

生命科学における オープンデータの理想と現実

科学技術振興機構
バイオサイエンスデータベースセンター

八塚 茂

ORCID: 0000-0002-6891-5229

自己紹介

- 八塚 茂（やつづか しげる）
 - 学部は人文系
 - たまたまSEになる
 - たまたまバイオ系システムの部門に配属される
 - IT&バイオバブルの栄枯盛衰を経験
 - **運命的にDBCLS⇒NBDCの担当になる**
 - 生命科学系データベースアーカイブの立ち上げから参画
 - 色々あってNBDCの「中の人」になる
 - 現在の担当：生命科学系データベースアーカイブ

「バイオ系研究データ流通業」

生命科学系データベースアーカイブ

生命科学系データベースアーカイブ

<https://dbarchive.biosciencedbc.jp/>

Life Science Database Archive
LSDB Archive

~あのデータベースが、丸ごとダウンロード可能に!~
生命科学系データベース アーカイブ

[Japanese | English] [寄託者専用サイトログイン](#)

アーカイブ内を横断検索

● 相同性検索 ● 画像検索

[ホーム](#) [アーカイブの説明](#) [寄託応募要領](#) [更新履歴](#) [利用状況](#) [ヘルプ](#) [お問い合わせ](#)

いくら良質なデータベースでも、説明が十分でない、利用条件が明確でない、ダウンロードできないなどの理由で十分に利用され、引用され、相応しい評価をうける機会を逃していることがあります。

生命科学系データベースアーカイブは、国内のライフサイエンス研究者が生み出したデータセットをわが国の公共財としてまとめて長期間安定に維持保管し、データ説明(メタデータ)を統一して検索を容易にすると共に、利用許諾条件などの明示を行うことで、多くの人が容易にデータへアクセスしダウンロードを行えるようにするサービスです(詳細説明)。

データを長期にわたり保全し、データベース作成者のクレジットを明示する一方、公的機関や民間等様々なユーザが利用しやすい状況にすることで、それぞれの研究の生命科学へのいっそうの貢献を支援します。データベースの寄託を随時募集しています(寄託応募要領)。

新アーカイブ情報

2016/02/02「イネプロテオームデータベース」(農研機構 作物研究所 小松節子 上席研究員)を追加しました

2016/02/01「JSNP」(東京大学 医学研究所、国立研究開発法人 科学技術振興機構)を追加しました

2016/01/14「RED II INAHO」(農業生物資源研究所 菊池尚志 上級研究員)を追加しました

アーカイブデータベース一覧 (ヘルプ)

一覧内検索

全 111 件 (1 件から5件) 件を表示

最初へ
前へ
1
2
3
4
5
...
23
次へ
最後へ

データベース	データベース運用場所	代表者	データベースカテゴリ	生物種	要約(キーワードを太字表示)	利用許諾
 イネプロテオームデータベース ダウンロード 簡易検索 オリジナルサイト	農業生物資源研究所	小松 節子	プロテオーム	イネ	イネの各種器官や細胞内小器官を対象に2次元ゲル電気泳動を行い、そのスポットを収集したデータベース	CC表示-継承 詳細
 JSNP ダウンロード 簡易検索 オリジナルサイト	東京大学 医学研究所	-	遺伝子多型	ヒト	日本人の持つ約19万7000の多型データにアノテーション(遺伝子情報、位置情報、アミノ酸置換情報等)を付加したデータベース	CC表示-継承 詳細

<http://biosciencedbc.jp/>

4

アーカイブの特長

- データセットの内容に関する説明（メタデータ）が充実
- 粒度の異なるデータダウンロードが可能
 - データセット一括
 - 条件で絞り込み
- 基本的な検索が可能
 - 横断検索
 - 詳細検索
 - 配列検索
 - 画像検索
- 原則としてCC BY-SAで公開
- DOIを付与

Scientific Dataの推奨レポジトリに

SCIENTIFIC DATA

Publish with Scientific Data

Scientific Data is a peer-reviewed, open-access journal for descriptions of scientifically valuable datasets

Announcement
Better Science through Better Data 2017 (#scidata17)

Watch all of the talks from #scidata17 on the new Springer Nature Research Data community

