

Kenton McHenry, Luigi Marini, Rob Kooper, Liana Diesendruck  
National Center for Supercomputing Applications (NCSA), University of Illinois at Urbana-Champaign (UIUC)  
{mchenry, lmarini, kooper, ldiesend}@illinois.edu

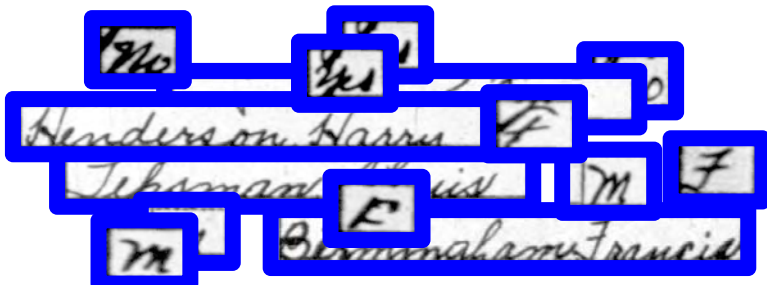
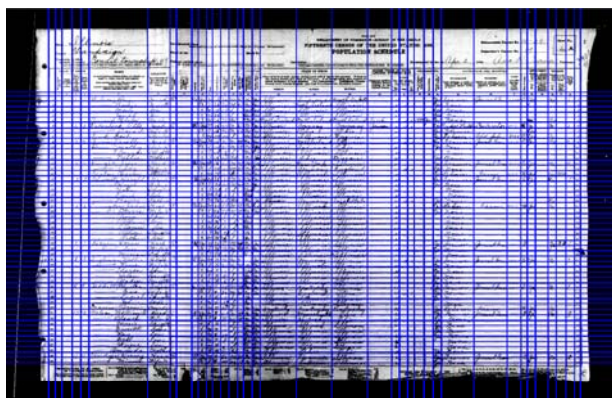
Digital versions of data have many advantages over contemporary analogue formats such as paper. However, with the move to digitize paper archives we move away from an old problem of providing everyone with access to raw shared information to a new problem of providing everyone with usable, or searchable, access to the content within that raw information. On April of this year the 1940s Census data will be released for the first time in a digital only form (a collection of approximately 4 million JPEG images). NCSA in collaboration with NARA have been working together to investigate means of attacking this new problem and providing searchable access to this next generation of archives. The approach illustrated here combines applied computer vision with passive crowd sourcing.

**The Problem:** Millions of images of handwritten forms (terabytes in size)



**An Applied Computer Vision Approach:**  
Word Spotting combined with passive crowd sourcing

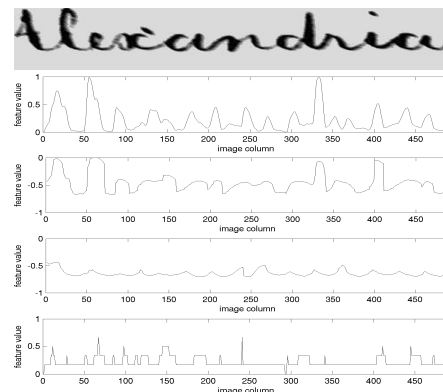
### Segmentation



71,740 estimated CPU hours to process 2.9 million images

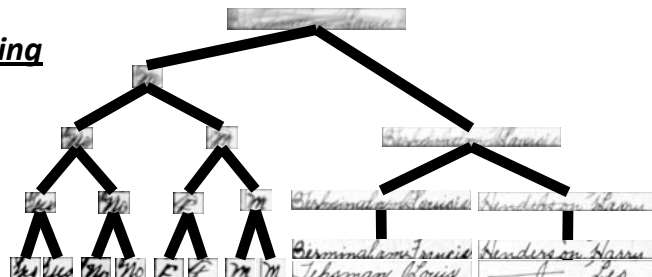
### Signature Extraction

Word Spotting (Rath, '04)



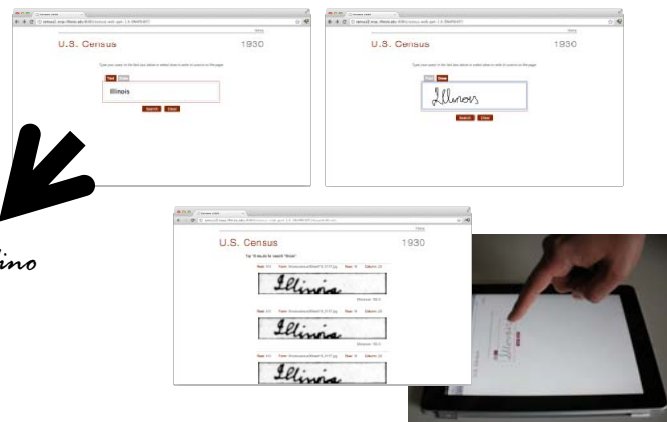
49,331 CPU hours to process 2.9 million images

### Indexing



2,085,048 CPU hours to process 5.5 billion sub-images

### Searchable Access over the Web



**Passive Crowd Sourcing:** As users use the system their queries are associated with viewed results allowing the system to improve over time.