ABSTRACT
Museums traditionally rely on trained cataloging professionals to create metadata for their collections. While this authoritative information is well-grounded, it is brief and limited in its description of the museum objects since the human cataloging task is time-consuming and expensive. New techniques provide an opportunity to expand subject-oriented explanatory metadata. Social tags and linguistic analysis of descriptive text holds promise, but there are many challenges to integrating these computational techniques for museum applications. In this paper, we present our initial investigations along these lines and discuss a research program to improve the integration of computational linguistics, human-computer interaction, and recommender systems to improve access to images in a museum context.

Keywords
Social tagging, computational linguistics, weighting algorithms, disambiguation, image search, museums and libraries, interdisciplinary collaboration, cultural heritage.

1. INTRODUCTION
Museums traditionally rely on trained cataloging professionals to create metadata for their collections. While this authoritative information is well-structured and grounded in cultural tradition, it is brief and limited in its description of the museum objects. New computational linguistic techniques provide an opportunity to expand the metadata and to create linguistically driven approaches to examining multiword descriptive phrases. Social tags and analysis of descriptive text holds promise, but there are many challenges to integrating these computational linguistic techniques for museum applications.

The work we present here comes from a collaborative, cross-disciplinary research project comprising academic researchers, digital librarians, and museum professionals. We explore the application of...
techniques from computational linguistics and social tagging to the creation of linkages between the formal controlled language of museums and the vernacular language of text descriptions. We use text mining algorithms, taxonomies, and lexical resources to identify suggested terms and thus aid users in tagging and retrieving images based on tags assigned from many different perspectives. We use the trust (Golbeck 2005, Golbeck 2008) a user places in particular metadata sources (e.g. other users or other sources) to infer a weighted set of results for their searches. Consideration of these weights in ranking algorithms—along with term relationships from lexical resources—has the potential to produce high-quality, focused and personalized retrieval of works from image collections.

2. METHODS
In this paper, we specifically address the methods and application of tag analysis and the integration of computational linguistic methods to identify meaningful phrases extracted from documents. Our ultimate goal is to use tags and text to identify images relevant to a user’s need, to cluster images according to similarity of terms from text and user-assigned tags used to describe them, and thus to develop new computational linguistic ranking algorithms in the process of achieving this application outcome.

Our partner and primary data source is the steve.museum project, a social tagging system designed for museums to gather annotations on their collections although we plan to extend this to other open source collections in the future.

Our data sets consist of:

1. A full set of social tags submitted by users in the steve.museum project over the last 5 years. These tags were collected through a web interface as part of the larger steve museum project. The initial steve dataset has 49,767 user-submitted social tags by token and 15,851 tags by type.

Recently, these tags have been augmented with nearly 300,000 additional tags arising from data collection using crowdsourcing. This substantial steve tagset thus enables realistic algorithm development and reliable evaluation.

2. A subset of the steve tagset specifically for the images in the collection of the Indianapolis Museum of Art (imamuseum.org), combined with tags assigned to these images by IMA visitors, along with

3. Descriptive text that accompanies each of the images for works of art in this museum, including paintings, sculpture, textiles, masks, etc.

As reported in Trant et al (2007), tags contributed directly by users might help bridge the gap between professional and public discourse by providing a source of terms not found in formal museum documentation. Note also that in this application area, there is no on-line "cnet" or other blog-type text data available from which informal language from the public can be collected or analyzed.

For each image in our dataset, we analyzed the set of user-generated tags, and then, using the CLiMB Toolkit (Klavans et al. 2008), extracted terms automatically from descriptive text. For example, consider the image in Figure 1, where a selection of tags is given on the right and an excerpt from the IMA handbook is also provided.

In the next section of this paper, we present an analysis of user-generated tags, using a basic part-of-speech tagger which takes a phrase as input, and assigns the probably part of speech to the words in that phrase. Next, we discuss more sophisticated computational linguistic methods using syntactic analysis for more complex phrase identification taking descriptive text as input in order to select representative terms. Finally, we present a novel method to assign weights to the tags and terms to better understand which of these inputs best reflects user judgments on ranked search terms.

3. TAG ANALYSIS
Identifying the types of tags that are used is important for understanding user tagging behavior and for planning applications that use the social tags. One of our first projects has been an analysis of tags that
users submit. Of these, 10,508 tokens and 7,317 types were multiword tags (21.1% and 46.2% respectively). Examples include “für trim”, “black and white photography”, and “dark skies”.

We were also interested in which of these multiword tags were lexicalized phrases or idioms. For example, the location “New England” refers to a single geographical area which is neither “New” (anymore) nor “England”. Similarly with an idiom, the meaning of that idiom cannot be constructed from its component parts; for example, when someone “kicks the bucket”, they are not generally able to kick at all.

