生命科学系公開データベースを対象とした クローリングワークフローの提案

A proposal of workflow for crawling public life science databases

大波 純一[1] 杉崎 太一朗[2] 牧口 大旭[2] 宮崎 敦子[1] 三橋 信孝[1] 畠中 秀樹[1] 川本 祥子[3] 高木 利久[1,4,5]

Jun-ichi Onami[1], Taichiro Sugisaki[2], Hiroki Makiguchi[2], Atsuko Miyazaki[1], Nobutaka Mitsuhashi[1], Hideki Hatanaka[1], Shoko Kawamoto[3], and Toshihisa Takagi[1,4,5]

[1] 科学技術振興機構
[1] Japan Science and Technology Agency, Japan
[2] 三井情報株式会社
[2] Mitsui Knowledge Industry Co., Ltd. , Japan
[3] ライフサイエンス統合データベースセンター
[3] Database Center for Life Science, Japan
[4] 東京大学
[4] The University of Tokyo, Japan
[5] 国立遺伝学研究所
[5]National Institute of Genetics, Japan

Abstract: Life Science Database Cross Search collects (=crawls) Japanese public databases for providing knowledge finding system. This database crawling needs specialized protocols to find data in deep-web sites, to select appropriate data range for user, and to make rule-based decision. This report shows methods and its results.

１．背景

Linked Open Data としてある一定の基準を満たしたデータは LOD cloud diagram [1]としてデータ間のクラウド図が公開されている。生命科学関連のデータベースはクラウドを構成する 295 件中、40 件という大きな割合を占めており（2011 年時点）、生命科学の研究分野では日々、ゲノム情報や各種実験情報の公開やデータベース間のリンク設定が精力的に進められている。多様な生命科学の公開データベースから利用者が必要とするデータを探し出し活用するためには、検索エンジンによる情報の探索が有効である。2008 年に大学共同利用法人情報・システム研究機構（ROIS）ライフサイエンス統合データベースセンター（DBCLS）は、文部科学省「統合データベースプロジェクト」のサポートにより「生命科学データベース横断検索」（以下、横断検索と記述）[2]を開発し公開した。これは、主に日本国内の生命科学系分野のデータベースに含まれるテキスト情報を横断的に検索することができるサービスである。2011 年にはこのサービスは独立行政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）へ移管され、現在も運用が継続されている。2014 年 7 月時点で、444 件の生命科学系公開データベースの 1000 万件以上のエントリを検索することができ、随時追加中となっている。

同様の生命科学系情報検索基盤としては、表 1 に示したサイトが存在するが、インターネット上に分散する生命科学分野のデータベースを数百件以上の単位で網羅的に検索できるのは本サービスのみである。また Web 上に散在する「データベース」のみを検索するサービスもあまり見られない。Google や Yahoo!の一般的な検索エンジンでは、ドメイン単位に絞った情報検索を行うオプションである Site

表１．主要な生命科学系データベースの横断検索サービス

	タイトル	URL	運用機関	検索対象の説明
1	National Center for Biotechnology Information（NCBI）ポータルサイト	http://www.ncbi.nlm.nih.gov/	National Center for Biotechnology Information	米国ファンディングで算出されたデータやゲノム配列情報を同サーバ内に集約して検索
2	Kyoto Encyclopedia of Genes and Genomes（KEGG）	http://www.genome.jp/kegg/	京都大学	京都大学、もしくは日本国内の機関と共同で構築した数十のデータを集約し検索
3	Sagace - 医薬基盤研究所統合検索データベース	http://sagace.nibio.go.jp/	独立行政法人 医薬基盤研究所	本プロジェクトと共同で管理・運用を行っているため共通するインデックスを使用した検索サービス
4	Medals 横断検索	http://medals.jp/xsearch/	独立行政法人 産業技術総合研究所・創薬分子プロファイリング研究センター	
5	JCGGDB 横断検索	http://jcggdb.jp/search/search.cgi?keyword=&lang=jp	独立行政法人 産業技術総合研究所・糖鎖医工学研究センター	
6	NIAS 横断検索	http://agrid.dna.affrc.go.jp/cgi-bin/sogo.cgi?class=719	独立行政法人 農業生物資源研究所	

