人文系研究資料・データの構築・公開ツールと国文学資料事例の紹介

- オープンソースソフトウェア Greenstone を利用して -

Tools for building Web System from Human Science research documents/information, and case study of Japanese ancient books

- Using Open Source Software Greenstone -

北村 啓子
Keiko Kitamura
国文学研究資料館，東京都品川区豊町 1-16-10, 142-8585
National Institute of Japanese Literature, 1-16-10 Yutaka-cho Shinagawa-ku, Tokyo, 142-8585

あらまし： 人文科学分野の研究資料・データからウェブ公開システムを構築するツールとして Greenstone を紹介し、国文学資料（和刻本研究情報）への応用事例を報告する。具体的に使用した研究情報データと構築の手順、実行例、特に外部 DB へのリンク参照方式の利用について説明する。さらに、Greenstone で構築したシステムが共同作業で DB を構築するツールとして利用可能であることについても述べる。最後に地理情報の利用可能性について簡単に触れる。

Summary: Greenstone is introduced as the tool for building Web System from Human Science research documents and information. The case study to develop the web system using Greenstone from the research information about Chinese ancient books published in Japan. The details are explained about what kind of information is used, how to build it, how it runs, especially how to link to external DB. Next the possibility of cooperative working to collect research information using this system at anywhere in this world is discussed. Finally it is also touched the possibility of using Geographic Information System in such system of Human Science shortly.

キーワード：人文科学，ウェブシステム，構築ツール，共同作業，オープンソースソフトウェア，Greenstone，電子図書館
Keywords: Human Science, Web System, Building Tool, Cooperative Work, Open Source Software, Greenstone, Digital Library

1. 人文科学研究情報の組織化・公開

人文科学分野での研究情報も、目録・書誌・調査研究情報から、翻刻などの全文テキストデータ、さらにデジタル画像へと、利用できる計算機技術の進歩、デジタルカメラの高性能かつ低価格化や記憶メディア・装置の低価格化に伴い、容易になってきた。

しかしながら、蓄積した研究情報を公開しようとすると同様に容易になてきたとは言い難い。初期のカード型 DB や RDB の時代は、商用のPCアプリケーション(DBMSなど)がサポートするウェブインタフェースを利用し、データ蓄積の延長上で公開が可能であった。全文データの作成が熱を帯びて来ると、データの交換が盛んになり、また文字列処理ツールが利用されるようになった。ウェブの普及に伴い、データの公開・ダウンロードが容易になり、さらに namazu に代表されるHP内全文検索を目的にしたフリーソフトウェアが多く出現し、全文検索は誰でも容易に利用できる当たり前の環境になってきた。デジタル画像作成が容易に廉価に可能になり、蓄積は急激に増加してきた。幸い記憶装置・メディアは大容量かつ低価格化の一途を辿っているが、大量の画像を如何に見る／見せるかについては、PCのブラウザやアルバムソフトなど、個人ユースには耐えるが、公開するとなると容易にウェブシステムを構築できるという訳にはいかない。

現状のモニタ精度やネットでのデータ転送スピードに耐える程度の品質の画像データであればアルバム式写真ホスティング(HTML での JPEG 配信)で可能であるが、研究利用に耐える精度の高品質(データ大)をネットスピードに耐え、かつ高い閲覧機能を持つシステムとなると、イメージサーバなどの利用が必要になる。さらに、書誌情報や全文情報と画像を融合して利用したいのは当然の要求であり、個別のシステム開発が必要になる。

(電子)図書館を始め大学や大きな公共機関では、様々なシステム開発の試みがなされてきている。しかしながら、人文科学分野では研究者が個人または研究グループなど小規模で質の高い研究情報を蓄積している例も多く、その体制での公開となると現実問題として難しい。

本稿では、蓄積されてきた人文科学研究資料・データをウェブ上に容易に公開するためのツールとして、オープンソースソフトウェア電子図書館システムの中でも代表的な Greenstone を紹介し、その利用事例として和刻本の『見返序跋刊記集成』に応用した例を報告する。

