

Museomics : 遺伝子データと博物館データの融合に向けて

仲里 猛留

NAKAZATO, Takeru

@chalkless



情報・システム研究機構 データサイエンス共同利用基盤施設
ライフサイエンス統合データベースセンター

Database Center for Life Science (DBCLS),
Joint Support-Center for Data Science Research, Research Organization of Information and Systems (ROIS)



2021/2/5
@オンライン

バイオインフォマティクスのデータ

GenBank (遺伝子データ)

```

LOCUS       AB055465                1692 bp    mRNA    linear   VRT 22-OCT-2004
DEFINITION  Tribolodon hakonensis mRNA for aquaporin 3, complete cds.
ACCESSION   AB055465
VERSION     AB055465.1
KEYWORDS    .
SOURCE      Tribolodon hakonensis (big-scaled redfin)
  ORGANISM  Tribolodon hakonensis
            Eukaryota; Craniata; Craniata; Vertebrata; Euteleostomi;
            Actinopteri; Actinopteri; Teleostei; Ostariophysi;
            Cypriniformes; Leuciscidae; Pseudaspininae; Tribolodon.
REFERENCE   1
  AUTHORS   Hirata,T., Nakazato,T., Nakazato,T., ...
  TITLE     Mechanism of adaptation of a fish living in a pH 3.5 lake
  JOURNAL   Am J Physiol Regul Integr Comp Physiol 284 (5), R1199-R1212 (2003)
  PUBMED   12531781
REFERENCE   2 (bases 1 to 1692)
  ..
FEATURES             Location/Qualifiers
     source            1..1692
                       /organism="Tribolodon hakonensis"
                       /mol_type="mRNA"
     CDS               72..9
                       /db_xref="db_xref:GeneID:151710"
                       /codon_start=1
                       /product="aquaporin 3"
                       /protein_id="BAB83082.1"
                       /translation="MGWQKAMLDKLAQTFRIRNKLLRQGLAECLGTLILVMFGCG...
                       ...NKDMEESLKLNDVTGKN"
ORIGIN
1  ggaattcgcg gccgcgtcga cacagttttt tttcacacag tcagaggaag acatccacac
//

```

ID/タイトル

生物種情報

Reference

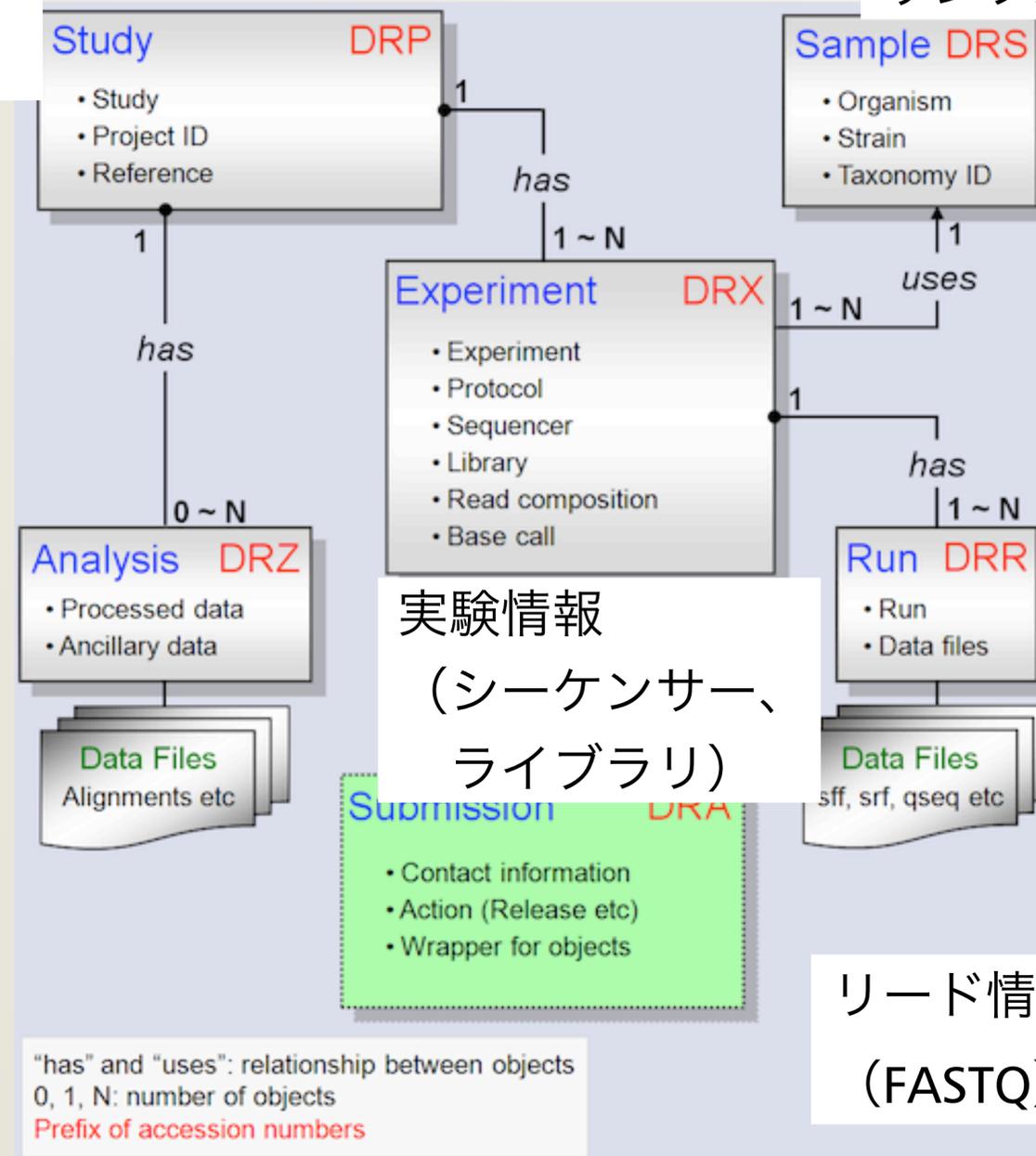
アミノ酸情報 (翻訳)

塩基配列 (FASTA)

SRA (NGS実験データ)

