の検出を試みたところ、そのシステムが元々対象としていた一般分野とは異なり、精度が著しく低下することが判明した。Decoref は対象となる言及の抽出と、それらの共参照関係の検出という 2 つのステップで構成されているが、精度低下の原因は、調査の結果、前者の精度が低下していることが大きな原因であることが判明した。我々は、作成したコーパス上で CRF モデルを用い専門用語の抽出システムを訓練し、このシステムによって抽出された専門用語情報で Decoref の言及抽出部分をフィルタリングすることにより（図５）、大幅な精度改善の実現に成功した。

現状では、抽出された専門用語以外の Decoref の言及候補は全て捨てられてしまっているので、これらを何等かの条件の下（C-value/NC-value 等）で追加できるようにすることで、更なる検出精度向上が期待できる。また、他のコーパス（ACL Anthology）に対しても有効であるかどうか検証し、本手法の汎用性を確認したい。

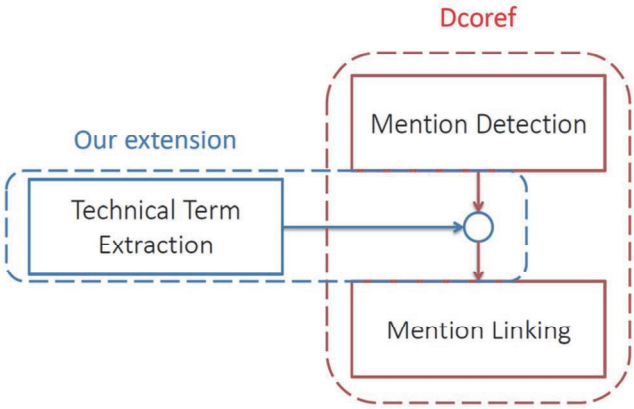


図5：抽出された専門用語による言及検出のフィルタリング

(3-b) 言語横断エンティティリンキングのための語義曖昧性解消

日本語の技術的な文章中の専門用語から、英語ウィキペディア項目へのリンクを付与（言語横断エンティティリンキング）することにより、専門用語の豊富な情報獲得を実現するために必要な語義曖昧性解消手法を開発した。また、手動で構築した評価データを用いて、提案手法の有効性を検証した。

エンティティリンキングは一般に、「用語抽出」と「曖昧性解消」の 2 つの手順からなる（図６）。

① 用語抽出： 対象テキストから候補語となるエンティティ記述（entity mention）を抽出
② 曖昧性解消： 候補語の曖昧性を解消して、リンク先として適切な知識ベースの項目を選定

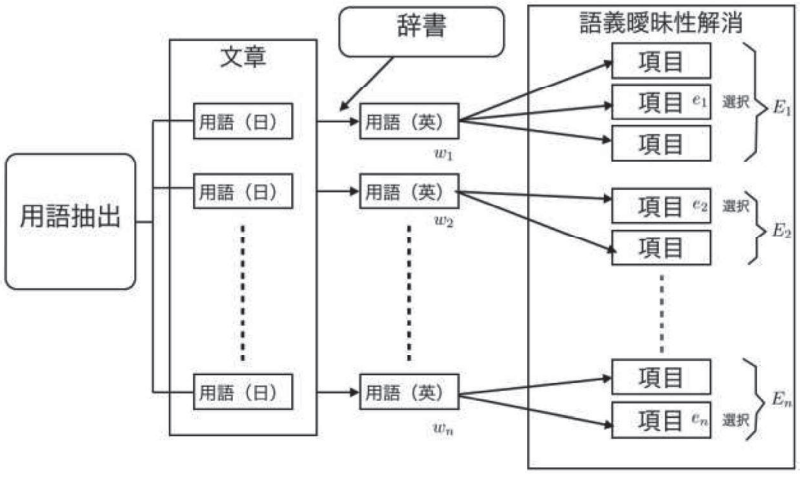


図6：言語横断エンティティリンキングの概要

本研究では前者の用語抽出は行われているものとし、曖昧性解消部分に焦点を絞る。対訳として得られた英用語に対応する Wikipedia の項目は複数存在し得るため、その中から正しい候補を選ぶことを目指す。提案する曖昧性解消手法は「局所的な曖昧性解消に基づくリンクの初期化」「 大域的な曖昧性解消に基づくリンクの再選択」の 2 つのステップからなる。

・　リンクの初期化

対訳として得られたそれぞれの英用語に対して、項目名が一致するウィキペディアのページの中から適切なものを選択する。具体的な選択法としては、Wikipedia の各項目のタイトルにつけられている補助情報を利用する単純なものと、テキスト中の用語集合を文書の象徴とみなし、日本語テキストとの間で、TF-IDF 重みづけで最もスコアが高いテキストを持つ候補項目を初期候補とするものを検討した。

・　リンクの再選択

全体の一貫性を考慮して、より適当なリンクが存在すれば、当該リンクを新たなリンク先として再選択する。これは、専門用語の分野特殊性から、初期リンク集合の間で、ある程度の一貫性が保たれていることを想定したものである。

図 7 は、再選択操作の例を示したものである。左には日本語テキスト中の用語に対する対訳が並んでいる。これら各用語に対し、リンクの初期化により、太枠の項目が初期候補として選択されている。この状態で、各英用語に対する候補項目それぞれに対して、他の英用語の選択候補との類似度を計算し合計した時に、最もスコアが大きい項目を候補として再選択する、ということを行う。類似度の選択には、初期リンク選択時と同様に TF-IDF 法を用いるものと、2 つの項目間で共通のリンク数を類似度スコアとして用いるものを検討した。

