Experimenting with a Machine Generated Annotations Pipeline
INTRODUCTION

WHO AM I AND WHY AM I HERE?
Introduction

Joshua Gomez, Head of Software Development & Library Systems

• Joined UCLA Library in Fall 2018
• Lecturer in Information Studies since Spring 2016
• Formerly at the Getty Research Institute

Tasked with:
1. Modernizing the software development process
2. Reorganizing the dev team
3. Improving the user experience of our digital systems
4. Bringing the systems portfolio under control
Modernizing Library Software Development

To improve our *quality* and *velocity*, adopt a DevOps culture, including changes to:

- **Structure**: small focused dev teams; embedded Ops and UX (a matrixed org)
- **Process**: Agile planning; testing (both code and usability)
- **Tools**: containerization & orchestration; CI/CD pipelines
- **Architecture**: evolutionary (event-driven & microservices)
- **Strategy**: experimentation & evidence-based decisions
MGAP EXPERIMENT OVERVIEW

WHAT’S IT ALL ABOUT?
MGAP Experiment

MACHINE GENERATED ANNOTATIONS PIPELINE

Research Question

*Can commercial image tagging services improve the digital library’s metadata?*

Assertion

*We will measure improvement via user testing.*

Hypothesis

*The tags will be of no use to scholars (expert users) conducting research, but they will be of use to casual (nonexpert) users just looking for interesting images.*
MGAP Context

UCLA DIGITAL LIBRARY

- New platform under development (https://digital.library.ucla.edu/)
- Uses IIIF protocol for images
  - See Project Mirador for demo of a IIIF-powered viewer
- Web Annotations is a sister protocol to IIIF
  - Enables anyone to annotate any web resource
  - Could be used for transcription, translation, crowd-sourced metadata, scholarly dialogue, etc.
MGAP Team

THE LABS – A SOFTWARE DEV SUBTEAM

• Mission: conduct experiments; build prototypes
  • Projects should be strategic!

• Team Members:

<table>
<thead>
<tr>
<th>Dev</th>
<th>Ops</th>
<th>UX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kristian Allen</td>
<td>Anthony Vuong</td>
<td>Tinuola Awopetu</td>
</tr>
<tr>
<td>Mark Matney</td>
<td></td>
<td>Sharon Shafer</td>
</tr>
</tbody>
</table>
MGAP Motivation

FIRST EXPERIMENT - MGAP

• Strategic Motivations:
  • Answer the question: Can tags improve the digital library?
  • Gain experience with Web Annotations (*for future user functionality*)
  • Gain experience with event-driven systems (*for future software architecture*)
  • Gain experience conducting usability tests (*part of UX Strategic Plan*)
    • Try out [usertesting.com](http://usertesting.com)
MGAP SETUP

HOW DID WE DO IT?
MGAP Setup

TOOLS

- **Queue processing**: Python, Celery, RabbitMQ
- **Image tagging**: Google Vision, AWS Rekognition, Clarafai, Azure CV*
- **Annotation storage**: Elucidate
- **Search Index**: Apache Solr
- **Search interface**: Blacklight
- **Deployment**: Docker, AWS
MGAP Architecture
MGAP Content

DIGITAL IMAGE COLLECTIONS

- LA Daily News Negatives
  - 5,172 images

- Will Connell Papers
  - 502 images

- Walter E. Bennet Photographic Collection
  - 79 images
MGAP TESTING

PART I: IN-PERSON
MGAP User Testing

IN-PERSON

• Participants: 9 UCLA staff (mostly from the library)
  • Note: It’s quite difficult to recruit students during finals week!

• Setup:
  • 5 instances of the digital library interface
    • 1 unaltered (base case)
    • 3 with tags from individual services
    • 1 with combined tags from all three services
  • Each user sees 2 browser windows:
    • Base case
    • 1 of the 4 alternatives
MGAP User Testing

IN-PERSON

• **Test 1: Specific search**
  • Participants given exact terms (*sunglasses, chinatown, cat*)
  • Perform search in both interfaces
  • Asked to rate relevancy of results

• **Test 2: Scenario**
  • Participants given topic (*wardrobe selection*)
  • Perform searches with their own terms
  • Asked to rate relevancy of results

• **Test 3: Free-form**
  • Participant performs a search of their own
  • Asked to rate relevancy of results
MGAP User Testing

IN-PERSON

• Findings
  • Users want to scan lots of images at once
    • *Will page through results, unlike text-based searching*
  • Users expect more semantic understanding by the system
    • *Synonyms should produce the same results (cat, feline)*
    • *Tags should be members of ontological hierarchies (cat, mammal)*
  • Users did not understand results ranking
  • Some tags were irrelevant
MGAP User Testing