プロジェクト情報

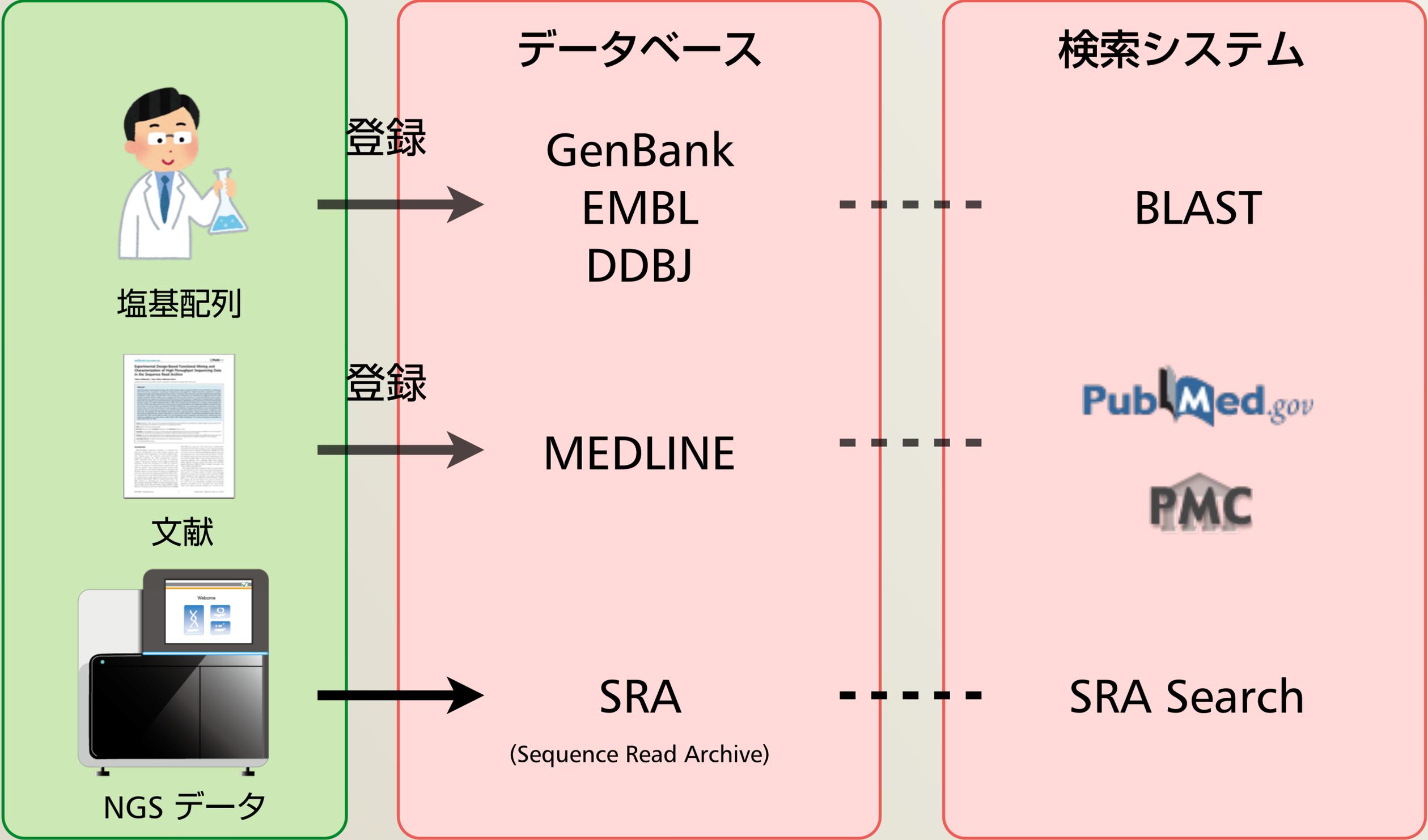
サンプル情報



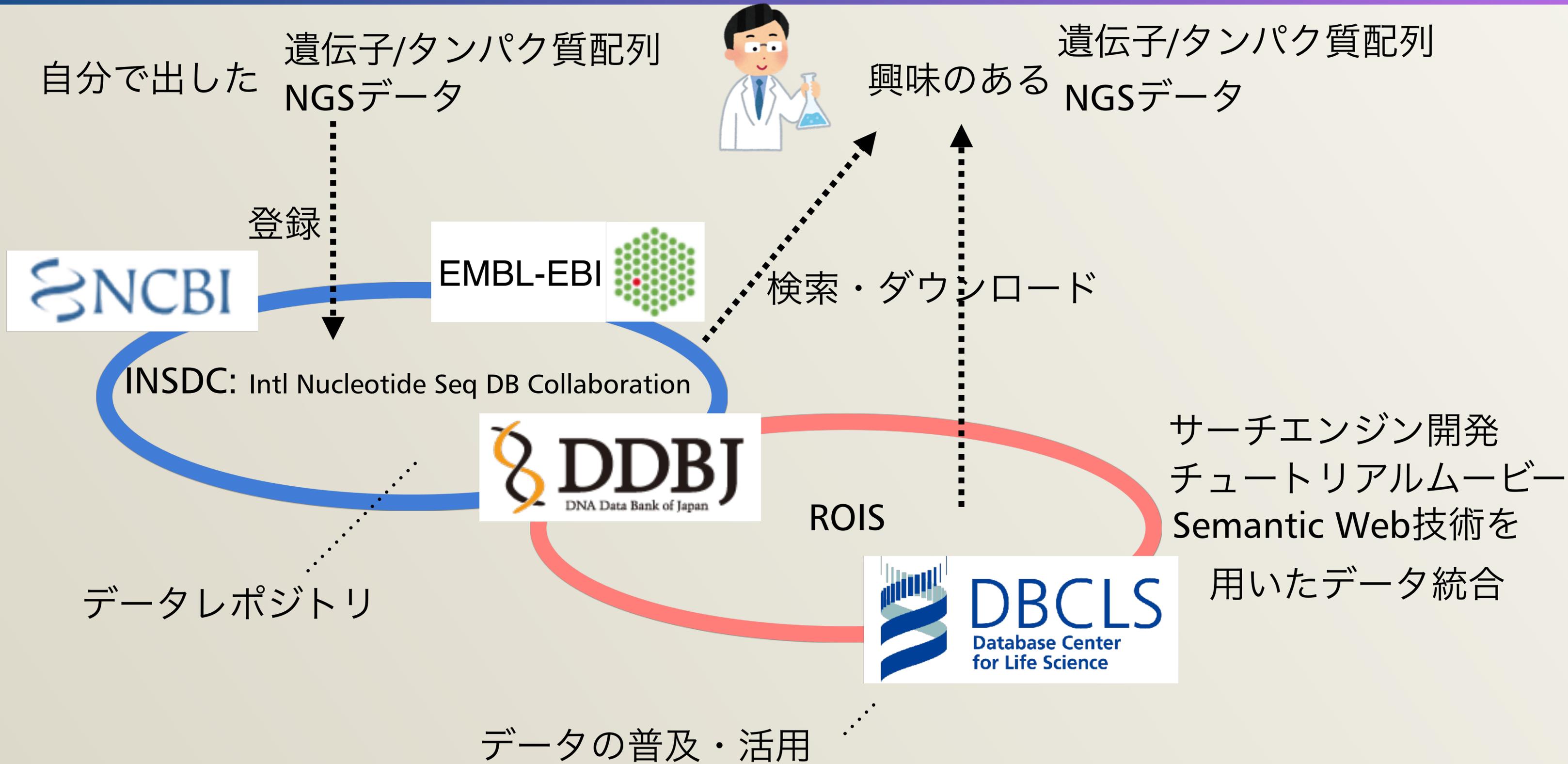
実験情報 (シーケンサー、ライブラリ)

リード情報 (FASTQ)

バイオインフォマティクスとデータベース



バイオインフォマティクスのデータベースの関係



バイオインフォマティクスのデータ量

Literature	
Bookshelf	853,965
MeSH	348,145
NLM Catalog	1,623,926
PubMed	32,102,481
PubMed Central	6,787,319

Genes	
Gene	30,620,745
GEO DataSets	4,369,122
GEO Profiles	128,414,055
HomoloGene	141,268
PopSet	357,567

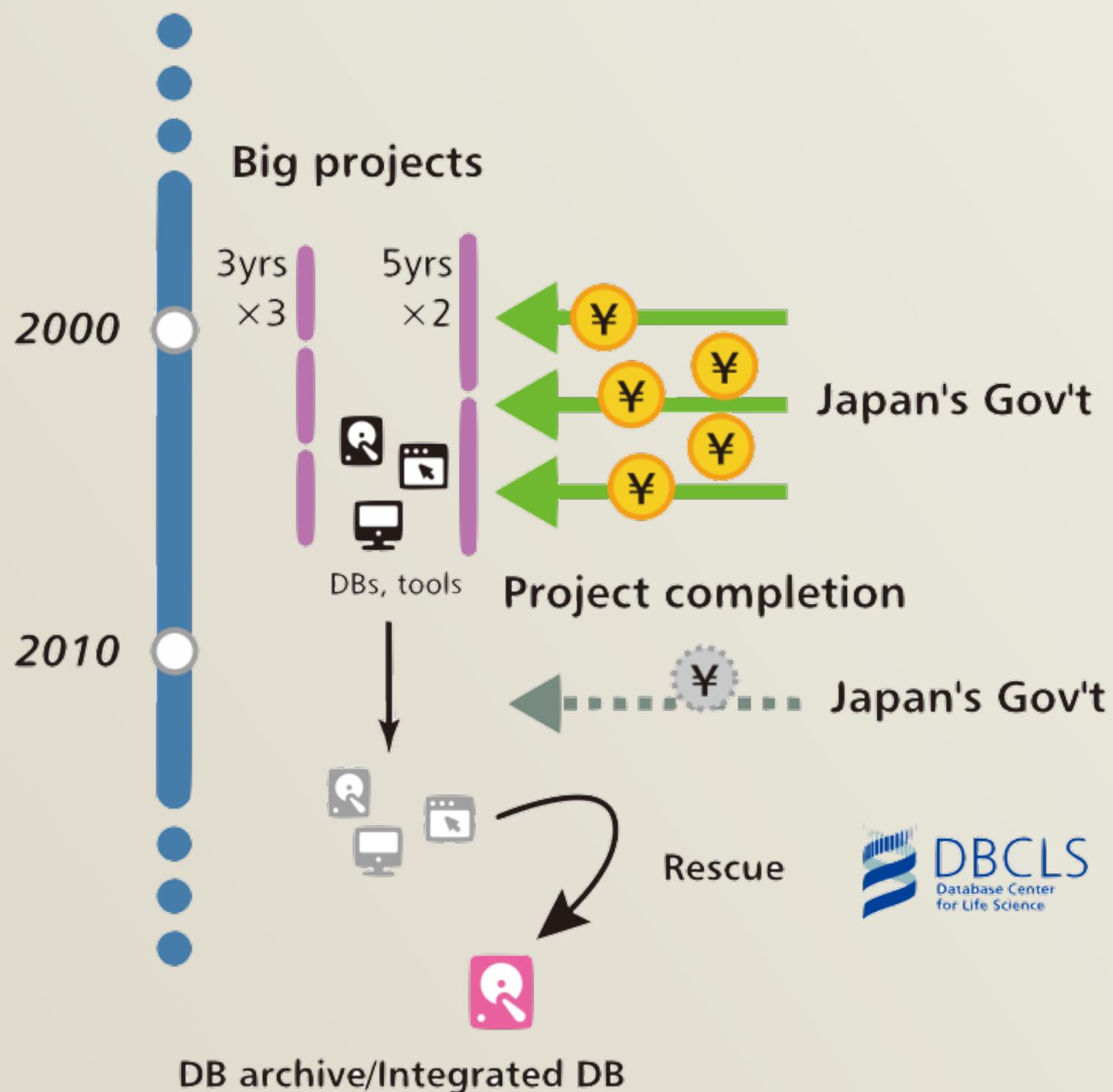
Proteins	
Conserved Domains	59,951
Identical Protein Groups	364,929,571
Protein	903,151,951
Protein Clusters	1,137,329
Protein Family Models	198,658
Structure	172,575

Genomes	
Assembly	931,351
BioCollections	8,445
BioProject	490,515
BioSample	16,312,648
Genome	59,064
Nucleotide	444,341,456
SRA	13,135,006
Taxonomy	0

