כנס מדעי הרוח הדיגיטליים - תקצירי הרצאות

Digital Humanities Conference – Selected Abstracts

22.05.19

Alicia Fornés

The combination of Computer Vision, Crowdsourcing and Gamification for speeding up the transcription of historical manuscripts.

Name: Alicia Fornés
Affiliation: Computer Vision Center, Universitat Autònoma de Barcelona, Spain.

Short bio:
Dr. Alicia Fornés received the Ph.D. degree in 2009 from the Universitat Autònoma de Barcelona (UAB). Her Ph.D. work on writer identification of old music scores received the best thesis award 2009-2010 by the AERFAI (Spanish Branch of the IAPR). She has published more than 100 papers in international conferences and journals. She received the IAPR/ICDAR Young Investigator Award in 2017 for outstanding contributions in the recognition of handwriting, text and graphics, with high impact to the field of Digital Humanities. She is currently a Senior Research Fellow at the Computer Vision Center (CVC) and the UAB.
For more information, please visit: http://www.cvc.uab.es/people/afornes/

Longer bio + picture here: http://www.cvc.uab.es/people/afornes/

Abstract:
At the present time, there are still many historical manuscripts in archives waiting to be transcribed and indexed. Since a manual transcription is time consuming, the automatic transcription through computer vision techniques is desired. Although the recent handwritten text recognition techniques based on deep learning architectures have shown great performance, the automatic transcription of historical manuscripts is far from perfect. Therefore, the incorporation of the human in the loop seems the path to follow. Crowdsourcing (i.e. splitting a big task into many small tasks that are distributed among many users) is a good strategy for transcribing a massive amount of historical manuscripts. However, the interest of transcribers usually decreases along time. For this reason, gamification (i.e. the application of game-design elements and principles in non-game contexts) applied to crowdsourcing has emerged as a promising solution to keep the users' interest and offer an engaging experience. In this talk we will describe the NETWORKS project, in which document image analysis techniques are combined with crowdsourcing platforms and gamification experiences to ease and speed up the transcription of demographic historical manuscripts, creating a historical social network. For more information: http://dag.cvc.uab.es/xarxes/

Micki Kaufman

Micki Kaufman, doctoral candidate in US History

"Quantifying Kissinger: Data Visualization and Historical Interpretation in Digital History"

Bio

Micki Kaufman is a doctoral candidate in US History at the Graduate Center of the City University of New York (CUNY). Her dissertation, "'Everything on Paper Will Be Used Against Me:' Quantifying Kissinger" has been a recipient of the CUNY Graduate Center's Provost's Digital Innovation Grant. In 2015 Micki was awarded the ACH and ADHO's Lisa Lena and Paul Fortier Prizes for best Digital Humanities paper worldwide by an emerging scholar. From 2015-2017 she served as a Virtual Fellow with the Office of the Historian at the US State Department, and currently serves as an elected member of the Executive Council of the Association for Computers in the Humanities (ACH).

Abstract:

More than any other former United States Secretary of State or National Security Advisor, the public 'celebrity' of Dr. Henry Kissinger is uniquely well known in American and global popular culture. Fixed in the public consciousness for his leadership of the Nixon and Ford administrations' foreign policy from 1968-1977, the contradictions between Kissinger's well-documented public persona and his more secret conduct still mystify, polarize and fascinate historians and the general public alike. Given the vast amount of material available for study, historians grappling with the complexities of his actions and character soon encounter a second problem – one of scale. As detailed on the Web site (http://blog.quantifyingkissinger.com), the project is an historical interpretation of the National Security Archive's Kissinger Collection, a focused but substantial subset of the torrent of material generated by Kissinger that comprises over 18,000 declassified meeting memoranda ('memcons') and teleconference transcripts ('telcons'). The project combines novel computational text analysis, data visualization and interpretive methods to provide new insights into complex historical subjects and to validate and promote the use of such methods in 'big data' historical research.