Data Descriptor | 14 November 2017 | OPEN
Early meteorological records from Latin-America and the Caribbean during the 18th and 19th centuries
Fernando Domínguez-Castro, José Manuel Vaquero [...] Marcos Villacís

Data Descriptor | 07 November 2017 | OPEN
Quantification of sensitivity and resistance of breast cancer cell lines to anti-cancer drugs using GR metrics
Marc Hafner, Laura M. Heiser [...] Peter K. Sorger

<https://www.nature.com/sdata/>

FAIRsharing.org standards, databases, policies

Search all of FAIRsh: Standards Databases Policies Collections Add/Claim Content Stats Log in or Register

Life Science Database Archive

Abbreviation: LSDB Archive

General Information

If a database is inadequate in terms of its description, unclear with respect to the terms of use, or is not downloadable, it may not be fully used, cited or rightly acknowledged by the (research) communities. This is even true for databases with high-quality datasets. The Life Science Database Archive maintains and stores the datasets generated by life scientists in Japan in a long-term and stable state as national public goods. The Archive makes it easier for many people to search datasets by metadata (description of datasets) in a unified format, and to access and download the datasets with clear terms of use (see here for detailed descriptions). In addition, the Archive provides datasets in forms friendly to different types of users in public and private institutions, and thereby supports further contribution of each research to life science.

Homepage <http://dbarchive.biosciencedbc.jp/index-e.html>

Developed in Japan

Created in 2009

Taxonomic range

All

Scope and data types

Data Sharing Life Science

Collected/Recommended By

SCIENTIFIC DATA

This record is maintained by [yatsuzuka](#) [ORCID](#)

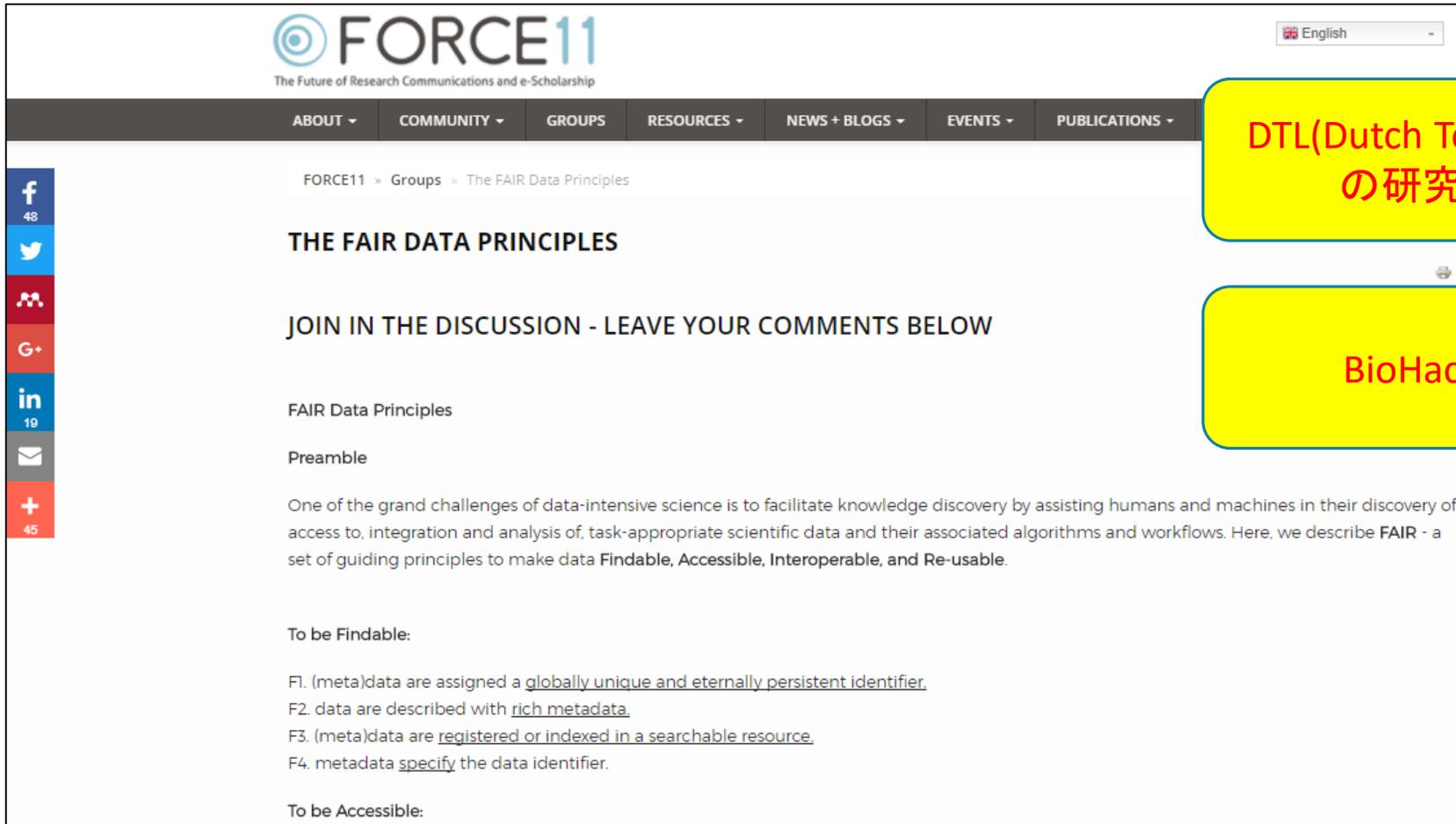
Record added: Aug. 17, 2016, 3:47 a.m.
Record updated: Aug. 31, 2016, 1:34 a.m. by [The FAIRsharing Team](#).