search が利用可能であるものの、検索対象データの網羅性[3]や結果の定常性[4]が十分ではないため、学術的な利用には難があると考えられる。

このように世界にあまり類を見ない横断検索でも、一般的な Web 検索エンジンの仕組み[5]と同様に、インターネット上に公開されている各データベースを発見し、含まれる全データをサーバ内にダウンロード（＝クローリング）する作業が必要となる。しかしインターネット上の学術論文のクローリング[6]や画像を含むブログのクローリング[7]、特定言語サイトのクローリング[8]等に対する考察は存在するが、データベースのクローリングに関する検討はあまりなされていない。これは一般的な Web ページのクローリングと比べ、後述の 3 つの理由から困難であるためと考えられる。１つ目の理由は各エントリの URI が動的なサイトにいくつかのパラメータを加えた形式であることが多く所謂ディープウェブ[9]の位置にあり確認することが難しいためである。また 2 件目は、１つのデータベースが持つエントリが非常に多い場合（データベースによっては数百万のエントリを持つ）、そのクロールがそのままサーバへの高負荷に繋がることがあることから、サーバ管理者にとってクローリングは望ましくないと判断され、サイトごとの規約や robots.txt に検索エンジンによるクローリングを控えるよう記述されることがあるためである。また 3 つ目の理由は、生命科学系データベースの形式は、データベースごとに有限個の対象（例えば生物種や細胞の種類など）を扱うものだけでなく、実験測定データの予測数値のような無限に近いデータの組み合わせが導き出されるものなど、多様なデータの範囲や内容について総合的に判断できる専門知識が必要となるためである。以上のような「A.技術的な条件（エントリ URI の位置不明）」、「B.法的な条件（クロールの制限）」、「C.学術的な条件（生命科学分野の知識の要求）」を漏れなく解決するため、横断検索では以下に示す独自の手順で検索対象を検討し、検索対象の追加作業を行っている。対象データベースや検索範囲の選定基準と、実際に行われているワークフロー内容について紹介する

２．検索対象データベースの選定

プロトコルの説明の前に検索対象のデータベースについて記述する。生命科学分野では、主に日本国内のデータベース情報を網羅的に収集し解説が記述されている Integbio データベースカタログ[10]が利用可能であるため、これに記載されたデータベースを検索対象とする方針としている。検索する対象のデータエントリの 1 例を図 1 に示す。図 1 の内容は本論文の例示のために作成した架空の情報である。このデータに対し、利用者は例えば「肝硬変　マウス　蛍光強度」などの複数キーワードで検索を行うことがあり得ると考えられる、また「nbdctestid0110」など ID を既に知っていてその情報に直接アクセスすることを目標として検索が行われる可能性がある。また結合配列や作成者の情報などでデータにアクセスしたいこともあるだろう。このようにデータベース内部を探す際に必要なユニークデータは、図 1 で示されるデータのテキスト部分が該当する。また「データベースエントリ ID」という項目を全文検索対象

に入れてしまうと、利用者が「トリ」（鳥）で検索した時に大量の項目名がノイズとして検索結果に混入する恐れがある。よって項目名はテキスト検索対象としない方針だが、ユニークデータを規定する属性名の情報として取得する。例えば「サンプル生物種」だけでは検索対象としての意味は薄いが、「サンプル生物種：*Mus musculus* X 系統」のような組のデータになっていれば、セマンティックな検索のために有用となる。また図１の「画像」項目として示したようなバイナリデータはユーザのテキスト一致検索で取得することはできない。また横断検索では画像類似検索等も実装はしていない。この一方で検索結果を視覚的に分かりやすくするサムネイル画像を表示させるような機構が組み込まれているため（図２中央部左側参照）、このような画像項目は「画像の URL 情報」を横断検索内のデータストアに取得している。結合配列の項目で示したような塩基配列（A、T、G、C、N などから成る文字列）やアミノ酸配列（各種アルファベットから成る文字列）は、生命科学分野の多くのデータベースで含まれているが、これらの配列は生物個体ごとに頻繁に突然変異が起こるため完全一致の検索を行う場合は少ない。これらの遺伝情報の配列に対しては BLAST などのある程度類似性のある配列を探索するプログラムで確認するのが一般的である。横断検索では現在 BLAST は実装していないがこれらの配列も検索インデックスに含めている。以上のように、横断検索では検索対象データとして、「エントリごとのユニークデータ（テキスト（自然言語、数値）・バイナリデータ）」を取得している状態である。

