"Digital History/Humanities Research Methodologies in the Study of Environmental Changes and Illness in Taiwan"

- DRAFT -

First World Congress of Taiwan Studies
Academia Sinica, Taiwan, April 26-28, 2012

Joseph Wicentowski

I am very excited to have the opportunity to speak here at the First World Congress of Taiwan Studies, a conference that marks the maturation and globalization of this interdisciplinary field of Taiwan Studies. This gathering can serve two important functions: (1) showcasing the latest research in the field and cultivating the many exciting research projects on display here, and (2) identifying ways that we can advance the field. By advancing the field, I mean educating ourselves about the latest advances in scholarly practices, embracing methods that will help raise the profile of Taiwan Studies in international academic circles and that will help continuing to attract new talent with diverse perspectives to the field. With these goals in mind, I would like to focus my presentation on an important opportunity before Taiwan Studies today: digital humanities.[1]

The field of digital humanities[2] has developed in academic circles as a response to the promise of digital technologies for enhancing research and teaching methods and

[1] What I have to say about the digital humanities applies equally well to the study of environmental change and illness in Taiwan as to any other subfield of Taiwan Studies, so I will speak to the entire field of Taiwan Studies rather than to any specific subfield. Public health researchers such as Hans Rosling have been at the forefront of using data visualization tools to convey complex data to audiences, even those new to the field. See Rosling's website and tool, "GapMinder," http://www.gapminder.org/.

[2] As digital methods come to each discipline, we see "digital" or "computational" prepended, e.g., digital history, computational statistics. The term "humanities computing" predated "digital humanities." For one discussion of the history of digital humanities and evolution of the community, see Patrik Svensson, "The

scholarly communication. Digital humanities broadly encompasses the work of humanities scholars who use digital sources and tools in their research, teaching, or publishing. Since these methods and tools are in their infancy compared to traditional scholarly methodologies, many digital humanists actively work to share their digital data, methods, and toolsets with others to encourage their spread and the understanding of their impact. Digital humanities has been gathering institutional steam. In the last 5-10 years, many universities have begun hiring faculty in the field of digital humanities, conferences are organized around digital humanities, universities are establishing centers for digital humanities research to serve as hubs of research and teaching, scholarships and post-doctoral fellowships are being offered, and government institutions like the National Endowment for the Humanities and private foundations like the Andrew W. Mellon Foundation have established grant programs to support digital humanities research projects.

Scholars and projects involved in digital humanities typically use the following kinds of techniques:

- data mining techniques from computational linguistics or genomics to find patterns and correlate features in vast amounts of text or other source material
- data visualization techniques from the statistical sciences to represent complex data in visual forms, either simplifying the data or revealing otherwise unknown features

Landscape of Digital Humanities," Digital Humanities Quarterly, Summer 2010, Volume 4, Number 1, http://www.digitalhumanities.org/dhq/vol/4/1/000080/000080.html.

- information retrieval techniques to sort vast numbers of search results according to relevance algorithms
- data modeling techniques to represent complex information such as textual documents
- digital tools to foster intellectual collaboration and dissemination of research.

Having characterized a number of the techniques used in digital humanities, let us examing some examples. One promise of digital humanities research is to help us deal with the challenge of so-called "big data": Imagine a dataset on the scale an entire library, an entire archive, or the entire world wide web. Datasets so large and complex have never been feasible for use in research, because the data is so big. New tools and methodologies are becoming available for making sense of such data and answering research questions. For example, the Google's Google Books project has been scanning hundreds of thousands of books from university libraries in the United States.[3] To date, Google has scanned an estimated 4% of all books ever printed. This is literally more data than a researcher or research project could ever hope to ingest and process in a lifetime. But now that this data has been organized, researchers have begun to use it for fascinating research. Researchers have used a free, online tool called the Google Books NGram Viewer to challenge long-held hypotheses about changes in the concepts of science and religion in Victorian England.[4] They have also demonstrated the real effects of propaganda and censorship during the 1930s in Germany and in 1989 following the Tiananmen Square Massacre, convincingly exposing terms that were systematically

[3] "History of Google Books," http://books.google.com/googlebooks/history.html.