<https://fairsharing.org/biodbcore-000801>

理想

「フェア」であるべきなのは
スポーツだけではない

The FAIR Data Principles



The screenshot shows the FORCE11 website header with the logo and tagline 'The Future of Research Communications and e-Scholarship'. A navigation menu includes 'ABOUT', 'COMMUNITY', 'GROUPS', 'RESOURCES', 'NEWS + BLOGS', 'EVENTS', and 'PUBLICATIONS'. The main content area is titled 'THE FAIR DATA PRINCIPLES' and includes a call to action: 'JOIN IN THE DISCUSSION - LEAVE YOUR COMMENTS BELOW'. The text describes the FAIR Data Principles as a set of guiding principles to make data Findable, Accessible, Interoperable, and Re-usable. It lists four principles (F1-F4) under the heading 'To be Findable:'.

DTL(Dutch Techcentre for Life Sciences)
の研究者などを中心に策定

BioHackathon2015でも議論

<https://www.force11.org/group/fairgroup/fairprinciples>

FAIR Guiding Principles (1) Findable

To be Findable:

見つけられるために

F1. (meta)data are assigned a globally unique and eternally persistent identifier.

(メタ)データが、**グローバルに一意で永続的な識別子 (ID) を有すること。**

F2. data are described with rich metadata.

データが**メタデータによって十分に記述されていること。**

F3. (meta)data are registered or indexed in a searchable resource.

(メタ)データが**検索可能なリソースの中に、登録もしくはインデックス化されていること。**

F4. metadata specify the data identifier.

メタデータが、**データの識別子 (ID) を特定していること。**

日本語訳は、NBDC大波研究員による翻訳を一部改変

FAIR Guiding Principles (2) Accessible

To be Accessible:

アクセスできるように

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
標準化されたコミュニケーションプロトコルを使って、(メタ)データを識別子(ID)から入手できること。

A1.1 the protocol is open, free, and universally implementable.

そのプロトコルは公開されており、無料で、制約なしに実装できること。

A1.2 the protocol allows for an authentication and authorization procedure, where necessary.

そのプロトコルは必要に応じて、利用権限や承認方法を提供できること。

A2 metadata are accessible, even when the data are no longer available.

