

DOI登録事例紹介 ～生命科学データの場～

国立研究開発法人科学技術振興機構

バイオサイエンスデータベースセンター(NBDC)

八塚 茂



生命科学系データベースアーカイブ

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

個々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

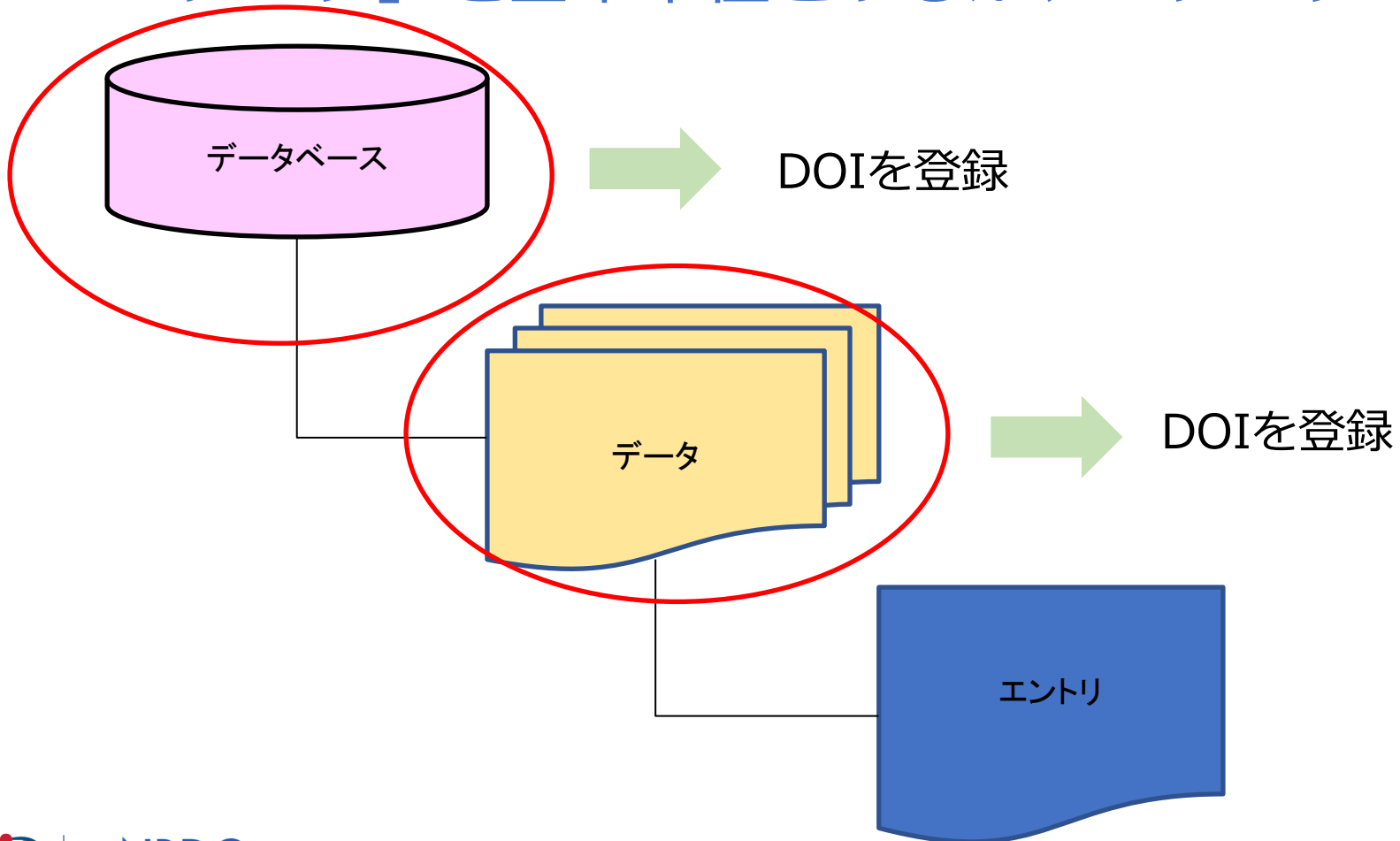
1. 長期間・安定的にまとめて維持・保管する
2. 十分な説明（メタデータ）をつけて発見しやすくする(Findability)
3. 容易なアクセスとダウンロードを可能にする(Accessibility)
4. 汎用的なフォーマット(TXT, CSV, JPGなど)で提供し、相互運用性を確保する(Interoperability)
5. 利用許諾条件を明示することで再利用を促進する(Reusability)

The screenshot shows the Bioscience Data Archive website interface. At the top, there are navigation links for Research, Services, Contact, and About. A search bar is present with the text "アーカイブ内を横断検索". Below the navigation, there are buttons for "ホーム", "アーカイブの説明", "寄託応募要領", "更新履歴", "データ一覧", "利用状況", "ヘルプ", and "お問い合わせ". The main content area features a diagram illustrating the archive's workflow: "データ生成/収集" leads to "データベース構築", which then leads to "データベース公開". A sidebar on the right contains "新アーカイブ情報" with updates from 2021/11/09 and 2021/08/31. Below this is a section for "当アーカイブの使い方(動画)" with a video player. At the bottom, there is a "アーカイブデータベース一覧(ヘルプ)" section with a search bar and a table of databases.

