Making History: an Emergent System for the Systematic Accrual of Transcriptions of Historic Manuscripts

(in proceedings of the *8th IEEE International Conference on Document Analysis and Recognition (ICDAR), Seoul, South Korea: August 29-Sept. 1, 2005*)

Abstract

To date, only a small percentage of the earth's historical written records have been tapped into to contribute to our knowledge of our past.

Our project aims at leveraging the work of the thousands of frequent or occasional visitors of the world's archives to accumulate and disseminate the raw transcriptions produced by the army of researchers who are digging into this treasure-trove of information from our antiquity. Since written records of our past are by definition finite, our system will put in place a mechanism that promises to capture – once and for all – the written records that our ancestors left behind. This paper describes an emergent system architecture that promises to eliminate redundancy by gradually yet systematically creating a cumulative, distributed, electronic repository of digital transcriptions of the written records from our past, that can be shared and searched without duplications and overlaps. The application is completely based on open-source software tools, which are made available to all users for free on the web. Testing and evaluation of the system will begin in the summer of 2005.

1. Introduction

There are three fundamental sources of historical knowledge: (1) oral traditions; (2) physical artifacts, and (3) written records. The latter are invaluable as keys to reconstructing crucial events of the distant past. Printed books are by now more numerous than manuscripts, but the latter hide some of the more arcane and revealing secrets of yore. Yet, despite the fact that there are over 5500 on-line repositories of primary sources in the world [1] containing millions of manuscript pages, perused by tens of thousands of scholars and visitors every day, and although millions of dollars are spent yearly to fund the transcription of significant historical documents, the world's archives still contain a largely-untapped goldmine of historical knowledge waiting to be unearthed.

The system described herein represents an attempt at streamlining the process of "unearthing" these treasures as transparently and as efficiently as possible, by making it easier for scholars to transcribe and share ancient manuscripts, thus avoiding redundant transcriptions through the creation of a virtuous cycle of automatically-accruing and ever-improving transcriptions of the contents of the world's archives.

This document describes the open-source architecture of this fledgling system, which is slated to be released in its beta-version in the summer of 2005. Only a limited, in-house evaluation of the system's performance has been conducted to date. The basic concept and design of the emergent system are introduced herein, which will later be subjected to rigorous testing in the years ahead.

2. History of the Project

The project started in the late 1990's, spurred by Prof. Reinhold Mueller (Univ. of Venice), Dr. Giovanni Caniato (Venice State Archive), and Dr. Stefano Piasentini, who frequented the *Archivio di Stato di Venezia*. The initial goal of the project was to produce an easy-to-use application that would facilitate the painstaking job of historic transcriptions. As soon as researchers began to bring laptops into the archives' reading rooms, they immediately realized the limitations of plain word processors in the laborious work of transcriptions and began to clamor for some custom-designed tools to make the thankless job as easy as possible. With this goal in mind, the author and Prof. Stanley Selkow (Worcester Polytechnic Institute – WPI –, Computer Science Dept.) began to sponsor undergraduate research projects at WPI in 1998.

The *Transcription Assistant*™ (TA) application began to take shape over a three-year period, as the users' needs were gradually translated into working computer tools [2][3]. In 2001, the Transcription Assistant was ported from Visual Basic™ to Java™ and the current interface was developed (see next page) [4]. In the same year, the author read Steven Johnson's *Emergence* [5], which inspired the design of a true emergent system, wherein the original Transcription Assistant tool is now merely the user-interface – and in a sense the "hook" – through which scholars are enticed into sharing their transcriptions