３．データ取得のワークフロー

データベースを横断検索へ取得する場合、実際に検索エンジンのインデックス情報として使用するため図３で示すようなワークフローでの判断を実施する。この 17 ステップは、１．背景で説明した特殊な対応が必要となる箇所「A.技術的な条件」、「B.法的な条件」、「C.学術的な条件」への対応が必要な部分を、定型化しフローとしてまとめたものである。各項目の作業は、必要な専門知識を有する横断検索の担当者数人が、キュレーターとして実施する。以下 17 ステップの各プロセスについて記述する。

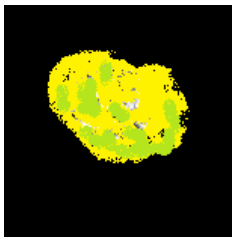
データベースエントリ名	肝硬変マウスの肝臓組織における、ある配列プローブの蛍光強度
データベースエントリ ID	nbdctestid0110
作成日	2014/6/1 12:00JST
更新日	2014/7/1 12:00JST
測定日	2014/5/1 12:00JST
サンプル生物種	*Mus musculus* X 系統
サンプル組織	肝臓
測定機器名	A社製B型測定器 a
最大蛍光強度	0.6
画像	
結合配列	ATGCATGCATGCATGCATGCATGC
作成者	C山D太郎
関連疾患	肝硬変
測定者所見	ミトコンドリア周辺域に偏在
別データベース ID	nbdcbetsuid2201
言及論文	Pubmedid:XXXXXX
備考	本サンプルはEE大学にて凍結保管され追試可能

図１　生命科学系データベースエントリの一例

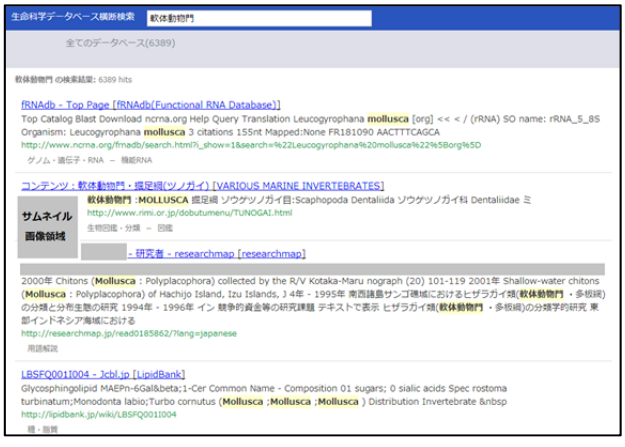


図２　横断検索の検索結果の一例
「軟体動物門」をクエリとした横断検索の結果（http://biosciencedbc.jp/dbsearch/result.php?lang=ja&phrase=軟体動物門）を示す。本研究の内容と関連しない研究者の個人名・所属・データベース情報については画像の中央部灰色領域に修正を入れた。

