IPKB: A Digital Library for Invertebrate Paleontology

Yuanliang Meng[†], Junyan Li[†], Patrick Denton[†], Yuxin Chen[§],
Bo Luo[†], Paul Selden[‡], and Xue-wen Chen[†]

[†] Department of EECS, The University of Kansas, Lawrence, KS, 66045, USA
[‡] Department of Geology, The University of Kansas, Lawrence, KS, 66045, USA
[§] Department of Computer Science, Swiss Federal Institute of Technology, Zurich, Switzerland
{ymeng, j357l401, pdenton, bluo, selden, xwchen}@ku.edu, yuxin.chen@inf.ethz.ch

ABSTRACT

In this paper, we present the *Invertebrate Paleontology Knowledgebase* (IPKB), an effort to digitize and share the *Treatise on Invertebrate Paleontology*. The Treatise is the most authoritative compilation of invertebrate fossil records. Unfortunately, the PDF version is simply a clone of paper publications and the content is in no way organized to facilitate search and knowledge discovery. We extracted texts and images from the Treatise, stored them in a database, and built a system for efficient browsing and searching. For image processing in particular, we segmented fossil photos from figures, recognized the embedded labels, and linked the images to the corresponding data entries. The detailed information of each genus, including fossil images, is delivered to users through a web access module. Some external applications (e.g. Google Earth) are acquired through web services APIs to improve user experience. Given the rich information in the Treatise, analyzing, modeling and understanding paleontological data are significant in many areas, such as: understanding evolution; understanding climate change; finding fossil fuels, etc. IPKB builds a general framework that aims to facilitate knowledge discovery activities in invertebrate paleontology, and provides a solid foundation for future explorations. In this article, we report our initial accomplishments. The specific techniques we employed in the project, such as those involved in text parsing, image-label association and meta data extraction, can be insightful and serve as examples for other researchers.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information search and retrieval; E.2 [**Data Storage Representations**]: Linked representations

General Terms

Design, Documentation

JCDL'12, June 10–14, 2012, Washington, DC, USA.

Keywords

Digital library, Digitization, Paleontology

1. INTRODUCTION

The science of paleontology advances not only from discoveries of important new fossils but also through innovative ideas based on data accumulated over previous generations. The robustness of such studies depends on the quality of data. Paleontologists agree that the most authoritative compilation of data on invertebrate fossils is in the *Treatise on Invertebrate Paleontology*[1]. The *Treatise* was founded in 1948 and the first volume appeared in 1953. Since then, The Paleontological Institute of the University of Kansas has published 50 volumes, authored by more than 300 contributors worldwide. This encyclopedic work now occupies more than 1.3 meters of shelf space. For paleontologists (and geologists, biostratigraphers, etc.) the world over, the Treatise holds an almost biblical significance and is to be found in every good library.

Given the rich information in the Treatise, understanding and modeling paleontological data are significant in many areas, such as: understanding evolution; understanding climate change; finding fossil fuels, etc. The vast repository of paleontological data contained in the Treatise needs to be made available in electronic form for present and future workers to extract the greatest possible use from the work. In order to realize the maximum possible benefit from this landmark effort, there is a strong desire within the paleontological community to be able to readily access and use this data. However, it is non-trivial to convert large volumes of the Treatise into a structured, and easily accessible digital library. In particular, we face the following challenges: (1) the Treatise contains heterogeneous data objects that are not easy to be associated under a unified framework, in particular, linking text entries with fossil images; (2) although the editorial policies and procedures of the Treatise have always been of the highest standard, it consists of manuscripts from hundreds of paleontologists, which introduces significant inconsistency of styles, formats, and sometimes terminologies.

In this paper, we report *IPKB*, a first effort towards digitization, organization and integration of the rich information extracted from the Treatise. The ultimate goal is to provide scientists with a general framework that facilitates knowledge discovery activities in invertebrate paleontology. IPKB first extracts raw data from the Treatise, and processes textual and image data objects separately. We have designed an

[1] http://paleo.ku.edu/treatise/

ontology to capture fossil information, and explored ways to link fossil (genera) descriptions with images. Finally, a web interface is designed to present the information, and provide easy browsing and searching functions. With the help of our system, users can achieve some goals that are impossible with a PDF file alone. The following are examples of such scenarios:

Use case 1. A paleontologist wants to browse all genera of superfamily *Plectorthoidea* between geological times *Eifelian* and *Moscovian*.