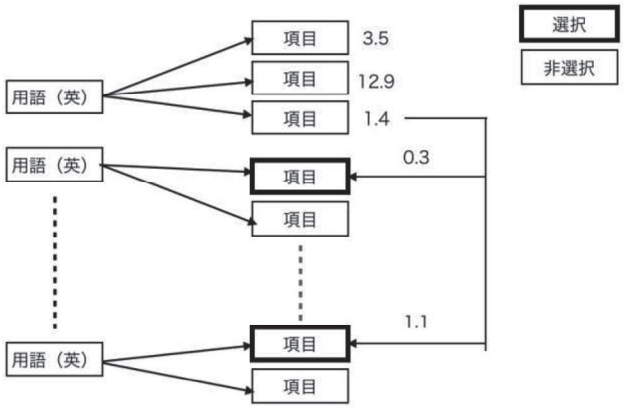


図7:一貫性による再選択操作

実験においては、知識ベースとして Wikipedia 記事をダウンロードし、記事およびセクションから主題、副題、URL、文章、内部リンクを抽出して、無効なレコードを削除した結果、248,027 レコードを得た。評価用データとしては、CiNii の論文抄録データから、英語対訳が 3 件以上存在する用語を一つ以上含む 30 の論文抄録をランダムに抽出した。

リンクの初期化については、TF-IDF を用いた場合の方が単純選択法よりも高い正答率を示した。また、リンク再選択では、TF-IDF を用いる場合、共通リンク数を用いる場合のいずれも、再選択をしない場合より高い正答率を示し、用語の一貫性に基づく大域的な曖昧性解消の有効性が確認された。

TF-IDF 法による初期化と再選択を行う手法について、失敗の大半は、候補の中に正解が存在しない場合であった。確認のため候補の中に正解が存在する語だけを使って同様に実験を行ったところ、正答率大幅に向上した。正解が候補の中に存在しない場合、単に該当する用語から誤った項目にリンクがはられてしまうだけではなく、一貫性の計算の際にも悪影響を及ぼすことが考えられる。したがって、正解候補の有無の判定を、一貫性の計算の前に適用する処理が有効であると考えられる。また、曖昧性解決ページの中に知識ベースの検索では見つからなかった正解項目へのリンクが存在するケースも観察されたため、候補とみなす範囲を広げることで、正解率が向上する可能性もあると考えられる。

サブテーマ２

1. はじめに

学術リソースのためのオープン・ソーシャル・セマンティック Web 基盤の構築プロジェクトにおいては、多様かつ分散的なデータを柔軟に連携させる仕組みについて Linked Open Data(LOD)の方法論に基づき研究を進めている。

Linked Open Data は、RDF などの言語を用いて記述されるシンプルで柔軟性がある仕組みで、多様なデータを記述することができる。そのため、欧米や米国では、既に新しい情報公開・共有の仕組みとして Linked Open Data が認知されつつあり、情報流通の仕組みとして普及しつつある。また、我が国においてもさまざまな研究や活動が行われている。

本プロジェクトでは、実際のデータを取り扱い、LOD 化し公開するという活動を行いつつ、そこで生じる問題を研究課題として解決することで、学術データの LOD 化を推進するという方法で進めている。これまでは、博物館情報の LOD 化、生物多様性情報の LOD 化について取り込んできた。これらも引き続き実施しているが、今年度はさらに統計情報の LOD 化についても取り組んだ。ここでは特にまず統計情報の LOD 化について述べる。次に昨年度の拡張として生物種データと絶滅危惧種データとの統合について述べる。

