Image Analysis and Infrastructure Support for Data Mining the Farm Security Administration – Office of War Information Photography Collection

Marcus Slavenas
NCSA
1205 W Clark St, 2002E
Urbana, IL 61801
217-244-0774
slavenas@illinois.edu

Paul Rodriguez
SDSC
10100 John Hopkins
La Jolla, CA.
prodriguez@sdsc.edu

Alan Craig
XSEDE
P.O. Box 5020
Champaign, IL 61825
a-craig@illinois.edu

Elizabeth Wuerffel
Valparaiso University
1709 Chapel Drive
Valparaiso, IN 46383
1-219-465-7908
liz.wuerffel@valpo.edu

Jeffrey Will
Valparaiso University
1700 Chapel Drive
Valparaiso, IN 46383
1-219-464-6875
jeff.will@valpo.edu

ABSTRACT

This paper reports on the initial work and future trajectory of the Image Analysis of the Farm Security Administration – Office of War Information Photography Collection team, supported through an XSEDE startup grant and Extended Collaborative Support Service (ECSS). The team is developing and utilizing existing algorithms and running them on Comet to analyze the Farm Security Administration - Office of War Information image corpus from 1935-1944, held by the Library of Congress (LOC) and accessible online to the public. The project serves many fields within the humanities, including photography, art, visual rhetoric, linguistics, American history, anthropology, and geography, as well as the general public. Through robust image, metadata, and lexical semantics analysis, researchers will gain deeper insight into photographic techniques and aesthetics employed by FSA photographers, editorial decisions, and overall collection content. By pairing image analysis with metadata analysis, including lexio-semantic extraction, the opportunities for deep data mining of this collection expand even further.

CCS Concepts

•Applied computing → Arts and humanities, Optical character recognition • Computing methodologies → Image processing, Object recognition •Information systems → Data mining, Information extraction, Web interfaces, Web services.

Keywords

Data Mining; Image Analysis; Photography; Art History; Humanities; Linguistics; Computer Science; Interdisciplinarity

1. INTRODUCTION

Mining large-scale image collections could open up new avenues of exploration for digital humanities. The collection we analyzed includes images from the Resettlement Administration (1935-1937), the Farm Security Administration (FSA, 1937-1942) and the Office of War Information (OWI, 1942-1944), as seen in Figure 1. In this paper, the collection will be referred to as the "FSA collection" for brevity.

The FSA collection offers visual documentation and visual rhetoric of the plight of farmers and sharecroppers and rural rehabilitation efforts, and, in 1942-1944, the war effort. This photography program generated hundreds of thousands of images, 171,146 of which are digitized, held by the LOC, and accessible online to the public.

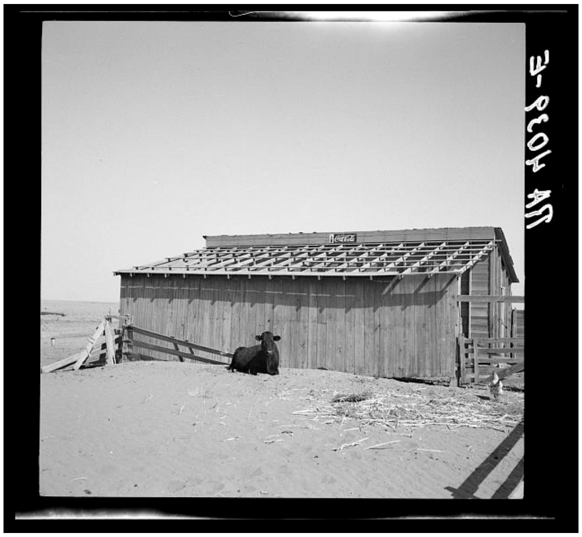


Figure 1 Image from the FSA collection and held by the LOC. Title: "This farmer took the roof off his barn to make a windbreak for his garden. There was no rain. Cimarron County, Oklahoma." Photographer: Arthur Rothstein. Date Created: 1936.

Our team is using existing algorithms and creating new ones to extract features and analyze the FSA-OWI image corpus held by the LOC and accessible online to the public. The analysis includes summary level properties of an image, such as mean grayscale, and the feature extraction includes both visual entities and metadata tagging. Development of these tools will not only be useful for analyzing this particular corpus, but also for other public image collections at the Library of Congress and elsewhere.

