文献公開を中心としたオンライン型デジタルアーカイブズについて

氷　野　善　寛

Public Access to On-line Digital Archives and Documents

HINO Yoshihiro

Recent public access to various on-line research databases in the humanities has facilitated access to the on-line university and non-university materials. This essay provides an outline to databases that primarily contain documents, digital archives, and digital data incidental to these two on-line sources. Consequently, it is my intention to question how documents should be preserved and digitalized for public access.

キーワード：データベース、アーカイブズ

はじめに

昨今オンライン上で様々な人文系の研究データベースが公開され、誰もが簡単にアクセスできるようになった。本稿ではオンライン上に公開されている東アジア文化交渉学に関連のある主に文献を扱ったデータベース及びデジタルアーカイブズ、またそれに付随するデジタルデータについて概観し、今後どのように所蔵する文献をデジタル化し公開していくのかについて議論を進めたい。

一　概　要

データベース、デジタルアーカイブズ、デジタルデータという言葉について、最初に定義しておく。アーカイブは本来書庫のことであり、デジタルアーカイブは、図書館や博物館や研究機関などの所蔵品、収蔵品といった具体的な物や無形のデータ類などをデジタル化し集積し、保管しておく場所のことを意味する。デジタル化されたデータにもいくつかの種類があり、画像、音声、動画、テキストなどのデータ形式が想定できる。デジタルアーカイブにはこのような様々なデータが集積されている。データベースはこれらのデータにアクセスするための仕組みであり、個別のデータの関連づけを行う仕組みであり、それらのデータの集合体を意味する。つまり、デジタルアーカイブズにある個別のデジタル化され

たデータにアクセスするために、データベースというツールを使ってアクセスすることになる。

最近ではデータベースを通じて文献目録を検索できるだけではなく、著作権が切れた文献がデジタル化され、画像が公開されていることもあり、ネットワークを通じて誰もが自由に目的の画像データやテキストデータなどにアクセスし利用することができる。それでは実際にどのようなデータベースがあるかというと、たとえば日本国内においては、図書館や研究機関などが一つないしは複数のデータベース及びデジタルアーカイブズを公開することが多い。一方、アメリカでは、Internet Archive やGoogle Books、Open Libraryといった文献のデジタル化プロジェクトが存在し、オンライン上に設置された複数の機関のデジタルデータを横断的に検索することが可能なサービスが見られる。そこで文献を中心とするデータベース・デジタルアーカイブズの類型について考えると、まず国会図書館の近代デジタルライブラリーや早稲田大学の古典籍総合データベース、オーストラリア国立図書館、大英図書館のように機関のコレクションを全てデジタル化し公開することを目標としている方式がある。これらの機関に共通するのはPDFやJPEG画像などの汎用性が高いファイル形式でデータを提供している点である。次に、関西大学の近代中国語コーパスのようにテーマやカテゴリーに基づいて文献画像を収集しデジタル化した上で、独自ビューワーによる画像データとテキストデータの提供、さらに全文検索ができるようにしている方式がある。ただし、テキスト化、文献内全文検索については英語で書かれた文献ではOCR（光学自動文字認識）を利用したテキストのデジタル化が比較的容易にできるが、中国語や日本語のように漢字を含む古典籍を対象としたOCRはまだまだ力不足のため、テキストデータの抽出には技術的な問題も多く、数は多くない。そしてこの形式のデータベースについては四庫全書や基本古籍庫、彫龍データベースなど製品版にいくつか見られるが、オンラインで無償で提供されているものは非常に少ない。三つ目にWEB泊園書院やWEB懐徳堂などのように一つの私塾を対象とし、所有する文献を含むデータだけではなく、その私塾を一つのテーマとしてより深く研究し、その一環として所蔵するデジタルデータを公開する形式がある。四つ目として、明治学院大学の『和英語林集成』デジタルアーカイブスのように一つのコンテンツを集中的に掘り下げていく形式がある。一方、文献画像を含まないデータベースに焦点を当てれば、国立公文書館式の文書庫としてのデータベース、シルクロードプロジェクトのように地理情報と連携したデータベース、モノの情報を提供している博物館、あるいはテキストデータに特化した言語データベース、誰でも編集可能な辞書データベースであるWikipediaなどもデータベースの例としてあげることができる。

