バイオリソースデータRDF化への取り組み

高月照江[1)]，臼田大輝[2)]，川本祥子[1) 3)]，桝屋啓志[2)]，川島秀一[1)]

[1)] 情報システム研究機構ライフサイエンス統合データベースセンター，[2)] 理化学研究所 バイオリソース研究センター，[3)] 国立遺伝学研究所

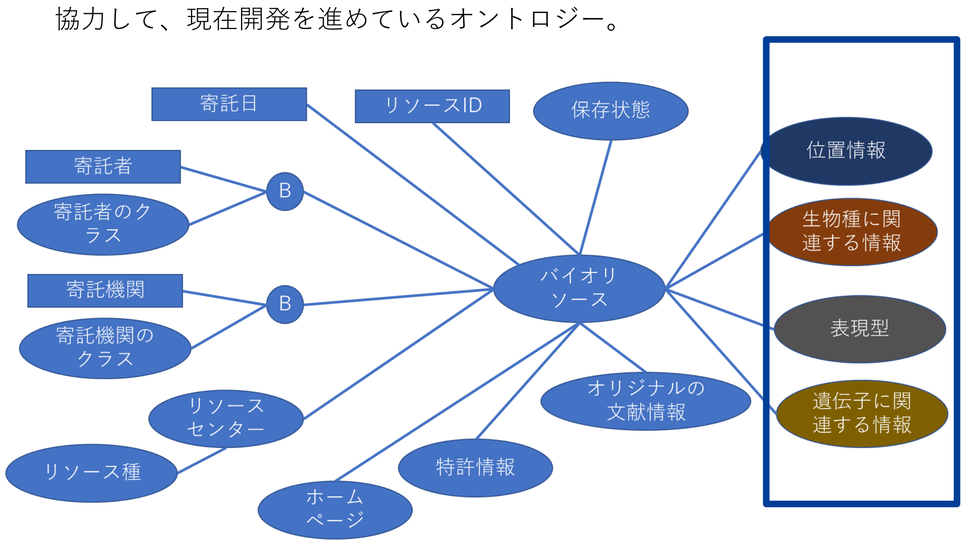
● 目的

ナショナルバイオリソースプロジェクト（NBRP）では、我が国独自の優れたバイオリソース（生物遺伝資源）に関して、所在情報、系統・特性情報、遺伝子情報などをデータベース化し公開することで、研究者が必要とするバイオリソース検索サービスの整備に取り組んできた。ただし、現状では生物種毎に個別のデータベースとして開発されていることから、生物種横断的に検索するようなことはできない。我々は、NBRPのデータベースをRDF化することで、バイオリソース横断的な検索が可能になると考えている。また、RDF化することで既存のRDF化されたデータベースやオントロジーとも統合することができるので、例えばオーソログ遺伝子や表現型類似性などの観点から、高度な統合検索も可能となる。本発表では、その第一段階として、異なるバイオリソース間で共通して記述できる項目を選び、RDF化に利用するための共通モデルを提案する。また、いくつかのバイオリソースに関して本RDFモデルを用いてRDF化したデータについても報告する。最終的には、NBRPが提供する全ての生物種について、共通した語彙およびモデルを用いたバイオリソースRDFを構築し、これまではできなかった高度なデータ検索を提供することを目指している。

バイオリソース共通スキーマの開発

●NBRPOの開発

NBRPOは、バイオリソース情報における共通項目部分をRDF化するために、理研バイオリソース研究センター、遺伝学研究所と協力して、現在開発を進めているオントロジー。



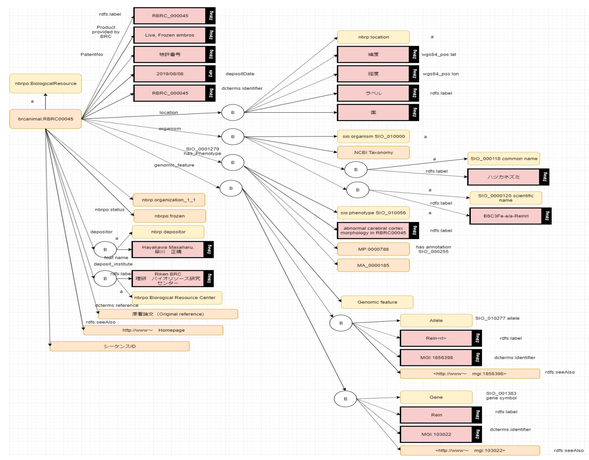
各バイオリソースが、個々に持つ情報については、大項目を分類。詳細な内容については、リソースの持つ固有の情報に沿って、記述を行う。

●バイオリソースを基準にして、共通項目を検討したスキーマ図

国内版バイオハッカソン BH19.7においてバイオリソースとして必要な情報の項目を整理し、スキーマを設計。（http://wiki.lifesciencedb.jp/mw/BH19.7）