Use case 2. In Japan, a geologist finds a brachiopod fossil which is subpentagonal and rectimarginate; and the foramen is permesothyrid. He wants to check the geological period it could correspond to.

In the following sections, readers will get to know how the system is designed and implementted. The rest of the paper is organized as follows: we introduce the general architecture of IPKB in Section 2, and present our text and image processing approaches in Sections 3 and 4. We then describe system integration, including accessing and searching components in Section 5. We summarize related works in Section 6, and finally conclude the paper in Section 7.

2. IPKB: SYSTEM OVERVIEW

The Treatise on Invertebrate Paleontology is a pandect of all invertebrate fossil genera, together with their taxonomic synonyms, stratigraphic ranges, geographic distributions, and illustrations of the type specimens. In the Treatise, tremendous amounts of data have been accumulated from different resources, and organized in old-fashioned ways. The Treatise is published in separate parts, *Part A* through *W*, and each part contains multiple volumes. Earlier volumes are scanned from printed books, while the newest volumes have PDF files from the publisher. We started from *Part H. Brachiopoda*, which has 6 volumes of paleontological data of brachiopods, a phylum of marine invertebrates with two valves (or "shells"). Fossils are classified in a hierarchical structure (from highest level to lowest): *orders* (e.g. *Lingulida*, *Orthida*); *suborders* (e.g. *Dalmanellidina*, *Orthidina*); *superfamilies* (e.g. *Plectorthoidea*, *Wellerelloidea*); *families* (e.g. *Allorhynchidae*, *Pontisiidae*); and finally *genera* (over 4,000 genera in Part H).

Although the recent volumes (e.g. all six volumes in Part H) are digitized [20], they are no more than electronic reprints of paper publications. As we know, the only way to search in a PDF file is using exact text matching. In our system, however, users are able to submit structured queries. As illustrated before, from a PDF file it is impossible to locate records of certain morphological features and/or geological distributions, across a variety of families. Our advanced search function, however, can find relevant records instantly and rank them according to relevancy. When reading a book, we often find it inconvenient to jump from one topic to a related one, not to mention some external sources. Our system provides a flexible interface for doing so, too. In addition, traditional publications do not sufficiently take users' experience into account. Modern systems like ours emphasize more on information layout and data presentation methods. Hopefully, information access turns out to be more enjoyable as well.

Figure 1 demonstrates the overall structure of the IPKB system. Please note that, for simplicity of the description,

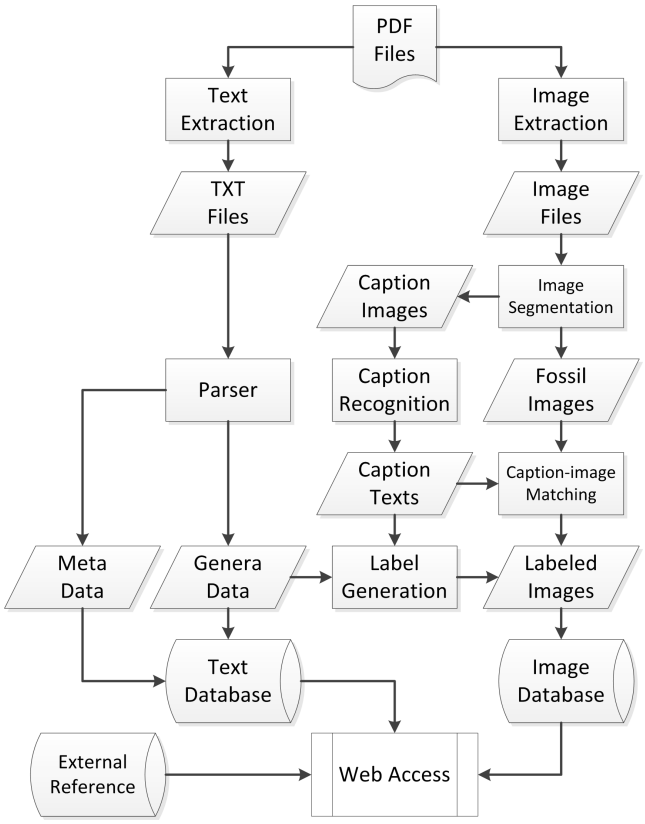

Figure 1: Overview of the IPKB system.