2. INFRASTRUCTURE

We have implemented a software stack with integration of tools such as DIBBs Brown Dog / NCSA Clowder on NCSA/ISDA[1] (Innovative Software and Data Analysis) resources. A main goal of the project was achieved by establishing a web interface to allow users to easily run queries on the corpus and automatically engage XSEDE HPC system (SDSC Comet) to run extractors which would automatically pull searchable information from image contents (Figure 2).

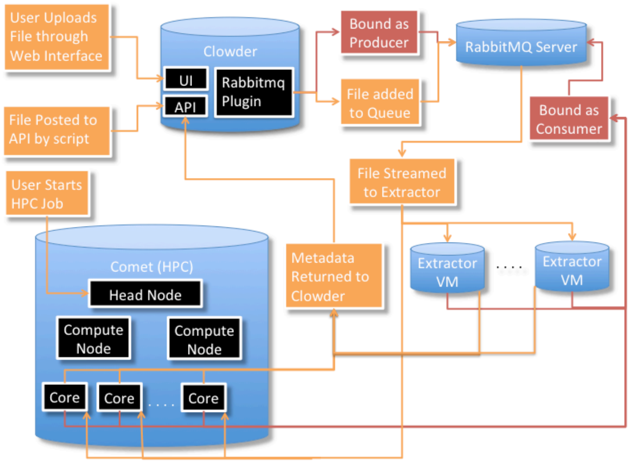


Figure 2 The setup (red) shows how Clowder is bound to a RabbitMQ server as a producer. The extractor VMs and the HPC resources are bound to the RabbitMQ server as consumers. Only one consumer is needed, but this diagram shows multiple consumers all of which will be used asynchronously given enough uploaded files. The extraction process (orange) starts by uploading a file into Clowder through the UI or programmatically through the API after which the file is then put into the RabbitMQ queue. The HPC job needs to be started by the user (future improvements will start the HPC job automatically when the RabbitMQ reaches a certain length). Each file is then streamed to the next available extractor, the file is processed, and the metadata is posted to the file in Clowder.

2.1 Components

2.1.1 Clowder Data Management System

A server VM was created on NCSA/ISDA resources at iarp.ncsa.illinois.edu and Clowder, a web based data management system targeting easing the curation and publishing of data [1][2], was installed. Clowder provides structures for storing file metadata and searchable tags, includes a frontend UI, a RESTful service API plugin that allows users to interact programmatically with the filesystem, and an extensible suite of data type specific extractors and previewers.

2.1.2 Rabbitmq, Extractors, and Automatic Extraction

An existing RabbitMQ server running on NCSA/ISDA resources was used and an IARP exchange was added to the virtual host. The Clowder instance on the IARP server was bound to the RabbitMQ server as a producer. Extractors are software routines that run on a file of a particular mime type (e.g. face detection on images or word count on text). Several extractors were deployed on NCSA/ISDA resources and bound to RabbitMQ as consumers. To summarize the extraction process: a file is uploaded to Clowder through the UI or posted programmatically through the API, the file is put in a RabbitMQ queue for appropriate mime type, the file is streamed to an available extractor, processed by the extractor, and the data is posted to Clowder and associated with the file as indexable metadata.

2.1.3 Extraction on HPC XSEDE Comet

On Comet, a SLURM job was created that runs a single instance of an extractor on each core within a node. The connectivity to Clowder has the same structure as with a VM. In this HPC case, the number of consumers registered to the RabbitMQ server is equal to the number of nodes (24 per core) requested for the job.

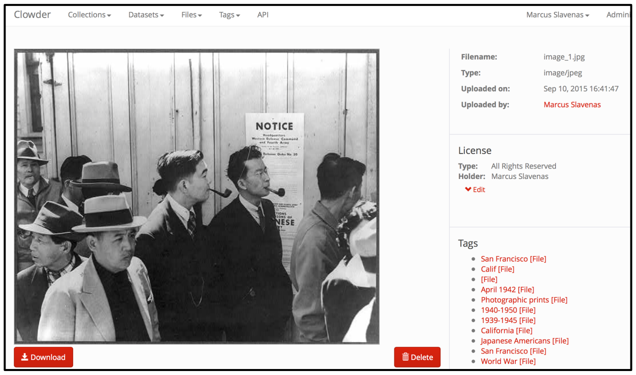
In non-HPC extractions, the extractors run on a VM continuously. However, that is not possible on HPC because of the time restrains on usage. For testing purposes, the RabbitMQ connection was instantiated by starting the job on Comet for a very short duration ~1 minute. Subsequently, the files are uploaded into Clowder and the queue is allowed to build. When the job is started manually on Comet - the files are streamed to Comet, passed to the next available extractor (core), and the extracted metadata is posted to Clowder. This continues until the queue is empty (or the HPC job times out).