データが利用不可能となったとしても、メタデータにはアクセスできること。

FAIR Guiding Principles (3) Interoperable

To be Interoperable:
相互運用可能となるために

11. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
それが示す知識情報を表現するために、(メタ)データは、**正式で、アクセス可能で、共有されていて、広く利用されている言語を使用すること。**
12. (meta)data use vocabularies that follow FAIR principles.
(メタ)データが**FAIR原則に従う語彙**を使っていること。
13. (meta)data include qualified references to other (meta)data.
(メタ)データは、他の(メタ)データへの**適切なリファレンス情報**を含んでいること。

FAIR Guiding Principles (4) Re-usable

To be Re-usable:

再利用できるように

R1. meta(data) have a plurality of accurate and relevant attributes.

メタ(データ)が、**関係する正確な属性情報を複数持つこと。**

R1.1. (meta)data are released with a clear and accessible data usage license.

(メタ)データが、**明確でアクセス可能なデータ利用ライセンス**と共に公開されていること。

R1.2. (meta)data are associated with their provenance.

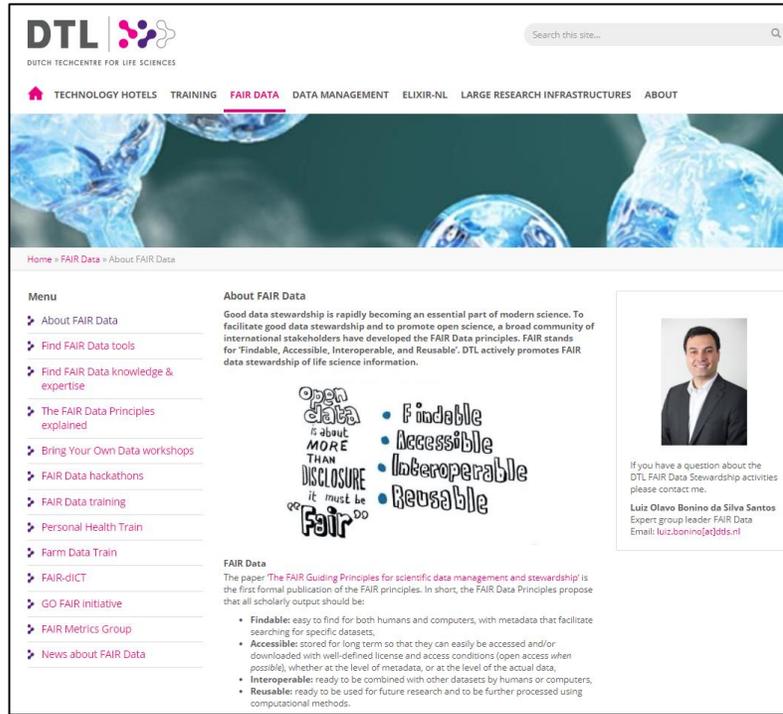
(メタ)データが、その**由来の情報**と繋がっていること。

R1.3. (meta)data meet domain-relevant community standards.

(メタ)データが、**ドメインごとのコミュニティの標準**に従っていること。

渡る世界はFAIRばかり

■ヨーロッパ



<https://www.dtls.nl/fair-data/>

旧名称: BioSharing



Integbioデータベースカタログとデータ交換

渡る世界はFAIRばかり

■アメリカ



Big Data to Knowledge

NIH Data Commons Pilot Phase Explores Using the Cloud to Access and Share *FAIR Biomedical Big Data

Program Snapshot

As biomedical tools and technologies rapidly improve, researchers are producing and analyzing a rapidly increasing amount of complex biological data called “big data.” The Big Data to Knowledge (BD2K) program, was launched in 2014 to facilitate broad use of biomedical big data, develop and disseminate analysis methods and software, enhance training relevant for large-scale data analysis, and establish centers of excellence for biomedical big data. The BD2K Program also supported initial efforts toward making data sets “FAIR” Findable, Accessible, Interoperable, and Reusable. Learn more about the FAIR principles.