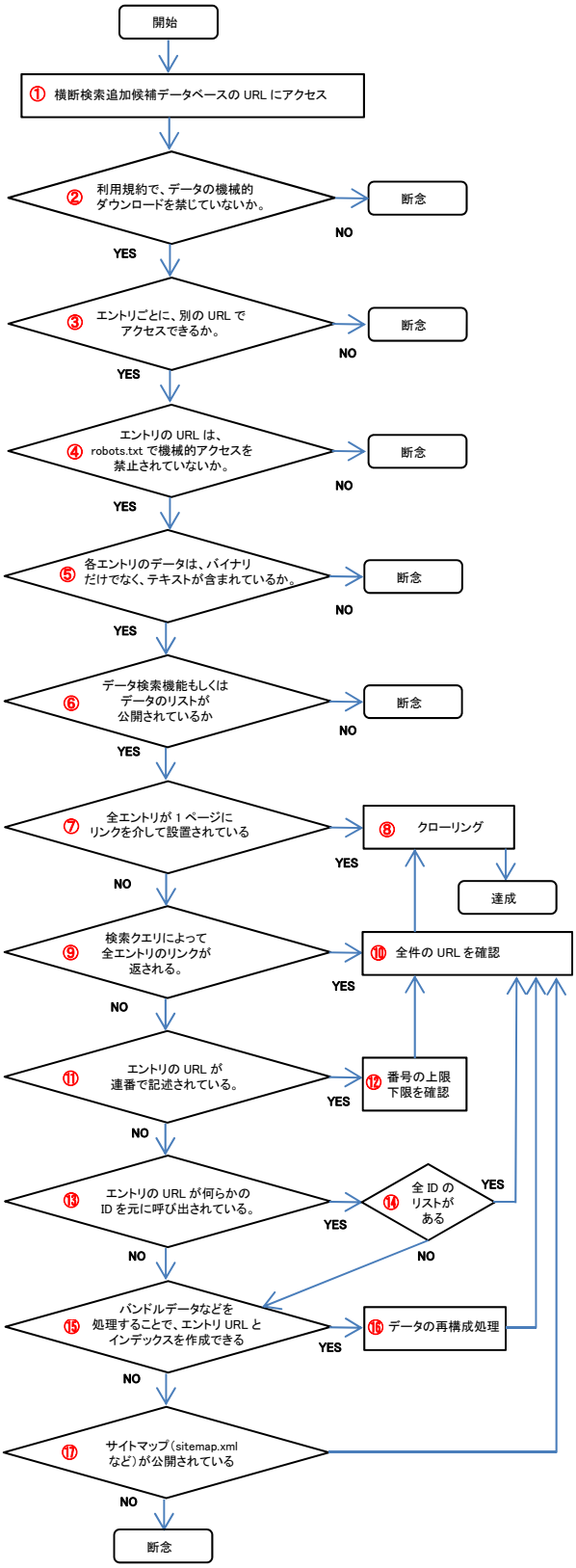


図３　17 ステップのデータ取得検討フロー

① 横断検索追加候補データベースの URL にアクセス

最初に Integbio データベースカタログ[10]の項目として登録された対象のデータベースサイトのトップページにアクセスし、データが公開されていることを確認する。Web ブラウザは一般的に広く使用されている Internet Explorer、Firefox、Google Chrome、Safari 等の最新版を使用する。データベースサイトによっては一部の Web ブラウザでしか表示されない場合もあるため、複数の環境を使用し、担当者が直接アクセスして状況を確認する。ここで対象が Web で公開されているという前提条件をクリアする。

② 利用規約で、データの機械的ダウンロードを禁じていないか

「B.法的な条件」への対応として、対象データベースのサイトにアクセス後、「利用規約」や「サイトポリシー」、「よくある質問」、「注意事項」などと記述されたページがあれば内容を確認し、データの扱いに関する基準を参照する。例えばここで「自動的または組織的なダウンロードは禁じます」などと記述されている場合はその文面に従い本横断検索の検索対象としないものとする。但し管理者と直接コンタクトを取り機械的アクセスの了承が得られている場合や運用機関が検索エンジンのサイト自身である場合はこの限りではない。図３のフローチャートではこの答えを YES か NO を辿り、NO だった場合は「断念」オブジェクトへ進みフロー完了とする。

データベースが著作物であるかどうかについては、その内容により法的判断は分かれる状態となっている[11]。著作権法第二条（定義）十の三 では、「データベース　論文、数値、図形その他の情報の集合物であつて、それらの情報を電子計算機を用いて検索することができるように体系的に構成したものをいう。」というデータベースの定義を元にし、著作権法の第十二条の二（データベースの著作物）「データベースでその情報の選択又は体系的な構成によつて創作性を有するものは、著作物として保護する。」と記述されている範囲の物は著作物であるとみなしている。本横断検索では念のため厳密な法解釈に従うため、対象データベースは基本的には著作物として扱っている。「著作権法施行規則の一部を改正する省令　（平成２１年省令第３８号）」では、著作権の侵害とならない範囲で公衆からの求めに応じ、検索エンジンサービスを提供することが可能だと判断可能である。またもしデータベース提供側が検索エンジンによるクロールを、法の基づけを受けて防ぎたい場合は、「著作権法施行規則　第七章　送信可能化された情報の収集を禁止する措置の方法 第四条の四」