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
NBDC NikkajIRDF ダウンロード オリジナルサイト	科学技術振興機構	-	化合物	-	日本化学物質辞書(日化辞)のデータを化合物情報のRDF形式で標準となっている オントロジー を用いてRDF化したデータベース	CC表示-継承 詳細
メタゲノム解析用リファレンス16S RNA配列データセット ダウンロード	広島大学 大学院統合生命科学研究科 グノム情報科学研究室	坊屋 秀雅	RNA配列データベース	-	Kraken2によるメタゲノムデータ解析で使える16S RNA配列データセット	CC表示 詳細
iPOSTデータベース ダウンロード 簡易検索 オリジナルサイト	iPOSTチーム	石濱 泰	プロテオミクス	-	iPOSTデータベースは、iPOSTリポジトリに登録されたプロテオームデータに対して統一された基準で再解析した結果を収録したデータベースです	CC0 詳細
LigandBox ダウンロード 簡易検索 オリジナルサイト	次世代天然物化学技術研究組合	福西 秋文	構造データベース-低分子	-	計算機によるバーチャルスクリーニング用の低分子化合物のデータベース	CC表示-継承 詳細

アーカイブの構成と DOIを登録する単位

- 「データ」を基本単位とするが、「データベース」も登録する



データベースメタデータ (ランディングページ)

NBDC NikkajiRDF

このデータベースについて

- データベースの説明
- ダウンロード
- 利用許諾
- このデータベースの更新履歴

データベースの説明

データベース全般			
名称	NBDC NikkajiRDF 		
名称の読み方	エヌピーディーシー ニッカジャーナルディーエフ		
別名	NBDC版日化辞RDF		
DOI	10.18908/lsdba.nbdc01530-02-000.V009		
バージョン	V9	10.18908/lsdba.nbdc01530-02-000.V009	更新履歴
	V8	10.18908/lsdba.nbdc01530-02-000.V008	更新履歴
	V7	10.18908/lsdba.nbdc01530-02-000.V007	更新履歴
	V6	10.18908/lsdba.nbdc01530-02-000.V006	更新履歴
	V5	10.18908/lsdba.nbdc01530-02-000.V005	更新履歴
	V4	10.18908/lsdba.nbdc01530-02-000.V004	更新履歴
	V3	10.18908/lsdba.nbdc01530-02-000.V003	更新履歴
	V2	10.18908/lsdba.nbdc01530-02-000.V002	更新履歴
	V1	10.18908/lsdba.nbdc01530-02-000.V001	-
作成者	作成者氏名: 国立研究開発法人 科学技術振興機構 作成者英名: JAPAN SCIENCE AND TECHNOLOGY AGENCY (JST) 作成者所属:		
連絡先	Email: support@biosciencedbc.jp		
データベース分類	化合物データベース		
データベースの説明	日本最大級の化合物データベースである日本化学物質辞書（日化辞）のRDF (Resource Description Framework)データです。		
データベースの特長・有用性・活用方法	NBDC版日化辞RDFは、日本化学物質辞書（日化辞）をChemical Information Ontology (CHEMINF) やSemanticScience Integrated Ontology (SIO) などの化合物情報のRDF記述で標準となっているオントロジーを用いてRDF化したデータです。オリジナルの日化辞では収録していないCanonical SMILESなどの情報を持っています。ユーザーはこのRDFデータをダウンロードし、各自の環境のトリプルストアに格納することで、RDFの問い合わせ言語であるSPARQLを使って検索が可能になります。日化辞を対象としたSPARQL検索は、J-GLOBAL knowledge (https://stirdf.jglobal.jst.go.jp/) でも実行可能です。		
データベースの利用許諾	CC 表示-継承 詳細		
学術的背景・プロジェクト	-		
論文等	文献名: 日化辞RDFデータの公開と化合物情報の統合 著者名: 木村 孝宏, 榎田 達矢 雑誌名/掲載年月/号: 情報管理 58(3) 204-212 (2015/06/01) http://doi.org/10.1241/johokanri.58.204 外部リンク: 		