次に具体的にデータベースについて今後人文系のデータベースを設計する上、あるいは利用する上で参考になりそうなものをまとめる。

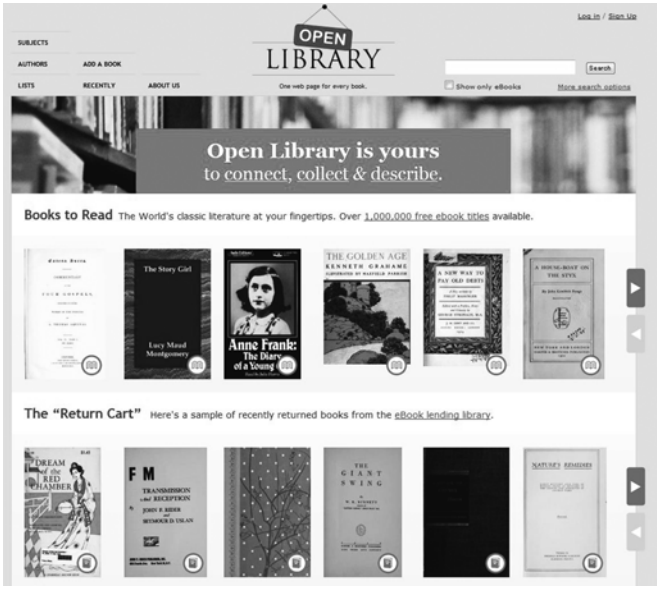
二　3つの文献目録型データベース

アメリカには横断的に文献データを検索できるサイトとしてInternet Archive、Open Library、Google Booksがある。

（Internet Archive）

まずInternet Archive[1)]はウェブやマルチメディアの資料のアーカイブ化を進めており、ウェブページ自体のアーカイブ化、音楽や動画、書籍などのデータを収集している。大部分のデータはパブリックライセンスで提供されており、誰でも自由に利用することができる。たとえば19世紀に利用された北京語の教科書『語言自邇集』を執筆した「Thomas F. Wade」の名前を入力して検索すると、31件の検索結果が表示される。結果一覧には、『語言自邇集』やWadeコレクションの目録、あるいはWadeについて書かれた書籍といったものが表示されており、さらに著者別、年代別、メディアの種類、コレクション別などで絞り込みができる。さらに個別のデータを表示すると、文献データの場合は、文献の書目情報のほか撮影したカメラの情報などデジタル化の際の環境情報なども記載されている。文献画像がある場合は、独自のビューワーソフトを通じて閲覧できるほか、PDF形式でダウンロードすることができるほか、Full Textデータや、Epub形式のデータなどダウンロードすることができる。ただし中国語が含む箇所についてはうまく文字認識がされていないことが多く、現時点では実用に耐えない。

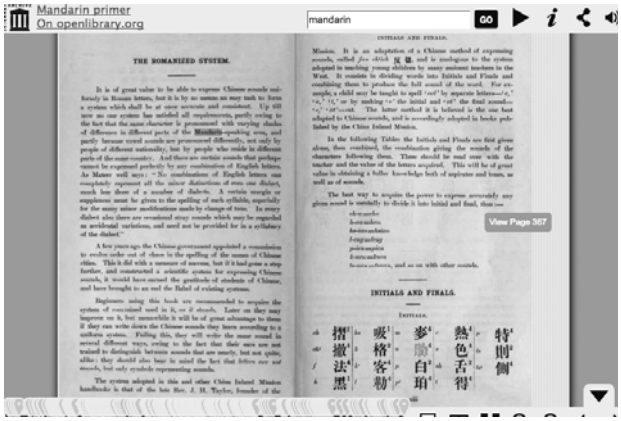
1） http://www.archive.org/



（Open Library）

次にOpen Libraryはこれまで出版された書籍のためのウェブページを作成することを目標としており、上述のInternet Archiveの一環として進められている文献デジタル化のオンラインプロジェクトである。非営利のプロジェクトで、カリフォルニア州立図書館とケール・オースチン財団の協力のもと進められている。