の記述に従い、サイトの規約や robots.txt（④で言及する）にクロール禁止の旨を記述することで対応することが可能である。また再利用許可の表示基準であるクリエイティブ・コモンズ[12]のライセンスがデータベースごとに設定されている場合もあるが、検索エンジンからの取り扱いについては先述のように著作権法が包含している範囲の案件であるため、影響はないものとしている。

③ エントリごとに､別の URL でアクセスできるか。
データベースの内部データへアクセスし、そのユニークなエントリがユニークな URL で別々に表示されることを確かめる。例えば、サイトの CGI オブジェクトに対しパラメータを POST して内容が表示されるデータベースは､1 エントリに対し 1 つの URL が対応する。もし複数のエントリが 1 つのページにまとめて記述されている場合は、エントリ単位の検索をメインとする本横断検索の基準から外れるため対象外とする。但し、1 ページ内にアンカーリンク（#が含まれる URL）が存在し、エントリが見分けられる場合は対象としている。また、Java アプレットのアプリケーションの動作で情報を表示するデータベースや FLASH コンテンツとしてデータを表示する場合もエントリごとの URL を見分けることができないため対象外とする。この結果を「A.技術的な条件」を解決する前提条件とする。

④ エントリの URL は、robots.txt で機械的アクセスを禁止されていないか
ドメインごとに記述される robots.txt を確認し、機械的アクセスを禁止されていないことを確認する。これは「B.法的な条件」の対応として実施する。robots.txt の規約の基準については、Google が公開するサイト管理者向けの情報[13]と、O'Reilly 社の Spidering Hacks[14]に記述されている情報を参照した。③で確認したエントリの存在するディレクトリが robots.txt の disallow で指示される範囲内に入っていないことを基本とし、コメント行に書かれた指示に従い、Crawl-delay:の項目が指定されていればその秒数を超えない頻度で機械的アクセスを行う。指示される内容は UserAgent ごとに区分けされているものもあるが、本横断検索では独自の UserAgent 名として「NBDCbot(biosciencedbc.jp/dbsearch)」を使用している。クロール先のデータベース管理者はログ情報などから UserAgent 名を確認できるため、クローラがサーバへ不必要な負荷をかけていた場合や、クロール方法について NBDC へ連絡が必要となった場合、連絡先を辿ることができる。②利用規約の件と同様、データベース管理者へクローリングの許可が得られている場合は、robots.txt の規定の限りではない。また、robots.txt が設置されていない、もしくは白紙状態で設置されている場合も機械的アクセスを禁止されていないと判断する。

⑤ 各エントリのデータはバイナリだけでなくテキストが含まれているか
横断検索ではユーザがテキストで検索を行うため、エントリにテキストが含まれていることが必須となる。検索データとして画像や動画しか無いデータベースはテキストで検索しアクセスすることはできない。但しページタイトルやファイル名でエントリを見分けられる場合はこの限りではない。また、Web ブラウザから通常見えない情報（HTML タグ、コメントアウトされた情報､メタタグで指示された情報）やドメイン名、ディレクトリ名、ファイル名もこの対象外とする。このプロセスは先述の A.B.C.の条件とは別に、横断検索の仕様に従うための条件として実施する。

⑥ データ検索機能もしくはデータのリストが公開されているか
検索対象として横断検索内に取り込むため、エントリが全部で何件あるか、どの範囲のデータが登録されているかを把握する必要がある。本横断検索は学術利用を目的としているため検索対象の網羅性確保への要求が高いためである。大抵のデータベースには利用者のために検索フォームやデータリストページが備えられているが、これらが存在しない場合は、データの範囲を知ることが難しい。このプロセスはエントリの単位やリストの網羅性を理解するために必要であり、A.C.の条件のための情報となる。