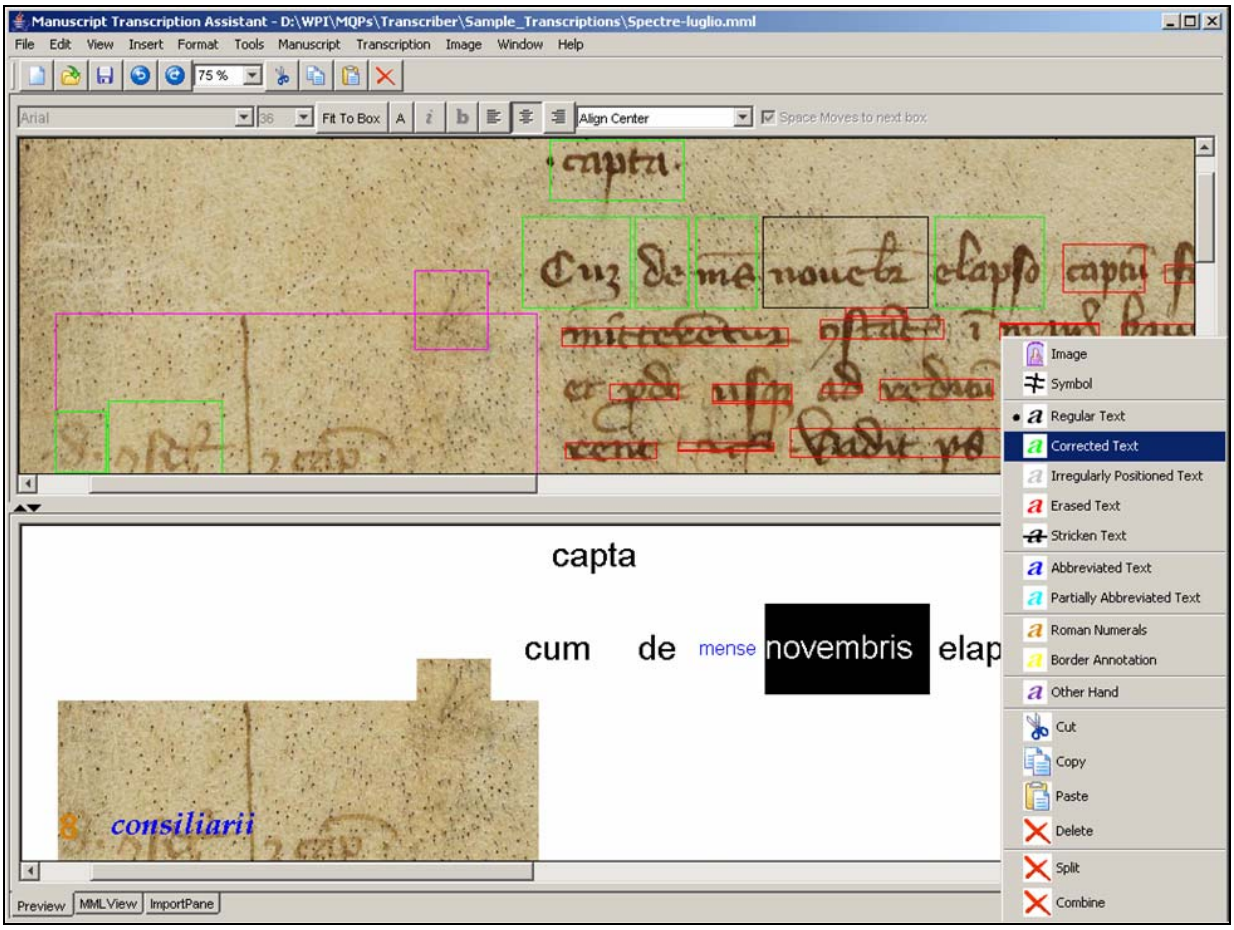

Figure 1. The Transcription Assistant ™ application interface.

with other researchers in a never-ending, self-propagating and self-correcting virtuous cycle. The new "emergent" system promises to greatly accelerate the production of historical studies, by making transcribed primary sources much more available and searchable thanks to the free, open-source, internet-based tools that we are currently developing as part of our *Emergent Transcriptions Initiative* at the Worcester Polytechnic Institute's *Emergent Systems Laboratory*.