Clinical	
ClinicalTrials.gov	0
ClinVar	862,168
dbGaP	1,397
dbSNP	720,643,623
dbVar	6,102,926
GTR	77,035
MedGen	322,319
OMIM	27,029

PubChem	
BioAssays	0
Compounds	0
Pathways	0
Substances	0

ライフサイエンスデータ統合化への道

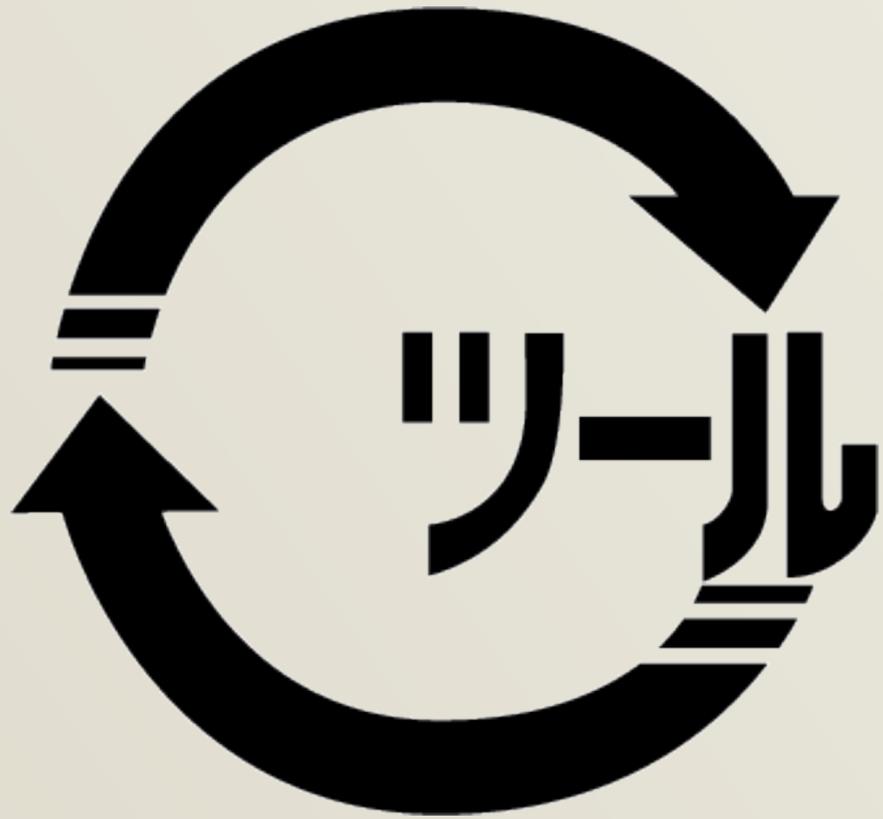


← ミレニアムプロジェクト
(研究費バカスカ)

← プロジェクト終了期
(研究費枯渇)

← パソコン壊れ、データも昇天

← 死に行くデータを永代供養！



bioinformatics と biodiversity informatics



生息環境
適応

捕食-被捕食
分布

疾患
表現系

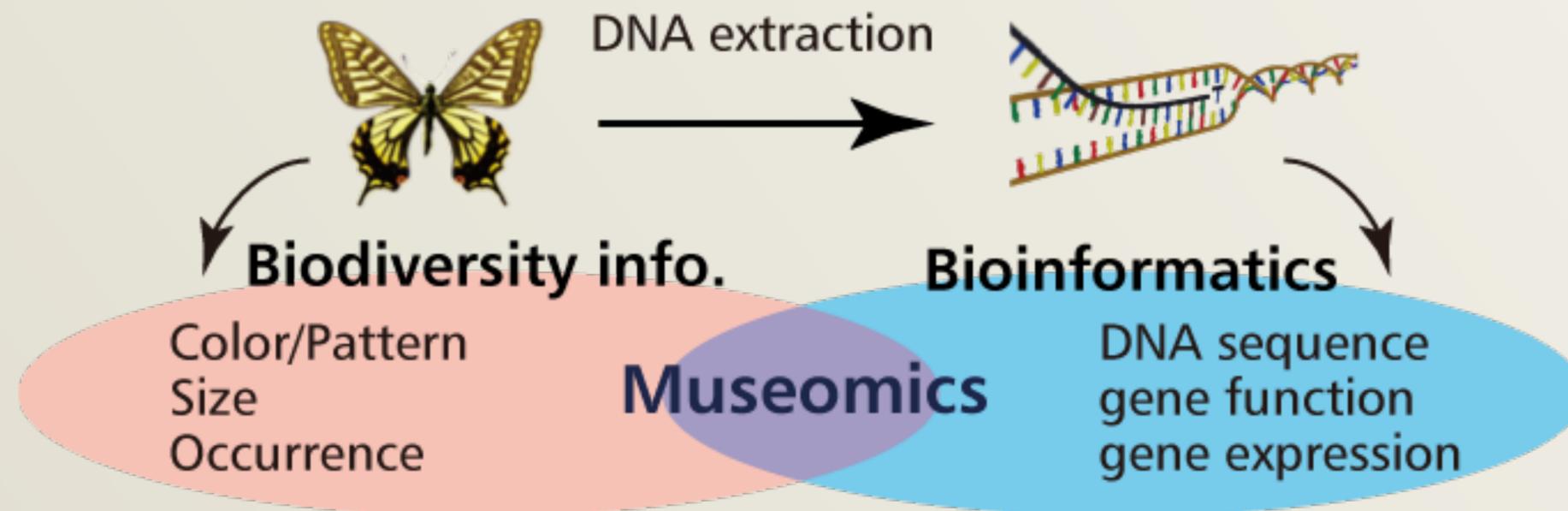
pathway
遺伝子発現

遺伝子
タンパク質

bioinformatics

biodiversity informatics
(生物多様性情報学)

Museomics



Museomics : 標本情報 (モノ) とシーケンス情報 (コト) の融合をめざす
(端的には、博物館標本をシーケンスすることで新たな分野を拓く)

2017年度 : 第8回ミュゼオミクス研究会 (生物多様性情報学研究会)

2018年度 : 第9回ミュゼオミクス研究会 (生物多様性情報学研究会)

2019年度 : Museomics : ライフサイエンスと博物館のデータを統合する
菌類博物館標本のゲノムシーケンシング解析

DBCLS : 仲里猛留

東海大 : 松前ひろみ

奈良先端大 : 武藤愛

東大 : 小寺正明

国立科学博物館 : 細矢剛

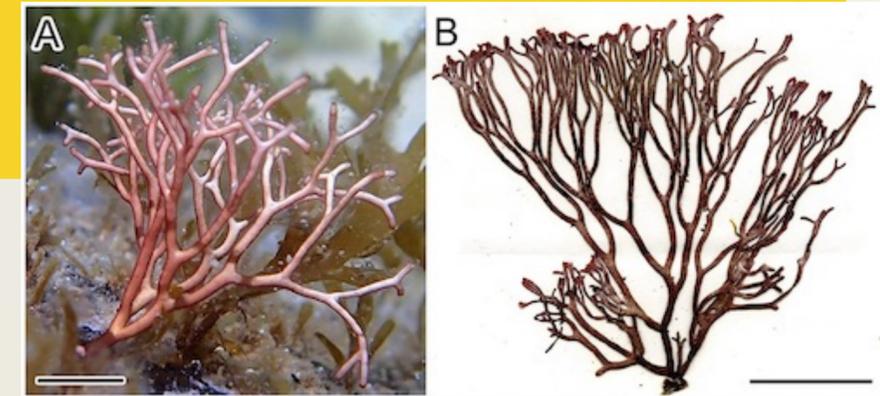
神保宇嗣

Museomics研究の例

Next-Generation Sequencing of an 88-Year-Old Specimen of the Poorly Known Species *Liagora japonica* (Nemaliales, Rhodophyta) Supports the Recognition of *Otohimella* gen. nov.

PLoS One. 2016 Jul 7;11(7):e0158944. doi: 10.1371/journal.pone.0158944.

採集した紅藻は絶滅種ヨゴレコナハダか？ 採集サンプルと標本をNGS



Evaluating the Phylogenetic Status of the Extinct Japanese Otter on the Basis of Mitochondrial Genome Analysis.

PLoS One. 2016 Mar 3;11(3):e0149341. doi: 10.1371/journal.pone.0149341.

ニホンカワウソの標本をNGSして大陸のものと比較。日本固有の種か亜種だった。

A partial nuclear genome of the Jomons who lived 3000 years ago in Fukushima, Japan.

J Hum Genet. 2016 Sep 1. doi: 10.1038/jhg.2016.110.