⑦ 全エントリが 1 ページにリンクを介して設置されている
⑦以降のフローはクローリング先のデータのパターンによって、A.の技術的な条件を解決するためのプロセスとなる。⑥でエントリのリストがあることを確認できれば、それぞれのリンク先に機械的にアクセスすることで全エントリデータを取得することができる。ここで全エントリデータが設置されていれば⑧へ､設置されていなければ⑨のフローへ進む。

⑧ クローリング
⑦でエントリリストのページから全エントリへ辿ることができる場合、クローリングを行うスクリプトやダウンロードアプリケーションにリンクを何回辿るかを設定し、全データをダウンロードさせることができる。本プロセス完了後､図３内では「達成」

オブジェクトに至り、検索のためのクローリングが可能な一つのパターンとなる。

⑨ クエリによって全エントリのリンクが返される

データベースの検索を実現するため、Web サイトにはテキスト入力ボックスが設置されることが多い。このボックスにユーザが要求するキーワードを POST することで、CGI スクリプト等からサーバ内のデータベース部へ SQL 文などで検索要求を内部で要求する形式となっている。データベース内では SQL 文では「*」はワイルドカードとして認識されるため、全データが返り値として得られる場合が多い。また「""」は無条件をリクエストする場合も全データが返される場合がある。さらに[0-9]や[a-z]を投げると、数字やアルファベットのワイルドカードとして認識され、全データに数字もしくはアルファベットが含まれている場合は全データが返される。但し、サーバ側のセキュリティ保持や安定性確保のためこれらのワイルドカードが入力をブロックされている場合もあり、その場合は本プロセスの実行は不可となる。特にエントリの件数が多い場合はサービス提供側の予期しない使い方になるため、相手方のサーバの状況に注意して実行する。

⑩ 全件の URL を確認

全件の URL を確認できれば、⑦と同様の結果が得られ、⑧のフローへ繋がる。

⑪ エントリの URL が連番で記述されている

⑩までの方法で全エントリの範囲が不明である場合、エントリの URL を見て POST するパラメータやデータの ID が連番の数字になっていることが分かれば、情報が得られる場合がある。例えば、あるサイトのエントリの URL が、「http://example/data/0002」や「http://example/data/0103」のようなものがある場合は、末端の 4 桁の数字がエントリの順番を示すものと推定できる。数字を増減させた URL にも同様にエントリが存在することが分かれば、帰納法に近い形でエントリ URL の範囲を確認することができる。

⑫ 番号の上限下限を確認

⑪でエントリごとの URL が連番になっていると推察できる場合、その番号を増加、もしくは減少させ、データが存在する番号の上限と下限を確認する。これを利用してエントリデータの範囲を推測する。この情報から⑩のフローに繋げることができる。⑪の 4 桁の数字の例では、0000 から 9999 までの範囲を実際にアクセスし確認する。

⑬ エントリの URL が何らかの ID を元に呼び出されている

エントリデータが⑪、⑫のように連番の数字ではなく何らかの ID で呼び出されていることを URL から判断する。例えば「http://example/data.cgi?id=xtU」のように記述されていれば、連番ではなく何らかの記号で情報が呼び出されていると推測される。

⑭ 全 ID のリストがある

⑬のエントリ呼び出しパラメータとして利用する ID がサイト内にまとめられているか確認する。これはサイト内のデータの関係を良く知るために C.の専門的な知識が条件として必要になる。この情報から⑩のフローに繋げることができる。生命科学系のデータベースでは、リファレンスとなる大規模遺伝子配列データベースや、蛋白質構造データベースが存在する。これらのデータベースのエントリには固有の ID が振られており、二次データベースや関連分野のデータベースではこの ID がそのまま呼び出しパラメータとして利用されている場合も多い。このためこの ID がそれらに該当するかは、専門家が確認して判断することができる。

⑮ バンドルデータなどを処理することで、エントリ URL とインデックスを作成できる

これまでのフローでエントリ URL を入手できない場合は、HTML ソースに記述された関連スクリプトからエントリや URL の情報を探す方法や、サイト内に置かれたバンドルデータの内容を処理して使う方法が考えられる。これらの情報を使い URL を構成できるか検討する。生命科学系データベースではデータを利用者がローカル環境でも使用できるよう、バンドルデータとして公開している場合が多い。

