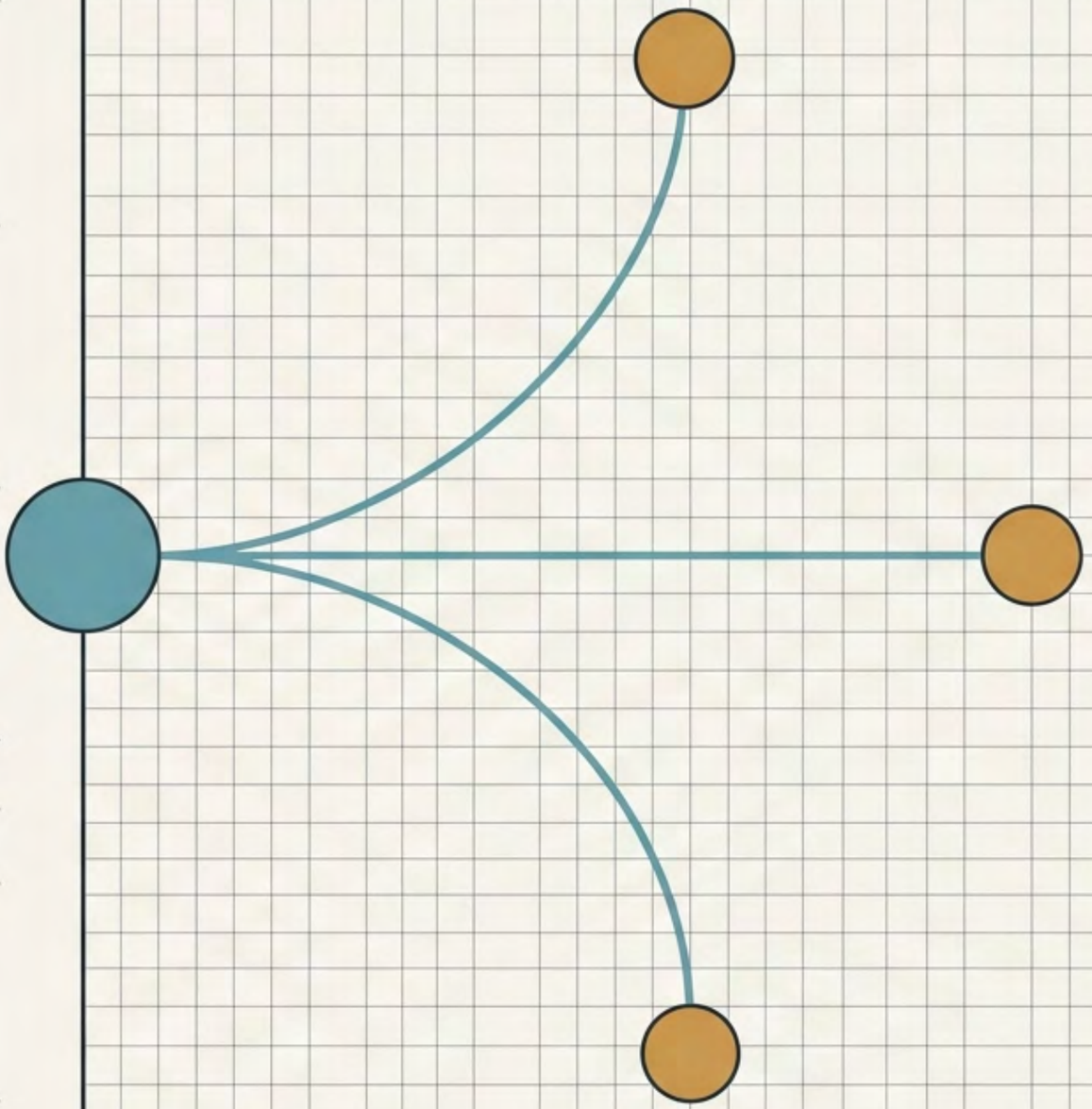
つながるデータのつくりかた

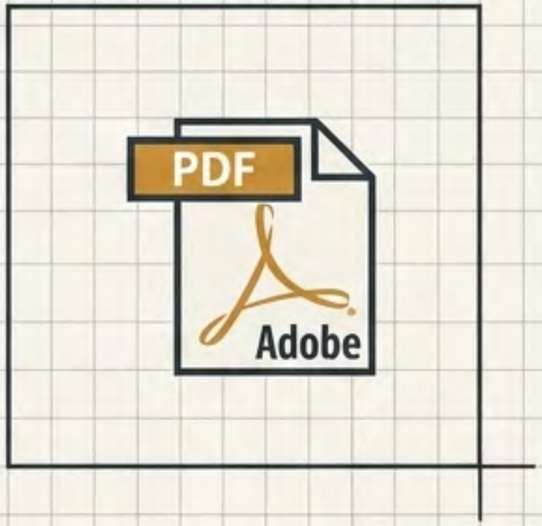
Iwao Kobayashi @ インフォ・ラウンジ株式会社

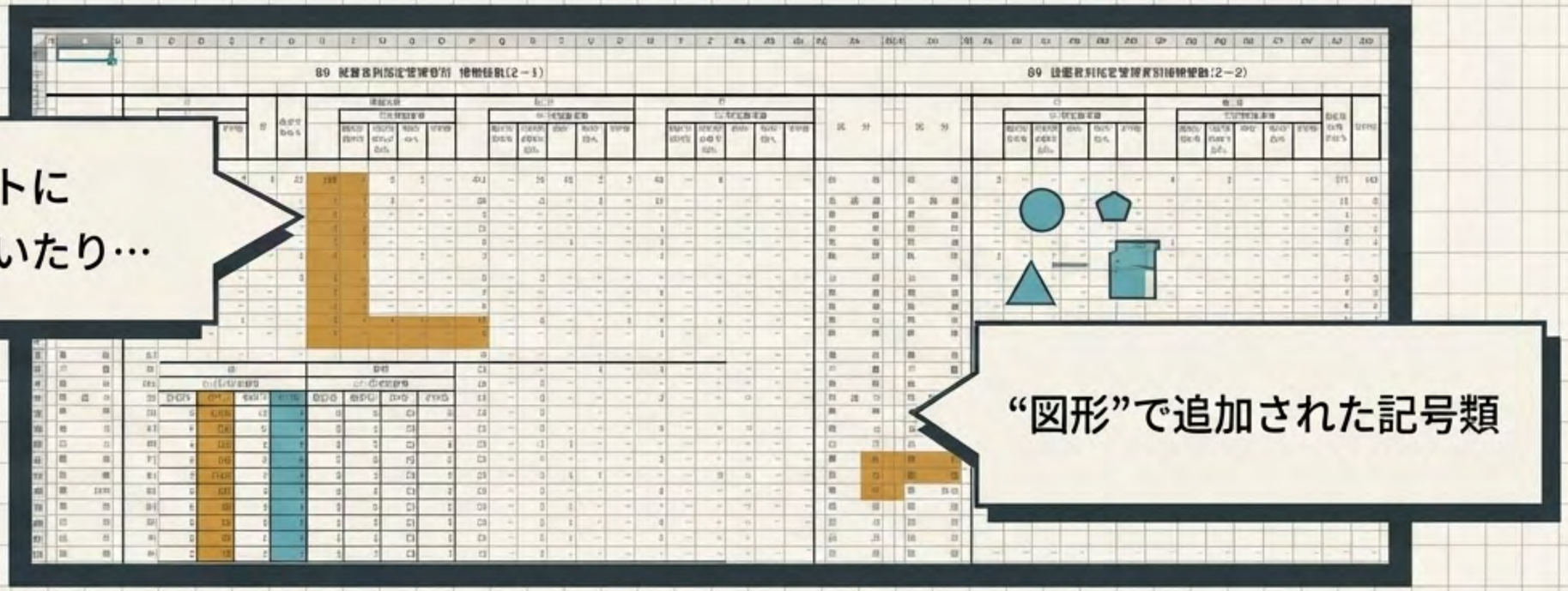


NotebookLM

つながらないデータ

情報が「ファイル」という箱に閉じ込められ、システム間で孤立している現状。人間には読めても、機械には処理できないデータの姿。





1つのエクセルシートに
2つの表がはいっていたり…

“図形”で追加された記号類

NotebookLM

人間の論理 vs. 機械の現実

私たちは同じ意味を意図していても、機械は全く異なるものとして認識してしまう。

日付形式

European Logic
DD.MM.YYYY

American "Logic"
MM/DD/YYYY

Japanese Logic
YYYY-MM-DD

平成29年10月29日 = 2017年10月29日

住所表記

東京都千代田区丸の内
一丁目9番1号

東京都千代田区丸の内
1-9-1

座標形式

35°41’16.23
130°22’06”