Future work would include allowing the RabbitMQ queue to build to a specified size and then start the HPC job automatically.

2.1.4 Fetcher Script gets LOC Images and Metadata

A Python script was written that pulls the Farm Security Administration images and all associated metadata from the LOC API. The script loads the images into Clowder, posts the metadata from the LOC, and parses the metadata into searchable tags (Figure 3). This process of uploading the images through the Clowder API initiates extraction and posting of the extracted metadata, which is not to be confused with the LOC metadata which is included with the file in the LOC collection.

[1] http://isda.ncsa.illinois.edu



Filename: image_1.jpg
Type: image/jpeg
Uploaded on: Sep 10, 2015 16:41:47
Uploaded by: Marcus Slavenas

License
Type: All Rights Reserved
Holder: Marcus Slavenas

Tags
- San Francisco [File]
- Calif [File]
- [File]
- April 1942 [File]
- Photographic prints [File]
- 1940-1950 [File]
- 1939-1945 [File]
- California [File]
- Japanese Americans [File]
- San Francisco [File]
- World War [File]

Figure 3 Example image streamed to Clowder from the LOC archive. In the lower right corner are the searchable tags that were parsed from the LOC metadata included with the image.

2.1.5 Analysis Web Application

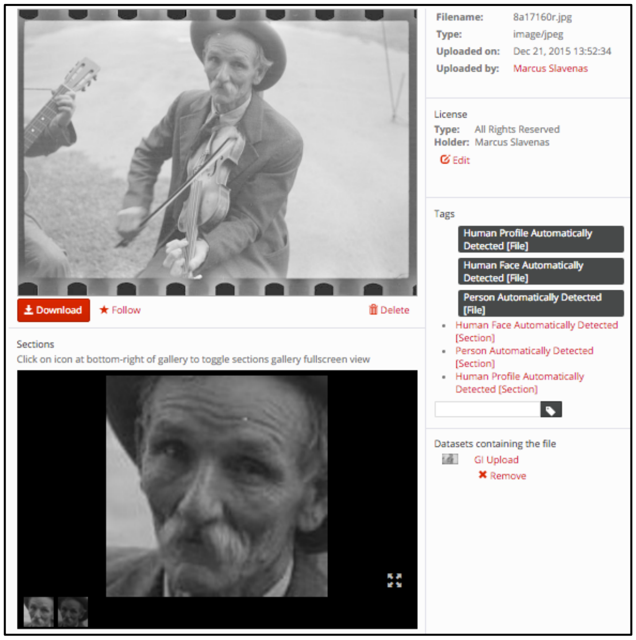
One of the main reasons for the Clowder API is to allow users to build applications that can access and process the files and metadata stored in Clowder. We have embarked upon this with a standalone web application for analyzing metadata across collections. Though the application has been successfully used for an example analysis (Figure 6), it is in a very rudimentary state. It will need further development to be used successfully on the entire collection and with greater flexibility.

3. FEATURE EXTRACTION: IMAGE ANALYSIS

The following sections will describe the feature extraction in more detail, as well as how users can use the data. Our ongoing work will begin to scale up the feature extraction analysis to the full set of images, as well as investigate methods to integrate different kinds of features and best practices to help investigators mine the data.

3.1 Preexisting Extractors

Clowder already has many functioning extractors that are at the disposal of this project. Several of these are installed on VMs and bound as consumers of Clowder uploads. These extractors include identifiers for person, face, eyes, and profile (Figure 4); OCR (Optical Character Recognition); and an object classifier based on the Caltech 101 library of tagged-object images.



Filename: 8a17160r.jpg
Type: image/jpeg
Uploaded on: Dec 21, 2015 13:52:34
Uploaded by: Marcus Slavenas

License
Type: All Rights Reserved
Holder: Marcus Slavenas

Tags
- Human Profile Automatically Detected [File]
- Human Face Automatically Detected [File]
- Person Automatically Detected [File]
- Human Face Automatically Detected [Section]
- Person Automatically Detected [Section]
- Human Profile Automatically Detected [Section]