2. 統計データの LOD 化

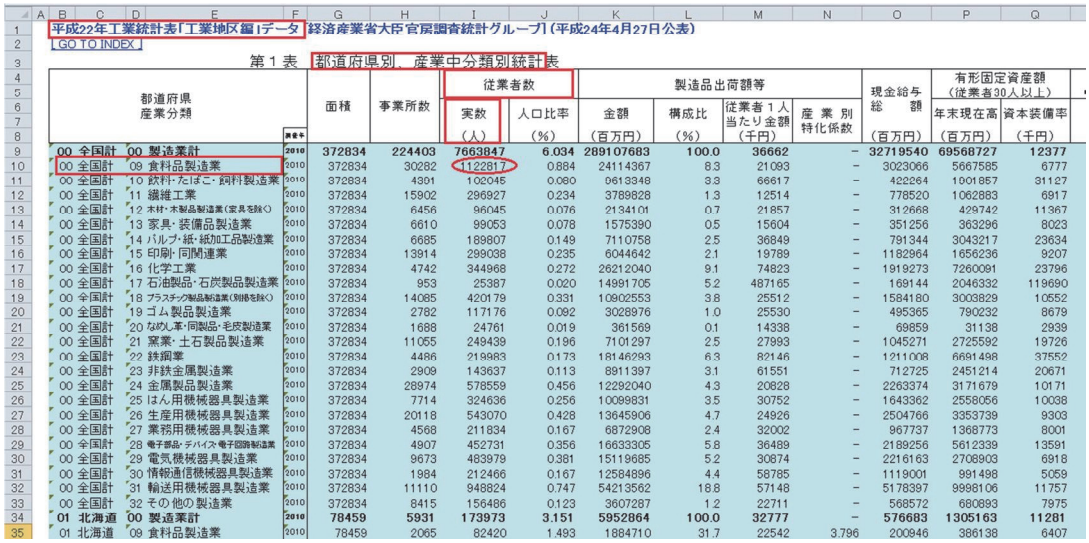
2.1 Open DATA METI

Open DATA METI はオープンデータ実証用サイトであり、前述の通り、経済産業省のデータを公開するサイトとして作成したものである。このサイトには二つの主な機能がある。一つはデータカタログであり、経産省のもつ公開データをリストアップして掲載することで、簡単に検索して、原データにアクセスできるようにするというものである。もう一つは、RDF 化されたデータに関して、SPARQL Endpoint を提供することで、LOD 化されたデータの多様な利用を可能とするものである。データカタログとしては、白書と統計データを中心に、200 件以上のデータセット、1 万件以上のリソースが登録されている。LOD としては、今回その一つである工業統計調査の一部のデータを試験的に変換して、利用可能としている。

2.2 工業統計調査データの LOD 化の論点

国勢調査のような調査活動の結果は、個票と呼ばれる調査結果から個人が特定されるような情報を除いて統計値が計算され、表形式にまとめられ、xls や csv のようなフォーマットで公開される。ここで具体的な統計数値の意味は、それに付随する各種属性のセットで表現されると考える。たとえば、「平成 22 年工業統計表『工業地区編データ　経済産業省大臣官房調査統計グループ』(平成 24 年 4 月 27 日公表)」の「第１表　都道府県別、産業中分類別統計表」(図 1) において、ある Excel セル I10 の意味は、「食料品製造業」という産業分類における全国合計の「従業者数」であるが、このとき表側および表頭にある「食料品製造業」や「従業者数」はもちろんのこと、正確には「全国計」、「実数（人)」に加えて、

他の表も同様な構造を持つために、「平成 22 年工業統計表『工業地区編データ　経済産業省大臣官房調査統計グループ』（平成 24 年 4 月 27 日公表）」や、「第 1 表　都道府県別、産業中分類別統計表」もこのセル I10 の数値の属性として考慮されなければならない。



平成22年工業統計表「工業地区編」データ　経済産業省大臣官房調査統計グループ（平成24年4月27日公表）

[GO TO INDEX]

第１表　都道府県別、産業中分類別統計表

都道府県 産業分類		面積	事業所数	従業者数 実数（人）	従業者数 人口比率（%）	製造品出荷額等 金額（百万円）	製造品出荷額等 構成比（%）	従業者１人当たり金額（千円）	産業別特化係数	現金給与総額（百万円）	有形固定資産額（従業者30人以上） 年末現在高（百万円）	有形固定資産額（従業者30人以上） 資本装備率（千円）
00 全国計	00 製造業計	372834	224403	7663847	6.034	289107683	100.0	36662	－	32719540	69568727	12377
00 全国計	09 食料品製造業	372834	30282	1122817	0.884	24114367	8.3	21093	－	3023066	5667585	6777
00 全国計	10 飲料・たばこ・飼料製造業	372834	4301	102045	0.080	0613348	3.3	66617	－	422264	1001857	31127
00 全国計	11 繊維工業	372834	15902	296927	0.234	3789828	1.3	12514	－	778520	1062883	6917
00 全国計	12 木材・木製品製造業（家具を除く）	372834	6456	96045	0.076	2134101	0.7	21857	－	312668	429742	11367
00 全国計	13 家具・装備品製造業	372834	6610	99053	0.078	1575390	0.5	15604	－	351256	363296	8023
00 全国計	14 パルプ・紙・紙加工品製造業	372834	6685	189807	0.149	7110758	2.5	36849	－	791344	3043217	23634
00 全国計	15 印刷・同関連業	372834	13914	299038	0.235	6044642	2.1	19789	－	1182964	1656236	9207
00 全国計	16 化学工業	372834	4742	344968	0.272	26212040	9.1	74823	－	1919273	7260091	23796
00 全国計	17 石油製品・石炭製品製造業	372834	953	25387	0.020	14991705	5.2	487165	－	169144	2046332	119690
00 全国計	18 プラスチック製品製造業（別掲を除く）	372834	14085	420179	0.331	10902553	3.8	25512	－	1584180	3003829	10552
00 全国計	19 ゴム製品製造業	372834	2782	117176	0.092	3028976	1.0	25530	－	495365	790232	8679
00 全国計	20 なめし革・同製品・毛皮製造業	372834	1688	24761	0.019	361569	0.1	14338	－	69859	31138	2939
00 全国計	21 窯業・土石製品製造業	372834	11055	249439	0.196	7101297	2.5	27993	－	1045271	2725592	19726
00 全国計	22 鉄鋼業	372834	4486	219883	0.173	18146293	6.3	82146	－	1211008	6691498	37552
00 全国計	23 非鉄金属製造業	372834	2909	143637	0.113	8911397	3.1	61551	－	712725	2451214	20671
00 全国計	24 金属製品製造業	372834	28974	578559	0.456	12292040	4.3	20828	－	2263374	3171679	10171
00 全国計	25 はん用機械器具製造業	372834	7714	324636	0.256	10099831	3.5	30752	－	1643362	2558056	10038
00 全国計	26 生産用機械器具製造業	372834	20118	543070	0.428	13645906	4.7	24926	－	2504766	3353739	9303
00 全国計	27 業務用機械器具製造業	372834	4568	211834	0.167	6872908	2.4	32002	－	967737	1368773	8001
00 全国計	28 電子部品・デバイス・電子回路製造業	372834	4907	452731	0.356	16633305	5.8	36489	－	2189256	5612339	13591
00 全国計	29 電気機械器具製造業	372834	9673	483979	0.381	15119685	5.2	30874	－	2216163	2708903	6918
00 全国計	30 情報通信機械器具製造業	372834	1984	212466	0.167	12584896	4.4	58785	－	1119001	991498	5059
00 全国計	31 輸送用機械器具製造業	372834	11110	948824	0.747	54213562	18.8	57148	－	5178397	9998106	11757
00 全国計	32 その他の製造業	372834	8415	156486	0.123	3607287	1.2	22711	－	568572	680893	7975
01 北海道	00 製造業計	78459	5931	173973	3.151	5952864	100.0	32777	－	576683	1305163	11281
01 北海道	09 食料品製造業	78459	2065	82420	1.493	1884710	31.7	22542	3.796	200946	386138	6407

