# SMART-GS: 文献研究のためのソフトウェアツール

大浦真[*], 林晋[†], 久木田水生[‡]
京都大学大学院 文学研究科

## 1　はじめに

SMART-GS は、歴史学、古典学、文献学、言語学などのテキストの研究のために利用できるソフトウェアツールである。

この十年ほどのコンピュータの進歩に伴い、今まで紙やマイクロフィルムで保持していた手書きの史料や古い印刷物を画像ファイルとして保存できるようになった。安価なデジタルカメラやスキャナが普及し、画像を簡単に取り込むことができるようになり、それらを保存するためのハードディスクドライブなども大容量化している。研究者は大量の史料をコンピュータに保存できるようになった。

しかし、研究対象とする手書きの史料や古い印刷物を、コンピュータ上で検索できるようにするためには、文字を電子的なテキストに変換する、つまり、翻刻を行う必要がある。きちんとした活字で印刷された文書であれば OCR を利用できるが、OCR でサポートされていない言語や、手書きの文書や古い印刷物の場合、OCR は十分に機能しない。また、文献研究のためには、アノテーションとして、史料にマーカーなどで書き込みをしたり、付箋を付けてメモをしたりすることがある。さらに、史料内の特定の部分を別の部分と結び付け、それについて考察を行うこともある。この作業は、コンピュータ上で行うには、いささか困難があり、紙に印刷された史料であれば、直接紙に書き込むことになる。

SMART-GS はこれらの翻刻やアノテーションの作業をコンピュータ上で行うために開発されたツールである。翻刻は画像と同一のウィンドウの中で行うことができ、史料に対するアノテーションは、元の画像を変更することなく行うことができる。

当初、その開発は、2006 年ごろ林を中心に京都大学文学研究科の情報・史料学専修で始められ、実装は、小林和晶の 2006 年度修士論文研究の一環として行われた。そして、2010 年夏より、大浦真と久木田水生が、非常勤特別研究員として、このプロジェクトに参加して、開発が進められている。この SMART-GS は、GNU General Public Lisence v2 [*1]に基づいて開発されているオープンソースのプログラムである。開発は、SourceForge.JP 内の SMART-GS プロジェクト[*2]で行われており、ここからダウンロードもできる。開発言語は Java であり、複数の OS で作動できるように考慮している。

現在、この SMART-GS は、林を中心とする京都大学の田辺元研究会で用いられていて、田辺の講義準備メモの分析に利用されている。また、京都大学の永井和教授らの研究グループによる『倉富勇三郎日記』の翻刻[*3]などにも用いられていて実績を挙げている。

この SMART-GS の主な機能としては以下のものがある。

- 画像とテキストの統合: 画像とそれに付随するテキストを同一の画面上で編集、表示することができる。
- 画像やテキストのマークアップ (アノテーション): 画像にマーカーやメモなどのマークアップを付けることができ、テキストにも、reference、comment といったマークアップを付けることができる。
- マークアップ間のリンク: 画像やテキストに付加したマークアップ間にリンクを付けることができる。
- 画像に基づくサーチ: 画像の類似性に基づいてサーチを行うことができる。
- 辞書機能: 画像の部分とその読みを辞書として保存することができる。

以下、本稿では、マークアップを中心に SMART-GS の

[*] mohura@ling.bun.kyoto-u.ac.jp
[†] susumu@shayashi.jp
[‡] minao.kukita@gmail.com

[*1] http://www.gnu.org/licenses/gpl-2.0.html (2013 年 1 月 15 日閲覧)
[*2] http://sourceforge.jp/projects/smart-gs/ (2013 年 1 月 15 日閲覧)
[*3] http://nagaikazu.la.coocan.jp/kuratomi/kuratomi.html (2013 年 1 月 15 日閲覧)

使い方を説明する。

## 2　SMART-GS の利用方法

### 2.1　テキストの翻刻

図 1 は、SMART-GS を起動した画面である。(図は、「SMART-GS マニュアル」[1] から引用)