<https://commonfund.nih.gov/bd2k>

2017-2020の予算総額: 約1億ドル

現実

データは右から左へ受け流せない

データを作成する・利用する

■データを作成する

- 実験機器などから産生されたデータを、作成者が自分の観点でまとめる

■データを利用する

- データの作成者が利用する
- データの作成者以外の人を利用する
 - 観点は作成者と必ずしも同じではない

再利用

メタデータ

■ 定義

- メタデータとは、データについてのデータ。あるデータそのものではなく、そのデータを表す属性や関連する情報を記述したデータのこと。
(「IT用語辞典 e-Words」より)
- データ (セット) の意味
- データセットの設計・構成図

データ作成者と利用者をつなぐ架け橋

アーカイブ作成という名の戦場



国立国会図書館デジタルコレクション 『平治物語』
doi:10.11501/1287476 より

アーカイブ作成：3つのプロセス



解体：データ流通に一役買いたい

- アーカイブに寄託されるデータベースの多くは、専用のWebサイト（オリジナルサイト）で公開されている（いた）
 - 作成者の観点で組み立てられている
- アプリケーションのためだけの項目（フラグなど）が多く含まれている
- データの利用者（再利用者）の多くは「建築家」
 - データベースを解体して、わかりやすい「素材」に再生する必要がある

お城のようなもの

ネジや釘のようなもの



解体前のデータのイメージ

main

ID	名称
P1	タンパク質A
P2	タンパク質B

sequence

ID	配列
P1	111-111.fsa
P2	human.fsa

配列そのものではなく、ファイル名のような

このフラグの意味は？

reference

文献ID	文献名	フラグ
123456	文献1	1
234567	文献2	0

mainのIDと同じなのだろうか？

このフラグの意味は？

ex_cond

ID	実験条件	フラグ	文献ID
P001	条件1	00	123456
P002	条件2	01	234567

ex_res

ID	実験結果	文献ID
P001	結果1	123456
P002	結果2	234567

mainのIDと同じなのだろうか？

「考古学」的考察： データベースの在りし日を偲んで

- メタデータ（＝データの意味、データセットの設計・構成図）が失われている or そもそも存在しないことが多い
- 作成者もいなくなっている or 忘れていることが多い
- データの意味やデータ間の関係を推定
- 関連論文や資料などを探す
- 時にはWebアーカイブも参照