[4] The tool is available online at http://books.google.com/ngrams. The results are discussed in Patricia Cohen, "Analyzing Literature by Words and Numbers", New York Times, December 3, 2010, http://www.nytimes.com/2010/12/04/books/04victorian.html.

excluded from publications in the countries with censorship, while in other countries during the same periods, those terms remained in use.[5] The team behind these research efforts have coined a new term to describe their methodology: culturomics. They define culturomics as "the application of high-throughput data collection and analysis to the study of human culture."[6]

On one hand, the Google Books project has limited applicability to the field of Taiwan Studies, because the corpus of the Google NGram Viewer tool includes a relatively small set of materials in Chinese, and simplified Chinese at that. The corpus could certainly become more useful to Taiwan Studies as Google scans incorporates more material from Taiwan, and even now, the tool could be useful for tracing international usage of terms relevant to Taiwan and helping test hypotheses for certain research questions. At the same time, I raise this tool because it illustrates not only the potential for the "big data" approach to digital humanities projects, but also the requirements for digital humanities research to be possible: 1. Digital humanities research requires investment in digitizing texts, 2. Data must be open to researchers, 3. Researchers need tools (or need to build tools) for exploiting the data. These are the real lessons of the Google Books project for Taiwan Studies.

The first point is important because some funding is required for any research and particularly cutting edge digital work. Like any aspect of scholarly endeavors, digitization requires resources. And the resources needed go beyond the cost of scanning. Scanning is a key step in the digitization of historically significant texts, but it is far from

[5] "Quantitative Analysis of Culture Using Millions of Digitized Books", Science, December 16 2010, DOI: 10.1126/science.1199644, http://www.sciencemag.org/content/early/2010/12/15/science.1199644.
[6] "Culturomics", http://www.culturomics.org/.

sufficient. Taking a digital image and making it "searchable" can be an expensive and difficult task. Optical character recognition technology is viable only in cases where the source text is in very clean shape. The bulk of historical manuscripts are not in such condition.[7] They require people to painstakingly manually enter the text into a computer. While commercial services do exist that perform this work at an increasingly affordable rate, many digitization projects simply lack the funding to pay for the digitization of the full text.[8] Many digitization programs stop once they have scanned the data, and simply post the images and some limited metadata about the image, such as the date of the original manuscript. Google was able to scan the contents of many university libraries because the universities provided this access to Google for free; Google scanned the books for free and made its NGram dataset available because it has a business plan for monetizing the data it scans (and, even so, remember that it is giving away the NGram dataset but not the raw data for all of the books it has scanned, not even to the universities whose libraries supplied the books). So how can digital humanities researchers and projects, who are not doing this work to make a profit, have any hope at succeeding?

Luckily there are other techniques and models for funding digitization projects, including seeking funding from public sources or private foundations and lowering costs using open source software and crowd sourcing. Governments or private foundations are wise to fund digitization projects, providing that the right technologies are selected, and

[7] Simon Tanner, Trevor Munoz, and Pich Hemy Ros, "Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive", D-Lib Magazine, July/August 2009, Volume 15 Number 7/8, ISSN 1082-9873, http://dlib.org/dlib/july09/munoz/07munoz.html.

[8] An example of the economics of using a commercial text scanning and digitization services: Most services charge at a rate "per thousand characters" delivered by the vendor after scanning and preparation of the text. A printed book page might contain 2,500 characters, and the XML tagging that many projects require might add 20% to the number of characters, so at $0.35 per thousand characters, the cost would run about $1 per page.

that the data is made freely available to researchers. But sometimes public funding is inadequate. Is the only alternative licensing the data to commercial publishers, who will then retain sole rights and charge all users fees for access? Two alternative techniques for lowering costs include open source software and crowd sourcing. Open source software is software that is free to download and use, and whose source code is completely open so that it can be improved by its users.[9] All digital humanities projects should explore using open source software, because it is cost effective, provides comparable or even better results than commercial alternatives, and the when projects use open source software, the improvements they make are shared by all.[10] Another alternative to make funding stretch further is to use crowd sourcing. The "Transcribe Bentham" project at University College London proves that crowdsourcing is a viable means of digitizing scanned images.[11] The project publicly solicited volunteers from around the world to transcribe Jeremy Bentham's folios—texts that are very difficult to read. Experts reviewed and edited the transcriptions, and it proved that volunteer labor is a viable method for digitizing texts.[12] The United States National Archives and Records Administration has embarked on a new crowd sourcing initiative to solicit help from the public in transcribing scanned archival sources.[13]

[9] Among many sources describing the benefit of open source software, see Katherine Noyes, "10 Reasons Open Source is Good for Business", PCWorld, November 5, 2010, http://www.pcworld.com/businesscenter/article/209891/10_reasons_open_source_is_good_for_business.html.