Datasets containing the file
GI Upload

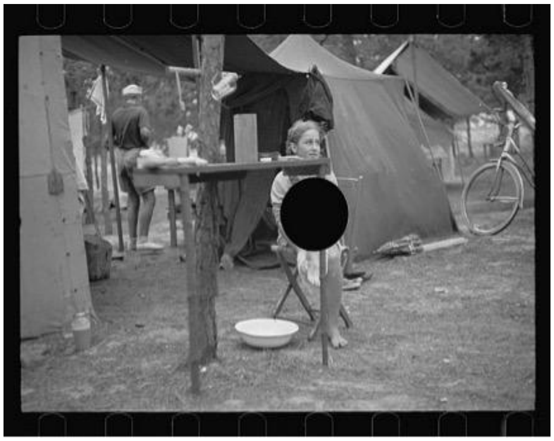
Sections
Click on icon at bottom-right of gallery to toggle sections gallery fullscreen view

Figure 4 Feature extraction example of face, person, profile; with tags listed alongside image.

3.2 Project Developed Extractors

3.2.1 Kill Punch Extractor

The kill punch extractor, which was developed for this project, locates solid black circles in images that have been identified as a punch through the negative that 'kills' it – meaning Roy Stryker, manager of the FSA photography program, determined it was not to be used. The extractor also identifies the original frame border of the negative, which could be used to help identify film or camera type, as seen in Figure 5.



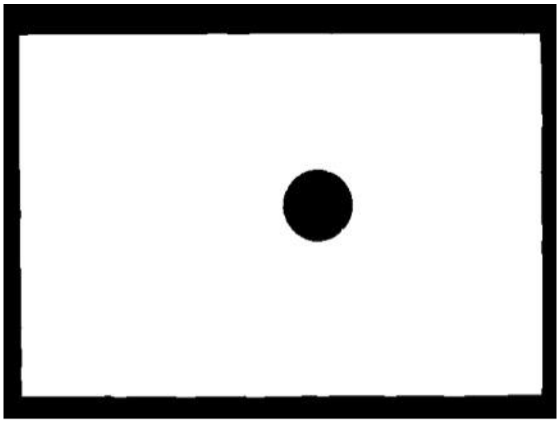
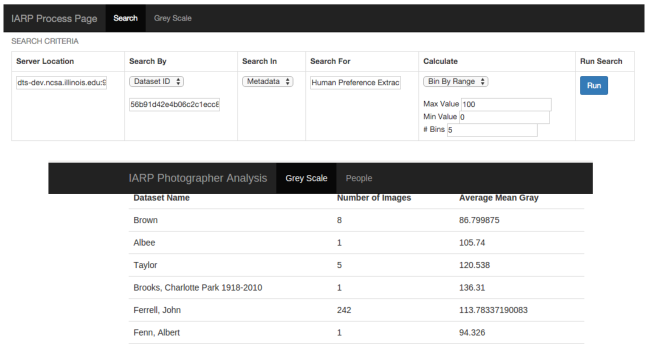


Figure 5 Extraction of "punched" images indicating Stryker's rejection of a photograph.

3.2.2 Mean Gray Analysis

The first completed analysis used the mean gray value that was generated for each image. In this case, the analysis web application gets the mean gray for every image from a photographer, calculates the average mean gray for that photographer, and displays the results in the UI.



SEARCH CRITERIA

Server Location	Search By	Search In	Search For	Calculate	Run Search
dts-dev.ncsa.illinois.edu:9	Dataset ID 56b91d42e4b06c2c1ecc8	Metadata	Human Preference Extrac	Bin By Range Max Value 100 Min Value 0 # Bins 5	Run

Dataset Name	Number of Images	Average Mean Gray
Brown	8	86.799875
Albee	1	105.74
Taylor	5	120.538
Brooks, Charlotte Park 1918-2010	1	136.31
Ferrell, John	242	113.78337190083
Fenn, Albert	1	94.326

Figure 6 The upper clipping shows the interface for the analysis web app. The user can select a server where Clowder is running and various search parameters. The lower clipping shows the results of a query for average mean-gray-scale of images from respective photographers.

3.3 Extractors Under Development

Two further extractors are under development. These will find smiles and identify the presumed gender of persons found within an image. These computer vision routines already exist within OpenCV documentation in C++, but currently Clowder only has wrappers in Java and Python. Development of C++ wrapper has ensued, libraries have been chosen for a RabbitMQ interface, a JSON parser, and an HTTP request. Much of the basic functionality is finished, but more work is necessary.

4. FEATURE EXTRACTION: METADATA ANALYSIS

The metadata from the LOC that accompanies the pictures includes city-state information, author, date/time, as well other contextual information, such as subject and captioning. The amount of information available varies tremendously, as some images have no captions and/or no subject. One of our objectives is to integrate this information, if present, and make the information available to the user alongside the visual feature extractions and image property analysis.