いわゆる縄文人ゲノム

生物多様性情報のデータとデータベース・その1

GBIF

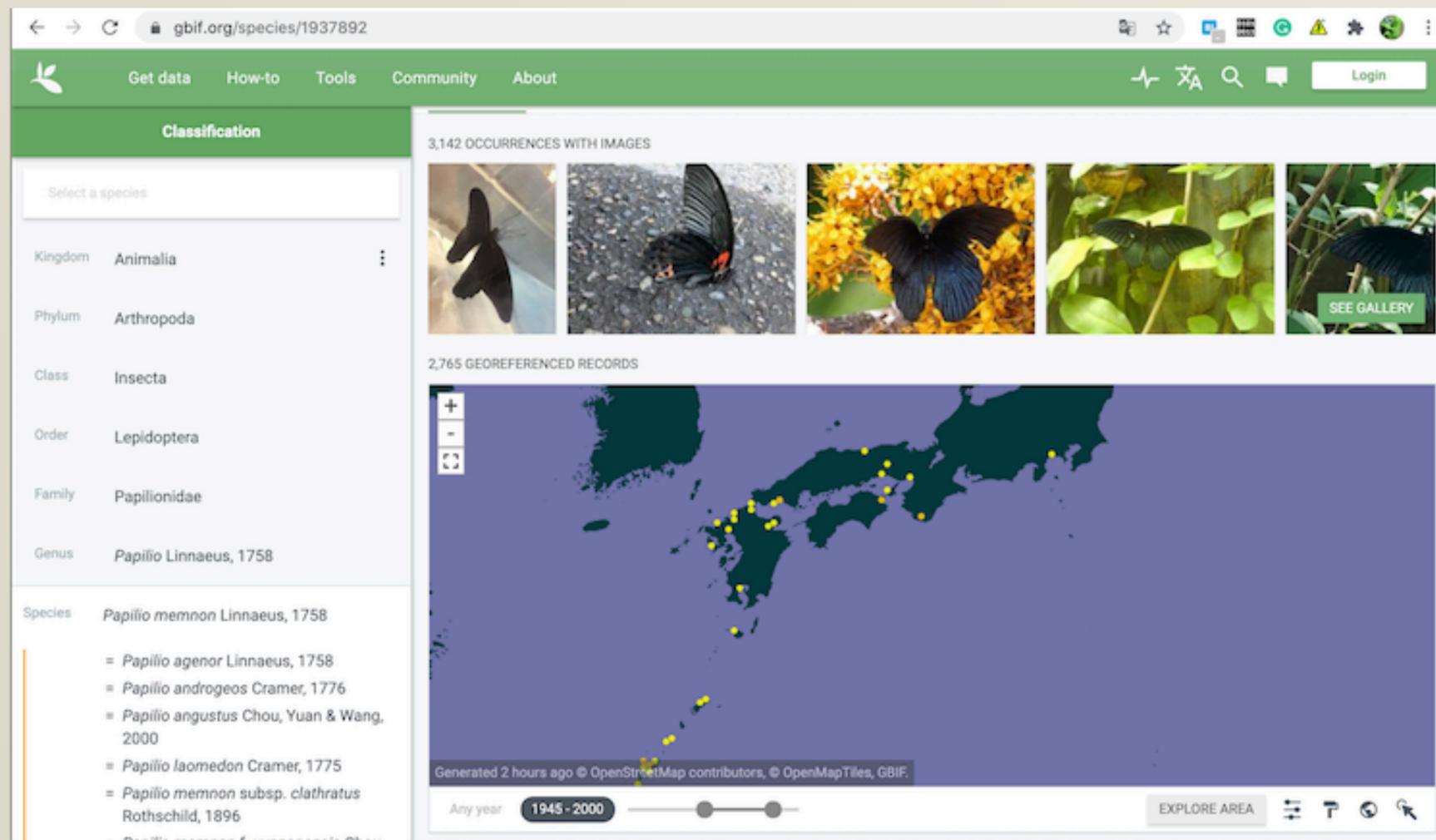
Global Biodiversity Information Facility

<https://www.gbif.org/>

occurrence情報（目撃情報・標本情報）

観測地点 [緯度、経度、高度、深度、地名]

観測者情報、同定情報 [種名、属名、...]



The screenshot shows the GBIF website interface for the species *Papilio memnon*. On the left, there is a classification sidebar with the following details: Kingdom: Animalia, Phylum: Arthropoda, Class: Insecta, Order: Lepidoptera, Family: Papilionidae, Genus: *Papilio* Linnaeus, 1758, and Species: *Papilio memnon* Linnaeus, 1758. Below the species name, a list of related species is provided, including *Papilio agenor*, *Papilio androgeos*, *Papilio angustus*, *Papilio laomedon*, *Papilio memnon* subsp. *clathratus*, and *Papilio memnon* f. *yunnanensis*. The main content area features a gallery of 3,142 occurrence images with a 'SEE GALLERY' button, and a map of 2,765 georeferenced records. The map shows a distribution of yellow dots across Japan, with a concentration in the southern regions. A time slider at the bottom of the map is set to 'Any year' with a range from 1945 to 2000, and an 'EXPLORE AREA' button is visible.

16.5億件のオカレンス

「年を追うごとに分布が北上」
という解析も可能

生物多様性情報のデータとデータベース・その2

BHL

Biodiversity Heritage Library

<https://www.biodiversitylibrary.org/>

主に昔の書籍、図鑑、図版をスキャンして公開



ファーブル昆虫記（第1巻）・1923年

当該生物種の登場する文献

BHL Biodiversity Heritage Library

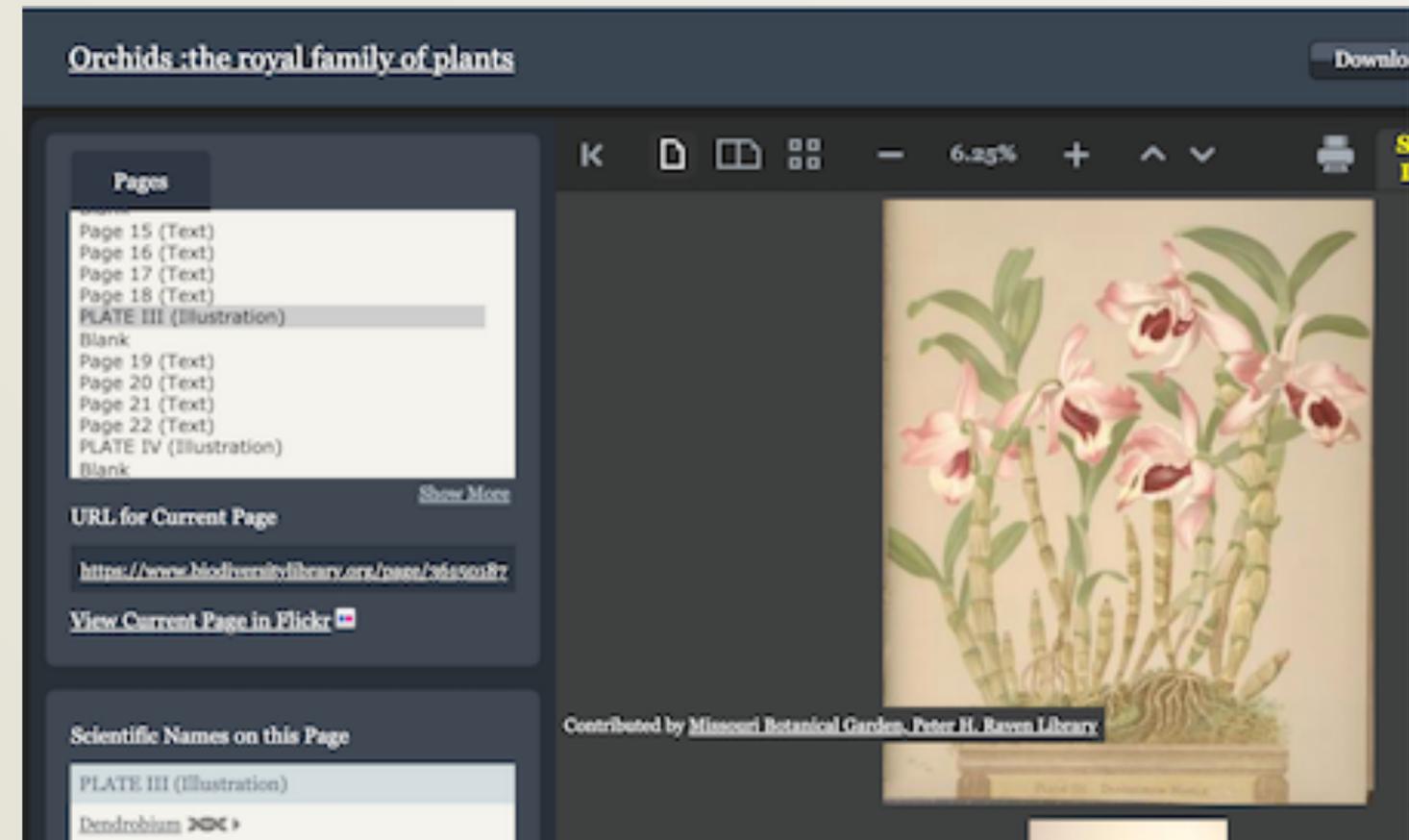
Browse by: Title Author Date Collection Contributor