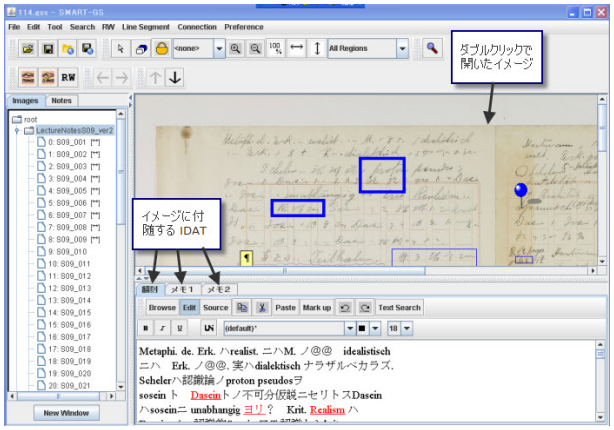


図 1　SMART-GS の起動

標準の状態で、ウィンドウの左側に画像のリスト、右上に画像、右下に画像に付随するテキストが表示されている。画像に付随するテキストは、IDAT(Image Document Attached Text) と呼ばれている。この IDAT は画像ごとに三つ用意されており、画像と同一のウィンドウで史料の翻刻などを入力することができる。IDAT は、HTML のテキストを保持する形になっていて、箇条書きや外部リンクなど通常の HTML タグを利用することができるようになっている。

また、IDAT 以外にも、Note と呼ばれる画像と独立したテキストも用意されている。Note には、史料全体に関するテキストなどを保存しておくことができる。

### 2.2　画像に対するマークアップ

SMART-GS では、図 2 のように、画像に対してマークアップを行うことができる。具体的には、以下のようなマークアップがある。

Rectangle　マウスをドラッグすることにより、長方形の描くことができる。

Marker　マウスをドラッグすることにより、特定の幅の線を引くことができる。

Polygon　頂点をクリックすることにより、多角形を描くことができる。ダブルクリックすると多角形の描画が終了する。

MemoPad　付箋のように画像の上にメモを付けることができる。

Bookmark　画像上の特定の場所に、目印となるピンを置くことができる。その場所は記録され、自由に移動できる。

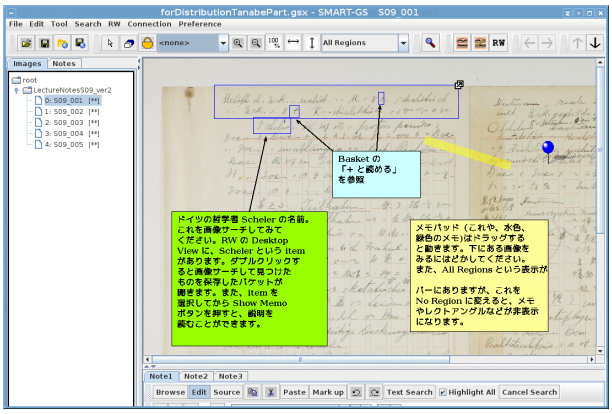


図 2　画像に対するマークアップ

図 2 は、画像上に Rectangle や MemoPad などのマークアップを行ったものである。

これらは、メニューでマークアップを選択した上で利用する。これにより、従来、紙の史料に直接書き込んで行っていたような作業をそのままコンピュータの画像上で行うことができる。さらに、これらのマークアップは、一度作成してから色を変更したり、場所を移動させたりすることもできる。もちろん、後からマークアップを削除することも可能である。紙の史料では行えない柔軟なマークアップを行うことができる。

なお、これらのマークアップの情報は、画像とは別に、gsx ファイルと呼ばれるファイルに保存されており、元の画像自体は変更されない。gsx ファイルには、前述の IDAT の内容やマークアップの種類や座標などが XML 形式で保存されており、画像については、その画像が保存されているパスが記載されているだけである。したがって、オリジナルの史料を別に保存しておく必要もなく、他の研究者と協働で研究を行うことも容易である。つまり、参照している画像とその置き場所さえ共通に持っていれば、マークアップなどは、サイズの小さい gsx ファイルだけをやり取りすればいいということになる。