⑯ データの再構成処理

⑮でバンドルデータなどから URL 情報を再構成できると判断された場合、データの再構成を実施する。

⑰ サイトマップ（sitemap.xml など）が公開されている

robots.txt と同様にドメインの直下に sitemap.xml が配置されている場合がある。ここには検索エンジンの機械的アクセスのためにサイトを構成するページの URL リストが記述されているが、ここにデータのリストが記述されている場合もある。記述する範囲のデータはサイトごとの裁量によるため、データベースのエントリ URL のみが記述されていない場合もある。

４．検索エンジンへのデータ設定

以上のプロセスでクローリングすることが問題ないと判断された場合、クローリングを実施する。クローリングは基本的にはデータベースごとの HTML タグの構造に合わせてエントリごとのユニークデータを取得する。クローリングで使用するスクリプトやアルゴリズムは、wget や各種ライブラリなど O'Reilly 社の Spidering Hacks[14]にまとめられている手法を参考とする。データのパースには昔から使用されている Web ラッパの手法を使用する[15]。クローリング完了後、取得したデータを分析・整理して検索システムへの設定を行う。検索システムは検索時に使用するインデックスが必要となるため、そのフォーマットに合わせた形式とする。横断検索では検索エンジンとして Hyper Estraier[16]を使用しており、そのフォーマットに合わせてクローリングデータを処理している。

５．まとめ

本プロセスを実施した結果、横断検索では多数の生命科学系公開データベースをクローリングし検索対象とすることができた。対象の中には Google などで検索できないディープウェブの範囲のデータも多く、学術目的の利用に耐える網羅性・信頼性を確保している。上記プロトコルでは、自動的な手法だけでなく、多様なデータベースに対しては、専門知識のある人間が別個に判断することで得られる利点を示すことができた。提案するワークフローは経験則や前例に従う部分も多く、定型化や自動化をできる部分はまだ残っていると考えられるものの、人間が専門知識やフレキシブルな判断力を以てこういった情報資源を活用するための作業は今後も必要と考えられる。こういった学問分野の作業を研究者や技術者とも違う目線で判断力を遂行する「キュレーター」の需要は、今後ますます増大するものと考えられる。

また、本プロセスを別分野の公開データベースに適用させ同様の横断検索を構築することも可能だと考えられる。分野ごとの専門家が詳細な情報を収集するフォーカストクローラ[17]を作成し、多数の横断検索を作成することで、一般的な Web 検索エンジンとは別方向の情報の信頼性を有する検索基盤が組みあがるかもしれない。

本ワークフローでは 3 件の課題が残されている。1 つは人手による作業が多く、必要なリソースが過剰であること、また担当者による主観が入る恐れがあることである。リソースの削減、効率化については前提として定型化された本ワークフローの実施が有効であると考えられるが、今後もより効率的な作業が可能となるよう改善を進めていくものとする。主観が入る作業については、1 つのプロセスに対し複数の手法で実行し、複数人で 1 つの作業を実施して結果をつきあわせることで、より客観的な内容とすることが可能と考えられる。2 つ目の課題は対象のデータベースサイトの更新や閉鎖、ドメインの引越し、大幅なリニューアルがあった際の追跡方法が確立していないことである。Web サイト[18,19]や Web 上の論文データ[6]の約半分が 2 年でアクセスできなくなると言われており、データベースサイトも頻繁に更新が行われていることから、変更の追跡は容易ではない。Xpath の構造変化追跡[20]などを利用し今後最適な手法を検討する予定である。3 番目の課題は本ワークフローの評価が不十分である点である。データベースサイト 1 件 1 件を確認し、Google などの Web 検索エンジンではアクセスできず、スニペット情報が不十分なものが、本横断検索では問題なく利用できることを確認している。しかし全体的にどの程度検索結果の信頼性があるかを定量的に示すことができていない。検索件数や内容に関する評価方法も今後検討していく方針である。