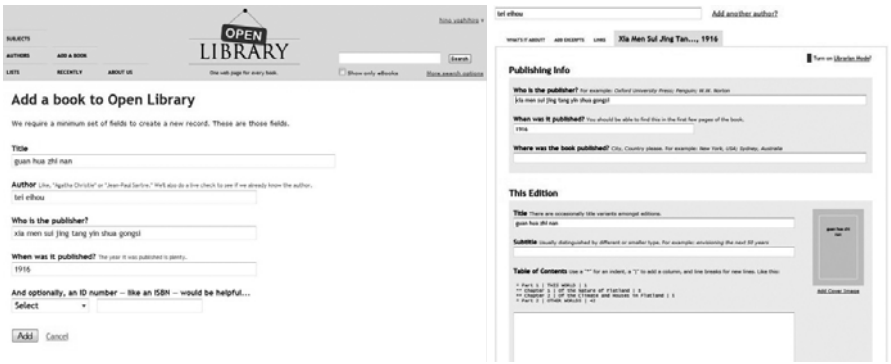
掲載されているフォーマットは独自のビューワーを通じてみるもののほかに、PDF、Plain text、DAISY、ePub、DjVu streaming、MOBI、Send to Kindleなどのファイルがあり、必要に応じてダウンロードして手持ちの端末で見ることができる。書籍についてはOCRされているデータであればテキストが埋め込まれているため、検索も可能である。



（全文検索の結果表示画面：該当箇所がハイライトされる）

F. W. Ballarの*A Mandarin Primer*を例にデータ構造を見てみると、書名と著者名のほかに「subject」という項目があり、Chinese language, Grammar, Readers, Mandarin, Dialects, Mandarin dialects, Accessible bookというデータが記載されている。次に表紙の画像があり、さらに８つの「edition」が

関連情報として紐付けられている。このedition情報では、1st 1894、8th 1911、14th 1933年といったデータが記載されているように、間に抜けているeditionもあり、全ての書目が記載されているわけではないということが分かる。そして一番下にHistoryという項目がある。このHistoryの項目こそがこのデータベースの要ともなる箇所で、この文献の情報を作成したユーザー、編集したユーザーの履歴が表示されているのである。この文献については2009年～2010年の2年間に3人のユーザーが情報を作成、追加していっていることが分かる。つまり、このOpen Libraryはユーザー登録しログインすることで、オンライン上で著作権が切れた書籍をユーザーがアップロードすることができる。また、ウィキエンジンを利用しているため、各種項目が編集可能となる。書名、著者、内容タグ以外にも不足する情報を随時追加することができる。



（文献情報の登録画面）

サインインすると右上に「Edit」という文字が表示され、それをクリックすると編集画面に切り替わります。そこでは関連人物、概要など様々な細かい情報を入力することができるようになっている。

最後にGoogle Booksは Googleが提供している書籍の全文検索サービスである。著作権が切れている文献画像で登録されているものについては画像を見ることができる。大学図書館と連携して文献のデジタル化を行っており、日本では慶応義塾図書館、アメリカではハーバード大学付属図書館、スタンフォード付属図書館、カリフォルニア大学付属図書館などの蔵書がGoogle Booksに提供されており検索可能である。

（Google books上で閲覧できる『語言自邇集』）

三　所蔵文献・所蔵物をオンライン公開している機関及びデータベース

以下、日本国内外でオンライン上でデータを公開しているデータベースについて概説する。データベース名称、概要、公開機関、開設日時、開設者、対象範囲、データ形式、利用方法、公開方法、ライセンス等の情報について順番に文献公開データベースを列挙する。

名称　　近代デジタルライブラリー

概要　　日本における唯一の国立図書館、立法補佐機関としてさまざまな役割を担っている図書館である国立国会図書館は、近年は近代デジタルライブラリーだけではなく、貴重書・準貴重書（洋書、新聞、雑誌を含む）をはじめとした江戸期以前の和古書、清代以前の漢籍などもデジタル資料として公開（http://dl.ndl.go.jp/）しているなど、電子図書館事業の拡充が行われている。