IN-PERSON

• Limitations
  • Participant pool may be "biased"
    • Library staff are not representative of library users
  • Image set was not very large
  • Did not specifically ask to compare across the two interfaces

• Challenges
  • Recruiting
  • Scheduling
MGAP TESTING

PART II: ONLINE
MGAP User Testing

ONLINE

• Platform: https://usertesting.com
  • Test creation similar to building a form

• Benefits:
  • Panel of ~1.5 million users/testers
    • No more recruiting!
    • Demographic filtering capability
  • Tests are unmoderated
    • No more scheduling woes!
  • Display and audio recording

• Drawbacks:
  • Tests must be very well defined
  • Cost $$$$
MGAP User Testing

ONLINE

• **Participants:** 21 total
  • 1 dry run
  • 5 for each of the 4 test setups

• **Setup:**
  • Same as in-person setup
  • 5 instances of the digital library interface
    • 1 unaltered (base case)
    • 3 with tags from individual services
    • 1 with combined tags from all three services

• Each user sees 2 browser windows:
  • Base case
  • 1 of the 4 alternatives
MGAP User Testing

ONLINE

• Test 1: Specific search
  • Same as in-person Test 1
  • *Added: Compare relevancy between the two interfaces*

• Test 2: Compare tags
  • Ask user to look at tags for a given image in all four interfaces
  • Compare relevancy across the sets of tags

• Test 3: Scenario
  • Participants given topic (wardrobe selection)
  • Perform searches with their own terms
  • Asked to rate relevancy of results

• Test 4: Free-form
  • Participant performs a search of their own
  • Asked to rate relevancy of results
MGAP User Testing

Task 1a "Sunglasses" Comparisons

- AWS
- Clarafai
- Google
- Combo

# of users

- Neither
- Equal
- Tagged
- Control
MGAP User Testing

Task 1b "Theater" Comparisons

- AWS: 3 users with Equal, 2 users with Control
- Clarafai: 4 users with Equal, 1 user with Tagged
- Google: 5 users with Equal, 1 user with Control
- Combo: 4 users with Equal, 1 user with Control

# of users
MGAP User Testing

Task 2 - LBJ riding a horse

- **Accuracy**
- **Usefulness**

<table>
<thead>
<tr>
<th></th>
<th>AWS</th>
<th>Clarafai</th>
<th>Google</th>
<th>Combo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating</td>
<td>3.0</td>
<td>2.5</td>
<td>3.5</td>
<td>3.2</td>
</tr>
</tbody>
</table>

avg. rating (1 to 5)
MGAP User Testing

Task 3 - Wardrobe Designer

<table>
<thead>
<tr>
<th></th>
<th>AWS</th>
<th>Clarafai</th>
<th>Google</th>
<th>Combo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neither</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Equal</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Tagged</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Control</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

# of users
MGAP User Testing

Task 4 - Free Form Search

- AWS
- Clarafai
- Google
- Combo

#of users

Neither | Equal | Tagged | Control
---|---|---|---
5 | 4 | 2 | 2
MGAP User Testing

ONLINE

• Challenges
  • Volume of respondent data is large
    • 15-45 minutes of video for each user (20x)

• Limitations
  • Image set was not very large and narrow in scope
  • Participant size was too small for statistics
  • Poor choice of instrument
    • We tried to get quantitative data from user testing, which is really a qualitative research method
CONCLUSION

WHAT'S NEXT?
MGAP Outcomes

• Findings
  • Inconclusive, but results suggest our hypothesis was correct
  • Tags do not provide a *strong* benefit; may actually confuse users
  • Of the 3 services, Google seems to perform best and AWS worst

• Decisions
  • Do not add CV tags to the digital library (*yet*)
  • Alternative tag-based image portal may be useful to casual users

• Achievements
  • The *experiment* was inconclusive, but the *project* was a success!
    • *Achieved strategic goals of training team on desired skills*
    • *Discovered UX problems during testing; led to enhancements on production site*
Future Work

- **AI & Libraries**
  - Try more metadata enhancement experiments

- **Event-driven systems**
  - PubSub for all repositories (catalog, Dataverse, Dig. Lib.)
  - Aggregated index and single search for all library data

- **IIIF & Web Annotations**
  - Custom collections
  - Scholarly annotations
  - Time-based media support

- **Library UI improvements**
  - Website redesign
  - More user testing!!!
Thank You