We have developed a textual analysis that can provide insight into the cultural and social dimensions of the collection. The main steps include: convert location information to geo-coordinates; extract subject categories, if available, and extract lexical-semantic features from captions (i.e. part of speech tagging, syntactic labeling, semantic categories, and sentiment qualities).

We have run a metadata analysis on a subset of 3965 images as described here.

1. Preprocessing: First, metadata is extracted from the image data, and images that have duplicate or untitled captions are noted and removed.
2. The 'subject/location' metadata field is parsed and translated into geographic location. For example, the following 'subject/location,' *United States--Illinois--Mercer County--Aledo.::Stockyards—Illinois* is split into a longitude and latitude (41.20078,-90.741629) for either the county or the city, in this case Aledo, and the abstract subject, if available, which in this case is 'Stockyards.' Out of a sample of 3965 images, 3796 subjects were found, 2487 locations were found for a city-state text, and 1438 locations were found for state-only text.
3. Lexico-Semantic Extraction: One aspect that makes this work especially relevant to image analysis for humanities is that the metadata was not written to only identify what is in a photograph, but rather to give some framing and cultural context. The image captions do not directly describe objects in the image – they do not say, "two men and a slab of concrete," but instead say things like "new homes under construction." The objective of this processing is to separate location information, extract words from the caption, and provide candidate parsing information and semantic categories for user review. An example of the summary level textual analysis of captions appears in the word cloud in Figure 7.

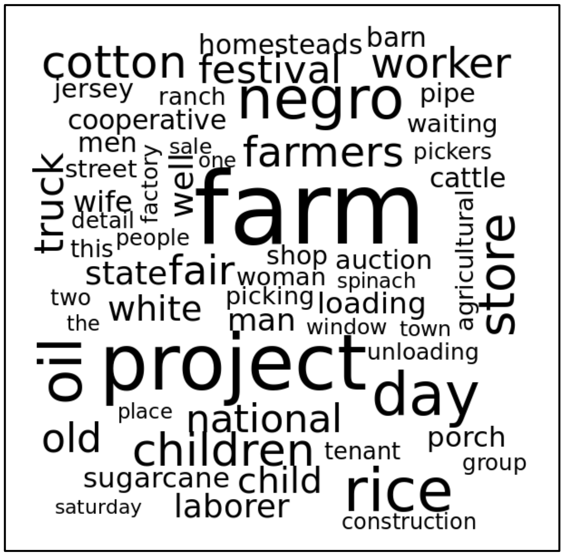


Figure 7 Word cloud for caption words for one photographer.

All captions are run through another python script that extracts lexico-semantic features. This script uses the Stanford Natural Language Processing module for parsing, Natural Language Toolkit 3.0 for part of speech tagging, and Wordnet ontology for semantic interpretations. Consider the following example (Figure 8) from photographer Russell Lee, Clowder Image 20950. Table 1 shows the lexical-semantic features extracted and part of the full data schema (other fields include position in sentence, lower level semantic categories, actual frequency counts for a word sense from the Wordnet corpus). Note that the actual image shows that there are no "persons," but there is a "structure." All the features are taken as possible interpretations. Part of future work will be to establish expectations and guidelines for using these features in this domain.

The database of such features would enable users to explore cultural and social perspectives embodied in the collection. For example, in the 4K image set, a query ran that asked: "Select all image captions where there is a word in the semantic category 'animal.'" We found that about 7% (275) have some mention of an animal or animal related topic, with 73 different subject categories such as "Farms," "Auctions," "Small towns," "Spinach workers," and so on. We are also evaluating such queries on manageable subsets to determine whether animals are actually depicted or are part of the larger photoset in that location.

Table 1 Lexical-semantic features from sample image metadata.

Word	Part of Speech	Semantic Category	Sens Freq.	Parse Depth
'type'	Noun	Person	Low	1
'dwelling'	Noun	Structure	High	1
Jersey Homesteads	Named Entity	Organizatio n	n/a	1
'construction'	Noun	Abstract	n/a	0

5. ONGOING WORK TOWARD DATA MINING RESULTS

From a digital humanities perspective the images are a rich and deep source of material. However, with over 171,000 images in the FSA collection, it is difficult to achieve an overall perspective.