この近代デジタルライブラリーでは、国立国会図書館が所蔵する明治期から昭和前期までに刊行された図書の本文を、デジタル画像で閲覧することができる。2011年７月現在で収録されている資料は、明治期刊行図書約164000冊、大正期刊行図書約96000冊、昭和前期刊行図書約310000冊となり、著作権処理が完了していない資料は国会図書館内の端末でのみ閲覧可能である。一方インターネットで閲覧可能な資料は明治期刊行図書約129000冊、大正期刊行図書約41000冊、昭和前期刊行図書約70000冊となっている。このデジタルライブラリーに収録されている図書の多くは、国立国会図書館の前身である帝国図書館の蔵書だったもので、基本的には日本語の図書資料が中心となるが、近代の中国語学習書、中国で利用された日本語の教科書なども収録されている。

公開機関　　国立国会図書館

開設日時　　2002年10月１日

開設者　　国立国会図書館

対象範囲　国会図書館前身の帝国図書館の蔵書本のうち、明治から昭和初期までに刊行された文献を順次公開。
データ形式　JPEG、PDF（画像は全て白黒）
利用方法　ウェブサイトの専用ビューワーによる閲覧、10コマ単位でPDFへの書き出しも可能。
公開方法　ウェブ形式、無料公開
ライセンス　国立国会図書館ウェブサイトからコンテンツの転載（画像、文書、記事、データ等の転載、放映又は展示）を行う場合には、転載依頼フォームにより、あらかじめ国立国会図書館に申込みが必要となる。その際使用した画面が国立国会図書館ウェブサイトから転載したものであることを明示することと、使用した画面を使用する目的以外の目的に使用しないことが注意として国会図書館のウェブサイトに掲載されている。
設置アドレス　http://kindai.ndl.go.jp/

名称　近代中国語文献資料データベース（近代汉语文献资料数据库）
概要　「近代漢語」に関連する文献資料を収集することを目指す目録型のデータベースと文献画像を蓄積したデジタルアーカイブから構成されている。目録及び文献画像については順次整理、公開が行われている。2011年４月の段階では試験的な公開として関西大学増田文庫及び内藤文庫が所蔵する一部の書籍の書誌情報と画像データの公開を行っている。2011年４月の段階では50タイトル159冊、撮影中のものが多くあり随時データを追加していっている。
公開機関　関西大学文化交渉学教育研究拠点、アジア文化研究センター
開設日時　2011年４月
対象範囲　関西大学個人文庫所蔵本（長澤文庫、増田文庫等、内藤文庫）、内田慶市個人蔵書、尾崎実旧蔵書の中において清代から民国初期に刊行された関連文献の原資料。
データ形式　ウェブサイトの専用ビューワーによる閲覧。
利用方法　データベースには「ゲストユーザー」でログインして利用することができる。ログイン後、文献一覧から必要な文献を選択して、文献の個別の目録データにアクセスし、そこから画像アーカイブである「近代中国語文献アーカイブズ」にアクセスすることができる。このアーカイブではFlash Playerを利用した専用の独自ビューワーを通じて文献画像を閲覧することができる。ダウンロードはできない。
ライセンス　特に記載はないが論文や研究など個人的な利用については特に届け出の必要無く利用できる。アーカイブ画像を利用した刊行物や図録作成などデータ本体の二次的な利用については許可していない。
設置アドレス　http://www.icis.kansai-u.ac.jp/ModernChinese-db/

名称　近代中国語コーパス（近代汉语语料库）
概要　このデータベース設置の目的の一つは近代中国語における「官話」と呼ばれるものの全

体像を明らかにすることにあり、『語言自邇集』は初版から第３版までの全版本や『官話指南』、『官話類編』などの清末の官話教本や漢訳聖書といった欧文資料、朝鮮資料、琉球資料、唐話資料といったいわゆる「周縁」地域の資料を中心に収集、公開している。このデータベースに登録されているデータについては画像データだけでなく、テキストデータも提供しており、画像とテキストの双方を比較しながら閲覧することができる。なおテキストデータについては登録文献全体に対する全文検索が行え、検索結果から個別の文献にジャンプすることができるのも本データベースの特色の一つである。2011年４月の段階で約100タイトルの文献の全文検索および画像の閲覧が可能。データについては現在も追加作業が行われている。

