

SMART-GS プロジェクト

歴史的文献研究のためのソフトウェアシステムの開発[†]

久木田水生[§]

名古屋大学「人類文化遺産テキスト学研究センター」公開セミナー

2014年9月15日

1. 概要

SMART-GS は歴史的史料の研究を支援するために、京都大学の林晋教授（文学研究科情報史料学専修）を中心に開発されているソフトウェアです。「GS」は「Geschichte Studie」（ドイツ語で「歴史研究」の意味）の略です。SMART-GS は画像ビューワーとテキストエディターが統合されたもので、画像とテキストにマークアップをつけ、それらの間にリンクをつけることができるのが特徴です。

SMART-GS は GPL2 ライセンス¹の下で公開されており、誰でも自由に利用・配布でき、またそのソースコードを改変する事も自由です。

SMART-GS は SourceForge.JP のサイトで公開・配布されています。URL は以下の通りです²。

<http://sourceforge.jp/projects/smart-gs/>

マニュアルは以下で読むことができますので、使い方について詳しくはこちらを参照してください。

日本語：<http://smart-gs.sourceforge.jp/manual/ja/index.html>

英語：<http://smart-gs.sourceforge.jp/manual/en/index.html>

2. 機能

2.1. 画像の閲覧とテキストの編集

SMART-GS を利用している画面は Figure 1 ようになります。この図のように史料画像を表示しながら、テキストを編集することができます。テキストにはそれぞれの画像に付随するものと画像とは独立したものがあります。例えば文書全体についてのデータや翻刻プロジェクトの紹介などは画像とは独立したテキストとして記しておくといでしょう。

[†] この資料は 2014 年 9 月 15 日に開催された名古屋大学大学院文学研究科附属「人類文化遺産テキスト学研究センター」公開セミナーにおいて配布した資料を若干修正したものです。

[§] 名古屋大学大学院情報科学研究科。minao.kukita@is.nagoya-u.ac.jp

¹ 使用者がソフトウェアを自由に使用、改変、複製ができることを保証するライセンス。フリーソフトウェアの思想に基づく。

² 現在のバージョンは 0.9.3 で、これは 2013 年 7 月 12 日にリリースされたものです。今月末に大幅に機能を拡張した新しいバージョンがリリースされる予定です。本日の説明とデモは新しいバージョンに基づいています。

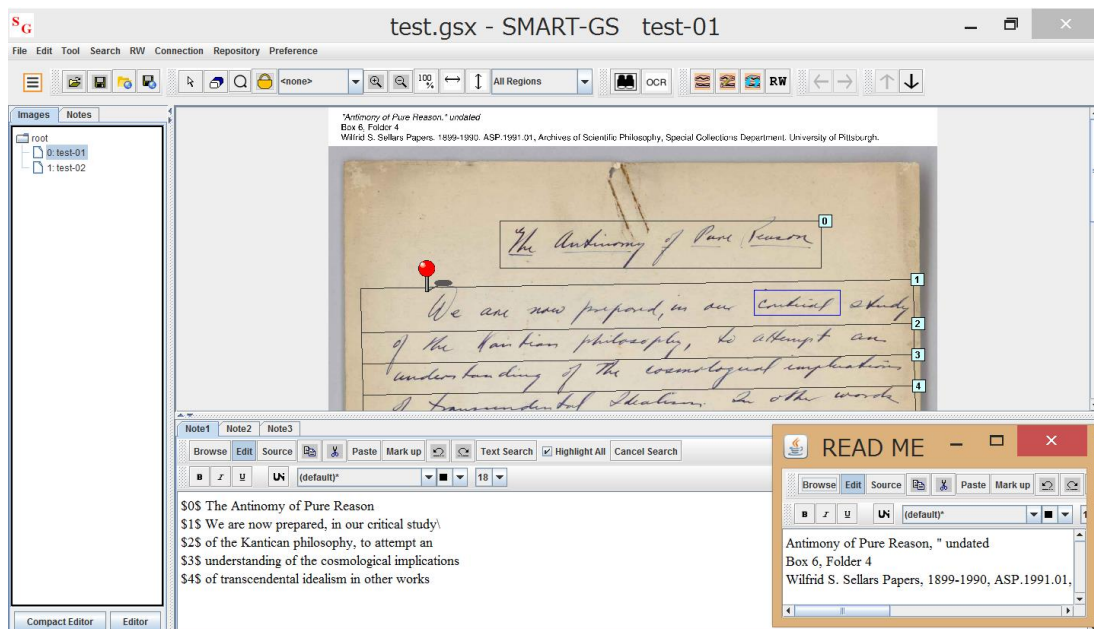


Figure 1. SMART-GS の利用画面

2.2. マークアップ

テキストと画像にもマークアップをつけることができます。私たちが紙に書かれた文書に対して行うこと（線で囲む、マーカーでハイライトする、メモを書き込む、しおりを挟むなど）を電子テキストにもできるように工夫されています。