Full-text Catalog ADVANCED SEARCH

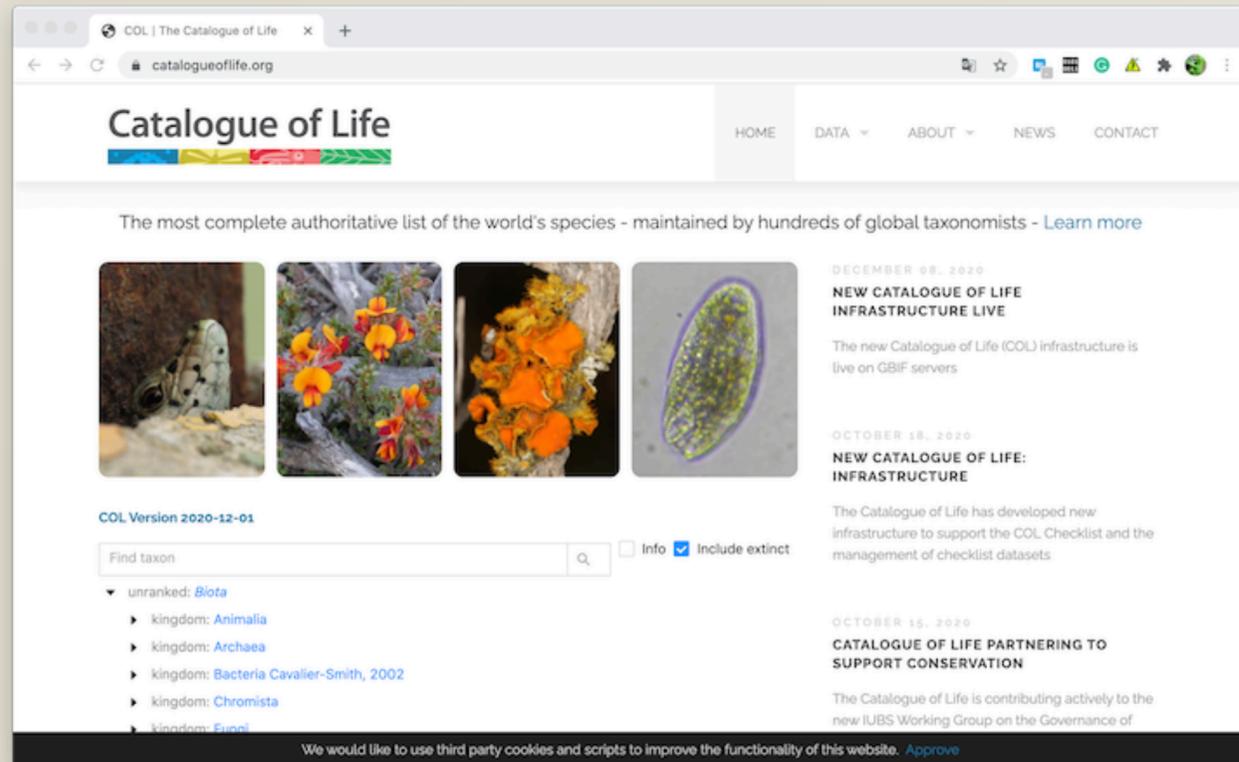
Search the catalog and full-text

Bibliography for Copris by Page

Title #	Authors	Volume	Date	Page #	View
[Two papers on coleoptera].	Kirby, William,	(go to volume)	1818	Page 396	
[Two papers on coleoptera].	Kirby, William,	(go to volume)	1818	Page 397	
1986 IUCN red list of threatened animals.	IUCN Conservation Monitoring Centre. Global Environmental Monitoring System.	1986	1986	Page 99	
1988 IUCN red list of threatened animals.		1988	1988	Page 145	
1990 IUCN red list of threatened animals.		1990	1990	Page 182	
1994 IUCN red list of threatened animals.		1994	1994	Page 251	
1994 IUCN red list of threatened animals.		1994	1994	Page 257	
Abbildungen und Beschreibungen merkwürdiger Insekten.	Wolf, Johann,	text	1818	Page 79	



生物多様性情報のデータとデータベース・その3



CoL

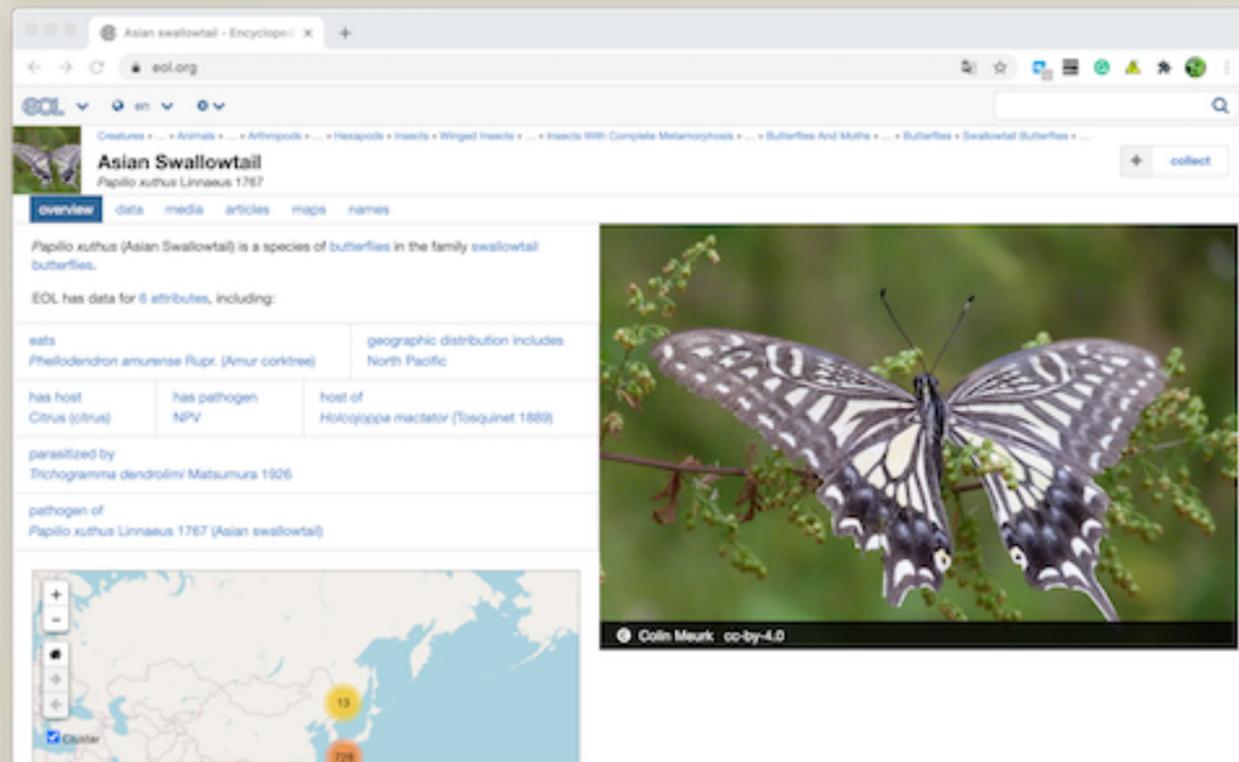
Catalogue of Life

<https://www.catalogueoflife.org/>

生物種名のデータベース

NCBI taxonomyとは体系が違う！

というかNCBI taxonomyが謎体系な時も



EOL

Encyclopedia of Life

<https://eol.org/>

各生物のさまざまな情報を集めたデータベース

種名、別名、生息地、写真、

"co-occurs with", "eats", "has pathogen", ...

生物多様性情報の標準仕様： Darwin Core (DwC)

<u>startDayOfYear</u>	イベント開始日。1は1月1日、365は12月31日を示す。	"1", "365"
<u>endDayOfYear</u>	イベント終了日。1は1月1日、365は12月31日を示す。	"1", "365"
<u>year</u>	イベント発生前	"2008"
<u>month</u>	イベント発生月	"1", "10"
<u>day</u>	イベント発生日	"9", "28"
<u>verbatimEventDate</u>	イベント発生日時（フリーテキスト）	"spring 1910", "Marzo 2002", "1999-03-XX", "17IV1934"
<u>habitat</u>	生息地	"oak savanna", "pre-cordilleran steppe"
<u>fieldNumber</u>	フィールドにおけるイベントに割り振られるID	"RV Sol 87-03-08"
<u>fieldNotes</u>	論文のURIやイベントに関するメモ	"notes available in Grinnell-Miller Library"
<u>eventRemarks</u>	イベントに関する補足説明	
<u>locationID</u>	採集地ID	
<u>higherGeographyID</u>	採集地ID	"TGN: 1002002" for Prov. Tierra del Fuego, Argentina
<u>higherGeography</u>	Locality要素の情報よりもおおまかな標本の採集地名。	"South America; Argentina; Patagonia; Parque Nacional Nahuel Huapi; Neuquén; Los Lagos" with accompanying values "South America" in Continent, "Argentina" in Country, "Neuquén" in StateProvince, and Los Lagos in County.
<u>continent</u>	標本採取地・観測地の大陸	"Antarctica", ISO3166 Continent Codeの使用推奨。
<u>waterBody</u>	標本採取地・観測地の大洋	"Indian Ocean", "Baltic Sea", Getty Thesaurus of Geographic Namesの使用推奨。
<u>islandGroup</u>	標本採取地・観測地の島群	"Alexander Archipelago". Getty Thesaurus of Geographic Namesの使用推奨。
<u>island</u>	標本採取地・観測地の島	"Isla Victoria". Getty Thesaurus of Geographic Namesの使用推奨。
<u>country</u>	標本採取地・観測地の国	"Denmark", "Colombia", "España". Getty Thesaurus of Geographic Namesの使用推奨。

生物多様性情報

観測地点 [緯度、経度、高度、深度、地名]

観測者情報

生物同定情報 [科、属、種]

生物の状態 [雌雄、生死、卵/幼生/、...]