35.696175
130°22’06.00

機械処理のエラー

NotebookLM

NotebookLM

柱1：データ構造

データ構造の進化

テキスト

吾輩は猫である。
名前はまだ無い。
どこで生れたかと
とんと見当がつか
ぬ…

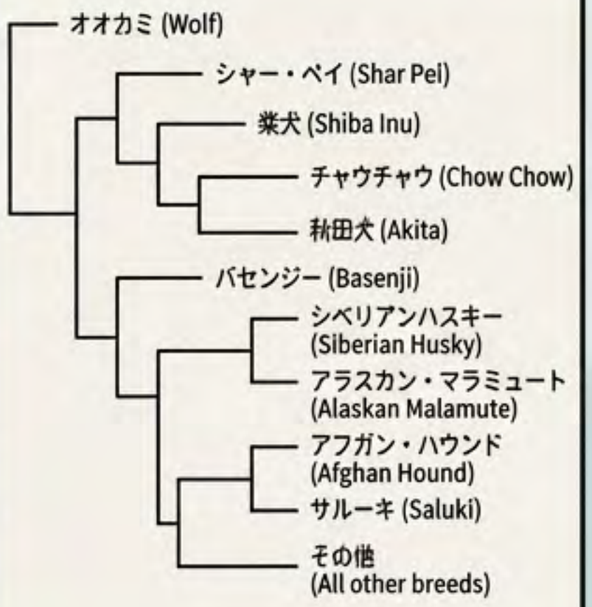
人間には読めるが、
構造がない

表

N7001	金沢さくら保育園…
	神奈川県横浜市金沢区
	…

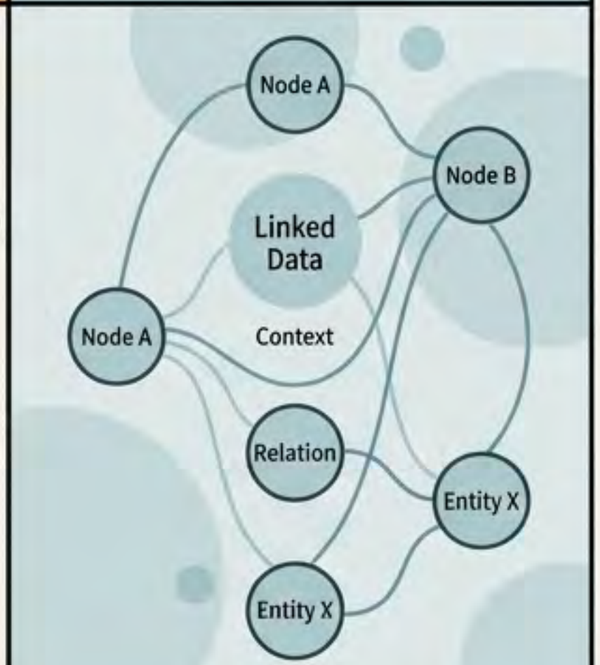
リストには向くが、
関係性は表現できない

ツリー



階層化はできるが、
横のつながりが硬直的

グラフ



多方向の文脈と関係性を
表現（Linked Data）

NotebookLM

柱2：意味の解釈

文脈を与え、機械が「意味」を誤解しないようにする。

schema:Factory

Plant

bio:Plant

機械にはどちらの意味か判別不能

NotebookLM

柱3：識別子

モノ・コトを明確に区別するための記号としての条件。

一意であること

ある範囲の中において、
他と重複しない

永続性が保証されていること

将来にわたって記号
が変わらない

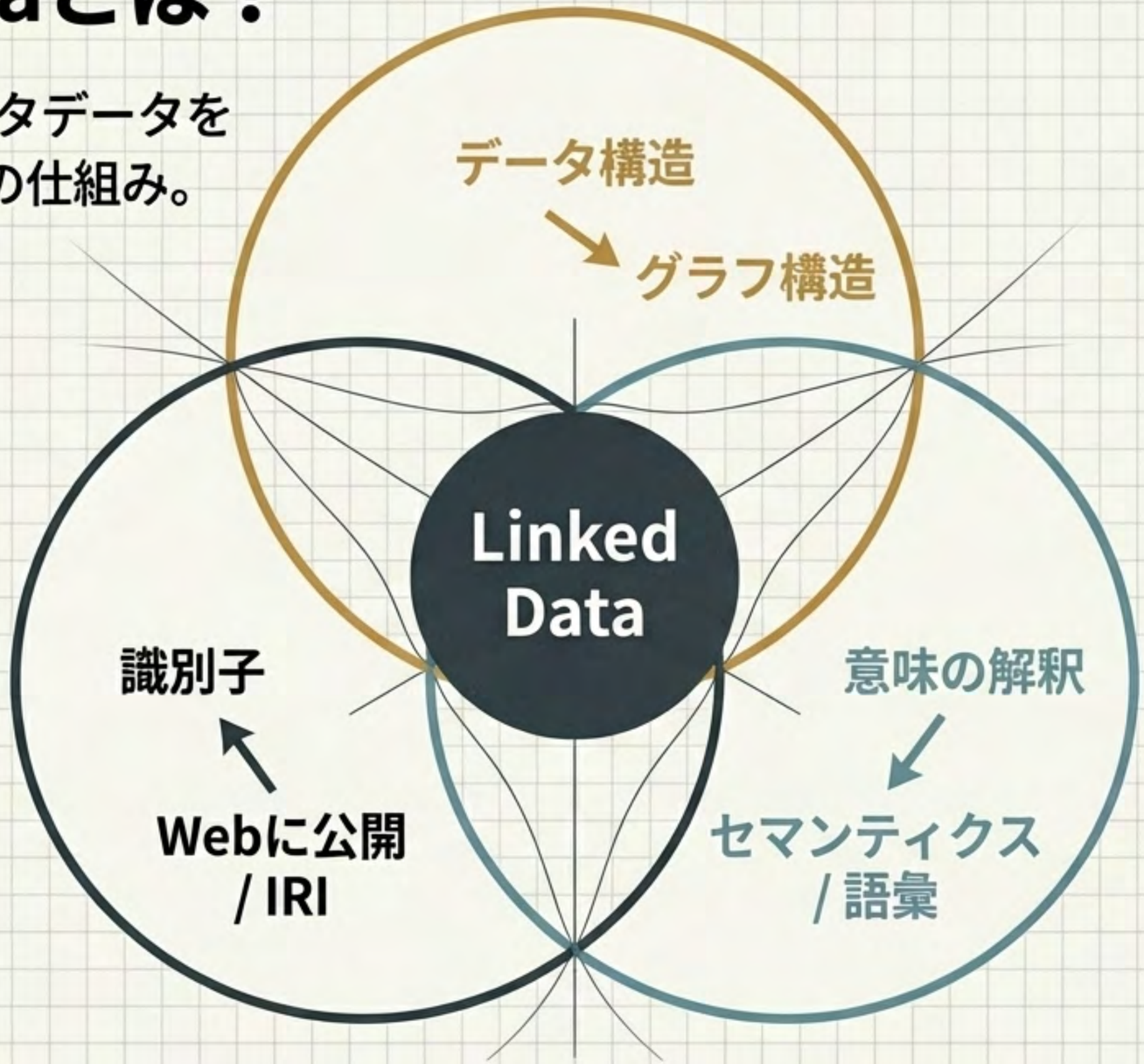
参照できること

ネットワーク経由で
アクセス可能である

NotebookLM

Linked Dataとは？

Webの標準技術でメタデータを
公開・共有するための仕組み。



NotebookLM

Webの原則 から Linked Dataの原則 へ

Webの原則
(Web of Documents)