2. Greenstone の紹介

Greenstone は GNU GPL(GNU General Public License)の下に配布されている電子図書館を構築、サービスするオープンソースソフトウェアである[1]。構築システムとサービスシステム、そして構築システムを簡便に使うためのインタフェース GLI(Greenstone Librarian Interface)[2][3][4[5][6][7]からなる。

「構築」は一般のアプリケーション開発に相当する。しかしながら Greenstone は開発というより自動生成に近い。特に初心者向けには、電子コレクションにしたい対象のドキュメントを指定すれば、ドキュメントの種類を判定しそれに合わせたウェブシステムを作ってくれる。

「サービス」は構築された電子コレクションをウェブ方式でインタネーットを通して広く利用に供する。一番簡易には Greenstone が持つウェブサーバ機能を使うことができ、Apach などの汎用のウェブサーバを使うことも可能である。

「インタフェース GLI」はコレクション作成の指定をビジュアルに行うことが可能で、設定変更を即座に反映し"再構築＆実行"を繰り返すプロトタイプ方式で効率的に開発できる。ユーザの技術力に合わせて4段階のモードが準備され、ソースドキュメントの指定だけで自動的に標準的なコレクションが生成される初心者用から複雑なドキュメントの構造に合わせた詳細なオプションを指定できるエキスパート用まで使い分けが可能である。

機能や特長を簡単にリストアップしておく。詳細は[8][10]を参照されたい。

・多種プラグイン
・検索用インデックスの自動作成
・分類(配列)機能
・自動/自由な表示フォーマット
・メタデータの自動抽出・設定・利用
・標準 Web Server - Web browser でアクセス可能
・マルチ言語対応
・マルチメディアドキュメント対応
・マルチ言語インタフェース
・マルチプラットフォーム対応(Windows Unix MacOSX)
・スタンドアロン環境～サーバ環境までサポート
・ノンストップ運用形態
・ソース/実行形式とも高いポータビリティー

3. 国文学研究資料の事例

和刻本で重要となる見返序跋刊記の記述を調査・収集したExcelデータを対象とし、Greenstone でウェブ公開システムを構築した事例を報告する。

3. 1 和刻本の『見返序跋刊記集成』

館蔵漢籍58点と当館がマイクロフィルムで所蔵する漢籍を合わせて1,600点について、その紙焼本を使用して見返・序・跋・刊記、全ての記述についてデータ収集を行った(館内『和刻本研究プロジェクト』による)。

データ例

番号: E10020
書名: 訳解笑林広記
著者名:清(遊戯主人)編 [遠山圓陀](一噱道人)點
刊:刊本
書肆:(江、玉巖堂和泉屋金右衛門)
見返序跋刊記情報:
見返○ 文政己丑(十二年)新鐫 遊戯主人纂輯・一＝道人訳解 笑林廣記 東都書房玉巖堂發兌
序○ 序 掀髯叟漫題于咲■■
跋×
刊記○ 三都書物問屋 京都寺町通松原下ル町勝村治右衛門・大阪心斎橋通北久太郎町河内屋喜兵衛・江戸日本橋通一丁目須原屋茂兵衛・同浅草茅町二丁目須原屋伊八・同日本橋通二丁目須原屋新兵衛・同所山城屋佐兵衛・同芝神明前岡田屋嘉七・同本石町十軒店英屋大助・同浅草廣徳寺前和泉屋庄治郎・同横山町三丁目和泉屋金右衛門
外字: ＝:D04403

当館所蔵の漢籍58点、約190冊については、画像データを参照できるよう全頁デジタル撮影を行っている。その他、典拠として長沢規矩也氏著『和刻本漢籍分類目録』をデータ化している。個々の書誌情報は当館の和古書目録DB、マイクロ資料目録DBを利用する。