図1：工業統計調査の結果の表の例

一方、LOD における RDF（Resource Description Framework）のモデルは RDF グラフと呼ばれるグラフ構造である。ここでグラフというのは、二つのノードと両者を結合するリンクを最小単位とする構造のことである。RDF グラフはラベル付き有向グラフである。すなわちリンクにも名前が付与されており、そのリンクには方向があるが双方向ではない。RDF のモデルはこのように単純なものであるため、表現方法として柔軟であり、根本的には表形式をこのようなグラフ構造に変換するのになんの問題もない。ただし、知識表現方法として原理的であればあるほど、どんな知識でも表現することはできてもその効率は悪くなる（表現に多くの記述を必要とする）というのが一般的であり、表形式から RDF グラフへの変換時の欠点として、表現効率の悪さが指摘できる。しかし、本来表形式に馴染まないデータで、表形式ではスパースになって 0 値ばかりが挿入されるようなデータでは、逆にグラフ構造のほうが表現効率がよくなる。

なぜ表形式になっているものをわざわざ RDF グラフにするのだろうか？ その利点はなにか？ 実は それは RDF の利点を生かして RDF としての利用を考えない限り、表になっているものをわざわざ RDF に直す必要性はない。一般に RDF の利点として次のような点があげられる。

a) URI を用いてグローバルに唯一な名前を定義し、その意味するところを定義できる。
b) 背後に厳密なモデル論、意味論があるために、誤解の余地がない。またそのため、機械の助けをかりて様々な自動処理が可能となる。
c) 様々なデータの関係を、組織を超えてリンクとして定義できるため、組織を超えたデータの利用が可能となる。

たとえば、市町村の様々なデータがすでに LOD として提供されていた場合、企業など各組織の所有するデータを市町村データとリンクすることで、市町村データを経由して政府統計データと民間のデータなど互いに当初無関係であったデータ間の関係も抽出することが可能となる。

2.3 論点 RDF Data Cube Vocabulary による統計データの LOD 化

統計データの LOD 化については、世界中で関心が高く、W3C の eGovernment Activity の中の Government Linked Data (GLD) Working Group で盛んに議論が行われている。その議論の中から、

The RDF Data Cube Vocabulary[Cyganiak、2013] が、Working Draft として提案されている。今回は基本的にこの語彙に従ってスキーマを設計した。

2.3.1 RDF Data Cube Vocabulary の概要

この仕様では、ISO 標準である統計データの交換規約 SDMX の考え方を取り入れ、それを RDF フォーマットと LOD で扱うための方法を提案している。表を多次元空間でとらえるデータキューブという考え方、コードリスト、データフローなどという用語は、SDMX からきているものである。

「RDF データキューブ語彙」仕様書では、RDF とリンクトデータの手法を用いて統計データを公開可能にすることについて次のような利点を挙げている。

(1) 個々の観測値や観測値のグループが、(ウェブ) アドレス可能になる。それにより公開者と第三者がこのデータを注釈づけし (annotate)、リンク付けすることが可能となる。たとえば、ある報告書が詳細な出典のトレースバックを考慮した特定の図を参照することが可能となる。

(2) データをデータセット横断的に、あるいは統計セットと非統計セットをフレキシブルに組み合わせることが可能になる (たとえば、宗教的寛容さに関連した国民的指標の高い値の国勢調査の領域で、すべての宗教的学校を発見するなど)。統計データはリンクトデータのより広範なウェブの不可欠な一部となる。

(3) リンクトデータとして公開することで、現在静的なファイルのみを提供しているような公開者には、フレキシブルな、かつ非プロプライエタリな機械可読可能な公開の手段を提供することになり、プログラムからアクセス可能なすぐに使えるウェブ API をサポートすることになる。

(4) 標準化されたツールやコンポーネントの再利用が可能となる。

データキューブでは、統計データセットは何らかの論理空間中の点とされる観測値の集まりから構成されると考える。一つのキューブは次元、属性、測度の集まりとして定義される。これらの各要素はデータキューブのコンポーネントと呼ばれる。

(1) 次元コンポーネントは観測値を同定するものである。次元コンポーネントの値の集合は一個の観測を同定する。たとえば一つの観測値には観測された時間や観測がカバーする地理学上の領域が含まれる。

(2) 測度コンポーネントは計測された値であり観察された現象を表現する。

(3) 属性コンポーネントは観測された値を限定し、解釈することを可能にする。それは測度の単位やスケーリングファクタを指定することを可能にし、どんなスケーリングファクタや観測値の状態 (推測値あるいは暫定値) のようなメタデータも指定することもできる。この語彙においては、