「考古学」的考察の結果

main

ID	名称
P1	タンパク質A
P2	タンパク質B

sequence

ID	配列
P1	111-111.fsa
P2	human.fsa

配列ファイル名
だった

0:独自のID
1:Pubmed ID

reference

文献ID	文献名	フラグ
123456	文献1	1
234567	文献2	0

mainのIDが0詰
めされていた

00:赤で表示
01:緑で表示
10:青で表示
11:黒で表示

ex_cond

ID	実験条件	フラグ	文献ID
P001	条件1	00	123456
P002	条件2	01	234567

ex_res

ID	実験結果	文献ID
P001	結果1	123456
P002	結果2	234567

mainのIDが0詰
めされていた

そして再生

Sequenceテーブルと統合

main

ID	名称	配列
P1	タンパク質A	MPLGLIGEKVGMTRVLLK
P2	タンパク質B	MYALLVISLYLQRFYNLSIIPQL

ファイル中の配列をテーブルに書き出し

フラグを廃止

reference

文献ID	文献名	URL
123456	文献1	https://www.ncbi.nlm.nih.gov/pubmed/123456
234567	文献2	-

Pubmedに含まれるものにはURLを記載

フラグを廃止

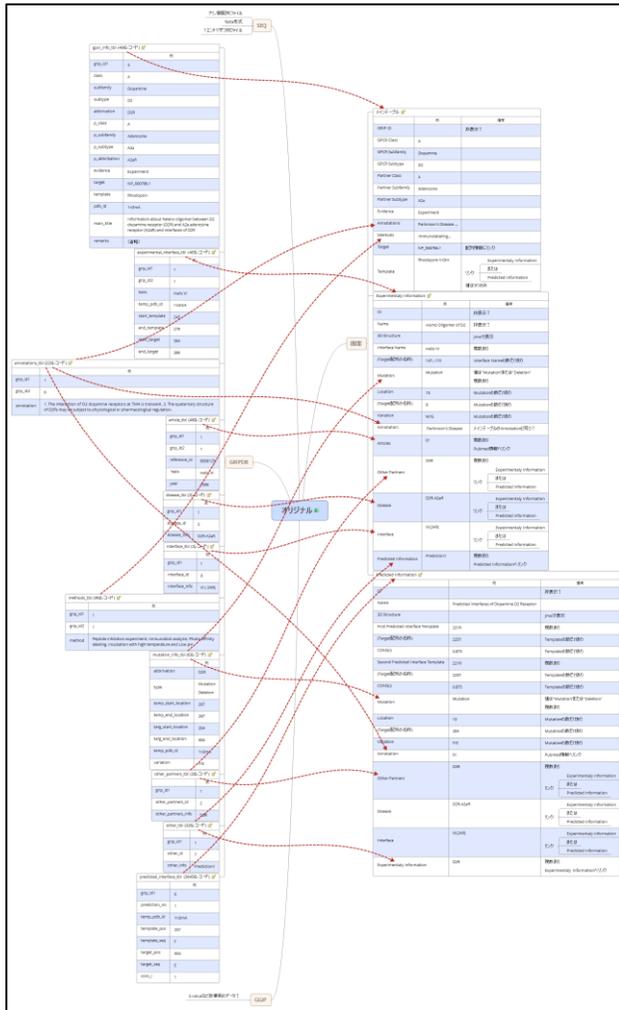
experiment

ID	実験条件	実験結果	文献ID
P1	条件1	結果1	123456
P2	条件2	結果2	234567

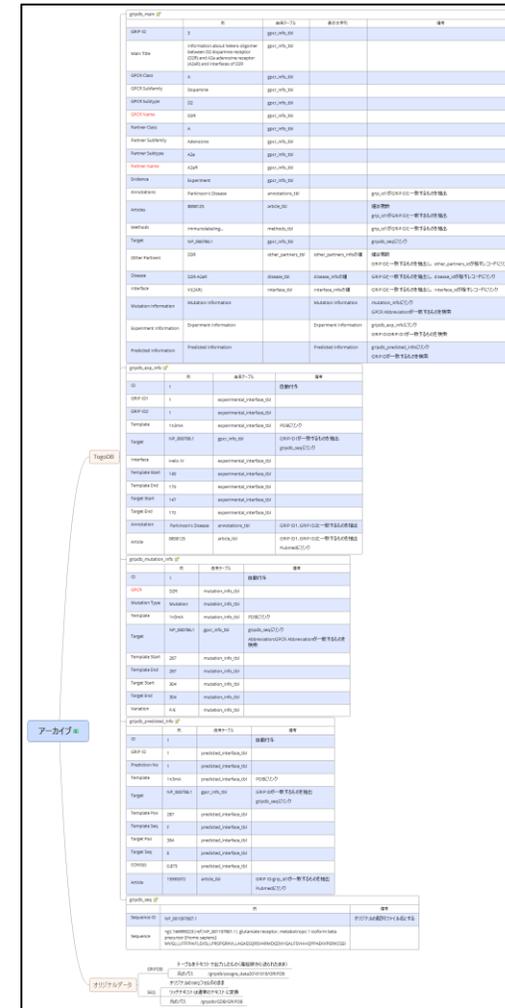
2つの"ex_..."テーブルを統合

IDをmainと揃える

実際の解体～再生の例



メタデータのない
14テーブル



メタデータ付き
5テーブル



もっと大きな問題

検索で「発掘」されるものが全て？



秘蔵？うっかり？



データバンクに登録して論文に記載したデータでも非公開のままのものがある

データバンク未登録、論文未記載のままの非公開データはどれくらいあるのだろうか？

<https://wideopen.bio/>

お願い

データ作成者の方へ

- 作成したデータが未公開であれば
 - 各種レポジトリへの登録をご検討ください
 - **生命科学系データベースアーカイブ**では、
完成形（いわゆる「データベース」）かどうかにかかわらず寄託を受け付けています
- 独自サイトで既にデータを公開済であれば
 - ダウンロード用データもぜひご用意ください
 - できればメタデータ付きで
 - **生命科学系データベースアーカイブ**からも公開できます
お気軽にご相談ください

ありがとうございました

- NBDC

<https://biosciencedbc.jp/>

- 生命科学系データベースアーカイブ

<https://dbarchive.biosciencedbc.jp/>

- アーカイブのご相談・お問い合わせ

dbarchive@biosciencedbc.jp

- 展示ブース

国際展示場 2号館 1階 BioDBコーナー

ぜひお越しください