3. 2 ウェブ公開システムの設計

今回使用するデータの関係を図1に示す。個々のデータの紹介と、Greenstone を使ってウェブ公開システムを構築する方法を説明する。

イ. 『見返序跋刊記集成』データは Excel データのまま ExcelPlug プラグインを使って Greenstone に取り込む。代わりに CVSPlug プラグインでの取り込みも可能である。

ロ. 書誌情報は和古書目録DB・マイクロ資料目録DBから抽出し、MetadataXMLPlug プラグインを使ってメタデータとして取り込む。CSV データとしての取り込みや、事前に『見返序跋刊記集成』にマージしておくことも可能である。

ハ. 典拠である『和刻本漢籍分類目録』は、事前に書名・刊年でのマッチング処理を行い、『見返序跋刊記集成』にマージしておくことにする。典拠DBとして独立して動かしておき、実行時にリンク参照することも可能である。

ニ. デジタル画像は、高精細 JPEG2000 を JuGeMu サーバに入れて、Greenstone から呼び出すことにより、高機能な外部アプリケーション(ウェブブラウザのプラグイン方式)の JuGeMu Player を使い高精細画像の閲覧可能とする方式をとる。精度は落ちるが軽い JPEG にして ImagePlug プラグインで取り込み、汎用ウェブブラウザ上で閲覧する方式も可能である。

ホ. 全國漢籍DB(京都大學人文科學研究所付属漢字情報研究センターと東京大学東洋文化研究所附属東洋学研究情報センターの共同作成、前者がウェブサービスを行っている)が著名である。35機関の所蔵、約62万レコードが蓄積されている。平成19年に、一部京都大学人文科学研究所附属漢字情報センター所蔵分 6,517 件がNACSIS－CATに収録され、全國漢籍DBへのリンク参照も可能となっている。

本システムからも、リンク参照は可能であり、今後『見返序跋刊記集成』データ収集を引き続き追加作業をする際、全國漢籍 DB を参照しながら情報採取をする方式も取れる。特に全国の司書・研究者との共同作業の形で行う場合には、強力なツールとなり得る。逆に、全國漢籍 DB から当方 DB へリンク参照が有益と考えられる場合は、歓迎する。