Susan Schreibman

Prof. Dr. Susan Schreibman

Professor of Digital Arts and Culture, Maastricht University

Bio:

Susan Schreibman is Professor of Digital Arts and Culture at Maastricht University, The Netherlands. ProfessorSchreibman has published and lectured widely in digital humanities and Irish poetic modernism. Her current digital projects include *Letters 1916-1923* and *Contested Memories: The Battle of Mount Street Bridge*.

Her publications include A New Companion to Digital Humanities (2015), *Thomas MacGreevy: A Critical Reappraisal* (2013), *A Companion to Digital Literary Studies* (2008), and *A Companion to Digital Humanities*(2004). She is the founding Editor of the peer-reviewed *Journal of the Text Encoding Initiative* and is a member of the Board of the National Library of Ireland.

Abstruct:

Participatory Engagement in the Digital Humanities as a Public Good

This paper will reflect on the differences between participatory engagement and crowdsourcing in digital humanities--a difference that resembles the distinction in the traditional humanities disciplines between collaborating with the public, on the one hand, and addressing it from a hierarchically superior position, on the other. It will explore this chiefly through the Letters 1916-1923 project, Ireland's first participatory engagement project which invites the public to become co-collaborators in, not only creating a wholly new collection (currently with over 100 contributory families and institutions), but in knowledge creation itself. The project had several other goals: to bring the archive to the public as the public so rarely ventures into the archive, provide opportunities for atypical audiences (eg secondary school students, disadvantaged groups, retirees) to engage deeply with primary sources, and to understand and value the work of humanities researchers. Through these goals, we have revised notions of expert knowledge as sharable rather than owned by a select few in the academy.

Gila Prebor

From authority data, to linked open data and Wikidata

The case study of a Hebrew manuscript catalogue

Gila Prebor

Traditionally, library catalogues have served as a tool to manage library collections and as a bibliographic tool for information retrieval. In the 20th and 21st centuries libraries focus not only on bibliographic record but also on data. The content of the catalog record has been standardized according to international rules and standard protocols such as AACR, RDA, MARC, and Z39.50 so it could be easily exchanged and duplicated. This enables the catalogues to be accessed from a distance both by human users and by machines. Library standards were intended to be used by librarians; because of this the catalogues serve only the library community. Eventually this caused library catalogues to be data silos (Chambers, (Ed.). 2013).

Today libraries are an important player in the linked data arena. Converting catalogues to large linked data enables large-scale analysis of cultural heritage Big Data. By applying linked data initiatives library data is open, available and reusable in the information space. Libraries can share their open metadata with non-library communities. RDA was published in the Open Metadata Registry (http://metadataregistry.org/) as a set of elements in the RDF standards model.

Several libraries have already taken the initiative to convert their catalogs to RDF-based triples and to linked data (Dunsire 2012). For example, the Swedish Union Catalog, LIBRIS (libris.kb.se), was one of the first catalogues that began sharing linked data in 2008. Other libraries using linked data are the British National Bibliography (http://www.bl.uk/bibliographic/datafree.html) and the Library of Congress (https://id.loc.gov/). The Getty vocabularies are now available as Linked Open Data (http://www.getty.edu/research/tools/vocabularies/lod/index.html) (Hastings, 2015).

The next stage is the integration of authority data from library catalogues to Wikidata. Wikidata is a collaboratively edited knowledge base hosted by the Wikimedia Foundation. It is one central database of human knowledge which contains structured and linked data. Wikidata offers a lot of advantages like: anyone can edit it, new items can be added to Wikidataby every user if something is lacking, it can be read by people and machines, it is multilingual, it is on the wiki platform,Wikidata items contain other data and are linked to Wikipedia articles and it is entirely in a free license (CC).

There are already several initiatives of this type in the world (Allison-Cassin , 2017 ; Forziati & Castro, 2018). In my lecture I will present the advantages of using Wikidata in terms of analysis, evolution and enrichment of catalogues by using the case study of a Hebrew manuscript catalogue as an example.