2.3.2 LOD 化の例

特定の領域における観測値の次元コンポーネント、測度コンポーネント、属性コンポーネントは、各々 qb:DimensionProperty、qb:MeasureProperty、qb:AttributeProperty のインスタンスであるプロパティを定義して利用する。例えば、あるノードが観測値を表現するものであり、次元コンポーネントとして産業分類を持ち、その次元の値が産業中分類「食料品製造業」であるとき、このノードからこの次元へのリンクをつけるために ktsh: refSangyoChuBunrui という名前のプロパティを次のように定義する。

```
ktsh:refSangyoChuBunrui a qb:DimensionProperty；
rdfs:label "日本標準産業分類(中分類)"@ja；
rdfs:range jsic:JsicConcept .
```

ここから ktsh:refSangyoChuBunrui は次元コンポーネントのためのプロパティであり、それは観測値の次元として jsic:JsicConcept のインスタンスを持つということがわかる。産業中分類「食料品製造業」を http://datameti.go.jp/scheme/jsic/2007/C09 (jsic:C09)という URI を持つノードとしたとき、こ

れは jsic:JsicConcept に型付けされる。

同様に、次元コンポーネントとして都道府県というものがあり、その次元の値として「全国計」を持つとき、ブランクノードからこの次元へのリンクをつけるために、ktsh-sac:refPrefecture という名前の次元コンポーネント・プロパティを次のように定義する。

sac:refPrefecture a qb:DimensionProperty；
rdfs:label "reference area (prefecture)"@en；
rdfs:label "都道府県"@ja；
rdfs:subPropertyOf sdmx-dimension:refArea；
rdfs:range sac:Prefecture；
qb:concept sdmx-concept:refArea .

ここで sac:refPrefecture は smdx-dimension:refArea のサブプロパティであることが述べてある。SDMX 標準には「内容指向のガイドライン（COG)」が含まれる。それは統計上の共通の概念とコードリストを定義しているが、RDF データキューブにおいても、この SDMX の概念を再利用可能とすることを目的に、以下のものが定義される。

```
kougyo:k6-data-j-2000t a qb:DataStructureDefinition；
    rdfs:label "工業統計表「市区町村編」データ（経済産業省大臣官房調査統計グループ）２．市区町村別、産業中分類別統計表(スキーマ)"@ja；
    # dimension
    qb:component [qb:dimension ktsh:refMunicipality; qb:order 1]；
    qb:component [qb:dimension ktsh:refSangyoChuBunrui; qb:order 2]；
    qb:component [qb:dimension ktsh:refYear; qb:order 3]；
    # measure
    qb:component [qb:measure ktsh:numberOfEstablishments]；
    qb:component [qb:measure ktsh:numberOfEstablishments_withBetween30To299Employees]；
    qb:component [qb:measure ktsh:numberOfEstablishments_with300OrMoreEmployees]；
（中略）
    # attributes
    qb:component [qb:attribute sdmx-attribute:unitMeasure; qb:componentAttachment qb:DataSet]；
```

図2：RDFS による表の定義の例

- sdmx-concept: COG 定義の各概念に対する SKOS 概念
- sdmx-code: COG 定義の各コードリストに対する SKOS 概念とそのスキーマ
- sdmx-dimension: 次元として用いられる各 COG 概念に相当するコンポーネント・プロパティ
- sdmx-attribute: 属性として用いられる各 COG 概念に相当するコンポーネント・プロパティ
- sdmx-measure: 測度として用いられる各 COG 概念に相当するコンポーネント・プロパティ

そこで観測値から数値（リテラル）へのリンクのプロパティとして ktsh:numberOfEmployee を qb:MesureProperty の実現として次のように定義する。

ktsh:numberOfEmployees a qb:MeasureProperty；
rdfs:label "従業者数(人)"@ja；
rdfs:subPropertyOf sdmx-measure:obsValue；
sdmx-attribute:unitMeasure ktsh:UnitOfPerson；
rdfs:range xsd:integer .

ただし単位はプロパティ ktsh:numberOfEmployees に定義してあることに注意されたい。もし「従業員数（人口比率)」のプロパティがほしければ、単位を変えたプロパティを別途用意する必要がある。

以上のコンポーネントを用いて、一つの表が定義される。図２に一つの表の定義の例を示す。

2.3.3 コード体系の LOD 化

統計調査においては、調査対象の分類の体系は重要である。統計調査ではそれらの分類にコード（英数字からなる文字列）を割り振ることが多く、コード体系（あるいはコードリスト）と呼ばれる。コード体系は特定の調査のみに出現する系もあれば、特定の統計調査とは別に定義されることもある。工業統計においては、前者の例は工業地区のコード体系であり、後者としては日本標準産業分類や標準地域コードはが例である。

Data Cube Vocabulary においてもコード体系は別に定義する。コード体系は基本的に SKOS (Simple Knowledge Organization System) を用いて表現する。すなわち、skos:broader/narrower で階層的な概念を関係づける。コード体系の要素（概念）は Data Cube Vocabulary の次元コンポーネントのプロパティの値となっている。コード体系の要素をあるクラスのインスタンス（4.2 節の例であれば jsic:JsicConcept）と宣言することで、それらの要素が次元コンポーネントのプロパティの値の候補になる。

2.4 工業統計調査の LOD 化

上記の方針の元に工業統計調査の LOD 化を行った。

工業統計調査とは、統計法に基づき行政機関が実施する統計調査のうち、重要なものとして総務大臣が指定した基幹統計調査の一つで、国が我が国の製造業の実態を把握するために行っているものである。今回はこのうち、平成 22 年（2010 年）の調査結果を利用した。