3. The Emergent Transcriptions Concept

Our system is predicated upon these assumptions:

- there are precious few researchers with the necessary paleographic skills who are able to produce reliable transcriptions of ancient manuscripts;
- despite this crucial bottleneck, these few capable individuals frequently duplicate efforts by re-transcribing the same exact manuscript that someone else has already worked on, often unbeknownst to each other;
- the constant manipulation of the primary sources (parchments and the like) renders them less and less legible as time goes by;
- very few manuscript transcriptions are published verbatim;
- the work put into transcriptions is subsumed into scholarly journal articles and books, thus it is rarely if ever seen or re-used by others;

Our solution to these conundrums is to put in place a system that will make the voluntary sharing of transcriptions simple and inviting, so that duplication can be virtually eliminated as more and better transcribed text is associated with each manuscript page in the world's archives, in a gradually expanding and self-regulating manner. Once a manuscript page is associated with a transcription, the next time that manuscript is requested by a scholar, the corresponding transcription would also be made available to the researcher, who can thus get a head start on his or her work with only minor overlap. The scholar who receives a pre-existing transcription could make corrections and expansions to it and resubmit it. The transcription will thus be steadily improved until it stabilizes into an well-accepted final version for the benefit of all subsequent examiners of that manuscript page.

A major element of our system is an incentive scheme that will entice users to share freely in exchange for some benefit that they value. One major incentive, right off the bat, is that our applications are downloadable for free as open-source programs.

Another *conditio sine qua non* is that the system that we make available to the users be truly useful to them, so that they will want to use our system in the first place and thus become (potential) participants in the global sharing of transcriptions.

Thus, a virtuous cycle is instituted whereby each transcriber can share his/her work and can in return benefit from the work of others. The gradual accumulation of transcriptions will greatly expedite the making of history in true "emergent" fashion.

4. The Emergent Transcriptions System

The system we are developing consists of three main applications:

1. The **Archive Assistant™ (AA)** that runs on a manuscript server and assists archive personnel in making manuscript images available to end users, with proper manuscript metadata;
2. The **Transcription Assistant™ (TA)** that runs on stand-alone or internet-connected end-user machines and facilitates the job of the researchers by making transcriptions easier to complete, track and manage as projects;
3. The **Contribution Accountant™ (CA)** that would run on a server to keep track of the "credits" and "debits" accumulated by each user.

All three applications are open-source and downloadable for free from the project's website (www.wpi.edu/~carrera/escripts.html).

4.1. The Archive Assistant™

The Archive Assistant™ (AA) is an open-source Java application that runs on a Linux™ machine running an Apache™ server with a MySQL™ database backend. The entire suite of applications is available for free on the internet from the respective open-source providers. The application is partly similar to other programs [6] in that it is meant to assist archives and libraries in scanning, cataloguing and disseminating their manuscript collections. It differs from other similar initiatives in the fact that it enables the institution to gradually receive, collate and accumulate the transcriptions that are returned from the end-users who use the Transcription Assistant™ application described below. Full technical details of the AA application will be provided in future technical papers.

An ineluctable premise upon which our whole initiative is predicated is that institutional archives will make available to transcribers the digital images of the manuscript pages in their holdings, so they can be downloaded via an internet search engine, as is becoming more and more customary at many institutions [1].

Our system is designed to work with any type of manuscript in any language, with any alphabet and of any age, though it is currently being tested on Venetian manuscripts from the XIII-XVII cent. from the Venice Archive and on early American manuscripts from the American Antiquarian Society.

Our emergent transcription system relies on the diffusion of digital images of manuscripts as the basis for the distributed asynchronous production of transcriptions. Scanned images are packaged together with the metadata of the manuscript that they depict, to create an XPG (eXtended jPeG) file. Appropriate metadata accompanies a manuscript to make it usable in a historical context. These metadata are generally already used in the manuscript catalogues in operation at libraries and archives. The XPG file type supports metadata and image packaging into a single XML file.

Our metadata sub-system currently consists of a superset of the MARC and Dublin-Core standards, allowing for the conversion from one standard to the other. We are currently working on AA functions to facilitate the bulk importing of existing MARC and Dublin-Core databases into our system.

More sophisticated components of the AA application are being developed to allow advanced searches on metadata and transcription text, with the possibility of expanding searches in the future to more sophisticated image-based algorithms such as those discussed in Rath *et al.* [7]. After a successful search, users will be able to browse the listed manuscript pages and select them for downloading into their own machine for use with the Transcription Assistant.

After the end-user has transcribed a manuscript page, the XPG file is augmented with an XML-based transcription section, according to the Manuscript Markup Language (MML) that we have developed for the occasion (technical details will be provided in a forthcoming paper). After an initial transcription is made, the manuscript page (manuscript metadata + image + transcription metadata + transcription) is packaged into an MML file from then on.