1. IRIを HTML文書の識別子として使う。
2. IRIには httpスキームを使う。
3. IRIを見ると HTML文書が返ってくる。
4. HTML文書には 他のIRIへのリンクがある。

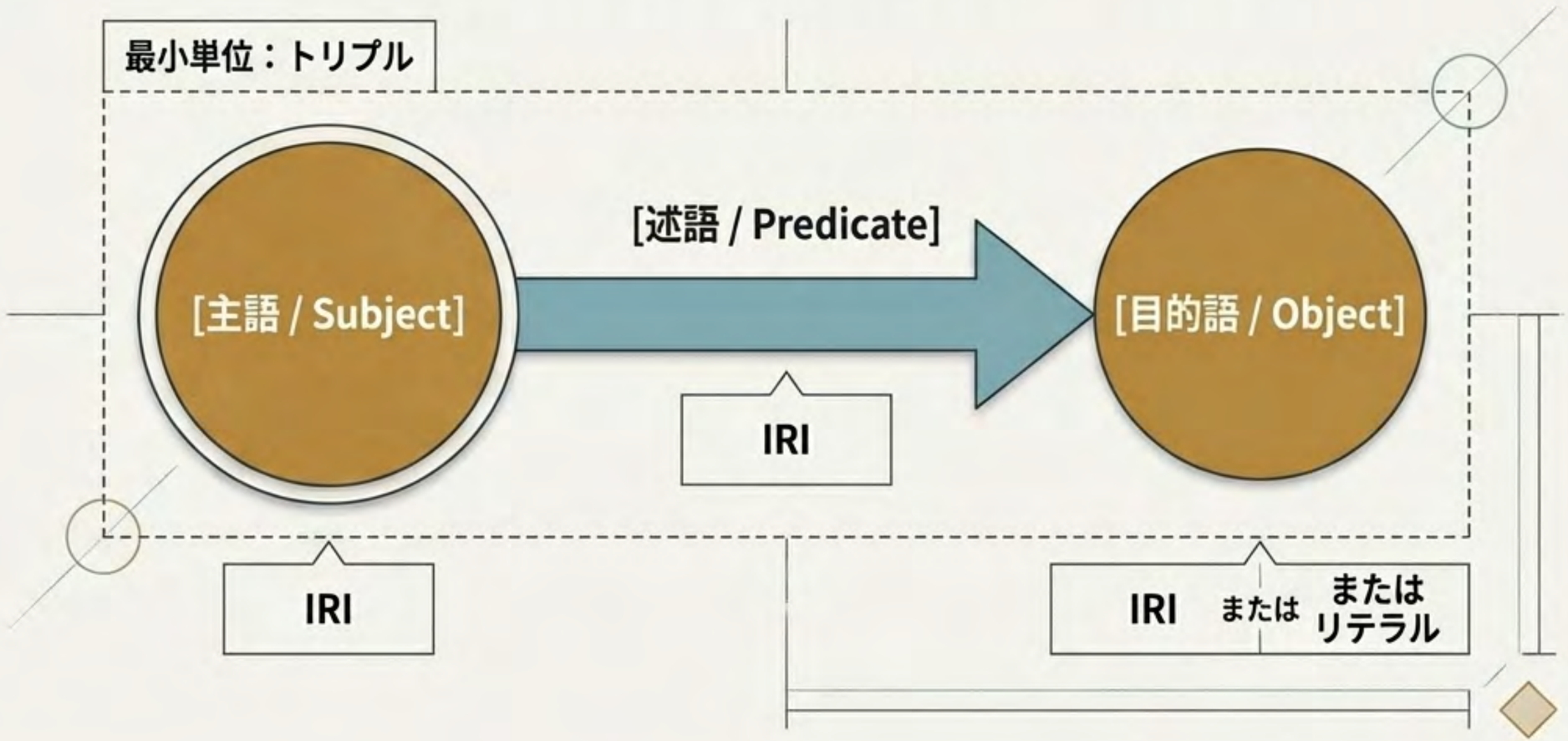
Linked Dataの原則
(Web of Data)

1. IRIを モノ(Thing)の識別子として使う。
2. IRIには httpスキームを使う。
3. IRIを見ると モノの構造化データが返ってくる。
4. データには 他のIRIへの型付リンクがある。

NotebookLM

フレームワーク：RDF (Resource Description Framework)

事物を記述するフレームワーク。
識別子にはIRIを使い、情報を構造化する最小単位「トリプル」で表現する。



NotebookLM

RDFの実践：東京都を記述する

NotebookLM

メカニズム１：参照解決

IRIを見ると、HTTPのGETリクエストで
機械可読な構造化データが返却される。

HTTP GET: http://geonames.jp/resource/神奈川県横浜市

人間が見る画面

機械が受け取るデータ（Turtle）

```
gnjp:神奈川県横浜市 a gn:Feature ;
    rdfs:isDefinedBy <http://geonames.jp/resource/神奈川県横浜市/about>
    gn:name "神奈川県横浜市"@ja ;
    gn:parentFeature gnjp:神奈川県 .
```

親データや子データへのリンク
が直接埋め込まれている

NotebookLM

メカニズム2：語彙

値の意味を解釈可能にするためのデータ定義。

独立したデータ作成者が
同じルールを共有する基盤

1. 意味の定義

<https://hoge.fuga/voc/apple> → = "コンピューター会社"

果物ではないことを明示

2. クラスとプロパティ

犬 (Dog)