標準仕様

Semantic Web技術対応（RDF＋ontology）

例：Excelでデータを記録するのに列ラベルにこれを使うべし
埋めるデータは可能な部分はontologyの用語から選択を

日本語での参照先：http://www.gbif.jp/v2/datause/data_format/index.html

生物多様性情報の標準化の動き

Darwin Core (DwC) : 生物多様性情報を記述するための標準仕様

※ Dublin Core : メタデータ (情報リソース) を記述するための標準仕様

Semantic Web技術 (RDF) を利用

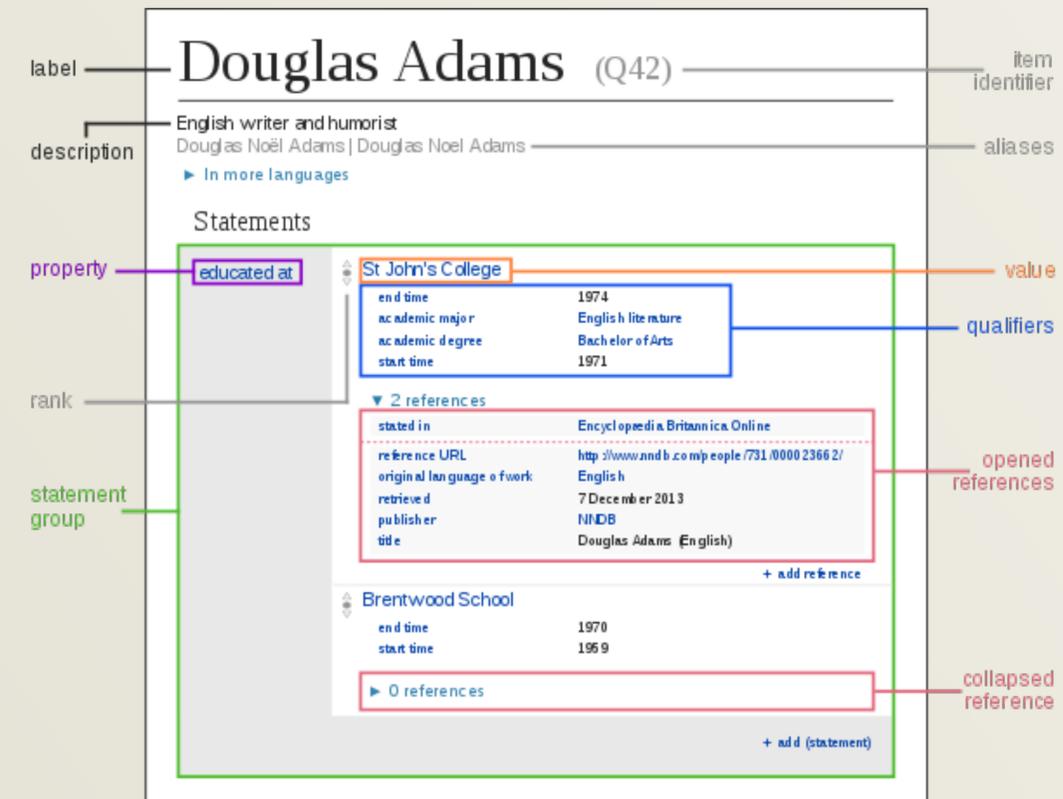
※ DBCLSではSemantic Web技術を用いてバイオインフォデータの統合を行っている

オカレンス情報 : GBIF

文献情報 : BHL、Wikidata

研究者情報 : ORCID、Wikidata

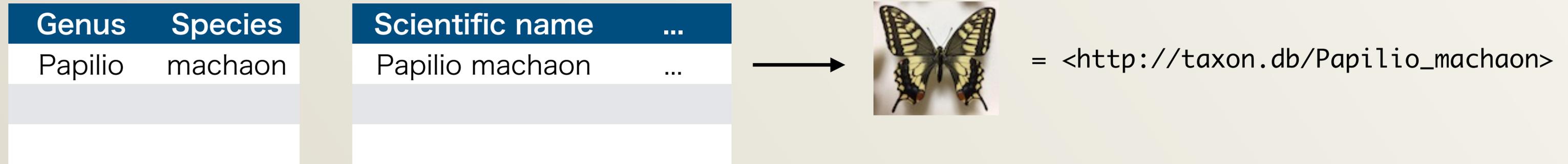
もろもろ : Wikidata



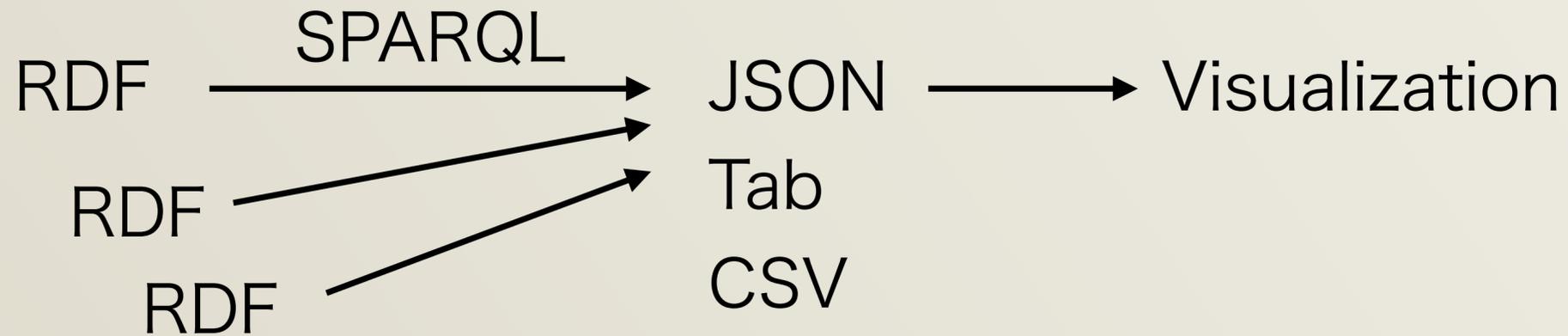
Wikidataのデータ構造

データ標準化・統合化のためのSemantic Web技術

- Standardization: Common keys and common value



- Usability: Easy to parse and easy to merge



- Reusable: No need to download and convert someone's DB to local data

```
@prefix taxon: <http://taxon.db/>
@prefix rel: <http://relation.org/>

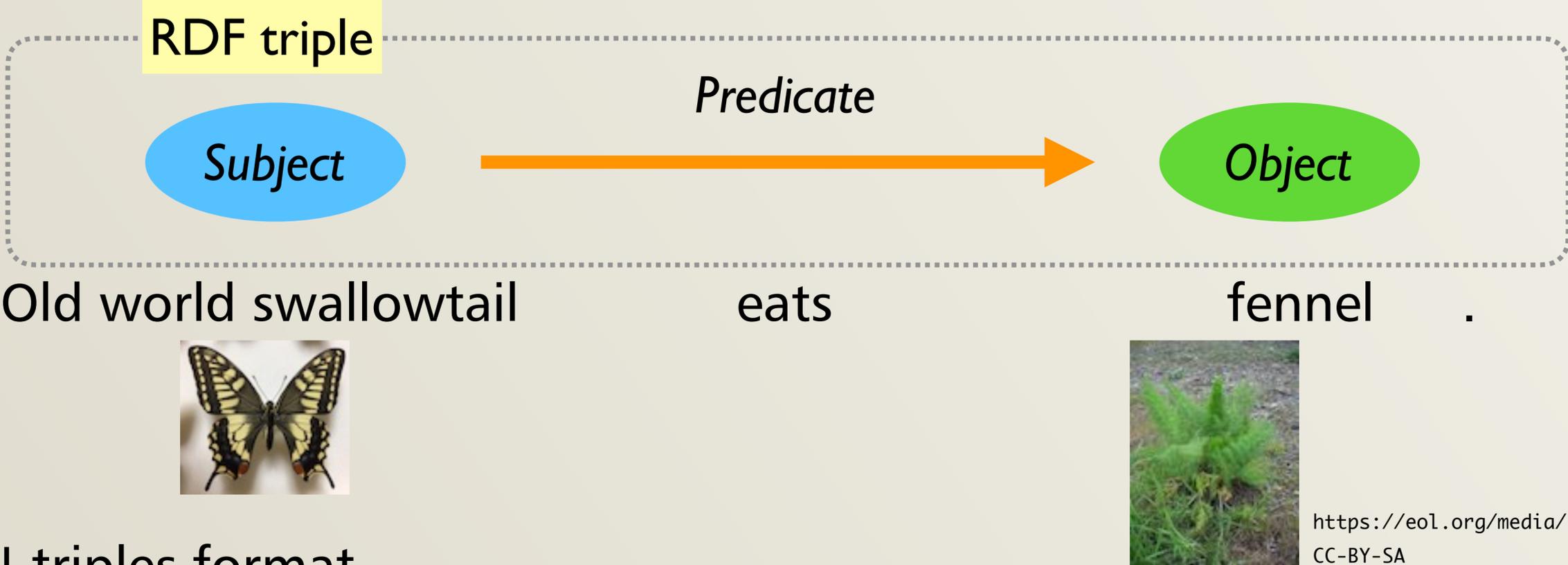
taxon:Papilio_machaon rel:eats taxon:Foeniculum_vulgare .
```

Someone already provide taxonomy DB

So only call DB with value

RDF (Resource Description Framework)

従来、表の形で記録されていたデータを2列間の二項関係にバラして管理+意味づけ



N-triples format

```
<http://taxon.db/Papilio_machaon> <http://relation.org/eats> <http://taxon.db/Foeniculum_vulgare> .
```

Turtle format

```
@prefix taxon: <http://taxon.db/>  
@prefix rel: <http://relation.org/>  
  
taxon:Papilio_machaon rel:eats taxon:Foeniculum_vulgare .
```

* These URIs are dummy

Ontology

The screenshot shows the OLS interface for the term **Papilio machaon**. The breadcrumb trail is: OLS > NCBI organismal classification > NCBITAXON > NCBITaxon:76193. The main title is **Papilio machaon** with the URL http://purl.obolibrary.org/obo/NCBITaxon_76193. The interface includes a tree view on the left, a graph view button, and a term info panel on the right. The tree view shows a hierarchical classification from root to Hexapoda. The term info panel lists various properties such as database cross reference, common name, and has exact synonym.

