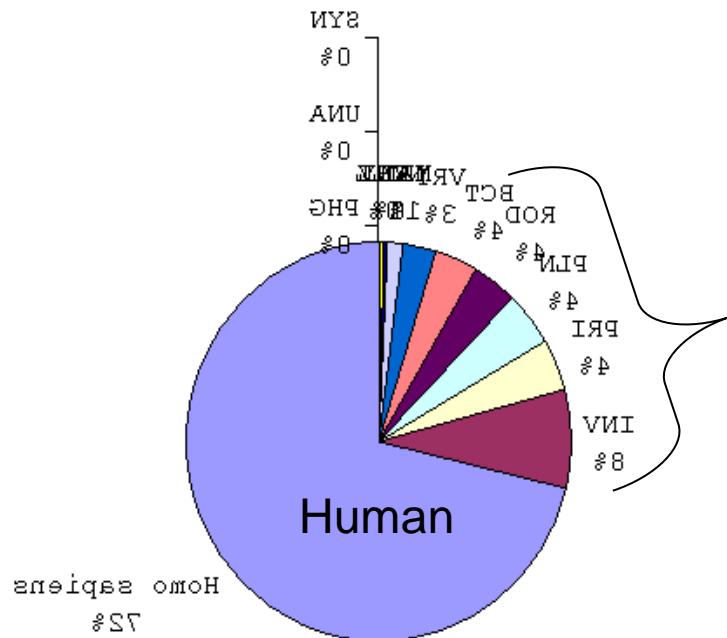


By submitting center

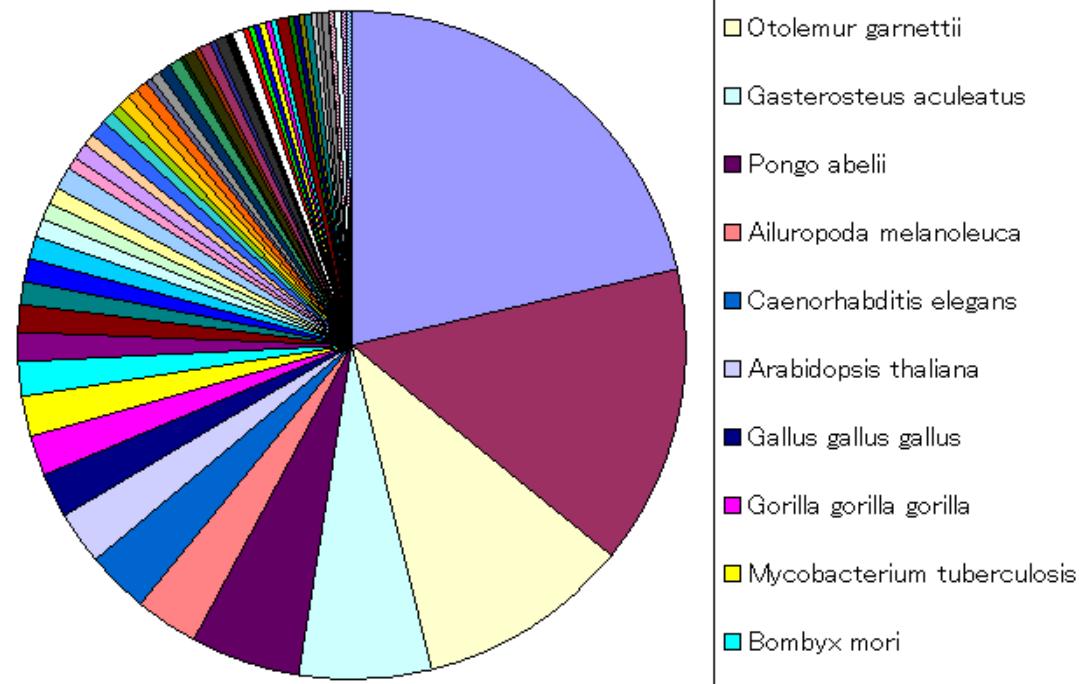


by species

ALL by sp. group



Non-human by species



- [Color Box] Drosophila melanogaster
- [Color Box] Mus musculus
- [Color Box] Otolemur garnettii
- [Color Box] Gasterosteus aculeatus
- [Color Box] Pongo abelii
- [Color Box] Ailuropoda melanoleuca
- [Color Box] Caenorhabditis elegans
- [Color Box] Arabidopsis thaliana
- [Color Box] Gallus gallus gallus
- [Color Box] Gorilla gorilla gorilla
- [Color Box] Mycobacterium tuberculosis
- [Color Box] Bombyx mori
- [Color Box] Saccharomyces cerevisiae
- [Color Box] Zea mays

