2019-2020 ACS Project updates

ACS awardees are sharing updates about their work to date roughly midway through their project cycles. Check them out below!

- A Half-Century of Illustrated Pages: ACS Lab Notes —
  ACS awardee: Stephen Krewson (Yale University)

  We've reached the mid-point of my Advanced Collaborative Support project, "Deriving Basic Illustration Metadata." Right now, sitting on a supercomputer named Big Red at Indiana University, is a rather remarkable dataset: *every illustrated page from every Google-scanned volume in the HathiTrust Digital Library for the period 1800-1850.* Although the image processing pipeline we are using is not new, working at this scale is.

- HTRC & Me —
  ACS awardee: Dan Sinykin (Emory University)

  In fall 2018, the study of contemporary literature changed. HathiTrust opened its vault of copyrighted material to computational scholarly use. Before then, a scholar wanting to study contemporary literature at scale needed to collect the rare publicly available texts and purchase the rest privately, creating an expensive boutique corpus (text dataset) that could not be shared for collective use or for others to test one's work's replicability: severe limits. Now scholars can access hundreds of thousands of contemporary texts.

- GlobalNames and the HathiTrust —
  ACS awardees: Dmitry Mozzherin and Matt Yoder (Species File Group, Illinois Natural History Survey, Prairie Research Institute, University of Illinois)

  Our team of researchers, the Species File Group, develop and use digital tools for biodiversity informaticians, those scientists who study the Earth's species. One of the things we focus on is locating information about the Earth's species via their scientific names, a project called GlobalNames. The idea is straightforward, find a biological name like *Homo sapiens* (humans), *Apis mellifera* (the Western honey bee), or *Anopheles gambiae* (a mosquito that transmits Malaria), and you may discover information important to scientists "nearby". In the context of the GlobalNames project finding a name means parsing digitized literature or datasets, small or large. Thanks to funding from the National Science Foundation (NSF ABI 1645959, 2015) initial tools developed by Dmitry Mozzherin and Alex Myltsev were developed and hardened against the large, free corpus of scientific publications in the Biodiversity Heritage Library (BHL). Within the BHL the diversity of data (e.g. different languages, publication types, general quality of parsed text), and its structure therein let us find and resolve many edge cases in the name detecting algorithms. While finding specially formatted latinized names is challenging, the results of this work are fairly simple: at their core, they are an index indicating that "*this* name was found *there*". From these simple data many downstream features and explorations emerge, for example the list of names found on any given page of the BHL (e.g. Scientific Names on this Page), is derived from our tools.

- Semantic Phasor Embeddings: Mid-Point Update —
  ACS awardees: Molly Des Jardin, Scott Enderle, Katie Rawson (University of Pennsylvania)

  Much recent discussion of quantitative research in the humanities has concerned scale. Confronted with the vast quantities of data produced by digitization projects over the last decade, humanists have begun exploring ways to synthesize that data to tell stories that could not have been told before. Our ACS project aims to make that kind of work easier by creating compact, non-expressive, non-consumptive representations of individual volumes as vectors. These vectors will contain information not only about the topics the volumes cover, but also about the way they order that coverage from beginning to end. Our hope is that these representations will allow distant readers to investigate the internal structures of texts at larger scales than have been possible before. But now that we've reached the midpoint of our work, our preliminary results have led to some surprising reflections about scale at much smaller levels.