- 足が4本
- 尻尾がある

機械に概念のルールを教える

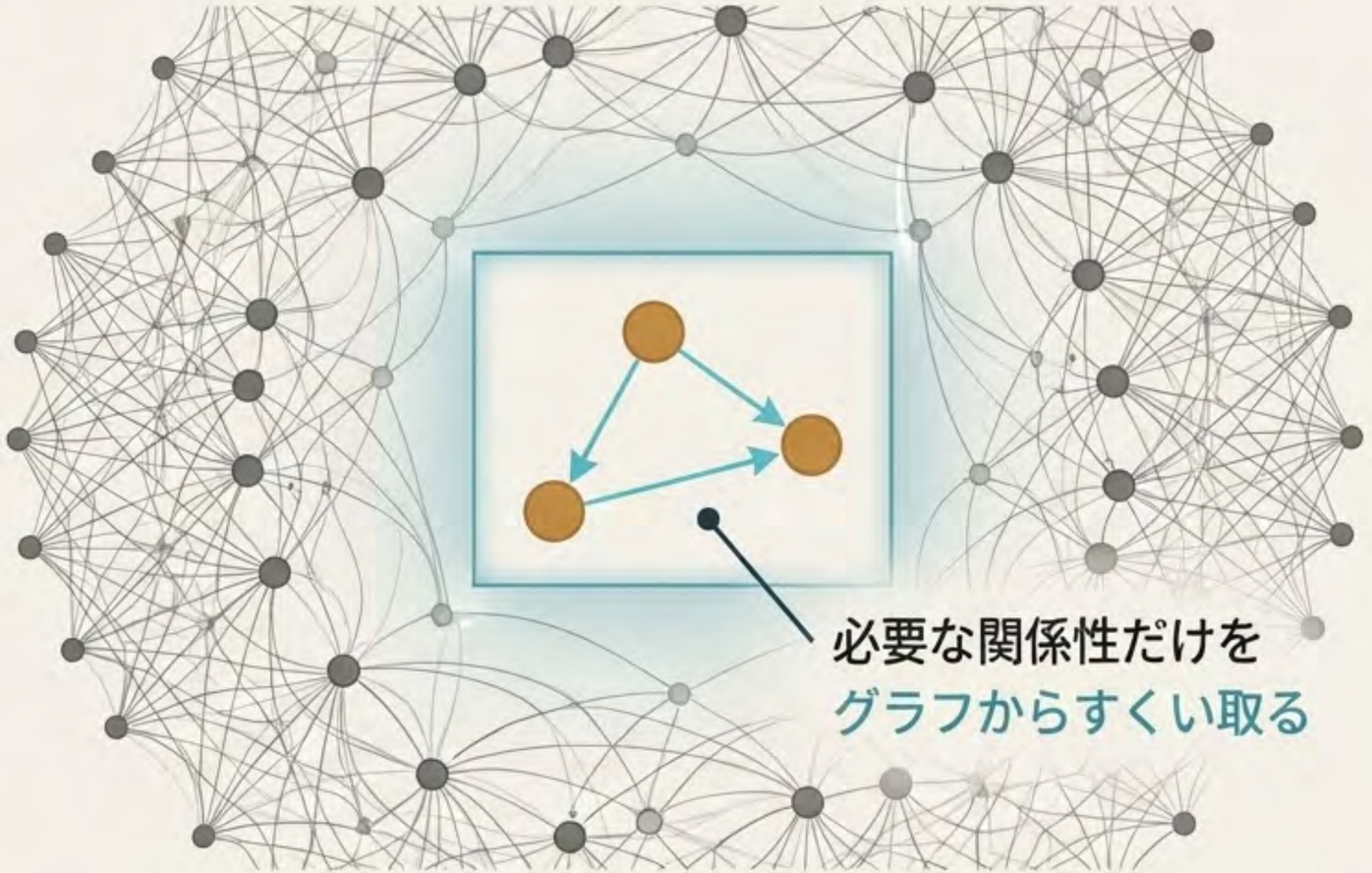
3. 階層と集合

NotebookLM

データを検索する：SPARQL

RDF用の標準クエリ言語（SPARQL Protocol And RDF Query Language）。

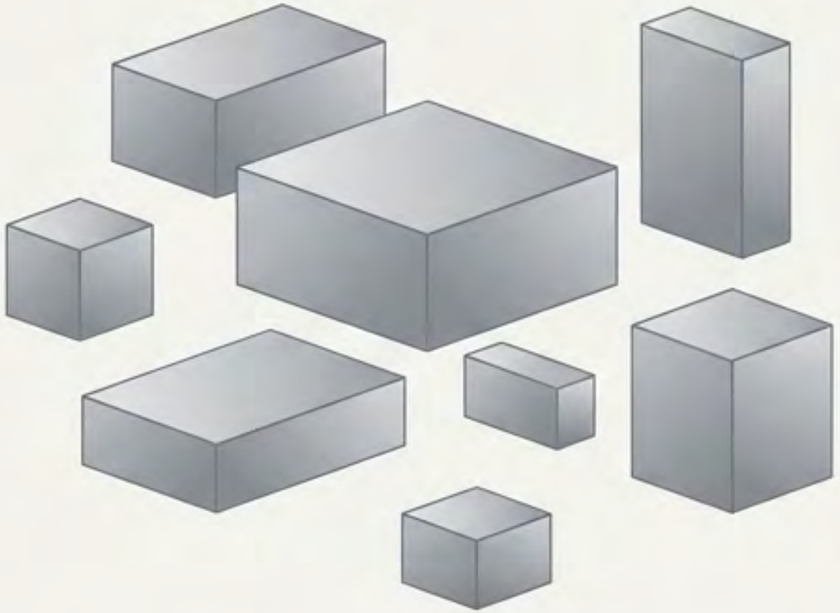
- W3Cで標準化（2008年 SPARQL 1.0 / 2013年 SPARQL 1.1）
- SQLに似た構造を持つ
- グラフ全体から特定の「トリプルパターン」を抽出する



NotebookLM

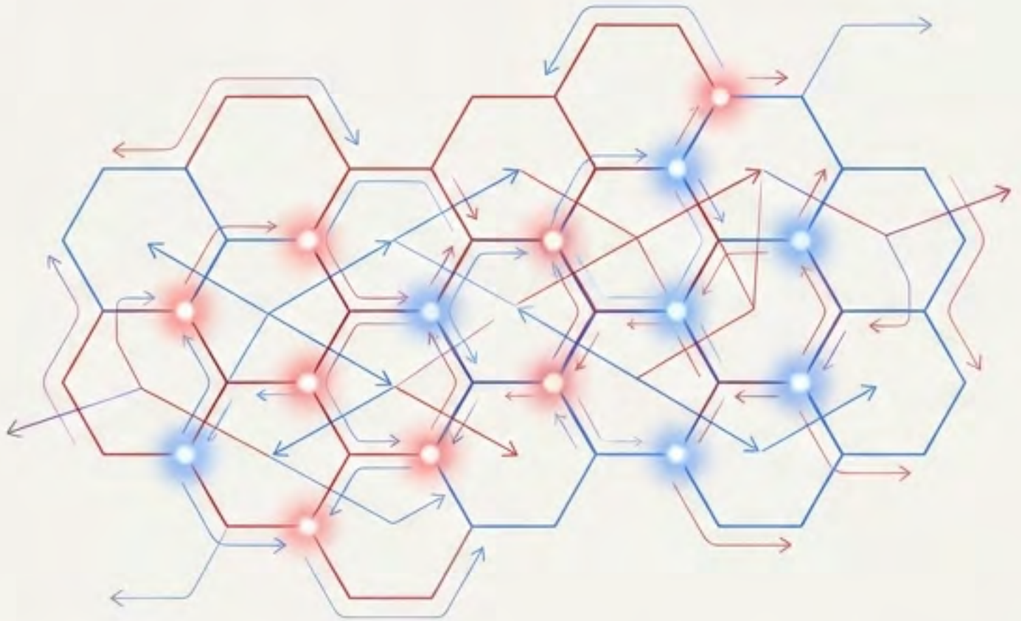
NotebookLM

データの孤島（Data Silos）



従来のデータ管理は、機関ごと・フォーマットごとに分断され、横断的な検索や機械的な連携が困難でした。

つながる知のウェブ（The Knowledge Web）



Linked Open Data（LOD）は、データに一意の識別子（URI）を与え、意味的なつながり（RDF）を持たせることで、世界中のデータを一つの巨大なデータベースとして繋ぎ合わせます。