we omit the details in the *Web Access* component for now. As shown in the figure, we first extract text and image data from PDF files, and process them separately. Text data is parsed to obtain structured genus records and other meta data (e.g. phylum introduction, glossary, etc). Images are segmented and image captions are recognized and associated with the corresponding segment. Text descriptions of the genus records are then linked with the images. Finally, processed text and image data is provided to the web access module, which includes browsing, searching and genus display as three main components. In the rest of the paper, we use *Part H. Brachiopoda* as examples to describe each module in details. However, the methods are applied to all volumes.

3. TEXT PROCESSING

Part H of the treatise consists 6 volumes with 3,226 numbered pages in total. All volumes are available in PDF format, which was directly generated from the typesetting software. There are 1960 indexed figures. In most cases, a figure contains a number of fossil images, and each image has a label. In IPKB, text and images are extracted and processed separately. We will discuss text processing in this section, and image processing in next section.

In brachiopod paleontology literature, the basic taxon is usually a genus. In the Treatise, each genus record consists of a paragraph of text description, and a few corresponding images. According to the editorial requirements, genus de-

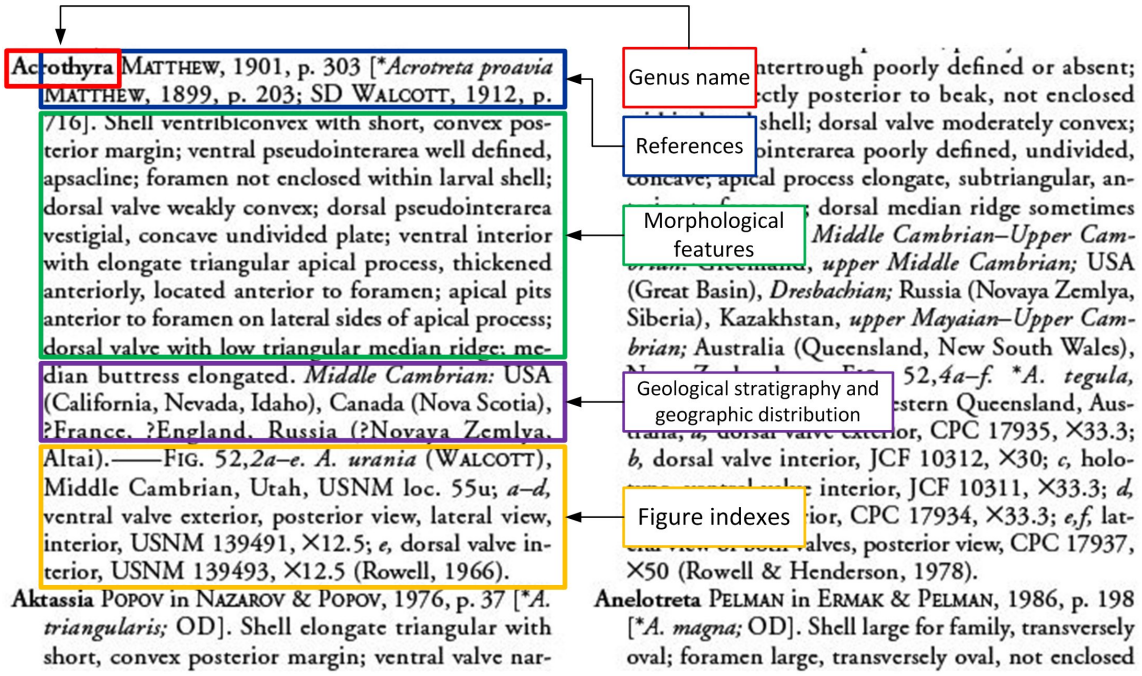

Figure 2: The description of individual genera in the treatise.