2.3. リンク

マークアップの間にリンクをつけて、リンクされたマークアップをすぐに参照できるようになっています。ウェブブラウザでサイト間のリンクをたどるのと同様と考えてください。ただしウェブサイト間のリンクが多対一になっているのに対し、SMART-GS のリンクは多対多になっています。またウェブサイト間のリンクは、リンク元からリンク先への移動しかできませんが、SMART-GS のリンクでは両方向の移動が可能です。

2.4. 検索

テキスト内の検索だけでなく、画像上の文字列を、形の類似性に基づいて検索ができます。形の類似性によってあるのでどんな言語でも検索ができます。縦書きにも対応しています。

画像検索ができるためには、画像に行が指定されていて、さらにその行情報に基づく DSC ファイルというものが作成されている必要があります。この説明はここでは割愛します。詳しくはオンラインマニュアルを参照してください。

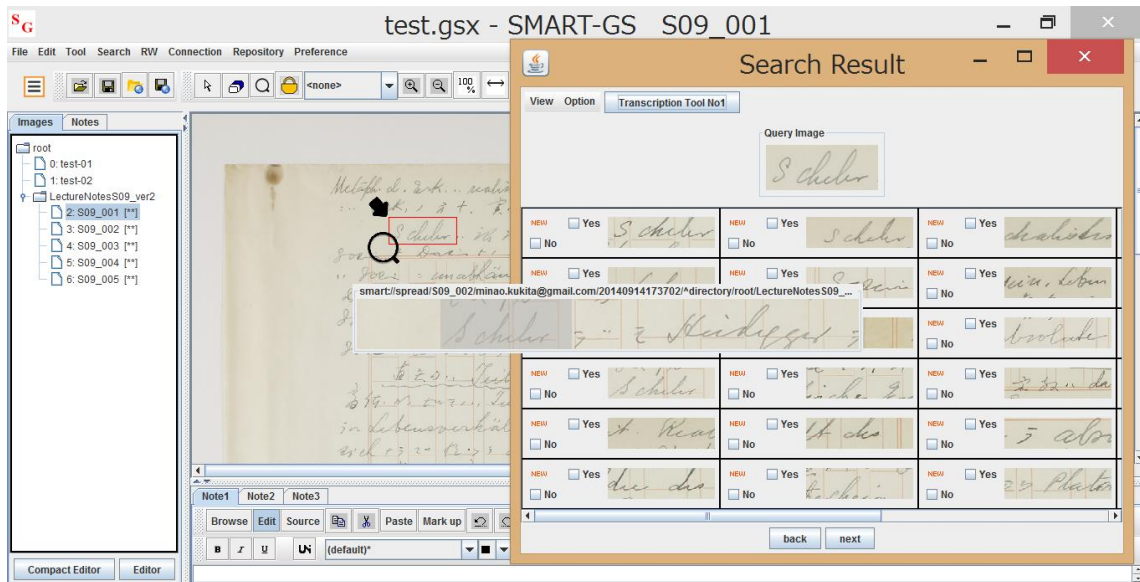


Figure 2. 手書き文字の検索

2.5. ファイル

SMART-GS では GSX ファイルという形式のファイルを使っています³。拡張子は「.gsx」です。このファイルは実際には ZIP ファイルになっていて、その中には XML ファイルが含まれています。XML には翻刻や注釈等のテキスト情報、マークアップやリンクの情報などが記録されています。史料画像そのものは含まれていません。画像ファイルは別のところに保存されているという前提です。これは「歴史史料（の画像ファイル）は変更されることがない」という考えに基づいた設計です。また画像検索をするために必要な DSC ファイル（2.4 参照）も GSX ファイルとは別の場所に保存されていることになっています。そのため GSX ファイル自体のサイズは小さくなっています。

2.6. TEI 対応

TEI は Text Encoding Initiative の略で、人文学系の資料を電子化する際のマークアップ方式を標準化することを推進している非営利団体、またはそのマークアップ方式のことです。従来の SMART-GS は、独自のマークアップを用いてきましたが、今後は TEI に対応したマークアップ方式を取り入れることになっています。

2.7. OCR

SMART-GS は手書きの原稿の翻刻を主眼に置いていますが、古いタイプ原稿などを対象とするときには OCR が有効です。次にリリースするバージョンでは、Tesseract という Google が開発している OCR ソフトを SMART-GS から呼び出して、文字認識をすることができるようになります（Tesseract のインストールはユーザーが別に行う必要があります）。OCR を使った時には自動的に行の情報も作成されます。