NotebookLM

Pillar 1: 巨大な知のハブ「ジャパンサーチ」

国の多様なコンテンツを束ねる分野横断プラットフォーム

概要

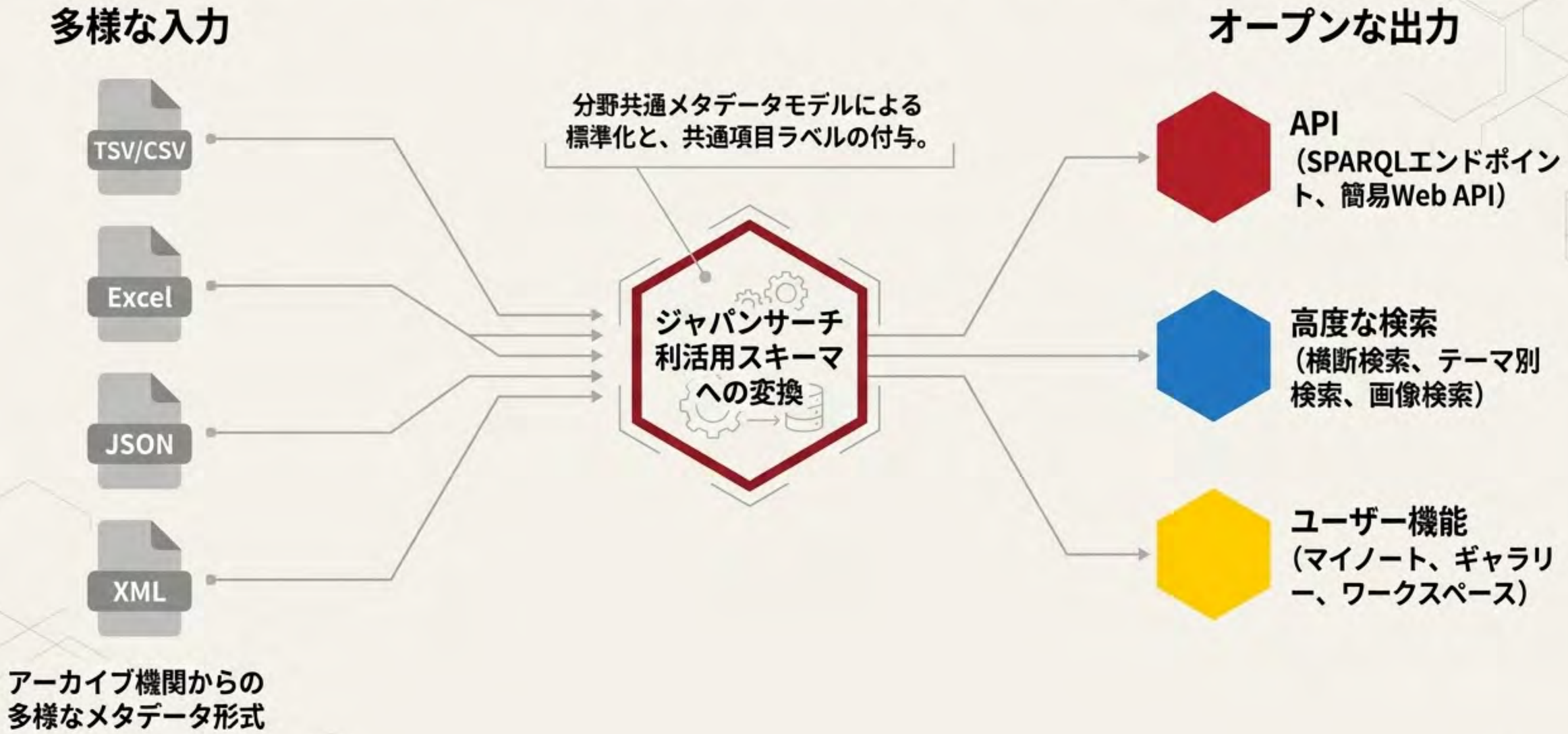
図書館、博物館、美術館、公文書館、学術機関などのデジタルアーカイブを連携する国の基盤プラットフォーム。

価値

メタデータとサムネイルをオープンに流通させることで、日本のデジタルアーカイブの「簡便」で「持続可能」な「活用」を実現します。

NotebookLM

データ・リファイナリー：多様な入力からオープンな出力へ



NotebookLM

Pillar 2: メディア芸術データベース（MADB）

日本のポップカルチャーを「構造化」する

対象	マンガ、アニメーション、 ゲーム、メディアアート。
提供形式	LODに対応したRDF形式で記述。SPARQLエンドポイントおよび 全データセットのファイル提供に対応。
目的	散在するメディア芸術作品の情報を体系化し、研究者や開発者が 誰でも自由に利用・分析できる技術標準を提供すること。