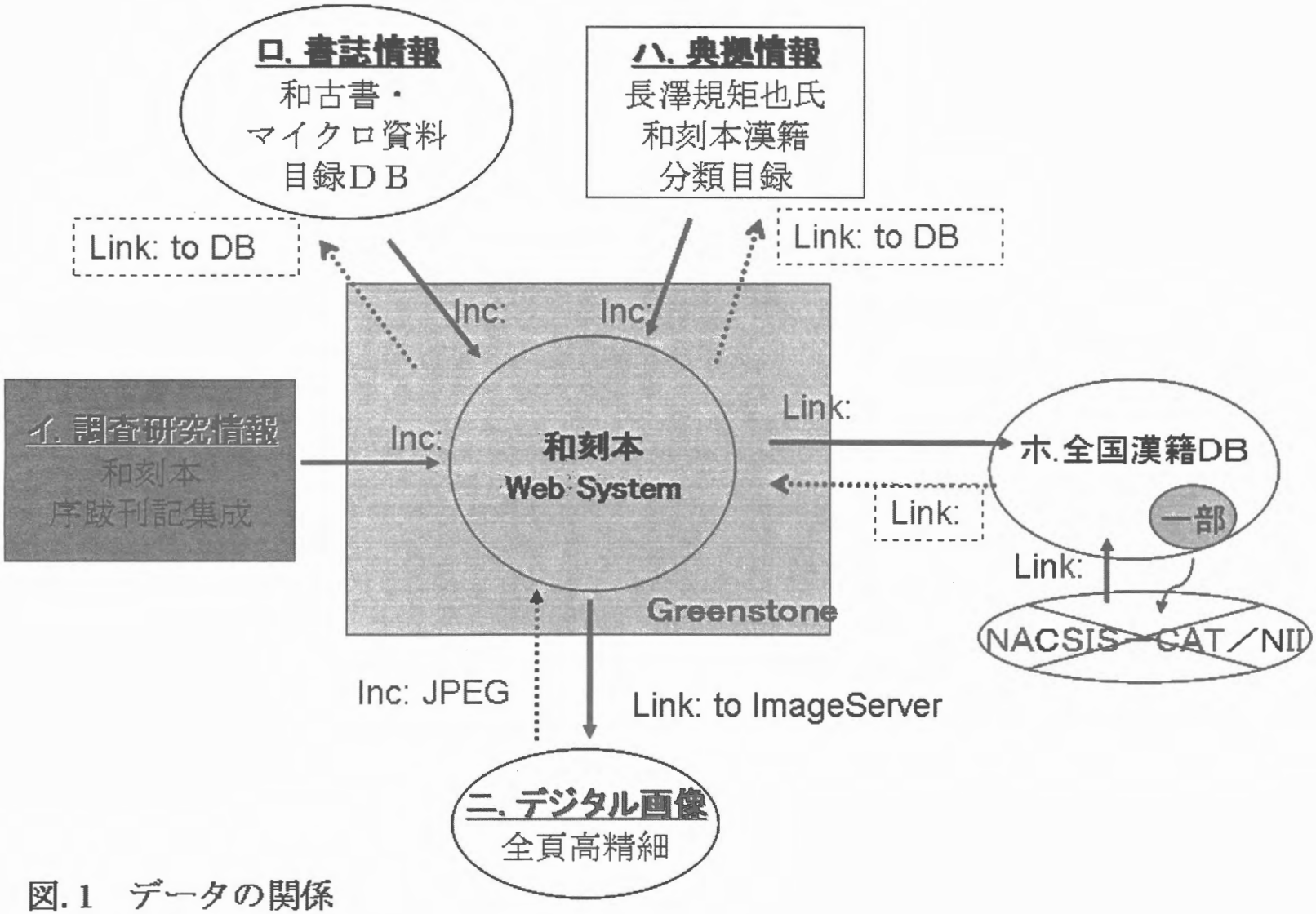
ロ. 書誌情報
和古書・
マイクロ資料
目録ＤＢ
ハ. 典拠情報
長澤規矩也氏
和刻本漢籍
分類目録
Link: to DB
Link: to DB
Inc:
Inc:
Inc:
Link:
イ. 調査研究情報
和刻本
序跋刊記集成
和刻本
Web System
ホ. 全国漢籍DB
Link:
一部
Link:
Greenstone
NACSIS-CAT／NII
Inc: JPEG
Link: to ImageServer
ニ. デジタル画像
全頁高精細

図. 1　データの関係

3. 3 ウェブ公開システムの構築手順

GLIを使って実際に指定する内容(ウェブシステムの仕様に当たる)を、構築の手順に従って説明する。

1. 対象ソースドキュメントの指定**(Gather)**:
   Excel データ1・XM メタデータファイル を指定
2. メタデータ追加**(Enrich)**: なし
3. 設計**(Design)**:
   - 3.1 プラグイン
     ExcelPlug を選択
   - 3.2 検索インデックス
     全文検索: **書名, 著者名, 書肆, 見返 序跋刊記情報** を指定
   - 3.3 ブラウジングの分類
     分類: **書名, 著者名, 刊年** を指定
4. 作成**(Create)**:
   - 4.1 取込オプション 既定値のまま
   - 4.2 構築オプション 既定値のまま
5. フォーマット**(Format)**
   - 5.1 全般
     トップ頁の説明文を入力
   - 5.2 検索 既定値のまま
   - 5.3 表示フォーマット
     書名分類: 書名-著者名(書肆) --刊年
     著者分類: 著者名(書肆)-書名
     刊年分類(刊年毎): 書名-著者名
     個々データ: 著者名 書名 刊年 ID 書肆 序跋刊記
     に変更 他は生成された既定値のまま
   - 5.4 固有マクロ なし