データベースのオリジナルサイト情報

データベース運用場所	国立研究開発法人 科学技術振興機構
オリジナルサイト	http://doi.org/10.18908/lsdba.nbdc01530-02-000
運用開始年月日	2005/03/31
最終更新年月日	2018/05/18
統括サイト	-
データの一括ダウンロードサイト	-
参照先データベース	<ul style="list-style-type: none"> ▪ 化学物質安全情報提供システム Kis-net ▪ 既存化学物質毒性データベース JECDB ▪ 有機化合物のスペクトルデータベース SDDBS ▪ 高分子データベース PoLyInfo
一覧表示	無し
クエリ検索	有り
Webサービス	無し
WebサービスURL	-
ユーザ登録	無し

データメタデータ (ランディングページ)

- データベース「Nikkaji RDF」には、データが8つある。その1つが「日化辞の本体のRDFデータ」である。

日化辞の本体のRDFデータ

データ説明			
データ名	日化辞の本体のRDFデータ		
DOI	10.18908/lsdba.nbdc01530-02-001.V009		
バージョン	V9	10.18908/lsdba.nbdc01530-02-001.V009	更新履歴
	V8	10.18908/lsdba.nbdc01530-02-001.V008	更新履歴
	V7	10.18908/lsdba.nbdc01530-02-001.V007	更新なし
	V6	10.18908/lsdba.nbdc01530-02-001.V006	更新なし
	V5	10.18908/lsdba.nbdc01530-02-001.V005	更新履歴
	V4	10.18908/lsdba.nbdc01530-02-001.V004	更新履歴
	V3	10.18908/lsdba.nbdc01530-02-001.V003	更新履歴
	V2	10.18908/lsdba.nbdc01530-02-001.V002	更新履歴
	V1	10.18908/lsdba.nbdc01530-02-001.V001	-
データ内容の説明	330万件以上の化学物質の日化辞番号、分子式、分子量、和名、英名など基本情報のRDFデータです。		
データファイル	データファイル名:	NBDC_NikkajiRDF_main.tar.gz	
	データのURL:	https://dbarchive.biosciencedbc.jp/data/nikkaji/LATEST/NBDC_NikkajiRDF_main.tar.gz	
	ファイルサイズ:	1.1 GB	
簡易検索URL	-		
データ取得方法	rdf, rdfs, owl, foaf, dc, ndlなど広く利用されているオントロジーや語彙を利用してRDF化しています。		
解析方法	-		
データ件数	89,929,862 トリプル		
データ詳細	-		

DOI登録のローカルルール

> ランディングページ

- > 原則として**英語ページ**をDOI登録する
 - > グローバル対応のため
- > 英語ページがない場合は、日本語ページをDOI登録する

> DOIのsuffix

- > **lsdba.nbdcXXXXX(-XX)-XXX.VXXX**
(Xは数字、カタログID (付番) +データ番号3桁+バージョン番号3桁)
 - > 「カタログID」は、integbioデータベースカタログ(<https://integbio.jp/dbcatalog/?lang=ja>)のID
- > 「データベース」のDOIは、データ番号を必ず「000」にする
- > **バージョン番号なし**のDOIも登録する
 - > バージョン番号を知らないユーザのため

> バージョン管理

- > 「**データ**」が更新された場合にのみバージョンを更新する
 - > メタデータのみ更新ではバージョンは更新しない
- > 「データベース」の中で「**データ**」が**1つでも更新**されれば、全「データ」のバージョンを更新する

日化辞の本体のRDFデータ

データ説明			
 データ名	日化辞の本体のRDFデータ		
 DOI	10.18908/lbdba.nbdc01530-02-001.V009		
 バージョン	V9	10.18908/lbdba.nbdc01530-02-001.V009	更新履歴
	V8	10.18908/lbdba.nbdc01530-02-001.V008	更新履歴
	V7	10.18908/lbdba.nbdc01530-02-001.V007	更新なし
	V6	10.18908/lbdba.nbdc01530-02-001.V006	更新なし
	V5	10.18908/lbdba.nbdc01530-02-001.V005	更新履歴
	V4	10.18908/lbdba.nbdc01530-02-001.V004	更新履歴
	V3	10.18908/lbdba.nbdc01530-02-001.V003	更新履歴
	V2	10.18908/lbdba.nbdc01530-02-001.V002	更新履歴
	V1	10.18908/lbdba.nbdc01530-02-001.V001	-

当該「データ」の更新はないが、他の「データ」が更新されたため、バージョン番号のみ更新している

少し困っていること

- ▶ 「データ」を更新する場合は
 - ▶ 新バージョンのDOI登録が必要
 - ▶ 旧バージョンのメタデータをWebインタフェースでコピーして利用したい
 - ▶ バージョン番号「なし」のDOIメタデータに更新日付を追加
 - ▶ 既登録のメタデータをWebインタフェースで編集したい
- ▶ メタデータ一括登録の場合、DOI登録に失敗してもどこが悪いのかがすぐにわからない