Relevant Bibliography:

Allison-Cassin, S. (2017). “Research Libraries and Wikimedia: A Shared Commitment to Diversity, Open Knowledge, and Community Participation.” Wikimedia Blog 10/04/2017. https://blog.wikimedia.org/2017/10/04/libraries-wikipedia-york-university-project/

Bermès, E. (2013). Enabling your catalogue for the semantic web .In Catalogue 2.0: The Future of the Library Catalogue, Chambers, S. (Ed.)., 117-142.

Dunsire, G. (2012). Linked data for manuscripts in the Semantic Web. Summer School in the Study of Historical Manuscripts, Retrieved August 05, 2015 from http://www.gordondunsire.com/pubs/docs/LinkedDataForManuscripts.pdf

Forziati, C., Lo Castro, V (2018) “La connessione tra i dati delle biblioteche e il coinvolgimento della comunità: ilprogetto SHARE Catalogue-Wikidata.” JLIS.it 9, 3 109-120. DOI: 10.4403/jlis.it-12488.

Hastings, R. (2015). Linked data in libraries: status and future direction. Computers in Libraries, 35, 12-16.

Sinai Rusinek

Introducing DiJeSt (Digitizing Jewish Studies)

Supported by the *Rothschild Foundation (Hanadiv) Europe*, *DiJeSt* aims to digitize not only materials and sources for Jewish Studies, but also the know-how and knowledge which are part of their practice. Jewish studies scholars read, organize, interpret, enrich and analyze sources in various scripts and languages. In DiJeSt we are preparing a pipeline that will guide scholars and students, from within the academy or outside it, to do it as joint effort, and in dialogue with the machine: starting with the know-how of reading: Training OCR for the various Jewish Languages, through transliterating, annotating and extracting structured information from texts, and finally visualizing and analysing the growing pool of data, making it into dynamic, collaborative knowledge. I will present the DiJeSt linked data repository and the challenge which we set for the Haifa Hackathon, to open it to updates, corrections, contestation and readings by different communities of interest.

"Hamapa"

Rabbi Eli Ficsher, Moshe Schorr

A Rabbinic Republic (or Oligarchy) of Letters

This project, HaMapah, began with a series of questions about how to understand rabbinic authority, especially in a "competitive" environment. How do sub-elites decide who to ask, and how do the prospective elites project themselves in ways that produce confidence, trust, and the aura of expertise.

One way to answer that question is by mapping the "metadata" of responsa literature. The names, places, and dates that appear in these correspondences outline the "sphere of authority" of a particular rabbi, both over time and across space. One of our goals is to create a comprehensive database of responsa by author, recipient, place of author, place of recipient, date, and other fields (before we even get beyond metadata and work with data—the content of the responsa).

To date we have mapped c. 20,000 responsa from about a dozen authors, but we have barely scratched the surface. There are millions of responsa, most of which do not contain metadata (and many of which are "pseudo-responsa"), but there is still a great deal of metadata to mine.

We don't do this by hand, of course. Once a corpus is in the public domain, we can automatically identify most places and names. We also have agreements with certain publishers who allow us to use their private-domain data. However, a good deal of material remains undigitized.

This is where crowdsourcing comes in. Hebrewbooks.org has over 2,000 volumes of responsa available for free, most of which have "bookmarks" that link to each individual responsum.

We aim to create a user interface that will show the beginning of a responsum and pose a series of simple questions that users can answer, for instance: Is this responsum written to an individual? Is a place mentioned? Is it dated? If theresponsum has metadata, users will be asked to enter it in Hebrew characters as written. This will greatly expand our database of individual responsa, and consequently enhance its overall research value.

A second version of our UI will be used to produce training data for machine learning. Users will answer a series of simple questions about the content of the responsa (i.e., whether it is permissive or prohibitive, what sources it cites, and what area of law it addresses). This will be the first step in our transition from metadata to data.