( )内はGLIで表示されるタブ名であり、以上の選択を順次指定していく。終了すると、作成**(Create)**タブの「構築**(build)**」ボタンを押す。指定された対象ソースドキュメントを読み込み、分析し、ドキュメントタイプに合わせて個々のデータ・メタデータを抽出する。引き続きプラグインの既定値とユーザ指定された作成指示(仕様)に従って、検索用インデックス、検索用プログラムと結果(個々のデータ)表示用HTML、各分類ビューでの表示用HTML等々のウェブ公開システムのパーツが生成される。

既定値のままで、オリジナルドキュメントの意味合いに沿った動くウェブシステムが作られる。実際に動かしてみて、変更したい所に手を入れ、再度動かして確認するという(修正→再構築→実行)プロトタイピング方式でシステムの構築・改造を行える。

[開発環境]
- WindowsXP
- 最新版 Greenstone2.7.5(Nov. 9th, 2007 リリース)
- Perl5.8.8+CPAN モジュール

3. 3 実行例

構築されたウェブ公開システムを実行した例のスクリーンショットを図2に掲載する。左上の画面が「書名分類」ビューの表示、その中の1件のデータを表示したのが左下の画面。さらに、画像ImageServer(JuGeMu Server)をリンク参照し、外部アプリケーション(JuGeMu Player)を利用し当該画像を閲覧しているのが右の画面である。

3. 3 共同DB構築ツールとしてのシステム

Greenstoneは、既存ドキュメントを公開できるウェブシステムとして構築(ほぼ自動生成)するという面で非常に強力なツールであるが、一方、コレクションにデータを追加していく、または作り込んでいくのにも強い威力を発揮する。Librarianをはじめ計算機非専門家にも容易に構築システムを使うためのビジュアルインタフェースGLI(Greenstone Librarian Interface)は、データの追加を容易に行え、特にメタデータを直接入力するインタフェースを持っている。更に、検索や分類項目の変更や、少しのHTML知識があればそれぞれの頁の表示内容やレイアウトの変更ができる。これは公開システムの改造にあたる。

今後、今回構築したウェブシステムを使って、和刻本の所蔵者(図書館司書や個人)や調査する地域研究者など外部の人も含めて、このDBを遠隔地から共同で構築していくのにも利用できる。

Greenstoneは、サーバークライアント形式でGLI-ClientからGreenstoneサーバへのリモートアクセス・コレクションの編集・構築が可能である。従ってインターネット上のどこの遠隔地からでも同じコレクションの構築作業に参加できる。共同でDBを構築するツールとして使えるのである。共同作業で利用するために、ユーザ管理(Group)機能を持ち、次の3タイプのグループが準備され、アクセスコントロールを行える。
- all-collections-editor
- personal-collections-editor
- <collection-name>-collection-editor:

GLI-ClientはJava appletで実現されており、ネットワークがあれば非常に軽装なマシンで利用可能である。

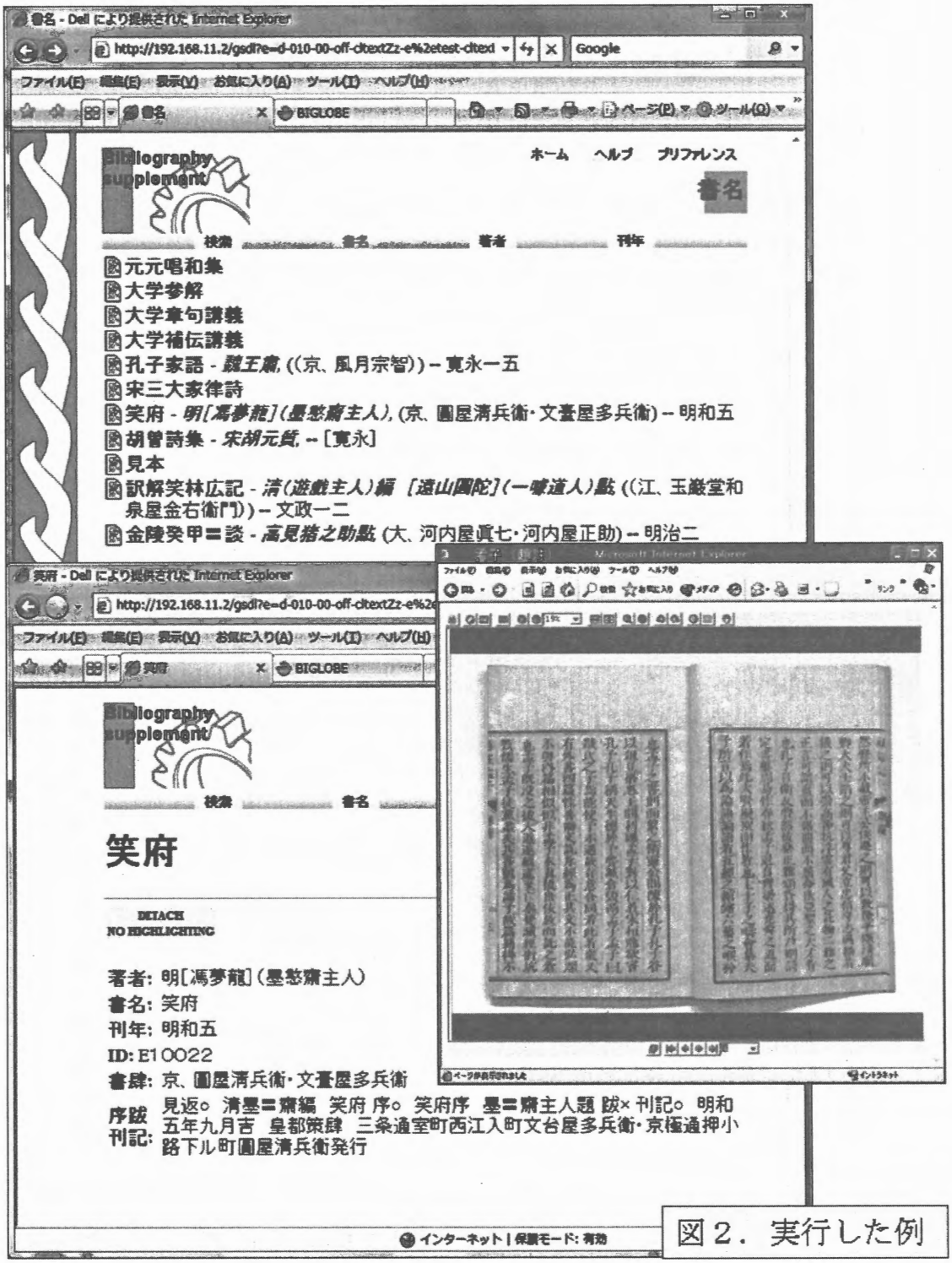
書名 - Dell により提供された Internet Explorer
http://192.168.11.2/gsdl?e=d-010-00-off-cltextZz-e%2etest-cltext
Google
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
BIGLOBE
ページ(P) ツール(O)
Bibliography supplement
ホーム ヘルプ プリファレンス
書名
検索 書名 著者 刊年
元元唱和集
大学参解
大学章句講義
大学補伝講義
孔子家語 - 龍王庸, ((京、風月宗智)) -- 寛永一五
宋三大家律詩
笑府 - 明[馮夢龍](墨憨齋主人), (京、圓屋清兵衛・文臺屋多兵衛) -- 明和五
胡曾詩集 - 宋胡元質, -- [寛永]
見本
訳解笑林広記 - 清(遊戯主人)編 [遠山圓陀](一噱道人)點 ((江、玉巌堂和泉屋金右衛門)) -- 文政一二
金陵癸甲二談 - 高見猪之助點 (大、河内屋眞七・河内屋正助) -- 明治二
笑府 - Dell により提供された Internet Explorer
検索 書名
笑府
DETACH
NO HIGHLIGHTING
著者: 明[馮夢龍](墨憨齋主人)
書名: 笑府
刊年: 明和五
ID: E10022
書肆: 京、圓屋清兵衛・文臺屋多兵衛
序跋 刊記: 見返○ 清墨二齋編 笑府 序○ 笑府序 墨二齋主人題 跋× 刊記○ 明和五年九月吉 皇都策肆 三条通室町西江入町文台屋多兵衛・京極通押小路下ル町圓屋清兵衛発行
インターネット | 保護モード: 有効