公開機関　関西大学内田慶市、関西大学文化交渉学教育研究拠点

開設日時　2006年

開設者　内田慶市、氷野善寛

対象範囲　現在のところ19世紀の官話関連資料、欧文資料に重点が置かれて公開されている。

データ形式　中易中標電子信息技術有限公司が開発した古典籍ビューワーを利用して閲覧、Internet Explorer/Firefoxとjava 6以上がインストールされている必要がある。Macの場合はSafari/Operaで利用可能。

利用方法　ユーザー登録が必要、ユーザー登録後アクセス権限によって閲覧できる資料の範囲や操作できる権限（検索、閲覧、印刷、画像保存）が異なるので、必要に応じて管理者に連絡が必要。

ライセンス　特に記載はないが論文や研究など個人的な利用については特に届け出の必要無く利用できる。アーカイブ画像を利用した刊行物や図録作成などデータ本体の二次的な利用については許可していない。

設置アドレス　http://www2.csac.kansai-u.ac.jp:8080/library/

名称　早稲田大学古典籍総合データベース

概要　早稲田大学図書館が所蔵する古典籍について、その書誌情報と関連研究資料、さらには全文の画像を公開したデータベースである。総数約30万冊、国宝２件、重要文化財５件を含むあらゆる分野の資料を、具体的詳細な書誌情報と、鮮明なカラー画像で提供している。早稲田大学図書館がすでに所蔵している約30万点の古典籍のデータ（書誌･画像）作成を目指して2005年度から作業を開始し、現在も作業を進めている。新たな資料の受け入れや、収載資料の充実につとめている。

公開機関　早稲田大学図書館

開設日時　2005年

開設者　早稲田大学図書館

対象範囲　主に、江戸時代以前に出版・書写された文献資料、清朝以前に出版・書写された漢籍及び、同時代の韓本、日本の古代から近代に至る一次資料（文書・書簡・原稿・拓本等）

が含まれる。また主として江戸時代の蘭学者（洋学者）が研究資料とした洋書を若干含んでいる。

フォーマット　JPEG画像及びPDF

利用方法　特別なビューワーソフトは不要、ブラウザのみで閲覧可能。また1冊単位でPDFファイルのダウンロードも可能。

ライセンス　このデータベースで公開している画像データを利用（現物の閲覧、ホームページ、雑誌への転載など）する場合には早稲田大学図書館特別資料室宛に事前に連絡が及び所定の手続が必要。ただし学校の授業等で一時的に利用する場合や研究資料として個人的に手元に保存する場合、学術目的のホームページへのリンクについては申請不要。

設置アドレス　http://www.wul.waseda.ac.jp/kotenseki/

名称　文化遺産のデジタルアーカイブ　ディジタル・シルクロード

概要　国立情報学研究所は、情報学という新しい研究分野での「未来価値創成」を目指す唯一の学術総合研究所として、ネットワーク、ソフトウェア、コンテンツなどの情報関連分野の新しい理論・方法論から応用展開までの研究開発を総合的に推進している。ディジタル・シルクロード・プロジェクト（Digital Silk Road Project）は、情報学と人文学の融合に基づく文化遺産のデジタルアーカイブを構築する研究プロジェクトとして、『東洋文庫所蔵』貴重書デジタルアーカイブ、古都北京デジタルマップ、中国石窟データベースをはじめとして複数のデータベースからなる。

公開機関　国立情報学研究所

開設日　時記載なし

開設者　国立情報学研究所

対象範囲　記載なし

フォーマット　JPEG、PDF、Google map、Google Earth等での閲覧データ

利用方法　ウェブサイトの専用ビューワーによる閲覧やJPEG、PDFによる画像閲覧のほか、地図情報についてはGoogle Map、Google Earthで閲覧も可能。

ライセンス　複数のデータベースから構成されているため、データベースごとにライセンスが異なる。たとえば『東洋文庫所蔵』貴重書デジタルアーカイブに掲載する史料に関する複写の手続きは、すべて東洋文庫が担っており、複写希望の場合は東洋文庫連絡が必要。詳細は個々のデータベースを確認のこと。