We have taken our tagset and processed all legitimate multiword phrases (excluding acronyms or tags with punctuation) with a part of speech tagger using the Natural Language Toolkit, a suite of open source libraries written in Python. We first ran the Morphy morphological analyzer (Lecius 1996) to compute raw overlap of tags and terms by type. Results of overlap on a random set of six images, one from each of the five types plus the image in Figure 1, is shown in Table 2. We first ran the Morphy morphological analyzer (Lecius 1996) to compute raw overlap of tags and terms by type. Results of overlap on a random set of six images, one from each of the five types plus the image in Figure 1, is shown in Table 2.

Table 2 shows that there is wide variety in the number of overlapping tags/terms, as shown in Column 2. Table 2 shows that there is wide variety in the number of overlapping tags/terms, as shown in Column 2. However, the average appears to be between 9-10% over the combined tag-term set. Note that the overlap

4. TEXT ANALYSIS

We have built and implemented a text-mining system which applies computational linguistic techniques to automatically extract metadata for image access as part of the Computational Linguistics for Metadata Building (CLiMB) research project. Our premise in this project is that automatic and semi-automatic techniques may help fill the existing metadata gap by facilitating the assignment of subject terms. Using the CLiMB Toolkit, we extracted terms from the handbook descriptions of 165 images from the collection of the Indianapolis Museum of Art (IMA) (imamuseum.org). There are 16,049 terms by token, and 4,788 unique terms by type. The tag-term collection constitutes a novel research dataset for computational linguistics and one which we will provide to the larger community for analysis and use.

Our hypothesis was that tags would be more informal, whereas the text, even in a handbook, would be more formal. We anticipated little overlap, and expected that the overlap would be in words of certain categories, i.e. color, shape and representation (e.g. red, box, and woman). In order to test this hypothesis, we selected images of five types - Asian Sculpture, Abstract Painting, Representational Painting, Costumes, and Biblical Work - from the larger set of 165 images. The rationale for selecting images of five types was to provide variety and balance in tags since different types of work enable different types of tagging. For example, abstract art rarely generates concrete noun tags, such as “girl” or “sisters” since there are no identifiable subjects in the images.

We first ran the Morphy morphological analyzer (Lecius 1996) to compute raw overlap of tags and terms by type. Results of overlap on a random set of six images, one from each of the five types plus the image in Figure 1, is shown in Table 2. Table 2 shows that there is wide variety in the number of overlapping tags/terms, as shown in Column 2. However, the average appears to be between 9-10% over the combined tag-term set. Note that the overlap

4 The CLiMB project was initially funded by the Andrew W. Mellon Foundation at the Center for Research on Information Access at Columbia University and then at the College of Information Studies and the University of Maryland Institute for Advanced Computer Science at the University of Maryland, College Park, MD.

5 The CLiMB Toolkit applies Natural Language Processing (NLP), categorization, and disambiguation techniques over texts about images to identify, filter, and normalize high quality subject metadata for the use of image cataloging professionals. More information about the CLiMB toolkit may be found at http://www.umiacs.umd.edu/~climb/ and Klavans et al. 2009.
for tags alone averages 20.2% whereas the average for terms is 15.2% since the term set tends to be larger. This comparison is found by looking at the average in Columns 7 and 8 of Table 2.

5. TERM WEIGHTING FOR TAGS AND TEXT

We have developed a novel algorithm that runs over tags and terms extracted from text documents to identify and filter terms which discriminate images for searchers.