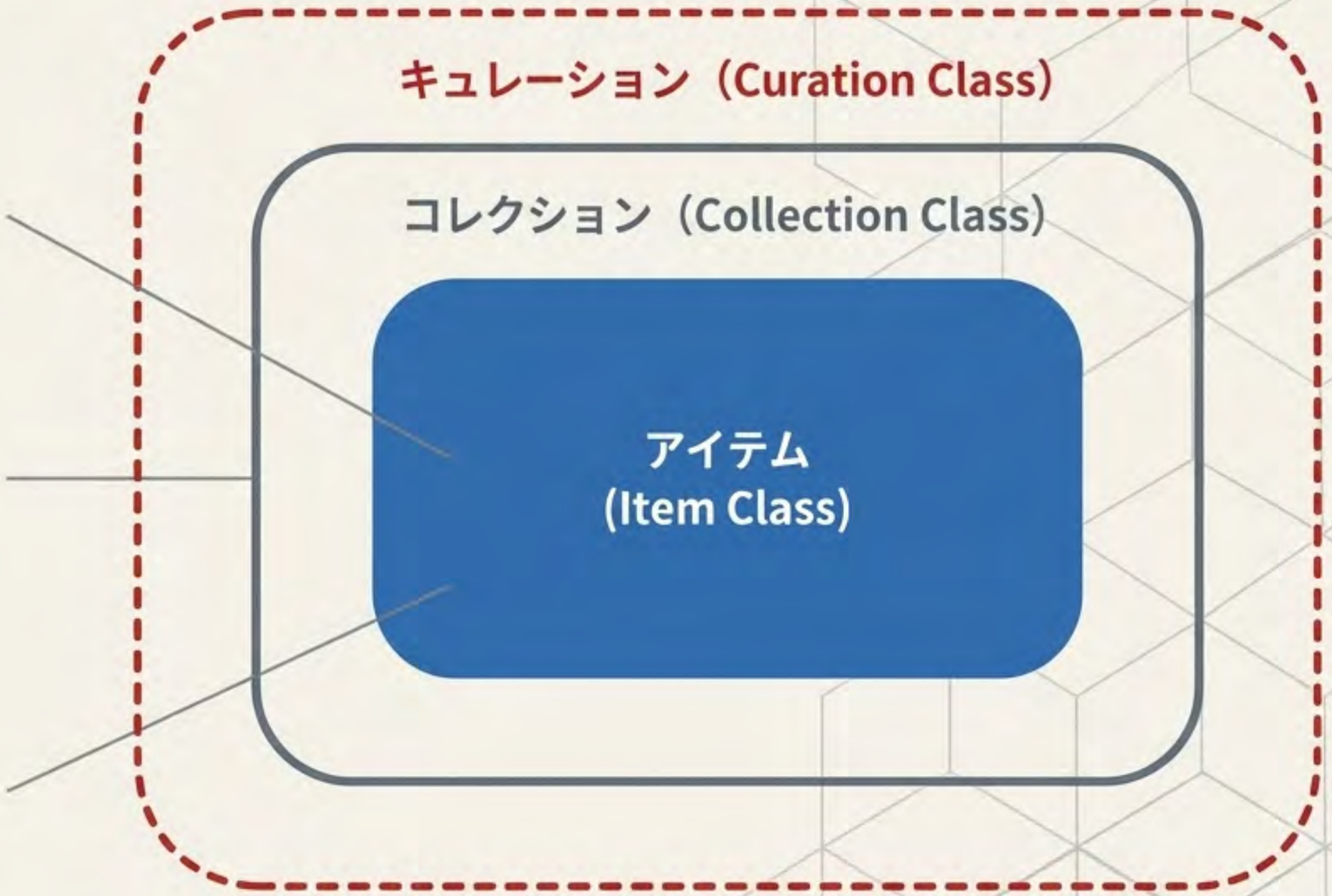
NotebookLM

MADBのデータ構造モデル

作品を体現する具体的な資料（（単行の（例：単行本の1巻 、ケーマ巻、ゲームの特定のパッケージ）。

同一分野のアイテムの集合（例：特定のマンガの全巻セット、シリーズ作品）。

複数の分野にまたがるアイテムの集合（例：特定のテーマに沿ったアニメ、マンガ、ゲームの横断的な展示セット）。

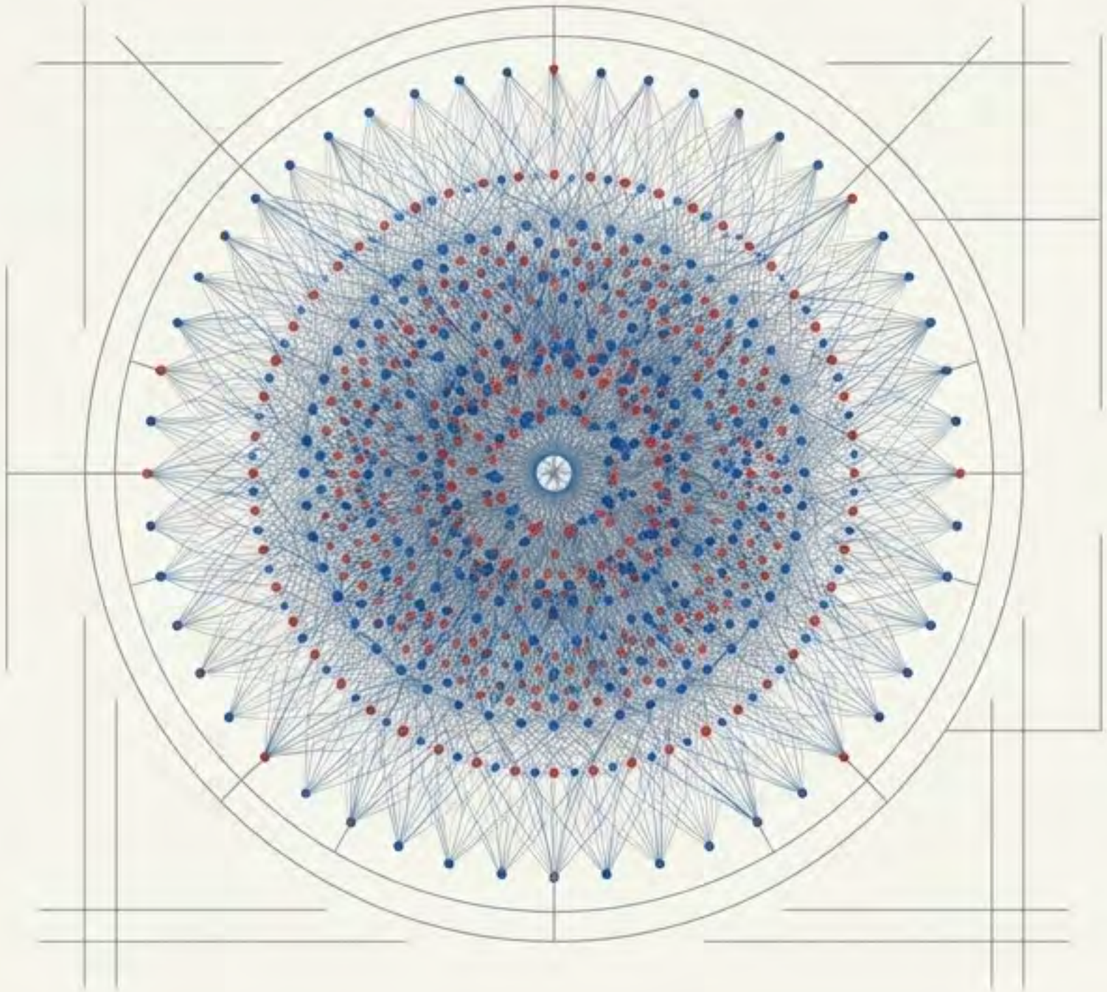


NotebookLM

Pillar 3: 国立国会図書館（NDL）LOD
書誌と典拠データの巨大な知識グラフ

提供データ

国立国会図書館サーチ、次世代デジタルライブラリー、Web NDL Authorities（典拠データ）などのメタデータ。



特徴

日本の出版物に関する最も包括的で権威あるデータベースを、機械可読なLODとして解放。

広がり

地名URIとのリンク(GeoNames.jp連携) や、外部アプリによる書誌データのリアルタイム活用を強力に推進。

NotebookLM

Pillar 4: 統計LOD

社会経済の基盤データをセマンティック・ウェブへ

総務省統計局

統計センター

概要

総務省統計局および統計センターが提供する、国勢調査などどの重要な統計データをLOD化するプロジェクト。

中核

統計に用いる標準地域コード（都道府県・市区町村コード）と、統計データそのもののメタデータを機械判読可能な形で公開。

効果

統計データと、外部の多様なオープンデータを「地域」や「時間」をキーとしてシームレスにリンクさせることが可能に。

NotebookLM

地域コードのLOD化が解決する「時間の壁」

廃置分合（市町村合併）の現実

自治体の合併や市制施行により、地域の境界と名称は常に変化する。従来の固定的なコードでは過去との連続性が絶たれてしまう。

LODによる時間軸の解決