[10] One excellent example of an open source project used by many digital humanities researchers and projects is eXist-db, an open source native XML database. This database allows users to store XML data of any type, freely query the data, and create websites to allow others to explore the data. See http://exist-db.org/.

[11] "Transcribe Bentham" homepage, http://www.ucl.ac.uk/transcribe-bentham/.

[12] Patricia Cohen, "Scholars Recruit Public for Project," New York Times, December 27, 2010, https://www.nytimes.com/2010/12/28/books/28transcribe.html. Also, see the forthcoming article in Digital Humanities Quarterly.

[13] "Citizen Archivist Dashboard," National Archives and Records Administration, http://www.archives.gov/citizen-archivist/. Previous initiatives at NARA have allowed citizens to scan

But the Google Books research and crowd sourced transcription projects are far from the only example of successful digital humanities research, and these are far from the only lessons for Taiwan Studies.

In contrast to the “big data challenge,” other digital humanities projects have focused on enabling scholars to annotate texts in ways that enable detailed analysis using those annotations. A case in point is the Text Encoding Initiative (TEI), a mature, open standard backed by a consortium of scholars that has developed a rich XML-based language for encoding texts and annotating them.[14] TEI can be used to capture books, manuscripts, poems, archival documents—any text. One of the best illustrations of TEI’s power is how it changes the discipline of annotating primary sources, like manuscripts or archival documents. Traditional scholarly printed editions of primary sources have had a handful of means for annotating texts, including: page layout, footnotes, glossaries, indexes, and prose introductions. Page layout refers to the layout of the text on a page, and it implicitly conveys the way the scholar believes the text should look. But the footnote is the tool with which the scholar explicitly explains a source: a footnote lets a scholar provide context or commentary directly to a point in the text.

TEI builds on these conventions of scholarly annotation and vastly enriches them. TEI retains footnotes but adds a rich vocabulary of analytical annotations, in the form of structured XML tags: tags for person, place, dates; handwriting, spoken text; prosopography, bibliography, manuscript descriptions; and so on.[15] Most powerfully,

public domain films held by the archives and make them available online; see “Citizen Archivists Making an Impact at the National Archives,” http://blogs.archives.gov/aotus/?p=1204.

[14] “Text Encoding Initiative Homepage”, http://www.tei-c.org/.

[15] The full vocabulary of tags, along with all of the rules for creating a TEI document and extending its vocabulary, is described in “The TEI Guidelines”, http://www.tei-c.org/Guidelines/. For more on how TEI

TEI annotations are written in a simple language that people can read and manipulate, and that XML databases can analyze. If people are key in your texts, you can tag every instance of a person (even if their name has changed, or they are referred to by a different name), and you can link these instances together. If you are interested in the history of medicine, then you know that it is common that many different terms have been used to refer to a single illness or condition. If you have many texts, you can tag each instance of an illness and link them together. If TEI happens to lack a tag or form of analysis that is important for your research, TEI allows you to extend the language and add your own vocabulary. TEI can also be shared among researchers, with each researcher adding their own layer of interpretation.

Since TEI is based on the XML standard, TEI is excellent from an archival standpoint: XML files are plain text files that can be read by any computer program, not just one program sold by a company that can go out of business or change its pricing scheme. I think we have all experienced the problem when we cannot open documents that we wrote ten years ago, because the company that makes the software is out of business. For this reason, using a plain-text based format, based on XML, and defined by an open consortium, is wise for any scholarly project.

Another virtue of TEI is that it is not a dead-end, single purpose format. It can be transformed into many different formats: TEI documents can be transformed into PDFs,

is used in documentary editions, see Joseph Wicentowski, "*FRUS* and the Outlook for Diplomatic Documentary Publishing in the Age of Open Government, E-Readers, and Digital Humanities", in Proceedings of the 11th International Conference of Editors of Diplomatic Documents, Jerusalem, September 19-22, 2011, http://www.archives.gov.il/ArchiveGov_Eng/Publications/ElectronicPirsum/EditorsConference2011/EditorsConferenceContents.htm.