2.4.1 対象データ

工業統計調査の結果は多数の表として公開されている。それぞれは調査票で集計したデータを様々な切り口で集計・集約したものである。大分類で品目編、産業編、用地・用水編、市区町村編、工業地区編、産業細分類別、企業統計編に別れ、その中に多数の表形式のデータがある。今回のその中の以下の 4 つの表を例題として取り上げた。これらの表の名称と次元と測度を示す 。

(1) 産業細分類別統計表 都道府県別産業細分類別統計表（kougyo_h22-k8-data-j-1003.ttl）
次元：都道府県、産業細分類
測度：事業所数、従業者数、現金給与総額、原材料使用額等、製造品出荷額等、生産額付加価値額、有形固定資産投資総額

(2) 市区町村編　市区町村別、産業中分類別統計表（kougyo_h22-k6-data-j-2000.ttl）
次元：市区町村、産業中分類
測度：事業所数、従業者数、現金給与総額、原材料使用額等、製造品出荷額等、粗付加価値額、有形固定資産年末現在高

(3) 産業編 都道府県別、東京特別区・政令指定都市別統計表 (2) 従業者 30 人以上の事業所に関する統計表 ②産業中分類別の在庫額、有形固定資産額及びリース契約による契約額及び支払額（kougyo_h22-k3-data-j-3220.ttl）
次元：都道府県、産業中分類
測度：在庫額、有形固定資産額、リース契約による契約額及び支払額

(4) 用地・用水編　第 1 部　事業所数、従業者数、製造品出荷額等、事業所敷地面積、建築面積及び延べ建築面積表　4．工業地区別、産業中分類別統計表(kougyo_h22-k4-data-j-1400.ttl)
次元：工業地区、産業中分類
測度：事業所数、従業者数、製造品出荷額等、事業所敷地面積、事業所建築面積、事業所延べ建築面積

このうち、次元においては「都道府県」(1、3)、「工業地区」(4)、「市区町村」(2)は階層的な関係であり、「産業中分類」(2、3、4)と「産業細分類」(1)も同様である。これらの次元はそれぞれコ

ード体系として表の表現とは別に用意される。
測度についてもいくつかの項目は共通である。例えば(1)と(2)は測度は同じである。(1)(2)と(4)でも共通の測度（製造品出荷額等）があるが、単位が異なる（前者が万円単位、後者が百万円単位）なので測度コンポーネントは別に定義する必要がある。

```
  #北海道の産業中分類別、有形固定資産土地（百万円）と従業員数
  PREFIX
ktsh:<http://datameti.go.jp/scheme/kougyou-toukei-schema/>
  PREFIX kougyo:
<http://datameti.go.jp/lod/kougyou-toukei/>
  PREFIX qb:                    <http://purl.org/linked-data/cube#>
  select distinct   ?sanchu_label ?total_jugyoin ?landprice
  where
  {
    {
      select distinct ?sanchu (SUM(?jugyoin) AS ?total_jugyoin)
      where
      {
        ?cell1          qb:dataSet kougyo:h22-k8-data-j-1003 .
        ?cell1 ktsh:refSangyoSaiBunrui ?sansai .
        ?sansho skos:narrower ?sansai .
        ?sanchu skos:narrower ?sansho .
        ?cell1 ktsh:refPrefecture
       <http://datameti.go.jp/scheme/standard-area-code/C01> .
        ?cell1 ktsh:numberOfEmployees ?jugyoin .
      } Group by ?sanchu
    }
    ?cell2 qb:dataSet kougyo:h22-k3-data-j-3220 .
    ?cell2 ktsh:refSangyoChuBunrui ?sanchu .
    ?cell2 ktsh:refPrefecture
  <http://datameti.go.jp/scheme/standard-area-code/C01> .
    ?cell2
  ktsh:valueOfTangibleFixedAssets_purchase_lands_byMillionY
  en   ?landprice .
    ?sanchu rdfs:label ?sanchu_label .
  }
```

図3：SPARQL Query の例

2.4.2　RDF 化とクエリ

上記の 4 つの表を RDFS を用いて定義して、データを RDF として表現した。スキーマの定義を含めて、全体で 2、827、017 トリプルとなった。このデータは Open Data METI サイトでダウンロードあるいは SPARQL Endpoint を通じてアクセス可能である。

このデータに対して行う SPARQL Query の例を図 3 に示す 。この例では異なる表のデータを一つのクエリの中で参照して、統合した解を返すようにしている。(1)の表から北海道における産業細分類別の従業員数のデータをとり、産業中分類に集約すると共に、それに対応する有形固定資産土地の値を(3)の表から集めている。

2.5 考察

今回、工業統計調査を例に取り、Data Cube Vocabulary の方法に則って統計データの RDF 化を行った。Data Cube Vocabulary の方法では個別のデータそのものをノードと表現し、そのプロパティとして次元や測度を表現する。このため表にまたがって次元を指定して検索したりといった柔軟性の高い利用方法が可能となる。例えばコード体系が共通なら、一つの統計に限らず複数の統計からのデータをつなげて使うことも可能になる。ただし、このためには次元や測度、コード体系を共通に利用できるようにきちんと定義する必要がある。