A yet-to-be-developed component will finally deal with the reception of returned MML files containing transcriptions produced by end-users, in conjunction with the Contribution Accountant and with the backend MySQL database where the XPG manuscript

images and MML files are permanently stored on the archive server.

4.2. The Transcription Assistant™

The Transcription Assistant™ (TA) is an open-source Java application that runs on an end-user machine through the Java Virtual Machine supported by the Java Application Interface.

With the TA, scholars will be able to create a project for each paper or research topic. A project will include several manuscript pages from a variety of collections that together contribute to the development of an historical paper on a specific subject matter.

For each manuscript used in a project, after an XPG image or an MML transcription is downloaded, the transcriber will use our *Transcription Assistant* (see interface in Figure 1) to transcribe all or part of the document. Once the transcription is done, transcribers are invited to submit their transcriptions back to the archive so they can be made available to other transcribers. Specific incentives are used to make this sharing inviting to the user (more on this below and in a separate paper).

The TA is designed to greatly facilitate the process of transcription. It consists of a main screen split into two windows (vertically or horizontally, according to the user's preference). On one window is loaded the manuscript image and on the other is visualized the transcription, either as a positionally accurate print preview, or as word-processable text, or in MML format (see tabs at bottom left of Figure 1).

The first step the TA takes, as soon as a new XPG image is loaded, is to automatically detect word boundaries in the manuscript image and create boxes around each word. The "autoboxdraw" function uses a sophisticated smearing algorithm, developed by one of our past projects [4]. The automatic boxing thresholds and settings are user-adjustable to fit different document conditions. Boxing is quite successful in the current version of the TA application, though a certain amount of manual adjustment is expected to be always necessary. Fortunately, the process of word-boxing needs to only be done once, so the time expenditure is worth the effort. We are currently working on making the manual act of box correction a rewardable, emergent feature of our system as well.

Once each word (or abbreviation) has been boxed, the user can begin the actual process of transcription. Un-transcribed boxes are red. The current box (in black in Figure 1) can be clicked on, to reveal a text field right above it where the transcription can be typed. To emulate the functions of a word processor, the user can move to the next box by simply hitting the space bar in between boxes. Once a transcription has been typed in, its box will turn green and its translation will be entered into the transcription MML and will appear in its exact relative location in the preview window (below the image in Figure 1).

The current version of TA allows the user to "right-click" on a box to annotate the transcription. A primary form of annotation has to do with differentiating graphics or symbols from text. If a box is tagged as an "image" or "symbol" a cropped piece of the manuscript will be copied into the preview window (and into the underlying MML) as shown in Figure 1 (lower left of preview window). Currently we provide the following other types of annotations for text boxes: (i) manuscript annotations such as for stricken text or corrected text; (ii) tagging of abbreviations; (iii) identification of numbers and (iv) identification of handwriting changes (different author). We foresee adding more of these annotations – such as the tagging of currency and marginalia – as well as second-order tags to identify proper names, names of places, professions, dates and the like.

The one-to-one pairing of word-boxes-to-text-transcriptions is broken only by abbreviations, wherein a single word box can be exploded into more than one transcription word. In any case, this pervasive one-to-one correspondence allows the accrual of handwriting-recognition capabilities, which are planned for future versions of the software. We foresee that when users will experience difficulties in transcribing a specific word, they will be able to ask for help by hitting a help key (such as F1).

Using the manuscript metadata for a bounded search – limited to manuscripts that are likely to have the "same hand" – the system will be able to pattern-match the handwriting in the box where the user is having trouble, with a storehouse of boxes from previously completed transcriptions from the same source, yielding transcription suggestions ranked by their different levels of matching. The user will thus be able to pick the suggestion that best fits the sentence being transcribed. We foresee making this advanced capability available "for a fee" in order to fuel our incentive program. We want to entice users to submit completed transcriptions in order to get the credit they need, so they can later spend their banked credits to "buy" services like this "transcription help".

Once a transcription has been completed, the user can save it and/or upload it back to the originating archive server for credit accounting. After several manuscripts have been worked on, the user can also save the project and wrap things up for the day.