Linked Open Data として利用可能な生命科学系データベースの中には表形式のデータだけではなく SPARQL で呼び出されるトリプル情報が公開されているものや、バイナリのデータベース、メタタグや RDF によってデータの定義づけがされているものも増加しつつある。今後も横断検索は様々な形式のデータベースに対応し、検索対象としていく方針である。

謝辞

生命科学データベース横断検索システム[2]は、ライフサイエンス辞書プロジェクト（http://lsd.pharm.kyoto-u.ac.jp/）より辞書データを、共立出版株式会社（http://www.kyoritsu-pub.co.jp/）より蛋白質核酸酵素の文献データを、さらに多数の生命科学系公開データベースより情報をご提供いただき構築された。

参考文献

[1] LOD cloud diagram, http://lod-cloud.net/, (2014.7.22 アクセス)
[2] 生命科学データベース横断検索, http://biosciencedbc.jp/dbsearch/, (2014.7.22 アクセス)
[3] 安形輝, 宮田洋輔, 池内淳, 上田修一,: 学術情報流通

における深層ウェブの実態--機関リポジトリに収録された文献を用いた調査, 情報学会研究大会発表論文集 2009 年度, pp. 37-40, (2009)
[４] 佐藤亘, 打田研二, 山名早人,: 検索エンジンのヒット数の信頼性に対する評価, 第３回データ工学と情報マネジメントに関するフォーラム, F6-1, (2011)
[５] Henzinger M.: Search technologies for the internet, Science. Vol.317, No. 5837, pp. 468-71, (2007)
[６] 石田栄美, 宮田洋輔, 池内淳, 安形輝, 野末道子, 上田修一,: 生存分析からみた学術論文 PDF ファイルのクローリング, 2008 年日本図書館情報学会春季研究集会発表要綱, (2008)
[７] 井原伸介，林貴宏，尾内理紀夫,: もぶろげっと：画像情報を含む blog 記事検索システム, インタラクティブシステムとソフトウェアに関するワークショップ(WISS2005)論文集, pp.69-74, (2005)
[８] 詹善斌, 山名早人,: 特定言語 Web ページ収集のためのフォーカストクローラの性能改善手法, 第２回データ工学と情報マネジメントに関するフォーラム, B2-1, (2010)
[９] Bergman, M. K.: White Paper: The Deep Web: Surfacing Hidden Value, Journal of Electronic Publishing Volume 7, Issue 1, August, (2001)
[１０] Integbio データベースカタログ, http://integbio.jp/dbcatalog/, (2014.7.22 アクセス)
[１１] 末吉亙,: データベースと著作権, 情報管理, Vol. 55, No. 2, pp.125-128, (2012)
[１２] クリエイティブ・コモンズ, http://creativecommons.org/, (2014.7.21 アクセス)
[１３] Google Webmaster Tools.:BLOCK URLS WITH ROBOTS.TXT, (2014.7.22 アクセス) https://support.google.com/webmasters/answer/6062608
[１４] Kevin H., Tara C., 村上 雅章,: Spidering hacks―ウェブ情報ラクラク取得テクニック 101 選, オライリー・ジャパン, (2004)
[１５] Giansalvatore M., Paolo A.: Cut and Paste, Journal of Computer and System Sciences, Vol. 58, Issue 3, pp. 453–482, (1999)
[１６] 全文検索システム Hyper Estraier. http://fallabs.com/hyperestraier/, (2014.7.22 アクセス)
[１７] Soumen C., Martin B., Byron D.: Focused crawling: a new approach to topic-specific Web resource discovery, Computer Networks, Vol. 31, No. 11-16, pp. 1623-1640, (1999)
[１８] Koehler, Wallace.: An Analysis of Web Page and Web Site Constancy and Permanence, Journal of the American Society for Information Science, Vol.50, No.2, pp.162-180, (1999)
[１９] Koehler, Wallace.: Web Page Change and Persistence-A Four-Year Longitudinal Study, Journal of the American Society for Information Science and Technology, Vol.53, No.2, pp.162-171, (2002)
[２０] 中野雄介, 寺西裕一, 西尾章治郎,: Web ラッパのアグリゲーションサービスへの適用と評価, 情報処理学会論文誌, Vol. 53, No. 8, pp. 2018-2027, (2012)