scription should provide the following pieces of information in a fixed order:

1. Genus name; e.g. *Acrothyra*.
2. References; linking to external publication(s), i.e. the original sources describing the genus.
3. Morphological features; e.g. "shell ventribiconvex with short, convex posterior margin; ventral ..." The editorial guidelines also requires the authors to describe features in a stated order, and use limited vocabulary. However, such guidelines are not strictly followed by all the contributors.
4. Geological stratigraphy and geographic distribution; e.g. "Middle Cambrian: USA (California, Nevada, Idaho), ..." Please note that fossils are usually found at multiple locations, and the locations are identified in highly inconsistent manner.
5. Figure indexes; e.g. "FIG. 52, 2a-e." Multiple images may be included for a genus.

Figure 2 shows an example of a genus record. In each paragraph of a genus, missing information is not uncommon and when something is absent, there is no explicit indicator. Currently, we are not exploiting the references, since all the related works are not digitally available, which prevents us from providing any hyperlink. All other pieces of information are crucial and are incorporated in the IPKB system. In addition, every genus is governed by some higher taxa as mentioned before. Names of the corresponding higher taxa are not given in the paragraphs, but they are found in the titles of corresponding chapters and their subordinate sections. These names are important information and are captured too.

We have implemented a Python program to extract the information and save the parsed text in an XML file. The algorithm is based on regular expression matching, which uses keywords and syntax of the text descriptions to split the text paragraph into structured data. Here the syntax does not refer to sentence structures, but merely indicates format patterns, including capital letters and punctuations and so on. Please note, other than pure textual data, we do not have any metadata (e.g. font, size, style, etc.), since we were unable to robustly extract typesetting data from the PDF (possibly due to the typesetting software used to produce the PDF file).

Genus names consist of a number of letters, with the first one capitalized. What follows is an author's name with all capital letters. This is a unique context and can be used in the extraction algorithm directly. The part of geological and geographical distributions starts with some reserved term for geological time, and ends with a long hyphen in the text. Therefore it can be obtained once a list of terms are provided. The beginning of the geological part is also the ending of the morphological description, so clear indicators of boundaries can be used in both ways, forward and backward. Such features also help us to identify and recover incorrectly formatted genus descriptions.

In order to get the names of higher taxa, we need to capture an order name first, and then assume all the genera belong to that particular order until a new order is captured. The same principle applies to suborders, super families, families, and subfamilies. Sometimes one or more taxon names are not available, and we just marked it as "UNKNOWN". Occasionally a taxon is "UNCERTAIN" because it is controversial in literature, and such cases are clearly indicated in the Treatise we work on.

We have designed an ontology to capture the genera data and meta data extracted in text processing. Genera data

includes entries of genus records as described above, while meta data includes other descriptive information that is included in the volumes. For instance, volume information, the glossary, an introduction to the phylum, and so on. Structured textual data are then stored in XML format, which will be used not only in the IPKB project.

4. IMAGE PROCESSING

Image extraction is the most time-consuming task of data preparation. Currently what we have processed is photos of fossils, corresponding to the entries of the genus records. Figure 3 (a) shows an example of the original figures extracted from the PDF files.

As shown in Figure 3 (a), the fossil images in the Treatise are manually selected, grouped and placed in a single image canvas. Each figure in the PDF file is a flattened bitmap image, without any metadata to split individual fossil images, and we do not have access to individual fossil images either. Each figure typically contains images of a few fossils, all in grayscale. The images with the same numeric index belong to the same genus. For instance, Figure 3 (a) *1a – e* all belong to the genus *Overtonia*. Different letters a, b, c and so on usually refer to various views of the same fossil, but they could also be different specimen from the same genus. As a result, the label of an image is typically a number followed by a letter, and occasionally only a number (if the genus has only one fossil photo).