### 2.3　リンク

さらに、SMART-GS では、画像とテキストに関して相互にリンクを貼ることができる。

元々、SMART-GS のテキストは HTML 形式であるので、HTML のタグとしてのリンクをテキスト間で張ることができる。それに加えて、画像のマークアップ間のリンクや画像のマークアップとテキストの間のリンクも張ることができる。

また、SMART-GS のリンクは、双方向的であり、かつ、1 対多対応になっている。双方向的であるとは、リンク先からリンク元に戻ることできるということであり、1 対多対応になっているとは、一つのリンク元から複数のリンクを張ることができるということである。これは、HTML のリンクにはない機能である。この機能を用いることにより、例えば、語句の索引を作ることも可能である。具体的には、その史料に現れる語句のリストを Note などのテキストで作成した上で、その語句から画像上でその語句が出現する場所のマークアップへのリンクを張る。リンクは 1 対多であるので、テキストから複数の場所へリンクを張ることが可能である。さらに、索引が完成すれば、画像上の語句からリンク元であるテキストに戻ることもできる。

さらに、リンクには、そのリンク自体の説明の記載することができる。これで、何のためにそのリンクを作成したか記録することができる。また、リンクには、作成時間と作成者の名前が記録されるので、協働研究の際に便利である。図 3 はそのリンクを Local View を用いて表示させたものである。

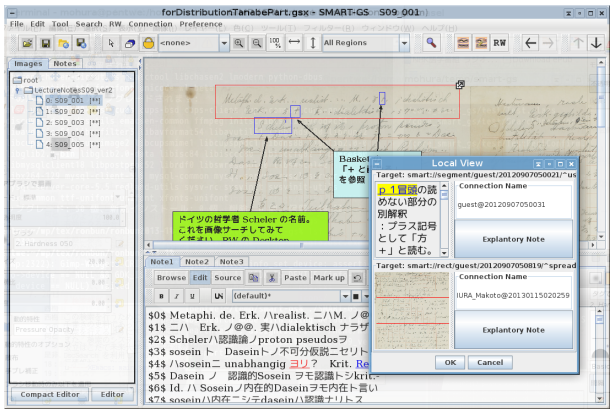


図 3　Local View でリンクを表示

さらに、図 4 のように、Reasoning Web という機能でリンクの一覧も表示できる。

## 3　画像に基づく検索

SMART-GS では、画像に基づく検索を行うことができる。

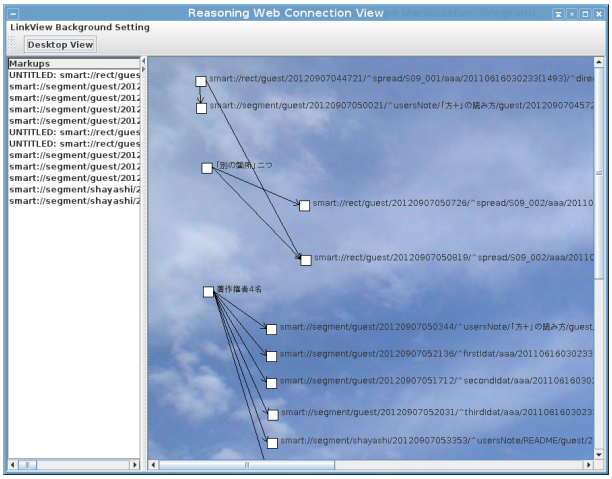


図 4　Reasoning Web でリンクの一覧を表示

手書き史料の場合、筆跡の悪さなどのため、その文字列がどう翻刻できるか判断することが困難なことがある。そのような場合、同じ史料の中で、同様の文字列がどこに出てくるか調べることが有効になる場合もある。また、判断が容易に可能であっても、同じ文字列がどこに出現するか手軽に検索できると便利である。そのために、SMART-GS では、特定の画像を Rectangle で選択した上で、その画像を検索できる。