Human Whole Genome Top 10					
tax	study_type	study_title	fastq	runs	DiskSht
Human	Whole Genome	1000Genomes Project Pilot 1	3.96TB	8895	17.5
Human	Whole Genome	1000Genomes Project Pilot 2	1.52TB	3214	6.7
Human	Whole Genome	Low coverage of the Japanese individuals	1.30TB	619	5.7
Human	Whole Genome	Low coverage of the Toscan individuals	1.23TB	1214	5.4
Human	Whole Genome	Low coverage of the CEPH individuals	0.67TB	383	3.0
Human	Whole Genome	Low coverage of the HapMap African ancestry individuals from SW US	0.63TB	202	2.8
Human	Whole Genome	Discovery of common Asian copy number variants using integrated high-resolution array comparative genomic hybridization	0.55TB	55	2.4
Human	Whole Genome	Low coverage of the Yoruba individuals	0.54TB	271	2.4
Human	Whole Genome	Low coverage of the Han Chinese individuals	0.45TB	74	2.0
Human	Whole Genome	The genetic structure of the indigenous hunter-gatherer peoples of Southern Africa	0.3TB	134	1.7
			11.2	15061	49.6
Human Other than Whole genome Top 10					
tax	study_type	study_title	fastq	runs	DiskSht
Human	Resequencing	1000Genomes Project Pilot 3	0.84TB	1017	3.7
Human	Epigenetics	UCSD Human Reference Epigenome Mapping Project	0.45TB	525	2.0
Human	Epigenetics	GSE1941B: Dynamic Changes in the Human Methylome During Differentiation	0.22TB	154	1.0
Human	Other	Complete Genome Sequencing and SH3TC2 Mutations Causing GMT1 Neuropathy	0.21TB	8	0.9
Human	Transcriptome	NULL	0.12TB	13	0.5
Human	Resequencing	Targeted capture and massively parallel sequencing of human exomes JS0001	0.09TB	61	0.4
Human	Other	Even Eichler Structural Variation Sequencing	0.09TB	6	0.4
Human	Transcriptome	GSE2011B: RNA-Seq of oral squamous cell carcinomas and matched normal tis	0.06TB	6	0.3
Human	Resequencing	The diploid genome sequence of an Asian individual	0.06TB	523	0.3
Human	Transcriptome	GSE19480: Understanding mechanisms underlying human gene expression variation	0.06TB	161	0.3
			2.21TB	2474	9.8

Major projects in Open Access SRA in NCBI (Mirrored in DDBJ) @ 2010 July

			Fastq	runs	DiskShare%
		all 1BB projects	22.6TB	40972	100
All NON-human projects collapsed by study type					
tax	study_type	study_title	Fastq	runs	DiskShare%
Non	Whole Genome	ex. <i>Escherichia coli</i> K12 MG1655	4.13TB	7357	18.3
Non	Transcriptome	ex. CAGE analysis of whole adult brain and whole embryo rat transcriptome	1.13TB	2674	5.0
Non	Epigenetics	ex. DNA methylation maps of pluripotent and differentiated cells	0.49TB	1286	2.2
Non	Resequencing	ex. Population genomics of domestic and wild yeasts.	0.37TB	539	1.6
Non	Population Genomics	ex. genetic diversity of picoplankton in subtropical coastal waters as revealed by 16S rRNA sequencing	0.19TB	201	0.9
Non	Other	ex. genetic variation detected in 206 <i>Escherichia coli</i> plasmids	0.11TB	461	0.5
Non	Metagenomics	ex. Bacterial carbon processing by generalist species in the coastal ocean.	0.05TB	6488	0.2
Non	Gene Regulation Studies	ex. GATA binding protein 1 (GATA1) binding sites in mouse embryonic stem cells	0.04TB	88	0.2
Non	RNASeq	ex. Maize B73 leaf developmental gradient	0.02TB	24	0.1
Non	Synthetic Genomics	ex. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis	0.00TB	8	0.0
Non	NULL	NULL	0.00TB	28	0.0
Non	Cancer Genomics	ex. Study of the cancer genome in T lymphoma showing a recurrent chromosomal rearrangement	0.00TB	2	0.0
			6.54TB	19156	28.92