### Table 2. Overlap of Tags and Terms over Six Sample Images

<table>
<thead>
<tr>
<th></th>
<th>Column 1</th>
<th>Column 2</th>
<th>Column 3</th>
<th>Column 4</th>
<th>Column 5</th>
<th>Column 6</th>
<th>Column 7</th>
<th>Column 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>(O):</td>
<td></td>
<td>(T+T):</td>
<td></td>
<td></td>
<td>(Tag):</td>
<td>(Term):</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tag and</td>
<td>Overlap</td>
<td>Unique</td>
<td>Tags +</td>
<td>Unique</td>
<td>Total</td>
<td>Total</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Term</td>
<td>between</td>
<td>Tags</td>
<td>Terms</td>
<td></td>
<td>Tags</td>
<td>Terms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Row 1</td>
<td>Image 1</td>
<td>9</td>
<td>91</td>
<td>36</td>
<td>64</td>
<td>9.9%</td>
<td>25.0%</td>
<td>14.1%</td>
</tr>
<tr>
<td>Row 2</td>
<td>Image 2</td>
<td>14</td>
<td>111</td>
<td>42</td>
<td>83</td>
<td>12.6%</td>
<td>33.3%</td>
<td>16.9%</td>
</tr>
<tr>
<td>Row 3</td>
<td>Image 3</td>
<td>13</td>
<td>140</td>
<td>82</td>
<td>71</td>
<td>9.3%</td>
<td>15.9%</td>
<td>18.3%</td>
</tr>
<tr>
<td>Row 4</td>
<td>Image 4</td>
<td>1</td>
<td>111</td>
<td>27</td>
<td>85</td>
<td>0.9%</td>
<td>3.7%</td>
<td>1.2%</td>
</tr>
<tr>
<td>Row 5</td>
<td>Image 5</td>
<td>19</td>
<td>121</td>
<td>68</td>
<td>72</td>
<td>15.7%</td>
<td>27.9%</td>
<td>26.4%</td>
</tr>
<tr>
<td>Row 6</td>
<td>Image 6</td>
<td>11</td>
<td>138</td>
<td>71</td>
<td>78</td>
<td>8.0%</td>
<td>15.5%</td>
<td>14.1%</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9.4%</td>
<td>20.2%</td>
</tr>
</tbody>
</table>

Since the algorithm combines weights over both tags and terms, we refer to this frequency as an inverse of document frequency, where “document” in this case refers to either the tag or term dataset. The normalized tag/term frequency by tag/term denominator is shown in Figure 2:

\[
\text{tag/term-frequency-h(t)} = \frac{2 \times \text{tag-frequency}(t) \times \text{term-frequency}(t)}{\text{tag-frequency}(t)} + \text{term-frequency}(t)
\]

**Figure 2: Weighting over Terms and Tags (WOTT)**

The tag/term-frequency-h(t) means that the average of tag and term frequency is computed by harmonic mean which is the same formula as F-score, while the tag/term-frequency(t) uses an arithmetic mean. Given the difference in tag and term sets, we have normalized to prevent bias for images with larger tag sets or for images with longer textual descriptions.

5.1 Experiments

Our goal was to identify which tags or terms might be useful to characterize the individual distinguishing features of images, and thus perhaps be helpful to users in differentiating and/or searching images. In order to test this hypothesis, we selected a random subset of the steve tagging data with 12,600 tags by token, and 4,000 unique tags by type. We varied the WOTT formula, generated results, and tested with a small set of human subjects. Subjects were asked to rate the tags and terms on a 5 point Likert scale where 1 indicates “highly useful” and 5 indicates “not useful”.

### Table 3. Rank of Tags and Terms over One Image

(1 = highly useful, 5= not useful)
Table 3 shows the results of user ratings of tags and terms that we have applied to a formative evaluation of how these judgments correlate with the WOTT metric.

Note that the highest ranked tags are ‘girl’, ‘vase’, and ‘drape’, whereas for terms, the highest ranked is ‘artist’. In this illustrative sample set, no single item received two top rankings of “1”, although many received more than one “5”, indicating that it may be more difficult to achieve popularity in ranking as opposed to skepticism.

With this evaluation data, we have been able to preliminarily compare user rankings with variations on WOTT to determine which algorithmic variations appear to match human judgments although more data is needed. Extending this evaluation is our next step for future work on this data.

6. RELATED RESEARCH

We are not aware of any research that addresses the issue of combining the use of tags and text mining for search over digital collections. Thus, we compared this task to the information retrieval task of determining which documents might be relevant to a given query by using basic metrics such as tf*idf or, in future work, Latent Dirichlet Allocation (Blei et al. 2003) or Latent Semantic Indexing (Deerwester et al. 1990). The notion of term frequency (tf) has a long history, starting with early research on summarization (Luhn 1958) to the exploration of inverse document frequency (idf) (Spärck Jones 1972), to relevance weighting (Salton & Buckley 1988), leading to many variations to characterize topic and thus determine relevance to a query.

7. CONCLUSIONS AND FUTURE WORK

In this paper we will present initial results in the analysis of social tags and the application of computational linguistics techniques to text in order to improve access to images in a museum context. We will present the results of several variations on the basic WOTT measure, and demonstrate the impact on output. We discuss evaluation of our results on the ability of the metric to reflect human judgments through experiments that illustrate the value of the approach.

In future work, we intend to measure the variance, to examine the types of lexical items which tend to overlap, and the relationship between these items and their frequency in a large corpus. The goal will be to gain further insights on the nature of the overlapping terms and tags and the images they describe. We will also address the issues of additional algorithms for measuring tag-term value, and specifically the complex issue of the relationship between language and the images described by language. Finally, we have started a related seeding project on multilingual terms and tags, to be able to access images regardless of the language or cultural barriers.

8. REFERENCES