現実的な課題としては定義ためのコストが高いことおよびデータ量を大きくなることがあげられる。

このため、柔軟性の高い利用（横断的利用など）が期待されるデータを中心に RDF 化を行うことが実際的な方策だと思われる。

3. 絶滅危惧種情報のデータ化と統合

3.1 はじめに

昨年度までに構築している生物種の LOD を実際に利用する例として、絶滅危惧種データとの統合を行った。

絶滅危惧種に関するデータとしてはレッドリストとレッドデータブックが有名である。それぞれ、国際機関、国、都道府県など異なるレベルで作られ、大抵の場合、名前と保全状況についての最低限の情報を速やかにレッドリストとして編纂し、後に詳細な情報を加えてレッドデータブックとして発行するという手順を踏んでいる。我々はまず、国レベル、都道府県レベルでのレッドリストのデータ化を試みた。具体的には環境省の生物多様性センターが 2012 年から 2013 年にかけて編纂した第 4 次レッドリスト（以下、環境省レッドリストと呼ぶ）と、2013 年に編纂された京都府レッドリスト[2]（以下、京都府レッドリストと呼ぶ）をデータ化した。

環境省は学名、和名、保全状況、その種のカテゴリーといった情報を公開している。そこで学名「Oceanodroma castro」を持つ種が、和名として「クロコシジロウミツバメ」を持ち、「絶滅危惧 IA 類（CR）」の保全状況で、「鳥類」にカテゴライズされていることは Turtle 形式の RDF で表 1 のように記述できる。前半は学名を主語としたトリプルであり、species オントロジの語彙を用いて和名とカテゴリを記述している。カテゴリに関してはデータ源の記載を尊重してそのまま記載している。例えば「汽水・淡水魚類」といったカテゴリが環境省レッドリストに存在するが、そもそも魚類というカテゴリ自体が、系統分類学的ではない用語であり、体系の異なる用語をマッピングすることは困難であるし、また LOD 化の段階でするべきではないと考えている。それを speciesOnto:hasSuperTaxon というプロパティで単に上位の分類群として登録することで、異なる分類体系の情報を共存させることを可能にしている。実際、この種については LODAC Species 内の既存の情報と統合され、NCBI や DBpedia といった他のデータベースへのリンクや、ミズナギドリ目ウミツバメ科に属するといった他の上位分類群の情報、[EOL] 等から取ってきた関連する写真が閲覧できるようになっている（図４）。また学名を主語としたトリプルの一つとして、保全情報オントロジ語彙の cnsvOnto:hasRedListEntry プロパティを用いてレッドリストのエントリを参照しており、その項目の情報を後半で記述している。このように情報を分けているのは、複数のレッドリスト情報が一つの種について存在し得るからである。そして、レッド

表１： レッドリスト１項目のデータ化例

```
<http://lod.ac/species/Oceanodroma_castro> a speciesOnto:ScientificName;
SpeciesOnto:hasCommonName <http://lod.ac/species/クロコシジロウミツバメ>;
speciesOnto:hasSuperTaxon <http://lod.ac/species/鳥類>;
rdfs:label "Oceanodroma castro";
cnsvOnto:hasRedListEntry redlist:jibis-redList2012_tyorui-17.
redlist:jibis-redList2012_tyorui-17 a cnsvOnto:RedListEntry;
rdfs:comment "クロコシジロウミツバメ"@ja;
cnsvOnto:ofSpecies <http://lod.ac/species/Oceanodroma_castro>;
cnsvOnto:currentStatus cnsv:CR;
cnsvOnto:ofArea "日本"@ja.
```

リストのエントリを主語としたトリプルで具体的な保全状況や指定されている地域などの情報を記述している。

京都府レッドリストについても同様であるが、そちらは学名が記載されていなかったため、和名を主語として情報を記述した。

統合に際しては、LODAC Species が名前ベースのアーキテクチャを採用し、各生物種に対応する URI が http://lod.ac/species/種名のような形式になっているので、種名が一致すれば自動的に統合されるようになっている。

3.2 結果と考察

3.2.1 統合に成功した比率

環境省レッドリスト 5690 件には学名と和名が存在するが、それぞれを用いて LODAC Species との統合を試みたところ、学名を介して 3294 件（57.9%）が既存のデータと統合された。一方、和名を介しては 4145 件（72.8%）の統合に成功した。その和集合は 4711 件（82.8%）となっている。京都府レッドリスト 1871 件については和名のみを用いたが、1598 件（85.4%）という比較的高い割合で統合に成功した。全体的に高い割合で統合に成功したのは、生物の種名が比較的ゆれがなく使われていること、LODAC Species のデータが広い範囲をカバーしていることを示している。和名の方が効率的に統合できた一因は、和名の方が表記ゆれが少ないことではないかと考えている。

3.2.2 失敗例の分析

元のデータと統合できなかったデータについて、統合すべきデータがそもそも入っていなかったのか、データがあるにもかかわらず何らかの理由で統合に失敗しているのか、を正確に知る術は無い。しかし、統合に失敗した名前について調査することで、より統合率を向上させるのに役立つ知見が得られたので、以下に報告する。

(1) 統合すべきデータが入っていなかったと思われるもの

例えば、京都府版レッドリストにある「ヨドゼゼラ」に関しては、2010 年に新種「Biwia yodoensis」の発見の論文が出ており、この論文の著者である細谷が琵琶湖生物多様性画像データベースに和名として「ヨドゼゼラ」を記載していることから、比較的新しい種であるために、元のデータの中に対応する種が含まれていなかったと考えられる。また、同じく京都版レッドリストにある「ルイスムネボソヨツメハネカクシ」については、「Boreaphilus lewisianus」という学名が付けられている種の発見自体は 1874 年と古いが、和名がつけられたのが柴田らによって 2013 年に編纂された『日本産ハネカクシ科総目録』においてであるため、元のデータの中に対応する種が含まれていなかったと考えられる。