Microsoft Word documents, web pages, and the newest format: ebook formats that can be read on ereader devices like the Amazon Kindle and Apple iPad.[16]

TEI could be very beneficial to Taiwan Studies in a number of ways. First, TEI offers Taiwan Studies a unified, common format for digitization of texts. Taiwan Studies research projects that use TEI would benefit from TEI's sophisticated language for annotating texts, its archival soundness, and its flexibility and ability to be transformed into many formats. The researchers that created the texts could use them for their own analysis, and other researchers could use the same texts to verify research findings and reproduce results, and they could add their own layers of analysis. If a critical mass of Taiwan Studies scholars and project could adopt TEI, the possibilities for advancing the field could be very significant. Let us examine a few examples where Taiwan Studies could benefit from TEI:

1. Scholarly editions of primary sources: Published editions of key, select primary sources are important for fields like Taiwan Studies to identify foundational documents, to provide readers with adequate context from subject matter experts, and to train students in how to read primary sources. In Taiwan Studies we have many examples of scholarly editions, including the multi-volume series of selected documents from the Sotokufu Archives.[17] These series are edited solely for book publication and do not have an online or digital counterpart. If a series like this were created in TEI, it would benefit

[16] My home institution, the Office of the Historian at the U.S. Department of State, publishes the *Foreign Relations of the United States*, the official documentary history of U.S. foreign relations. TEI is the master digital format for our publications, from which the web and ebook versions are derived. See the Office of the Historian website, http://history.state.gov/, and more information about the ebook edition, http://history.state.gov/historicaldocuments/ebooks.

[17] Taiwan zongdufu dang'an website, http://sotokufu.sinica.edu.tw/sotokufu/.

from the enriched set of tags, it could be published in many formats, and the richly annotated source text could be used for computer-assisted analysis.

2. Scholarly publishing: I would like to praise the Institute of Taiwan History here at Academia Sinicia for opening up access to its excellent journal, Taiwan Historical Research (Taiwanshi yanjiu).[18] This is one of the few journals (within Taiwan, within East Asia, and even beyond) that allows full and free access to its back catalog. Each article in each issue is available online as a downloadable PDF. But imagine if Taiwan Historical Research became a TEI-based publication. Becoming a TEI-based publication would allow the Institute to publish each journal article as a webpage, not just as a static PDF file. The journal contents could be searchable on the Institute's website, cross-references to other articles could be live links that would lead readers directly to the target article, and the Institute could build tools for analyzing the articles and trends in the publication. Each article could be downloaded as raw TEI, allowing other researchers to perform their own analysis on the contents. Researchers could even embed their source data (texts, spreadsheets), allowing others to reproduce findings and build on them. Also, we should not neglect TEI's ability to serve and even expand the journal's readership. The new generation of e-readers are making it incredibly convenient to read digital texts portably, but on screens of 9, 7, or 3.5 inches, static PDF articles designed for a full sized piece of paper are too small to read, or they require awkwardly zooming in and panning from line to line. If the article were written in TEI, it could be transformed into e-book formats, which can be displayed on any size screen. Indeed, as these e-reader devices

[18] Taiwan Historical Research website, http://www.ith.sinica.edu.tw/quarterly_01-en.php.

proliferate, readers will begin to expect that our publications are available in these formats, and we risk losing readers if we do not offer our publications in these formats.

Whether we are talking about digitizing primary sources or secondary research, choosing a format like TEI makes sense on many levels: scholarly rigor and precision, scholarly communication and interchange, archivability, and cost.

So far I have focused on promising research methodologies and the benefits of adopting good formats for our data and publications. But to successfully benefit from all that digital humanities has to offer Taiwan Studies, I contend that Taiwan Studies needs to take one more step: it needs to needs to fundamentally reorient its approach to digital data, and open up that data for research.

Taiwan has many exciting digitization programs underway. However, a number of the most rich digitization programs are subject to access restrictions that stifle the ability for researchers to use these resources, and thus, I fear, hinder the growth of Taiwan Studies. These restrictions take one of two forms: use can be restricted to the walls (or IP addresses) of certain institutions, or it can be restricted to holders of a ".gov.tw" or ".edu.tw" email address.[19] These restrictions hinder the ability of scholars who lack the affiliation or who lack the requisite email address from using the resource. Granted, many resources rely on subscription fees to fund maintenance and continued development of the resource, but often there is no subscription plan for an individual researcher. Perhaps some resources could be made available to paid members of