³ 以前のバージョンでは GS ファイルという形式でした。

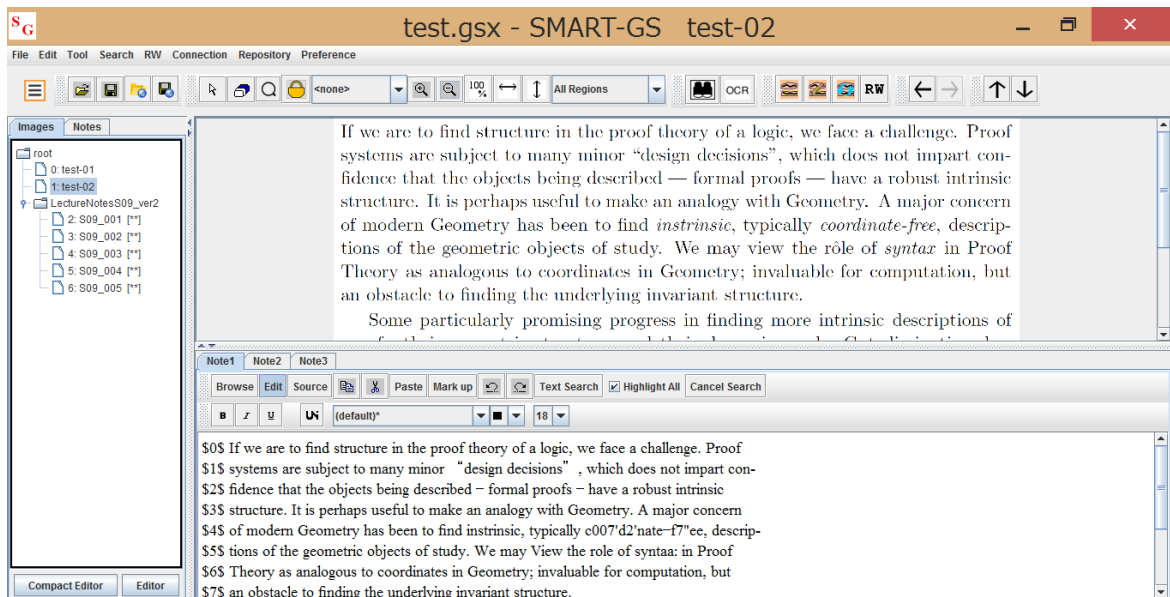


Figure 3. OCR による文字認識

3. 使用例

私たちが把握している限り、SMART-GS を利用した翻刻プロジェクトには次のようなものがあります⁴。

- 数学者ダーフィット・ヒルベルトの手稿研究（京都大学・林教授）
- 哲学者田邊元の手稿研究（京都大学・林教授）
- 倉富勇三郎日記翻刻プロジェクト（京都大学・永井和教授）
- 内海忠司日記翻刻プロジェクト（近畿大学・近藤正巳教授）
- テルグ語インド古典文献研究（京都大学・志田泰盛助教）
- ガリレオ手稿研究（京都大学・伊藤和行教授）

4. SMART-GS の特色

4.1. 文献研究を専門にするプロの人文学者のためのツール

ソフトウェアには、すぐに使えるようになるが機能が限定されているものと、十分に使いこなすには習熟を要するが高機能なものがあります。例えばメモ帳とエクセルのようなソフトを思い浮かべていただけるとわかりやすいでしょう。エクセルを十分に使いこなすには時間がかかりますが、ひとたび慣れてしまうとそれなしで作業をするのが考えられないようになります。SMART-GS はそのようなソフトウェアであることを意図して作られています。もちろん可能な限り直観的に使いやすいものであるように心がけてはいますが、しかしやはりその機能を十分に使いこなすまでのハードルは低くありません。それでもプロの研究者にとって本当に有用な機能が備わっており、翻刻の作業を大幅に効率化できるものを目指しています。

4.2. 協働翻刻の促進

SMART-GS は特に協働作業の場面において大きな力を発揮します。上述したように、SMART-GS が扱う GSX ファイルは、基本的に XML ファイル（を ZIP 化したもの）ですので、比較的サイズが小さくて済みます。各自が画像と DSC ファイルを別に持っておけば、GSX ファイルだけのやり取りで簡単に翻刻プロジェクトを共有することができるのです。また画像につけられたマークアップにはユーザーの ID や作成された時間を含む、一意的な識別記号 (URI) がつけられるため、誰がどのような編集をしたかを知ることができます。また SMART-GS

⁴ 橋本雄太「集合知で読む歴史史料—SMART-GS が実現するグループリーディング」、『人文情報学月報』、37号（前編）、2014年8月25日、http://www.dhii.jp/DHM/DHM37_smartgs