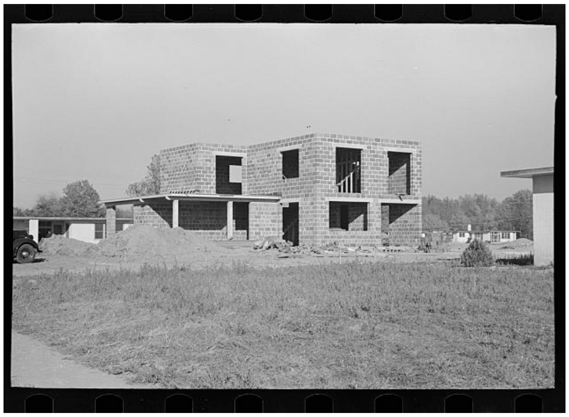


Figure 8 Image 20950: 'New type of two-story dwelling under construction, Jersey Homesteads, Hightstown, New Jersey.'

We are currently scaling up the initial work performed during our startup allocation by continuing to find existing analysis algorithms and creating new ones, running these algorithms on the full corpus, and reviewing results. The analysis runs are primarily embarrassingly parallelizable, where each visual extractor takes about 1 minute of 1 CPU core on Comet, and each image's metadata analysis (all parsing, semantic look ups, etc.) take 2-3 minutes of 1 CPU core on Comet.

Our result is designed to be flexible to allow many kinds of mining with respect to the artistic and socio-cultural-historical nature of the photo collection. There are a series of questions that have been guiding this work. As we scale up the analysis we will begin to address these questions and explore what is possible for the digital humanist. For example:

- How many persons per image?
- How many portraits (i.e. faces) are present versus total persons in images?
- Are there animals, and what types?
- Can we discern indoor/outdoor images?
- How many images are there per location?

Note that the above questions could be taken from a combination of visual or metadata analysis, and broken out by artist and/or date, with added information about image grayscale, location, and subject information.

For example, finding the mean grayscale value of a single image has value for categorization. To connect the grayscale value of a set of images to a particular photographer, location, or time offers deeper insight. Being able to search the collection for photographs that have human faces is helpful, as is being able to search for punched, or “killed,” photographs. Pairing average grayscale or facial recognition with Roy Stryker’s decision to punch or not punch a particular image provides a higher level of inquiry. Combining image analysis with lexio-semantic extraction from the image’s caption could provide incredible fodder for interdisciplinary research between, for example, visual rhetoric and rhetorical/compositional studies.

6. ACKNOWLEDGMENTS

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575 and was supported by BrownDog, NSF grant ACI-1261582. PI and Co-PIs thank Kenton McHenry, Paul Rodriguez, Marcus Slavenas, and Alan Craig for their assistance with image analysis, which was made possible through the XSEDE Extended Collaborative Support Service (ECSS) program.

7. REFERENCES

[1] Myers, J., M. Hedstrom, D. Akmon, S. Payette, B. Plale, I. Kouper*, et al.*, "Towards Sustainable Curation and Preservation: The SEAD Project's Data Services Approach," *Interoperable Infrastructures for Interdisciplinary Big Data Sciences Workshop, IEEE eScience,* 2015.

[2] Clowder - Research Data Management in the Cloud: 2016. https://clowder.ncsa.illinois.edu/. Accessed: 2016-04-28.

[3] Finkel, J. R., Grenager, T., and Manning, C. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 363-370.

[4] Klein, D. and Manning, C. D. 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423-430.

[5] Marini, L., R. Kooper, J. Futrelle, J. Plutchak, A. Craig, T. McLaren*, et al.*, "Medici: A Scalable Multimedia Environment for Research," *Microsoft eScience Workshop,* 2010.

[6] Smruti Padhy, Greg Jansen , Jay Alameda , Edgar Black , Liana Diesendruck , Mike Dietze , Praveen Kumar , Rob Kooper , Jong Lee , Rui Liu , Richard Marciano , Luigi Marini , Dave Mattson , Barbara Minsker ,Chris Navarro , Marcus Slavenas , William Sullivan , Jason Votava , Inna Zharnitsky , Kenton McHenry, 2015 IEEE International Conference on Big Data, pages 493-500, DOI=10.1109/BigData.2015.7363791

[7] Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., Hazlewood, V., Lathrop, S., Lifka, D., Peterson, G. D., Roskies, R., Scott J. R., Wilkins-Diehr, N. 2014. XSEDE: Accelerating Scientific Discovery, *Computing in Science & Engineering,* 16, 5, 62-74. DOI=10.1109/MCSE.2014.80

[8] Bird, S., Klein, E., and Loper, E. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.