図２．実行した例

4. 地理情報の応用について

最近の著しい地理情報技術の発展に伴い、Google Map、国土地理院の電子国土を始め地図製作機関の提供する地図情報は研究分野から日常生活にまで浸透している。ウェブ上での地図利用方法も公開されており、誰でもHPから容易に地図を参照可能になっている。

今回の和刻本の例でも、漢籍の出版地を地図上で参照したり、全体の分布を地図上に重ね合わせるなどの応用例が考えられる。先に述べた外部DBへのリンク方式と同じ方法で容易に実現可能である。ソースドキュメント中に、またはメタデータとして個々の本の出版された位置を座標または地名で入力しておき、実行時に該当場所の地図を参照する。また、全ての本の位置情報を使って重ね合わせデータを作成しておき、実行時に地図と重ね合わせることによって分布図を参照することもできる。

人文科学分野での研究対象資料は、「場所・時間」との関連が強く、種々の研究情報を検索、参照する時も、地理情報と連携しフィルタリングが効率良く行えたり、新しい関連やアプローチを発見することもあるであろう。逆に、情報提供する側にとっても、地理情報と連携させることにより、異なる多くの方向からのアプローチや示唆に富んだ情報の見せ方の可能性を広げてくれると考える。

5. おわりに

Greenston の紹介と和刻本研究資料を対象にしたGreenston によるウェブ公開システムの作成事例を紹介した。[10]では和古書の書誌目録コレクションを、[9][10]では原本画像と翻刻テキストの同期頁めくりの作成事例の報告を行った。人文科学分野の研究資料を対象とする情報の構造を考えてみると、

○目録情報
○目録＋調査情報テキスト＋画像
○画像＋翻刻テキスト(or 注釈)

で多くの場合をカバーできると思われる。

対象とする(取り込み)データも
BibTexPlug BookPlug EMAILPlug ExcelPlug FoxPlug HBPlug HTMLPlug ImagePlug IndexPlug ISISPlug LaTeXPlug MARCROPPlug MARCPlug METSPlug MP3Plug OAIPlug OggVorbisPlug PagedImgPlug PDFPlug PPTPlug ProCitePlug PSPlug RecPlug ReferPlug RogPlug RTFPlug SRCPlug TEXTPlug W3ImgPlug WordPlug ZIPPlug
と多種のプラグインが用意されているので、収集しただけ、文書入力しただけのデータから、既に DB に入れてあるもの、印刷物にしたものまで、ありとあるゆる種類の情報をウェブシステム化することが可能である。今回は紹介できなかったが、プラグインの仕様を自分用にカスタマイズしたり、新たなプラグインを作るなど、オープンソースソフトウェアならではの拡張性の高さにもここで触れておく。

本稿ではウェブ「公開」システムと呼んできたが、言うまでもなく、クローズドなメンバ内のみでので公開や個人利用も可能である。自分が検索利用しやすくするための情報整理ツールとしても有効である。