More technical details about the TA will be provided in a forthcoming paper.

4.3. The Contribution Accountant™

The Contribution Accountant™ (CA) is an open-source Java application that runs on a Linux™ machine running an Apache™ server with a MySQL™ database backend. This may or may not reside on the same server as the Archive Assistant (AA). This application is currently only in the planning stages.

We foresee an unobtrusive and semi-transparent accounting system that will keep track of the contributions made by each transcriber in order to attribute appropriate *credit* to the authors that submit new or improved transcriptions. Such contributions are rewarded by a sort of "monopoly money" that can be used to acquire advanced services from our system as discussed above. Transcribers would also be rewarded through an automatic academic citation system that recognizes an author's contribution as a bona-fide intellectual property which ought to be properly referenced whenever an author's transcription is used in a project or cited in a paper.

The credit system is predicated on the registration of each user and on the system's ability to unequivocally recognize legitimate users through passwords. After registering, users will receive a virtual "bank account" where their credits will be posted by the CA. Users will also immediately be assigned a "credibility rating" not too dissimilar to the "reputation management" system" used to rate *e-bay*™ sellers or *slash-dot*™ contributors [8].

The essence of the CA accounting system is based on before-and-after comparisons between what the AA sent out to the user and what comes back from the user for each manuscript page. A differential engine will quantify the number of changes made to the original transcription as well as to the word boxes in the manuscript image. The net change will be the basis for the credit awarded to the user. However, to avoid spurious submissions from users bent on "gaming" the system, the initial credit would be awarded on a provisional basis as a percentage of the total earned by the transaction. The "discounted" rate of the temporary credit will be based on the credit-worthiness of the user, which depends on his or her credibility, which is accrued over time, by indirect peer review.

Peers will in essence show peers their approval by implicitly confirming the quality of submitted changes, simply by not putting forward additional corrections to what the original contributor(s) had provided. A sort of silent consent will thus enhance the credibility of users whose submissions pass muster with subsequent users. In addition to this process of "dynamic accreditation", some archives may choose to start from an official list of fully accredited professionals to expedite the "reputation management" component of the system, using a form of "*a priori* accreditation"

Details of the exact functioning of the CA system are being worked out at this time and will be the subject of upcoming research projects at WPI. For the time being, we will assume that each archive will manage its own user accounts, but the aim is to arrive at a universal registration system that will maintain the distributed and emergent nature of the overall system without compromising the quality of the transcriptions that bubble up to the top of the heap after successive refinements by a variety of contributors.

Users will accumulate "transcription credits" and "word box credits" (possibly on different "pay scales") when they make contributions to the system. Credits will in turn be usable to pay for such services as: on-demand scanning of manuscripts, handwriting recognition assistance, remote storage of project files on the archive servers, automatic transcription processing, advanced searching, and others.

In the end, we plan on unleashing this self-propagating emergent system onto the internet by the end of 2005, complete with open-source software so that the entire system, including the three applications themselves, will grow on its own without central control. If successful, this system could really "make history" by exponentially expediting the production of historical research in the whole world.

5. References

[1] University of Idaho, http://www.uidaho.edu/special-collections/Other.Repositories.html. Last Accessed 1/20/05.

[2] James, Kimberly M., *Text Transcripting Tool*, WPI Major Qualifying Project, April 1998.

[3] Ho, Oliver, Kligman, Ricardo, Patel, Chirag, Patel, Ravi, *Manuscript Transcription Assistant*, May 2003.

[4] Calhoun, Shaun, Kumar, Raja, Tiscia, Michael, Turner, Terrence, *Manuscript Transcription Assistant Initiative*, April 2004.

[5] Johnson, Steven, *Emergence*. Touchstone, NY, 2001.

[6] Center for Technology and the Arts, De Montfort U., http://www.cta.dmu.ac.uk/projects/master/index.html, last accessed 2/28/05.

[7] Rath, T.M., Manmatha, R. and Lavrenko, V.: *A Search Engine for Historical Manuscript Images*. Proc. of the ACM SIGIR 2004 conference, Sheffield, UK, July 25-29.

[8] Wolf, Gary, *The New Multiple Personality Disorder*, Wired, 13.05, May 2005.