"Ilanot"

"Maps of God – Building a Portal to Visual Kabbalah"

Abstract:

An *ilan* (Heb., "tree") is a kabbalistic "Map of God," inscribed on a large parchment roll. Wherever there were kabbalists, they produced ilanot: from Germany to Kurdistan, in Italy, Iraq, Poland, Morocco, Yemen, and many other places. For some five hundred years, these artifacts circulated throughout the Jewish world, resulting in *ilanot* (pl.) that are amalgamations of traditions of knowledge visualized in the local environment and traditions of knowledge visualized in places and times far away. For a number of years, Prof. Chajes and Dr. Baumgarten have been devoted to undertaking basic research of this genre, resulting in numerous publications and a database of over six hundred artifacts. It has nevertheless become clear that the particular nature of these artifacts presents challenges on a variety of levels, beginning with basic research and culminating in their publication and presentation to the general public. Studying the ilanot requires attentiveness to their graphical elements — including a full repertoire of iconography and diagrammatic schemata — as well as to their textual elements, which must be analyzed philologically. Prof. Chajes and Dr. Baumgarten have teamed with SUB Göttingen to create the digital humanities platform required for the advanced study—as well as introductory presentation—of the *ilanot* to scholars and the broader public. "Maps of God" will allow researchers to track image-text units in their specificity and more complex compound configurations, while inviting users to explore the most stunning and significant artifacts of visual Kabbalah in precisely the ways that suit them: from close study of individual parchments, assisted by transcriptions, translations, and commentary, to broad searches for concepts as they have been graphically visualized over centuries and continents.

Computer experiments on the Khirbet Qeiyafa ostracon איתן לוי

The title is: "**Scrypt, a web application for computer-assisted decipherment of Old Hebrew and Proto-Canaanite inscriptions**".

Abstract:

The Scrypt application (www.ScryptApp.com) is an online tool for computer-assisted decipherment of
of Old Hebrew and Proto-Canaanite inscriptions. Its main aim is to enable users to encode readable and unreadable or partly readable graphemes using regular expressions. The tool then automates the search for lexemes in the Brown-Driver-Briggs Hebrew (BDB) Biblical Hebrew dictionary and uses automatic insertion of matres lectionis in order to allow finding the words in a Biblical Hebrew dictionary, even when they are written in the older, fully defective, orthography. They tool also enables other features, such as basic image processing, filtering words according to word classes, and saving/loading previous work. The talk will present the theoretical foundations behind the Scrypt tool, and a demo on a real-case inscription.

"Scribes of the Cairo Genizah"

Scribes of the Cairo Genizah is a crowd sourcing project aimed at classifying and transcribing the great treasure of the Cairo Genizah. The project is a collaboration of DH Judaica in Penn, Princeton Genizah center, Zooniverse and the eLijah-Lab at the university of Haifa. The project developed a multilingual (Hebrew, English, Arabic) platform of workflows for classifying, transcribing and spotting keywords in Genizah fragments. The lecture will present key stages in the development of the project and a reflection about the process of development as a first stage towards Sofer STAM - Systematic Textual Availability of Manuscripts.

"Tikkoun Sofrim"

Tikkoun Sofrim is a joint French - Israeli project aimed at combining crowd sourcing correction workflow with deep learning based automatic transcription of Hebrew Manuscripts (HTR). Tikkoun Sofrim is part of a network of projects (Ktiv, Sofer Mahir, eRabbinica, Scribes of the Cairo Genizah, Scripta-PSL) aimed at developing a future framework for comprehensive and systematic textual availability, editions and deep annotation of (Hebrew) manuscripts via a pipeline that combines HTR with crowdsourcing the corrections and validation by scholars. The project focuses on Tanhuma-Yelamdenu Midrashim, late rabbinic exegetical works without full scholarly critical edition. We trained models for four large manuscripts with CERs vary between 2.8% (BNF), 2.9% (Parma), 6.9% (Vatican) and 8.9% (Geneva). Our crowd sourcing platform offers a designated UI for mobile and desktop with emphasis on smooth and quick contribution experience. The project has attracted a large community of volunteers, contributing more than 700 correction lines/day. At least 5 people corrected each line, and their aggregated transcriptions reduced to error rate to 00.22 CER.