今回紹介した事例でも、外部 ImageServer へのリンクを使用している。メディアの異なるデータを融合して使うためには、それぞれのメディアに適した専用サーバを利用し、DBからDBへのリンクを使うことにより、既存のDBを容易に利用したり、軽い新たな開発コストでのウェブシステム構築を可能とするなど、一つの良い手段である。また、イメージ、ビデオなど新しいメディアについては、外部アプリケーションを利用する方が、汎用のウェブブラウザの制約を受けず、高機能で良い場合もある。

国文学はじめ人文科学分野でのDB、特に専門色の濃いDBの構築は、従来クローズドなスタッフの手で行われてきた。所蔵目録や研究情報などは、現物がないとデータ採取ができない種のものであり、例えば古典籍のように日本中、世界中に散在しているものの調査が必要な類は、所蔵者や地域専門家に参加してもらいインターネット上で共同構築していく形態が望まれている。その一例である当館古典籍総合目録構築システムの開発については[9]で報告している。

インターネット上では、Wikipedia, WIKIなど不特定多数が参加して大規模な知識・情報を作り上げていく文化が根付き、これらの共同作業用のシステムの中には利用できるものもあるが、DBには必ずしも適していない。何よりこの文化は、人文科学分野には馴染に難いかもしれない。

Greenstone はこの目的での使用も期待され、大規模DBの世界規模での構築に安定的に耐えるか否か、これから実験・評価をしたい。

12. 謝辞

本研究では、館内和刻本(五山版・近世初期刊本)の研究プロジェクト』とのジョイントの形で、研究成果の利用・公開を目的に Greenstone の応用事例としてに取り組ませて頂いた。プロジェクト代表の山崎教授を始め、プロジェクトメンバの入口助教、陳準教授、山田助手(平成 18 年度まで)に感謝致します。

13. References

[1] Witten, I.H., McNab, R.J., Boddie, S.J., and Bainbridge, D. (2000) "Greenstone: A comprehensive open-source digital library software system" Proc Digital Libraries 2000, 113-121, San Antonio, Texas, June.

[2] Witten, I.H. (2004) "Creating and Customizing Digital Library Collections with the Greenstone Librarian Interface." International Symposium on Digital Libraries and Knowledge Communities in Networked Information Society, University of Tsukuba, Tokyo.

[3] Witten, I.H., Bainbridge, D., and Boddie, S.J. (2001) "Power to the people: end-user building of digital library collections" Proceedings Joint Conference on Digital Libraries, 94-103, Roanoke, VA, June.

[4] Witten, I.H. and Bainbridge, D. (2003). How to Build a Digital Library. Morgan Kaufmann, San Francisco, CA.

[5] Witten, I.H., Bainbridge, D., Paynter, G. W. and Boddie, S. (2002) "The Greenstone plugin architecture." Proceedings Joint Conference on Digital Libraries, 285-286. Portland, Oregon.

[6] Witten, I.H., Moffat, A., and Bell, T.C. (1994) Managing gigabytes: compressing and indexing documents and images. Van Nostrand Reinhold, New York.

[7]David Bainbridge, Wendy Osborn, Ian H. Witten, and David M. Nichols, "Extending Greenstone for Institutional Repositories", Digital Libraries: Achievements, Chanllenges and Opportunities, 9th International Conference on Asian Digital Libraries ICADL 2006, pp303-312,(2006)

[8] 北村啓子, オープンソースソフトウェア Greenstone による古いマニュスクリプトコレクションの開発- Jawi 語(マレーシア国立図書館)と日本語のケーススタディ -,ディジタル図書館ワークショップ 第 28 回論文集,「ディジタル図書館」No. 28, ISSN 1345-9198 (冊子体: ISSN 1340-7287)(2005)

[9] ウェブ方式の古書目録DB用データ入力システムの開発- 外字問題を中心に -, 国文学研究資料館紀要, Vol.32, pp.1-10 (2006)

[10] オープンソース電子図書館システムの利用支援, 国文学研究資料館紀要, Vol.33 pp.1-18, (2007)