これらの例については、データベースや図鑑、目録といった形で発行される新しい情報を積極的に入力する他、そのフローを効率化するために各分野の分類学者と協力していく必要がある。

(2) 複数の和名を持つ種

京都府版レッドリストにある「イモリ」や「モモンガ」は族や科の名前である一方、種として「モモンガ」といった場合には「ニホンモモンガ」（「ホンドモモンガ」ともいう）のことを指し、種として「イモリ」といった場合には「アカハライモリ」を指す。例えば、Wikipedia の生物分類表テンプレートには「和名」という項目があり、それぞれの種について「アカハライモリ、ニホンイモリ、イモリ」「モモンガ、ニホンモモンガ、ホンドモモンガ」という形で名前が列挙されてい。また、イモリについては京都府がウェブ上に公開している 2002 年版レッドデータブック[11]において学名「Cynopus pyrrhogaster (Boie、1826)」が付されているため、「Cynops pyrrhogaster」という名前を持つアカハライモリであると推測され、それは環境省のレッドリストにおいても「準絶滅危惧（NT）」指定されていることがわかる。これらの例

については、DBpedia や他のレッドリストデータから同じ種を指す複数の和名を抽出し、それらを関連付けることで解決できると考えられる。

(3) ミススペルと表記ゆれ

前述の例の前半「Cynopus pyrrhogaster」と同様の学名を用いた論文、文書はウェブ上に他にも存在したが、タンパク質に関するデータベースにおいて、Cynopus は Cynops のミススペルであるとされている。また、発見者名や年号を後ろにつけるのは学名において多くみられる表記法であり、1700 万以上の学名についての情報を提供している Global Names Index では、「Cynops pyrrhogaster」「Cynops pyrrhogaster (Boie, 1826)」「Cynops pyrrhogaster Boie」の 3 つを表記グループ（Lexical groups）としてまとめている。一方、和名についてはこういった形の表記ゆれはないことが、前述のように効率的に統合できた一因だと考えられる。また環境省レッドリストにある「Aerobryum speciosum (Dozy & Molk.) Dozy & Molk. var. nipponicum Nog.」については、非常に近い表記の「Aerobryum speciosum (Dozyet Molk.) Dozy et Molk. var. nipponicum Nog.」が米倉らの作成した YList に記載されており、それが LODACSpecies にも含まれているが、記号表記「&」とラテン語表記「et」の差異やスペースの数の差異といった細かな差異によって統合に失敗している。これらの例については前節と同様、ミススペルや表記ゆれについて扱っているデータベースの情報を追加する他に、すでにあるエントリとの表記上での類似度を計算し、統合先として推薦するという手法が考えられる。我々はすでにオープンソースの検索エンジンである Apache Solr を用いて類似の文字列を検索するシステムを実装しており、今後異なる情報源からのデータの統合の際にこのエンジンを活用したいと考えている。

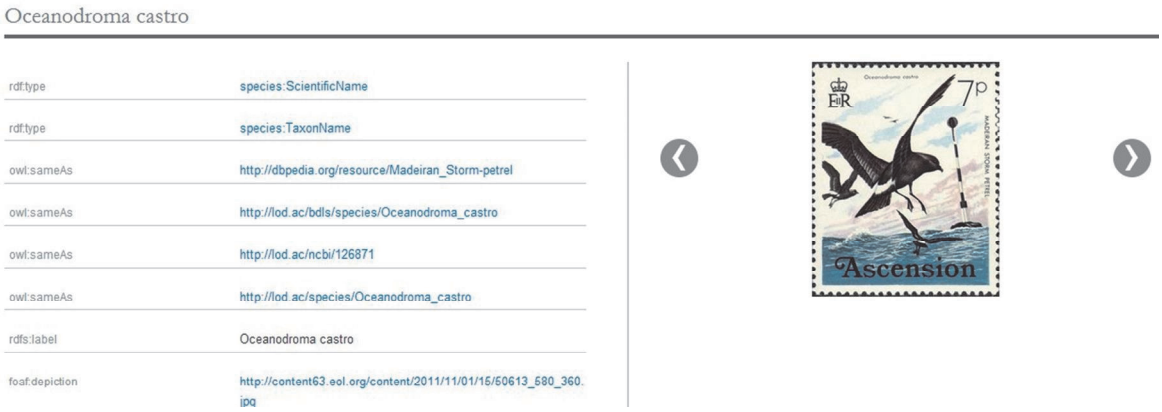


図4：生物種一項目の HTML 表示例

(4) 同名異種

統合できたものについては統合成功としているが、その統合が適切だったかどうかについては十分に検討できていない。異なる種に同じ学名が付けられることはほぼ無いと思われるが、和名については昆虫類のカマキリと淡水魚類のアユカケの別名であるカマキリが同名になってしまっているといった例が実際に存在し、それらを区別する仕組みが必要と考えられる。

3.3 おわりに

本研究では、生物情報を共有する LOD 基盤として構築を進めてきた LODAC Species へ、生物多様性に関する重要なデータである絶滅危惧種情報の統合を行った。学名や和名を手がかりとして多くの絶滅危惧種情報を既存の情報と統合できた一方、失敗例の分析を通して、より効率的な統合のためにデータの追加や統合手法の改善の必要性が示された。