- Controlled vocabulary (統制語彙、キーワード集)
 - 階層構造
 - 親子の用語の関係性も定義される
- 例: "part of", "is a", "subclass of".

<http://taxon.db/Papilio_machaon>

Ontology

Data value

A zoomed-in view of the tree view from the screenshot, showing the hierarchy: Papilionidae > Papilioninae > Papilionini > Papilio > Papilio machaon. The term **Papilio machaon** is highlighted in a blue box.

Bioinfo/Biodiv infoの融合例：DNA barcoding

GenBank

<https://www.ncbi.nlm.nih.gov/nucleotide>

```
LOCUS       DQ433192                599 bp    DNA     linear   VRT 14-JUL-2016
DEFINITION  Spizella atrogularis voucher MVZ:Bird:170231 cytochrome oxidase
            subunit 1 (COI) gene, partial cds; mitochondrial.
ACCESSION   DQ433192
VERSION     DQ433192.1
KEYWORDS    BARCODE.
SOURCE      mitochondrion Spizella atrogularis (Black-chinned sparrow)
  ORGANISM  Spizella atrogularis
...
FEATURES             Location/Qualifiers
     source            1..599
                       /organism="Spizella atrogularis"
                       /organelle="mitochondrion"
                       /mol_type="genomic DNA"
                       /specimen_voucher="MVZ:Bird:170231"
                       /db_xref="BOLD:CDMVZ038-05"
                       /db_xref="taxon:40208"
                       /country="USA: California"
                       /lat_lon="35.3578 N 120.305 W"
                       /collection_date="20-May-1984"
                       /collected_by="K. Ned"
                       /PCR_primers="fwd_seq: ttctccaaccacaagacattggcac,
                       rev_seq: acgtgggagataattccaatcctg"
     gene              1..599
                       /gene="COI"
     CDS               1..599
                       /gene="COI"
                       /codon_start=1
                       /transl_table=2
                       /product="cytochrome oxidase subunit 1"
                       /protein_id="ABK29354.1"
                       /translation="MVGTAALS..."
ORIGIN
1 atagtaggta cgcacctag ctcctcatt cgagcagaac taggccaacc cggagccctc
61 ...
```

BOLD (Barcode of Life Data System)

<http://boldsystems.org/>

BOLD SYSTEMS DATABASES IDENTIFICATION TAXONOMY WORKBENCH RESOURCES LOGIN

Record Details For CDMVZ038-05 [Back to Search: Records](#)

IDENTIFIERS

Sample ID: MVZ 170231 Museum ID: 170231

Field ID: Collection Code:

Deposited In: University of California, Berkeley, Museum of Vertebrate Zoology

SPECIMEN IMAGES: N/A

COLLECTION SITE:

TAXONOMY

Phylum: Chordata Subfamily:

Class: Aves Genus: *Spizella*

Order: Passeriformes Species: *Spizella atrogularis*

Family: Passerellidae Subspecies:

BIN ID: BOLD:AAF3044

UNITE - Species Hypothesis

https://unite.ut.ee/bi_forw_sh.php?sh_name=SH1511239.08FU#fndtn-panel1

Tuber huidongense Y. Wang (DOI: TH010993) | SH1511239.08FU

Distance to the closest SH: 1.5
No. of sequences in SH: 25

Older version(s) of this SH is/are available

SH code (Count/Total count): SH1511239.08FU (25/25)

Placement in the fungal classification
Fungi: Dikarya: Ascomycota: Pezizomycotina: Pezizomycetes: Pezizomycetidae: Pezizales: Tuberales: Tuber
Index Fungorum: urn:lsid:indexfungorum.org:names:383644

Reference sequence: DQ486032
Chosen by: Tine Grebens
Date: 2015-03-31 14:24

Statistics Taxonomy Ecology

Interacting taxa

Accession number	UNITE taxon name	INSD taxon name	Sequence source	Interacting taxa	Sampling area
AB553376	Tuber	Tuber (Tuber sp. 3 KA-2010)	sample	Fagaceae	Japan
AB553370	Tuber	Tuber (Tuber sp. 3 KA-2010)	sample	Quercus	Japan
AB553369	Tuber	Tuber (Tuber sp. 3 KA-2010)	sample	Fagaceae	Japan
AB553374	Tuber	Tuber (Tuber sp. 3 KA-2010)	sample	Fagaceae	Japan
AB553375	Tuber	Tuber (Tuber sp. 3 KA-2010)	sample	Quercus	Japan

菌類ITSはUNITEが便利

<https://unite.ut.ee/>

GenBank中の生物多様性情報の記述

```
LOCUS       DQ433192                599 bp    DNA     linear   VRT 14-JUL-2016
DEFINITION   Spizella atrogularis voucher MVZ:Bird:170231 cytochrome oxidase
             subunit 1 (COI) gene, partial cds; mitochondrial.
ACCESSION   DQ433192
VERSION     DQ433192.1
KEYWORDS    BARCODE.
SOURCE      mitochondrion Spizella atrogularis (Black-chinned sparrow)
  ORGANISM   Spizella atrogularis

...
FEATURES             Location/Qualifiers
     source           1..599
                     /organism="Spizella atrogularis"
                     /organelle="mitochondrion"
                     /mol_type="genomic DNA"
                     /specimen_voucher="MVZ:Bird:170231"
                     /db_xref="BOLD:CDMVZ038-05"
                     /db_xref="taxon:40208"
                     /country="USA: California"
                     /lat_lon="35.3578 N 120.305 W"
                     /collection_date="20-May-1984"
                     /collected_by="K. Ned"
                     /PCR_primers="fwd_seq: ttctccaaccacaaagacattggcac,
rev_seq: acgtgggagataattccaaatcctg"
     gene             <1..>599
                     /gene="COI"
     CDS              <1..>599
                     /gene="COI"
                     /codon_start=1
                     /transl_table=2
                     /product="cytochrome oxidase subunit 1"
                     /protein_id="ABK29354.1"
                     /translation="MVG TALS..."

ORIGIN
1 atagtaggta cgc cctaag cctcctcatt cgagcagaac taggccaacc cggagccctc
61 ...
```

← KEYWORDSにBARCODEとある

← 標本ID、BOLD ID

← 遺伝子名 (ただし記述の揺らぎあり)

COI, CO1, COX1, ...

実際のデータ統合は難しい

	BOLD (direct submission)	GenBank	BOLD (Mined from GenBank)
Record ID	JBOL054-11	AB63763	GBDP15012-14
BOLD ID	JBOL054-11	JBOL054-11	GBDP15012-14
BIN ID	AAV1368		AAV1368
GenBank ID		AB638763	AB638763
Museum ID	NSMT-I-Dip-6785	NSMT:I-Dip-6785	NSMT:I-Dip-6785
voucher status	museum vouchered (type series)		
tissue descriptor	legs	legs	
Country	Malaysia	} Malaysia: Selangor, Ulu Gombak	Malaysia
Province/State	Selangor		Selangor
Region/County	Ulu Gombak		Selangor, Ulu Gombak
latitude	3.32	} 3.32 N 101.75 E	3.32
longtitude	101.75		101.75
elevation	260 Meters		
seq length	636 bp	636	633 bp
publication		+	