[19] While we do not yet have a systematic review of Taiwan Studies resources online, one data point is that 5 of the 7 resources listed on the Institute of Taiwan Studies digital resources webpage are restricted to users with .gov.tw or .edu.tw email addresses (despite the claim on the website that these resources merely require an "application"). These are some of the most valuable online resources in Taiwan Studies. See http://archives.ith.sinica.edu.tw/en/en-resources.

professional associations, such as any of the regional associations of Taiwan Studies. In any case, such barriers do hinder the growth of Taiwan Studies. Researchers, particularly those active in digital humanities, are drawn to fields that make research materials readily accessible. If Taiwan Studies materials were more freely available online and in structured formats, then researchers who previously might not have considered pursuing research with an aspect about Taiwan would certainly consider it. Particularly if the material that is being digitized is using public funds, the greatest effort should be put on making the materials available for free, or failing that, minimizing the barriers to accessing the data. The digital humanities only succeeds if we share our data and tools.

This philosophy — releasing studies and data produced using public funds — is the basis for Open Access journals and the Open Government Data movement. For universities and institutions that have an Open Access mandate, the research of the faculty and staff (or the research funded by the grantor) must be made available for free online. Open Access does require changes in the business of scholarly publishing, but it means that research can be disseminated more quickly to a wider audience, and research findings can enrich scholarship and make their way into public debate and the classroom more quickly and broadly.[20] For national and local governments that have instituted an Open Government Data policy, these institutions go a step further, not simply committing to release scans or PDFs of government-funded datasets, but committing to release the raw data, in structured, computer readable formats that researchers can import directly

[20] Of the many open access journals, Public Library of Science (PLoS) is perhaps the most well known. See PLoS's statement on open access, "The Case for Open Access," http://www.plos.org/about/open-access/. For coverage of one university's transition to Open Access, see John Timmer, "MIT to make all faculty publications open access," Ars Technica, March 2009, http://arstechnica.com/tech-policy/news/2009/03/mit-to-make-all-faculty-publications-open-access.ars.

into their databases and toolsets, without costly conversion or adaptation.[21] If data and research are already paid for by taxpayers, governments should not put up barriers to accessing the data and research results; governments and the institutions they fund should make the data freely available.

In an age where talented researchers have an abundance of freely available, structured data upon which to perform their research, fields of study that hinder access or release data in formats unsuitable for ingesting into digital toolsets risk being sidelined. In light of this, Taiwan Studies would be wise to redouble its commitment and efforts to digitizing primary and secondary sources and releasing this data in as raw, structured a format as possible. Restricting resources could risk an unfortunate effect: the ability to perform original research in Taiwan studies will have a narrower and narrower base of researchers. Many would argue that if a digitization program is funded by a government or a private foundation, it should be open to all to research — both to citizens and the international community. And if a program simply cannot survive on non-commercial funding and must be done in cooperation with commercial partners, then the rights to the product should be retained by the public institution, and the digitized data should remain open to the public.

Beyond this issue of the production of digital datasets, this is an exciting time to be a scholar of Taiwan Studies. Whereas we once had to rely on journal articles,

[21] The open government movement in United States took a major step forward when President Obama announced the Open Government Initiative, which affirmed the principle that transparency is needed in good government and mandated that all U.S. federal government agencies release their data online in structured formats. The home page for this effort is http://data.gov. For one example of data that is released through data.gov, the Office of the Historian in the U.S. Department of State has released bibliographic data about its publications to the data.gov, and is preparing to release the raw TEI XML data for its publication. See “Office of the Historian: Open Government Initiative”, http://history.state.gov/open. The publication is available for free to an international readership.

monographs, university libraries, and annual meetings to disseminate and house our research, we live in an age where individual researchers can establish online voices and can create direct links to other researchers and to interested readers, including young people who will form the next generation of Taiwan Studies. The immediate feedback researchers receive can help them refine their arguments. Besides blogs, researchers should consider learning TEI and using it to post their primary sources with their annotations, or using other structured formats if they better capture the findings.

In summary, digital humanities offers the field of Taiwan Studies new, robust, open standards for capturing our source materials, new methodologies and tools for enriching, annotating, and analyzing our sources, and highly effective services for publishing our findings and connecting with peers and our audiences. Taiwan Studies has much to contribute back to this new interdisciplinary field of digital humanities, from its rich sources to the high caliber of its researchers and Taiwan's top-notch expertise in technology. Given all of these assets, Taiwan Studies has the opportunity to become a shining example in digital humanities.