の UI には、複数の研究者によるディスカッションを効率化するための工夫が色々と施されています。画像とテキストの行対応、ルーペ、一時的マークアップなどの機能は、実際に協働翻刻を進めていく上で、こういう機能があったら便利だという認識から生まれてきたものです。

協働翻刻の重要性は現在、様々な翻刻プロジェクトの中で認識されています⁵。私たちが推進している田辺元の手稿の翻刻プロジェクトでも、協働翻刻は大きな力を発揮しています。田辺の手稿は非常な悪筆な上に、新カント派、ドイツ観念論、現象学、数学、物理学の専門用語、日本語（の古いくずし字）、ドイツ語、フランス語、古代ギリシャ語などが混ざるために非常に読みにくく、これまでは日本哲学の専門家でも匙を投げてきました。しかし現在、日本哲学を専門家としない研究者チームの協働によってこのテキストの翻刻が進められています。

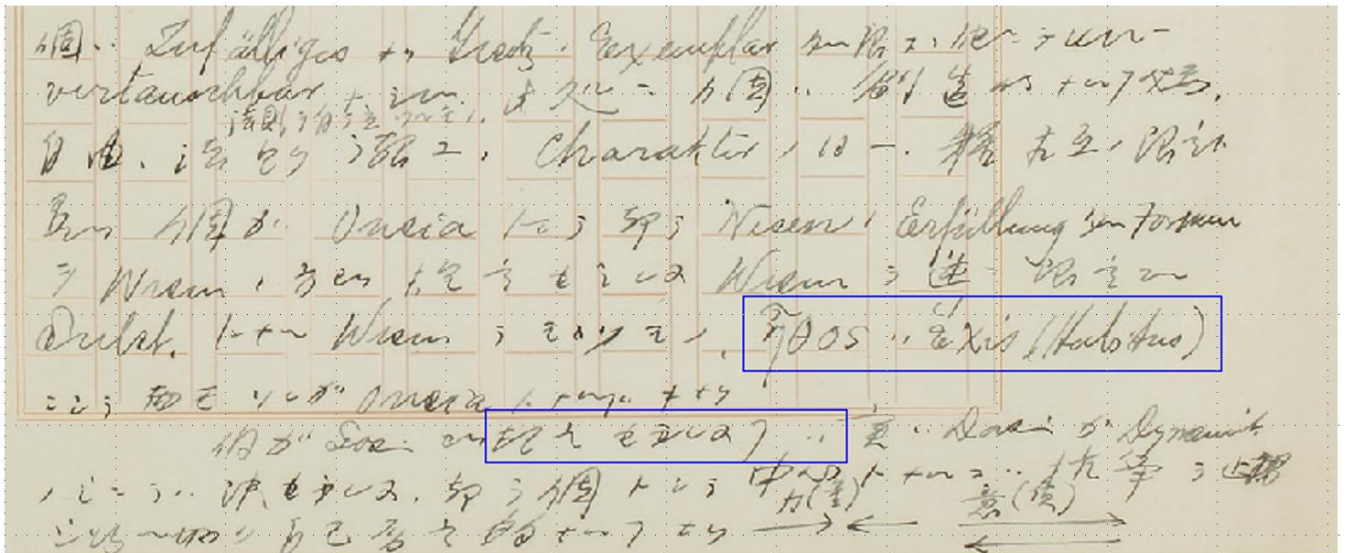


Figure 4. 田辺元の手稿。上の青い四角には古代ギリシャ語、下の四角には日本語の古いくずし字が含まれる

4.3. 開発のスタイル

SMART-GS の開発チームに特徴的な点の一つは、リーダーの林先生をはじめとしてその主要メンバーが人文学者で、かつ SMART-GS の一番のユーザーであるという点です。開発チームは週に一度、開発ミーティングと SMART-GS を使った田辺元の手稿の研究会を行い、そして次の週のミーティングと研究会までに各自がコードを書き、翻刻を進めるというやり方を取っています。このようにして協働翻刻の現場で何があれば便利かということが認識されると直ちにソフトウェアの機能にそれが反映されます。

人文情報学の一プロジェクトとして考えた時に、この SMART-GS の開発のスタイルは興味深いものかもしれません。というのも人文情報学ではツールの開発、実践、方法論の開発という三つのフェイズがいかに効率よく影響しあうかということが非常に重要だからです。一般に人文情報学のプロジェクトにおいては、ツールの開発者は工学者でユーザーは人文学者という組み合わせが多いと思います。そのため実践と開発の間にはどうしてもギャップが出てしまいます。SMART-GS の開発スタイルは、人文情報学の効率的な方法論の一例を提供するかもしれません。

⁵ 例えば University College London の Transcribe Bentham プロジェクト ([http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe Bentham](http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe+Bentham)) など。