1970年からの変遷履歴をURIでリンク。過去の「庄和町」のデータと現在の「春日部市」のデータを意味的に結びつけ、時系列でのシームレスな統計分析を可能にする。

NotebookLM

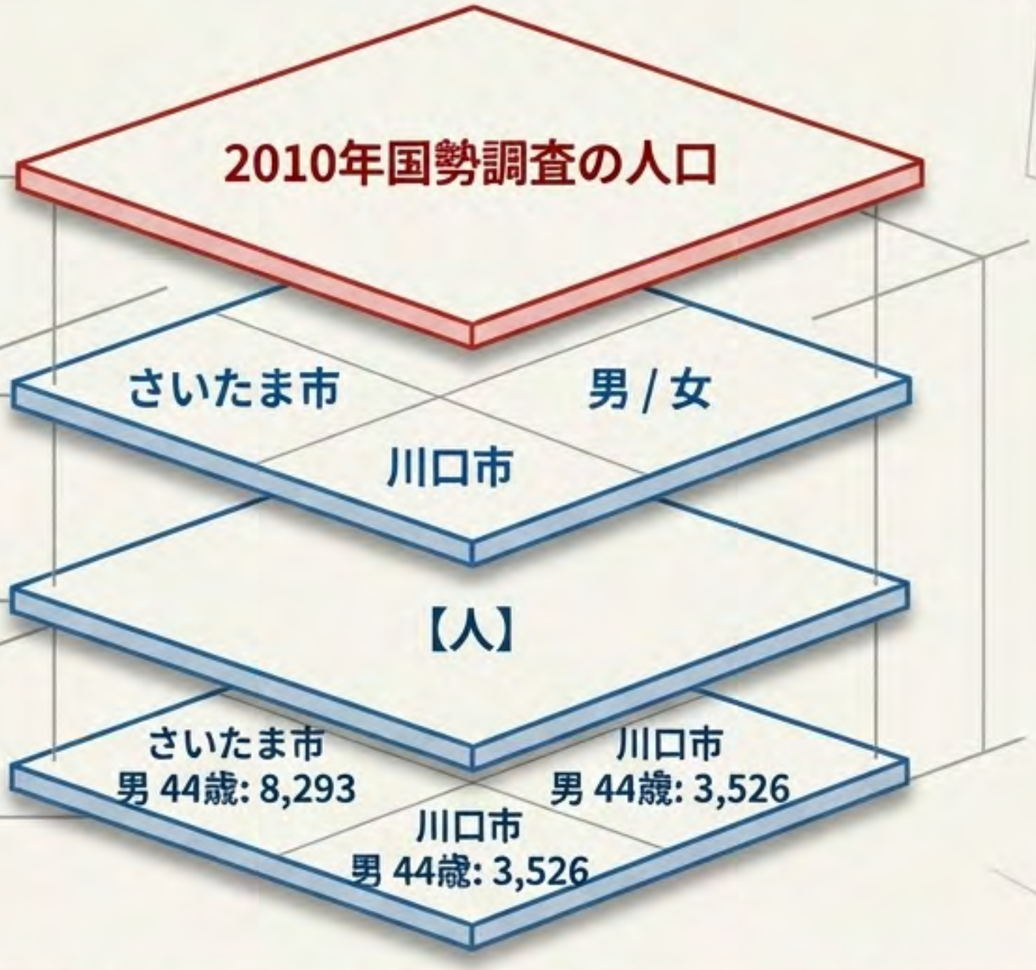
表計算から多次元のLODへ：RDF Data Cube Vocabulary

測度 (Measure)：
何を集計しているか。

次元 (Dimension)：
何で分類しているか。

属性 (Attribute)：
データの単位や状態。

観測値 (Observation)：
実際のデータ。



統計を単なる「表」ではなく、「次元と観測値のグラフ」として再定義することで、あらゆるシステムからの柔軟なクエリ展開を実現。

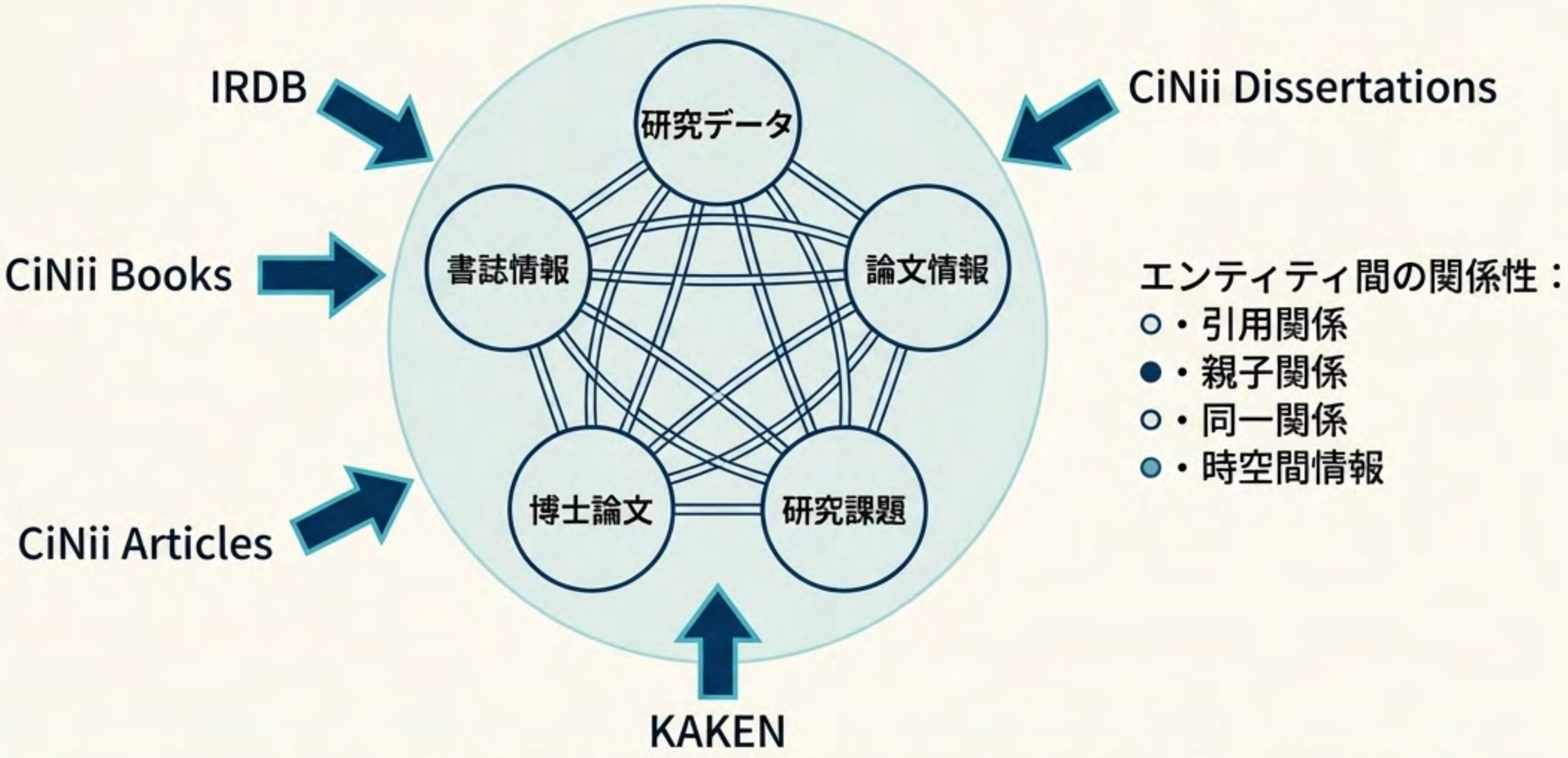
NotebookLM

学術・OAの統合ハブ：CiNii Researchの圧倒的スケール

2021年5月時点のCiNii Research検索エントリ数。既存の学術
サービス群を内部で完全に統合し、オープンアクセスの巨大なハブが誕生。

NotebookLM

エンティティ間の関係性を介した統合化

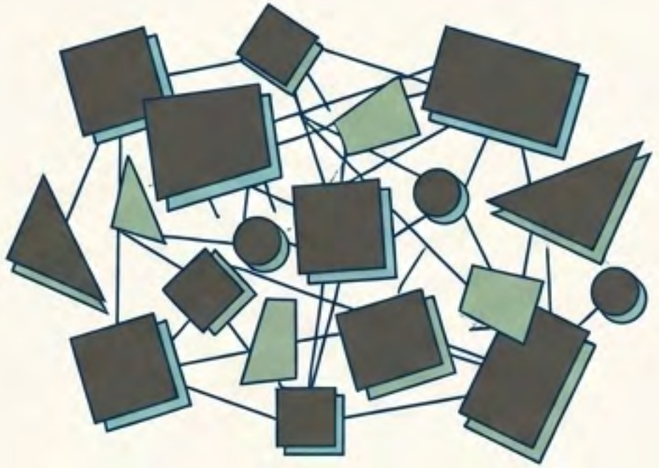


知識発見基盤としての一本化：単なるデータの集約ではなく、エンティティの「関係性」を定義することが次世代プラットフォームの真価である。

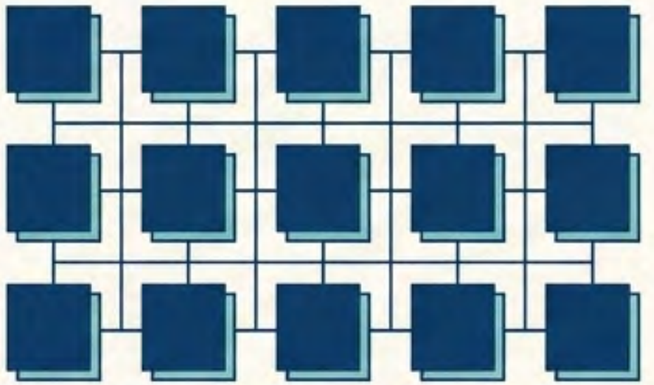
NotebookLM

相互運用性を支える「JPCOARスキーマ」

各大学・機関のリポジトリ
（非標準化データ）



メタデータの標準化と
国際的な相互運用性



- 日本の機関リポジトリのメタデータ規格を標準化
- CiNii Researchの共通スキーマ項目の基盤となり、学術成果の円滑なグローバル流通を実現

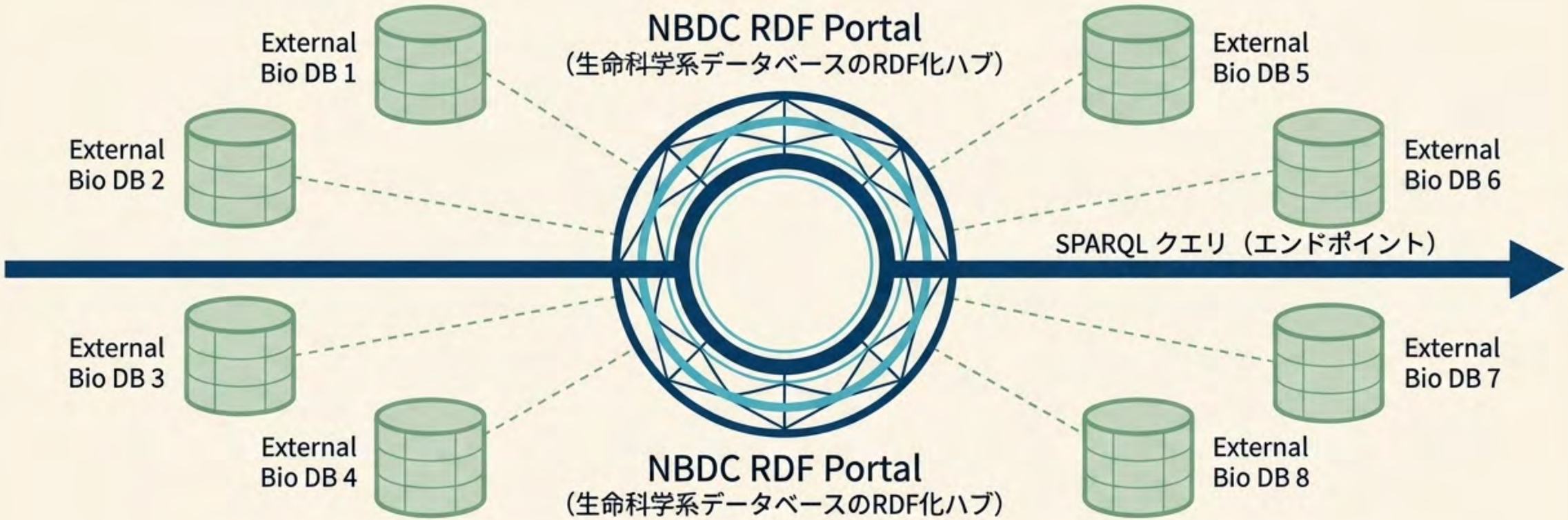
NotebookLM

最前線のドメイン：生命科学（Bio）分野への挑戦