バイオリソースRDF化までの流れ

国内で収集されているバイオリソースの各種データ（個々のデータベースで現在は管理されている）



生物種毎に、共通スキーマを用いて、必要項目の全体スキーマモデルを検討

（サンプルはRIKEN BRCのマウス）

共通化該当項目より、RDF化の作業を実施

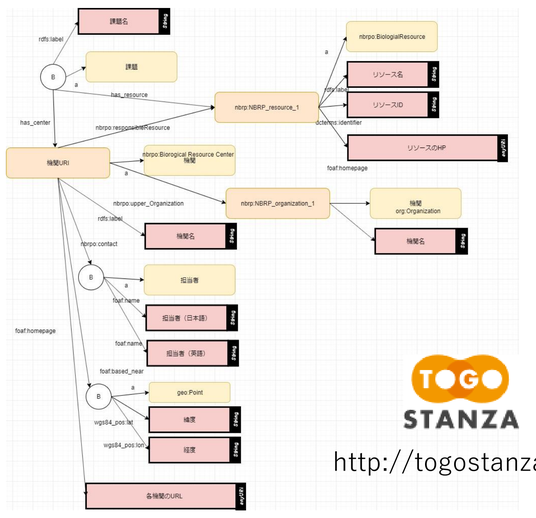
RDF化したデータは、NBDC RDF Portalから公開予定

各公共データベースでの利活用

バイオリソースセンター情報のRDF化

●NBRPに所属するリソースセンターの情報のRDF化

バイオリソースデータに関連付けを実施するため、NBRPに所属する、各種リソースの担当機関について、RDF化を実施。

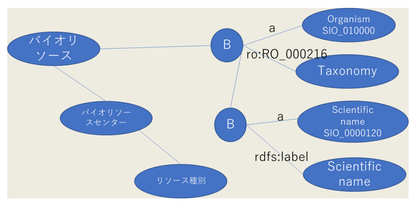


●リソースセンター用のスキーマ図

TOGO STANZA
http://togostanza.org/

リソースセンターのRDFデータを用い、“TOGO STANZA”を利用して、可視化のために、日本地図に各リソースセンターをマッピング。（表示は、マウスを管理する、理研バイオリソース研究センター）

共通スキーマを用いたバイオリソースデータのRDF化



共通スキーマの中から、管理しているリソースセンター、生物に関する情報（学術名、タクソノミーID）等の部分を、マウス、米、微生物について、部分的にRDFデータを作成。

	common_name	count
1	"Red rice"	"842"
2	" wild rice"	"351"
3	"Barth's rice"	"351"
4	"Australian wild rice"	"39"
5	" ヒゲナガノイネ"	"673"
6	"アマゾン野生イネ"	"163"
7	"African wild rice"	"351"
8	"ノイネ"	"673"

Showing 1 to 8 of 8 entries

稲のデータに登録してある一般名称について、その名称で呼ばれるリソース数をSPARQLでカウントした結果。

	Research_CenterID	center_name	count	resource_name
1	nbrp:NBRP_organization_1_1	"国立研究開発法人理化学研究所バイオリソース研究センター実験動物開発室"	"6376"	"マウス"
2	nbrp:NBRP_organization_1_5	"国立研究開発法人理化学研究所バイオリソース研究センター微生物材料開発室"	"16276"	"一般微生物"
3	nbrp:NBRP_organization_22_1	"国立研究開発法人国立環境研究所生物・生態系環境研究センター生物多様性資源保全研究推進室"	"2927"	"藻類"
4	nbrp:NBRP_organization_8_3	"大学共同利用機関法人情報・システム研究機構国立遺伝学研究所生物遺伝資源センター植物遺伝研究室"	"1730"	"イネ"

リソースセンター毎に、担当しているリソース名と、リソース数をSPARQLでカウントした結果。

Showing 1 to 50 of 7,119 entries

	resource_name	taxon_URL	count
1	"General microbes"	"http://identifiers.org/taxonomy/193462"	"10"
2	"Mice"	"http://identifiers.org/taxonomy/10092"	"10"
3	"Mice"	"http://identifiers.org/taxonomy/10090"	"10856"
4	"Rice"	"http://identifiers.org/taxon/83308"	"12"
5	"General microbes"	"http://identifiers.org/taxonomy/1630166"	"12"
6	"Rice"	"http://identifiers.org/taxon/29690"	"128"
7	"Rice"	"http://identifiers.org/taxon/4529"	"1346"
8	"Rice"	"http://identifiers.org/taxon/52545"	"14"
9	"Rice"	"http://identifiers.org/taxon/110450"	"14"
10	"General microbes"	"http://identifiers.org/taxonomy/1685415"	"16"
11	"Rice"	"http://identifiers.org/taxon/40148"	"198"
12	"General microbes"	"http://identifiers.org/taxonomy/53368"	"2"
13	"General microbes"	"http://identifiers.org/taxonomy/1303408"	"2"
14	"General microbes"	"http://identifiers.org/taxonomy/83265"	"2"

リソースの生物種について、生物種毎に登録してあるリソース数をSPARQLでカウントした結果

●共通スキーマを利用することにより、生物種を限らず、必要な項目のデータについて、検索および集計が可能となっている。

● 今後の展望

現在、共通化できる項目についてスキーマを検討し、RDF化への試みを実施しているが、その他にも、表現型や遺伝情報の部分についても、共通化できる生物種については、情報統合のため、共通スキーマの開発を行っている。共通スキーマを用いてRDFデータを作成することにより、遺伝情報や表現型からも生物種横断検索が可能となってくる。

それらのデータは、NBDC RDF Portalから広く公共に公開することにより、各、公共のデータベースにおける、幅広いデータの利活用が期待できる。