図 5 は、画像に基づく検索の結果表示である。上部に検索した画像が、下部に検索結果が表示されている。なお、この検索は画像の類似性に基づく検索であるので、検索結果から、それが正しい結果であるかどうかを人の目でチェックする必要がある。結果の表示の左側に「Yes」と「No」のチェックボックスが付いている。その上で、その結果を元に、さらに芋づる式に検索を続けることができる。

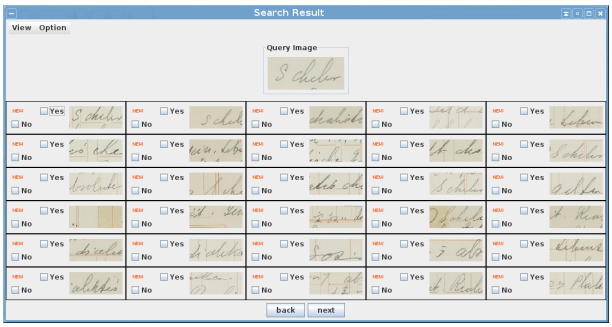


図 5　画像に基づく検索結果

この検索を行うためには、画像に行の情報が付け加えられている必要がある。手書きの史料であっても、大抵行の概念が入っていて、検索を行えるようにする

ためには、まず行に分ける作業をする必要がある。また、検索のシステム自体は、公立はこだて未来大学寺沢憲吾准教授が開発した画像検索エンジンの DscSearch[*4] を利用しており、行情報から、DscSearch の形式に変換する必要がある。

## 4 終わりに

SMART-GS の開発チームでは、今後の課題として、この画像検索の機能を用いた「手動 OCR」のようなものができないかと考えている。つまり、検索結果を人の目でチェックする際、検索結果が正しいと判断された画像に対応する翻刻を IDAT に自動的に表示できるようにすれば、OCR が使えないような史料であっても、それに近いものとして使うことができる。また、既存の OCR が利用できるような活字の印刷物に対しては、OCR の結果を IDAT に結び付けることにより、翻刻の補助にできるのではないかと考えている。

また、現状の SMART-GS は、テキストに対して、HTML の簡単なマークアップしか適用できないが、これを人文学の文書におけるマークアップ言語である TEI(Text Encoding Initiative)[*5] に拡張できないかと考えているところである。

なお、SMART-GS の開発チームでは、SMART-GS を実際の研究に利用してもらって、使い勝手や機能のリクエストなどをフィードバックしてもらうようにしている。最近では、「経済学史ヤングスカラーセミナー」[2] や「白眉センター&応用哲学・倫理学教育研究センター共催セミナー」[3] などで、SMART-GS の使い方に関する発表を行い、普及に努めている。

さらに、本稿ではあまり触れなかったが、SMART-GS は、史料を協働で研究し、翻刻するために使われていることが多く、今後は、ネットワーク上で協働作業ができるようなシステムに拡張していく予定である。実際、インターネットのサーバーにリポジトリを置き、プログラミングで用いるようなバージョン管理システムで、協働翻刻できるようなシステムがすでに動いている。SMART-GS プロジェクトでは、今後、これをさらに拡張していく予定である。

[*4] http://km.meme.hokudai.ac.jp/people/terasawa/WS/dscsearch.html (2013 年 1 月 15 日閲覧)

[*5] http://www.tei-c.org/index.xml

## 参考文献


[1] SMART-GS/HCP プロジェクト. 「SMART-GS マニュアル」. http://smart-gs.sourceforge.jp/manual/ja/ (2013 年 1 月 14 日閲覧).

[2] 林晋. 「Smart-GS による手稿解析——その実際」. 経済学史学会 ヤングスカラーセミナー, 大阪学院大学, 2012 年 12 月.

[3] 久木田水生, 林晋, 大浦真. 「SMART-GS 歴史的文献研究のためのソフトウェアツール」. 白眉センター&応用哲学・倫理学教育研究センター共催セミナー「デジタル・ヒューマニティーズの現在」, 京都大学, 2012 年 12 月.