論文と著者を繋ぐのは、知識のグラフの「入り口」に過ぎない。遺伝子、タンパク質、疾患、化合物など、エンティティの種類と関係性が爆発的に増加する生命科学分野こそ、Linked Open Dataの真価が最も試される究極のフロンティアである。

NotebookLM

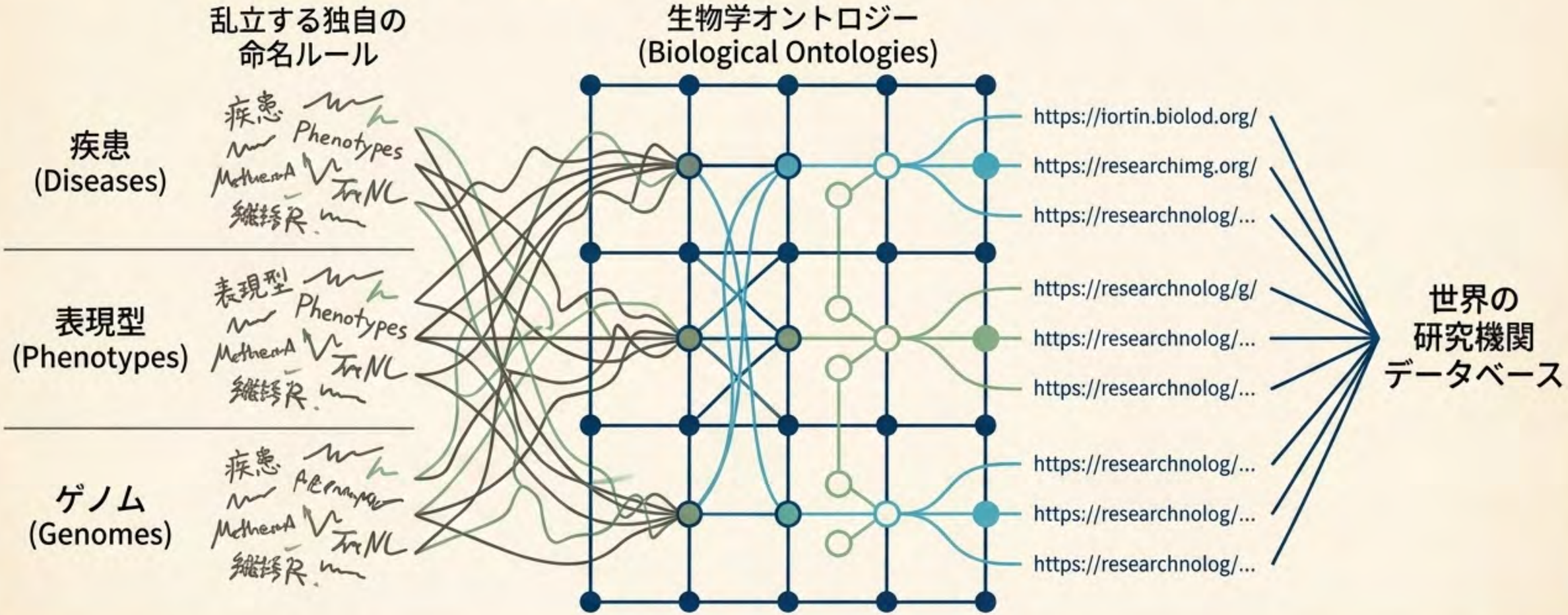
Bio系LOD事例 1：NBDC RDF Portal



従来：	現在：
各種データの個別ダウンロードと手元での面倒な統合	ウェブ上で直接、SPARQLによる意味的・横断的な一括検索

NotebookLM

Bio系LOD事例2：BioLOD.orgとオントロジー



独自の命名ルールが乱立しがちな生命科学用語を「概念体系（オントロジー）」を用いてLOD化。
世界中の研究データと「語彙」を揃え、シームレスなデータ結合を実現する。

NotebookLM

Bio系LOD事例 3：MicrobeDB.jpが促すセレンディピティ

セレンディピティ（偶発的発見）の創出

MicrobeDBを通じて、環境データと遺伝子データなど、全く異なる領域のデータがLODによって繋がり、研究者に新たな仮説構築のインスピレーションを与える。

NotebookLM

統合分析：LODドメイン比較マトリックス

	基盤・文化（NDL等）	統計（e-Stat LOD）	学術・OA（CiNii等）	生命科学（BioLOD等）
対象データ	書誌・文化資源	市区町村・人口動態	論文・プロジェクト・研究データ	遺伝子・タンパク質・疾患
主要スキーマ・語彙	NDLオントロジー	RDF Data Cube	JPCOARスキーマ	各種生物学オントロジー
目指す統合の形	典拠の確立と横断検索	空間・時間軸の機械可読化	エンティティ関係性を介した一本化	異分野データの超高密度リンク

対象や複雑性は異なれど、根本の設計思想はすべて共通。
URIとRDFを用いたLinked Open Dataが、
日本の知のインフラを根本から作り変えている。

NotebookLM