設置アドレス　http://dsr.nii.ac.jp/

名称　東洋文化研究所データベース

概要　東洋文化の総合的研究を目的として設置された研究所で、汎アジア部門、東アジア部門、南アジア部門、西アジア部門の4部門と2011年に設置された新世代アジア部門からなる大部門制を採用して研究を推進している。また1999年度に比較文献資料学と造形資料学

という2つの分野からなる東洋学研究情報センターが新設され、2009年度からアジアの社会調査資料を対象とするアジア社会・情報分野が増設され、旧東洋学文献センターの業務の中からデータベース作成や漢籍整理長期研修など重要なものを引き継いで実施している。『貴重漢籍善本全文画像』『東洋文化研究所所蔵漢籍目録』『近現代中国文学関係雑誌記事』や東洋文化研究所所蔵の漢籍善本全文影像資料庫、明代図書資料三才圖會データベース、サンスクリット語写本データベースなどを公開している『アジア古籍電子図書館』など複数のデータベースからなる。また『東洋文化研究所所蔵アジア写真資料集成データベース』には「清朝建築図様デジタルアーカイヴ」といった図様に関するデータベースもある。またTibetan-Sanskrit 構文対照電子辞書やサンスクリット語写本なアーカイブもあるなど多様化されたデータベース群が設置されている。

公開機関　東京大学東洋文化研究所
開設日時　記載なし
開設者　東京大学東洋文化研究所
対象範囲　記載なし
データ形式　JPEG画像など
利用方法　専用ビューワーによる画像表示
ライセンス　すべての画像等データについて、東洋文化研究所への承諾なしに、全部または一部を複製・再配布・販売することを禁じている。
設置アドレス　http://www.ioc.u-tokyo.ac.jp/

名称　国立公文書館アジア歴史資料センター
概要　アジア歴史資料センターは、国の機関が所蔵公開している歴史資料のうち日本とアジア近隣諸国等の歴史に関する資料をデータベース化してインターネット上で公開する役割を担っており、国立公文書館において運営されている。国立公文書館、外務省外交史料館、防衛省防衛研究所図書館所蔵の資料を対象としてデジタル化が行われたものから順次提供しており。平成13年11月のセンター開設以来、毎年約15万～20万件（200万～300万画像）の資料を公開しており、平成23年4月現在での資料公開数は約162万件・2246万画像となっている。
公開機関　日本、国立公文書館
開設日時　2001年
開設者　国立公文書館アジア歴史資料センター
対象範囲　国立公文書館、外務省外交史料館、防衛省防衛研究所図書館が保管するアジア歴史資料のうち、デジタル化が行われたものから順次、提供されている。
フォーマット　DjVu
利用方法　ウェブサイトから検索して表示
ライセンス　ダウンロードした資料を記事及び論文等の著作物に論拠、参考等として副次的に利用す

る場合は、センター長への申し出等を必要としない。但し、当該資料の出典は明らかにして掲載する必要あり。

設置アドレス　http://www.jacar.go.jp

名称　全國漢籍データベース

概要　全国漢籍データベース協議会は、このデータベースの作成事業に賛同している全国の図書館関係者によって2001年３月に組織された。国立情報学研究所、東京大学東洋文化研究所 附属東洋学研究情報センター、京都大学人文科学研究所 附属東アジア人文情報学研究センターの３者を幹事機関に選定し、幹事機関のもとに全国漢籍データベース作成委員会を組織してデータベースの作成事業を推進している。公開されているデータによると、日本の主要な公共図書館・大学図書館が所蔵する「漢籍」の書誌情報について、伝統的な「経・史・子・集」の四部分類（叢書部を加えて五部分類）に基づいて収集・登録した連合漢籍目録データベース。2011年の段階では69機関の漢籍目録と３機関の巻頭画像を検索することができる。また現代の目録法に準拠したWebcatでは検索しにくい漢籍を検索するのに非常に便利なデータベースである。