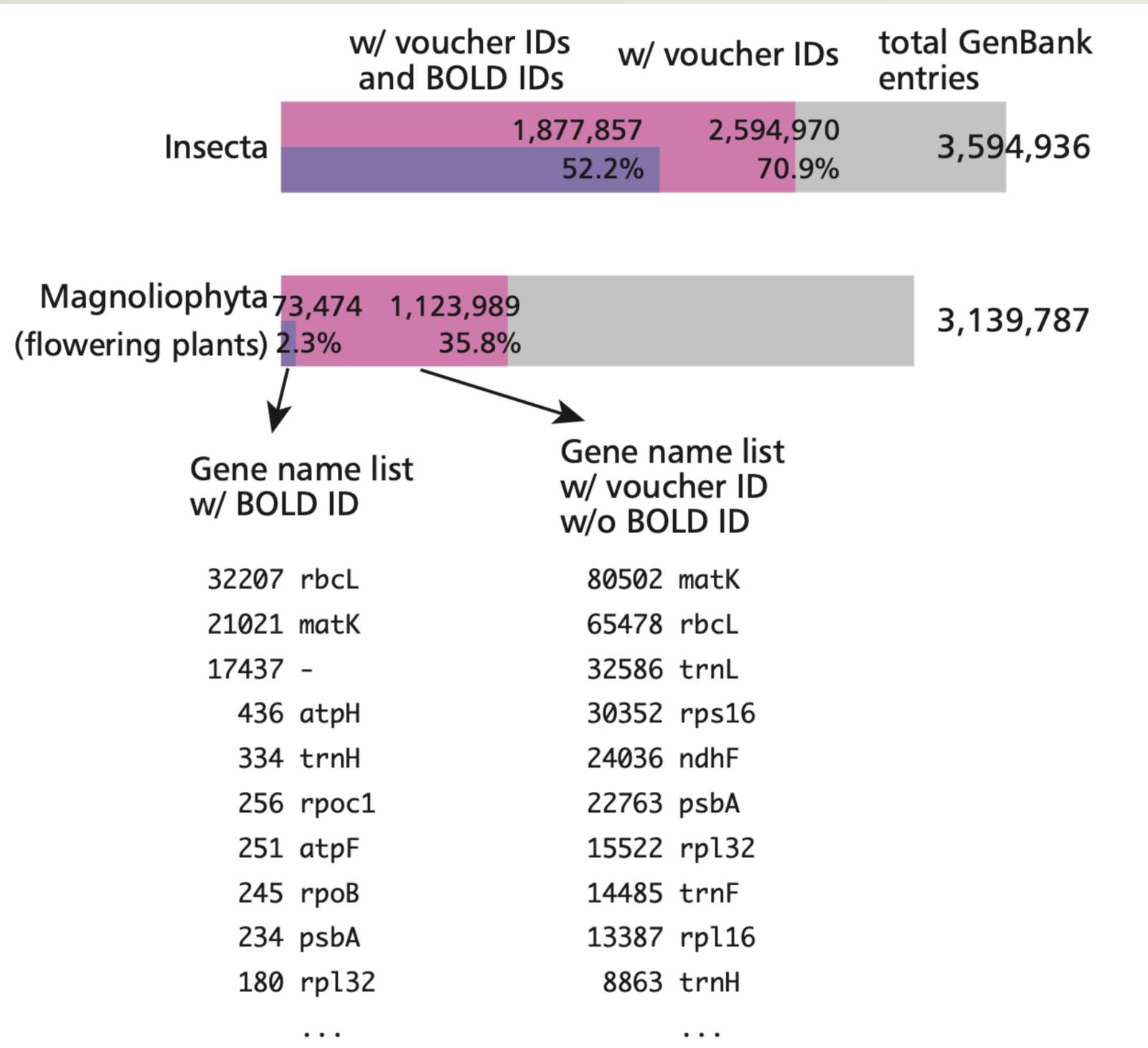
登録はBOLDでもGenBankでも可能

標本情報がrichだとBOLD
配列メインだとGenBank

BOLDはGenBankからデータの
取り込みも

二重登録になる場合も
そして各々 記述が違う場合も

GenBank中のDNA barcodingデータ (昆虫・被子植物編)



voucher ID: USNMENT0092157

Stenammina megamanni (Hymenoptera)

GenBank中に1510配列

2019年10月現在

GenBank中のDNA barcodingデータ (魚類編)

BOLD

273,426 エントリ

211,589 エントリ
w/GenBank ID

(登録として) 種数

18,952

種レベルで種数

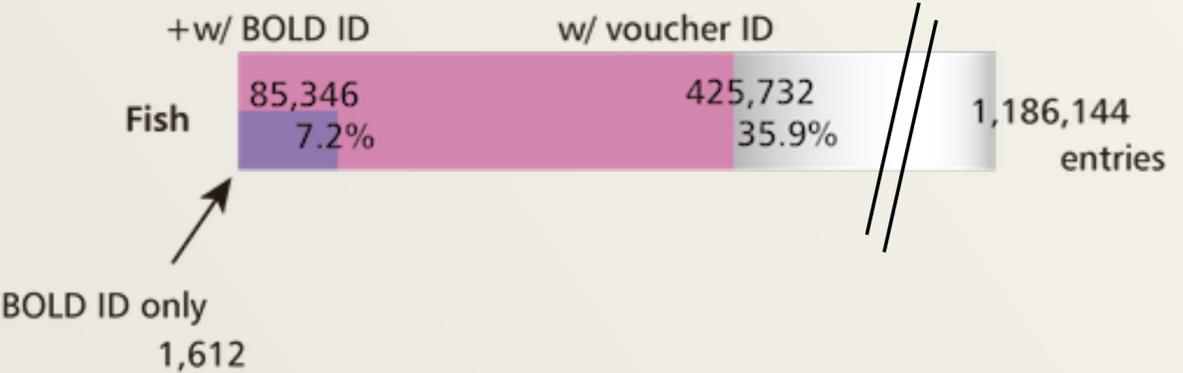
15,063

w/o sp./cf./aff./...

GenBank

"Mined from GenBank, NCBI"

121,748 エントリ (44.5%)



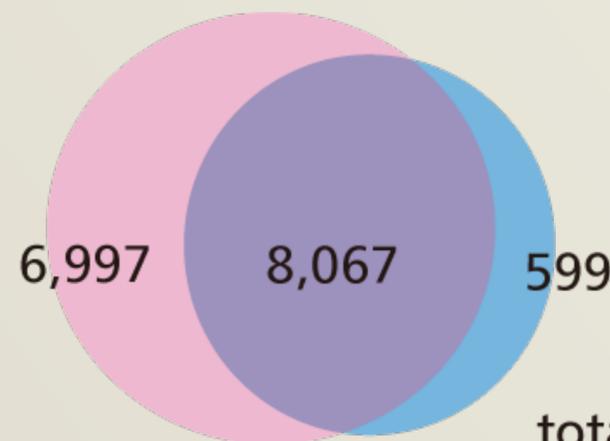
NCBI taxonomy

11,784

39,127

8,665

20,987



total: 15,663 species

BOLDの方がカバー率が良い
同じエントリでもGenBankの方が種までが多いが、その種同定が正しいとは限らない

菌類標本を用いたMuseomics研究



Wikipediaより

世界で最も一般的な菌類のひとつ
東南アジアでは食用にも
実は呼吸器疾患の原因にもなっている
遺伝的多様性が極めて高い?



野外採集株、科博の標本、患者からの採集株をシーケンス

ゲノムサイズ=38.5Mb

染色体数=14

他のグループがゲノム配列報告済（が、mapping率悪い）

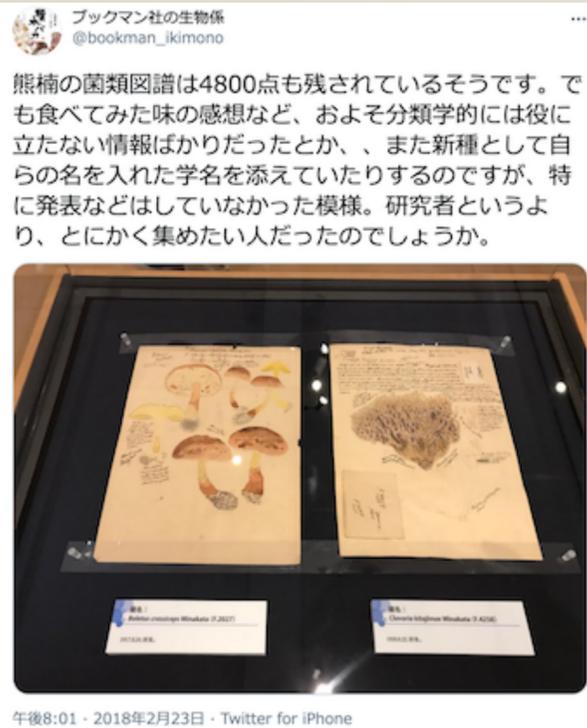
菌類のアセンブルは難しいものもあるらしいが、かなり順調

museomicsへのDS活動の応用例 (構想)

人文学オープンデータ



標本ラベル読み取り



日本の資料

菌類図譜

[南方熊楠]

ライフサイエンス
統合DB

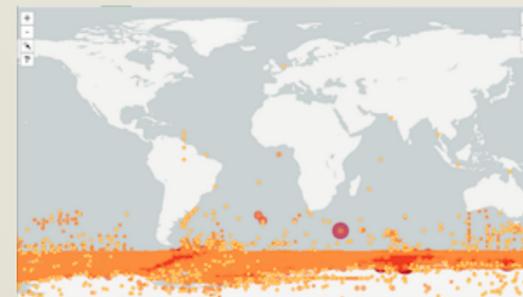
標本情報と遺伝子情報
DNAバーコーディング
メタゲノム

Museomics

極域環境

極域の生物分布 (オカレンス)

環境と生物



ゲノムデータ

博物館標本のシーケンシング
湿度や燻蒸によるDNA断片化
フィールドでのシーケンシング
(サンプル保存)

社会データ

市民科学 (SNS上での目撃情報)

データ同化

データとしての生物多様性情報
環境・生態系シミュレーション