公開機関　全国漢籍データベース協議会

開設日時　2001年３月

開設者　全国漢籍データベース協議会（京都大学人文科学研究所附属東アジア人文情報学研究センターが管理・運営）

対象範囲　日本の主要な公共図書館・大学図書館が所蔵する「漢籍」の書誌情報。また全国漢籍データベースには、民國22年上海商務印書館排印本を底本とした『四庫全書總目提要』のテキストデータベースが併設されており全文検索が可能となっている。2011年３月の時点では、漢籍目録、69機関、814203レコードと巻頭画像３機関、14122枚ある。

データ形式　テキスト

利用方法　検索、レコード表示形式

ライセンス　記載無し

設置アドレス　http://www.kanji.zinbun.kyoto-u.ac.jp/kanseki

名称　長崎大学電子化コレクション

概要　幕末・明治期日本古写真コレクション、医学和漢古書目録、近代黎明期翻訳本全文画像データベースなど複数のデータベースが公開されている。

開設日時　2008年８月

開設者　長崎大学付属図書館

対象範囲　長崎大学付属図書館のコレクション

データ形式　JPEG画像

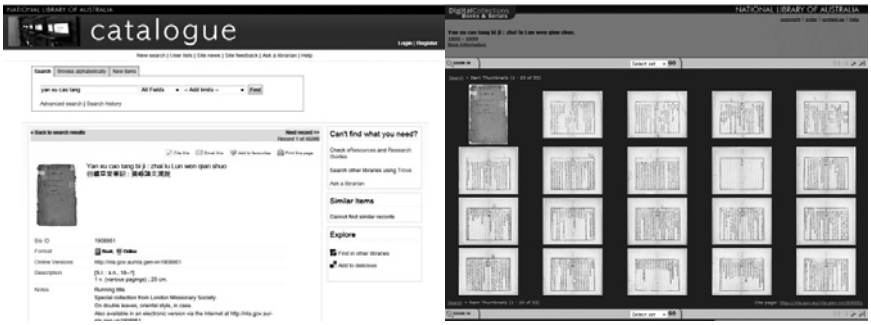
利用方法　ウェブサイトから検索して表示

ライセンス　古写真については記載あり、それ以外については図書館に直接連絡して確認。
設置アドレス　http://www.lb.nagasaki-u.ac.jp/search/ecolle/

名称　WEB泊園文庫
概要　現時点ではウェブページによる関西大学泊園文庫の解説のみ掲載されているが、同文庫が所蔵する手稿本や貴重書を撮影したものが公開される予定。
開設日時　2011年10月
開設者　関西大学、吾妻重二
対象範囲　関西大学泊園文庫所蔵本
データ形式　ウェブサイトの専用ビューワーによる閲覧方式
利用方法　ウェブサイトを通じてカタログを検索して個別の文献画像にアクセス
ライセンス　記載なし
設置アドレス　http://www.db1.csac.kansai-u.ac.jp/hakuen/

名称　『和英語林集成』デジタルアーカイブス
概要　明治学院大学図書館所蔵の幕末から明治時代の代表的英和・和英辞典の画像と年表をつけて、和英・英和辞典の総合的な情報を検索できるサイトである。
開設日時　2006年
開設者　明治学院大学図書館
対象範囲　明治学院大学図書館所蔵の幕末から明治時代の代表的英和・和英辞典
データ形式　JPEG画像
利用方法　ウェブサイトの検索ツールを通じて個別のデータにアクセス
ライセンス　記載なし
設置アドレス　http://www.meijigakuin.ac.jp/mgda/

名称　National Library of Australia
概要　オーストラリア国立図書館のコレクションにはロンドン伝道会、上海支部の図書が収められており、『衍緒草堂筆記』など官話を研究する上で貴重な資料が多数ある。
開設日時　不明
開設者　オーストラリア国立図書館
対象範囲　オーストラリア国立図書館の蔵書
データ形式　JPEG画像
利用方法　ウェブサイトを通じてカタログを検索して個別の文献画像にアクセス
設置アドレス　http://catalogue.nla.gov.au/



（オーストラリア国立図書館カタログの検索画面と画像ビューワー）

おわりに

以上、今回は目についた物を列挙しただけだが、非常に多くのデータベースが存在することが分かる。今後データベースの設計をしていく上で、これらのデータベースを大いに参考にしたい。