We first need to segment and extract all fossil images from the figures. In order to locate the images correctly, the label under each fossil (e.g. 1a) needs to be identified and recognized too. Moreover, we need to exploit the labels to match the images with the corresponding genus descriptions, hence link genus records with fossil images. Images are first extracted from PDF files using a function provided by Adobe Acrobat. The process is mostly automatic, while manual intervention is often necessary. Segmenting the images and recognizing the labels, however, are much more difficult. We implemented the algorithms with Matlab. The major steps are:

(1) Segment fossil images and label images;

(2) Detect contours of fossil images;

(3) Match fossil images with label images, using their mutual distances in the figure;

(4) Compare the label images with templates and determine the content;

(5) Name the fossil images with their label.

In the following we will describe the methods in more details.

4.1 Image segmentation

Part H of the Treatise documents Brachiopods, which are marine invertebrates with two shells, hence, the photos exhibit fossilized shells of various shapes. The vast majority of fossil images are presented with a blank background, such as those in Figure 3 (a). However, occasionally a fossil can been seen *in situ* i.e. in a rock, as Figure 4 demonstrates. Segmentation of such images is a hard computer vision problem, which is not handled at this stage of the project. In this case, we just treat the fossil with the rock as an entity and do not further segment.

We have implemented an approach that is similar to the

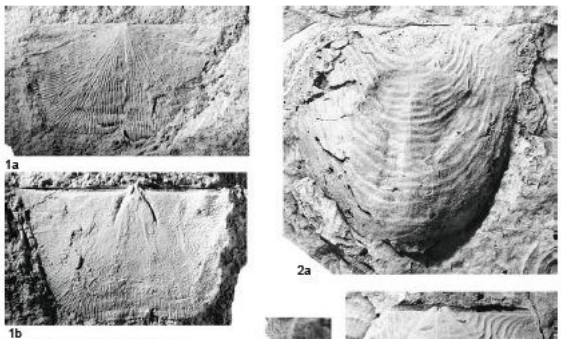

Figure 4: A figure from the Treatise. The fossils are embedded in rocks.

Marker-Controlled Watershed Segmentation method [10] to separate multiple fossil images in the same figure. First, we use the *opening-by-construction* approach to mark the objects i.e. fossil images as well as the background. This approach consists of an image erosion phase followed by an image reconstruction phase [32]. In image erosion, we choose a disk-like structuring element object with a diameter of 20 pixels. Anything smaller than 40 pixels in any dimension will be totally eroded. Since the labels and captions are usually about 30 pixels tall, such an operation can retain fossil photos and eliminate labels, markers, embedded captions and other noise. In the image reconstruction, the idea is to use the eroded image as the "seed", and use the original image as the mask. In our implementation, we reconstructed the background (which includes labels and captions), i.e. the reverse of the fossil photos. Figure 3 (b) shows the result of the reconstruction. The background is now bright (i.e. 1) and the fossil photos are dark (i.e. 0).

After the image fossils (i.e. large objects) are identified and the background is marked, we identify the labels (i.e. small objects) similarly. We first mask out the fossils from the original image, as Figure 3 (c) shows. Another image opening procedure is implemented to obtain the label blocks. In this case, we used an image erosion followed by an image dilation, and the size of the structuring element object is set to 10 pixels so that embedded text labels are kept. Among the identified text blocks, we retained the smaller ones since we are only interested in labels, not the captions (e.g. "Overtonia" in the figure). The identified labels are shown in Figure 3 (d).

Finally, we have employed the *watershed* approach to partition the fossil images in the figure, and later they are saved as separate JPG files. The watershed approach is to find local minima in an greyscale image, and cut them in pieces [17]. Assuming that pixel intensity values in a digital image represent the altitudes, we can draw an analogy between a grayscale image and a topographic relief. Therefore there are "peaks" with local maxima and "basins" with local minima. The idea of the watershed algorithm is to continuously add "water" into the basins to make them "flood". When the levels of water rise to the point where two water sources meet, a "barrier" is built. The resulting set of barriers are the watersheds, which separate the regions defined by basins. In our case, the fossils in the figure have been marked with dark pixels (i.e. basins), and the background has been marked with white pixels. The watershed method is used to determine

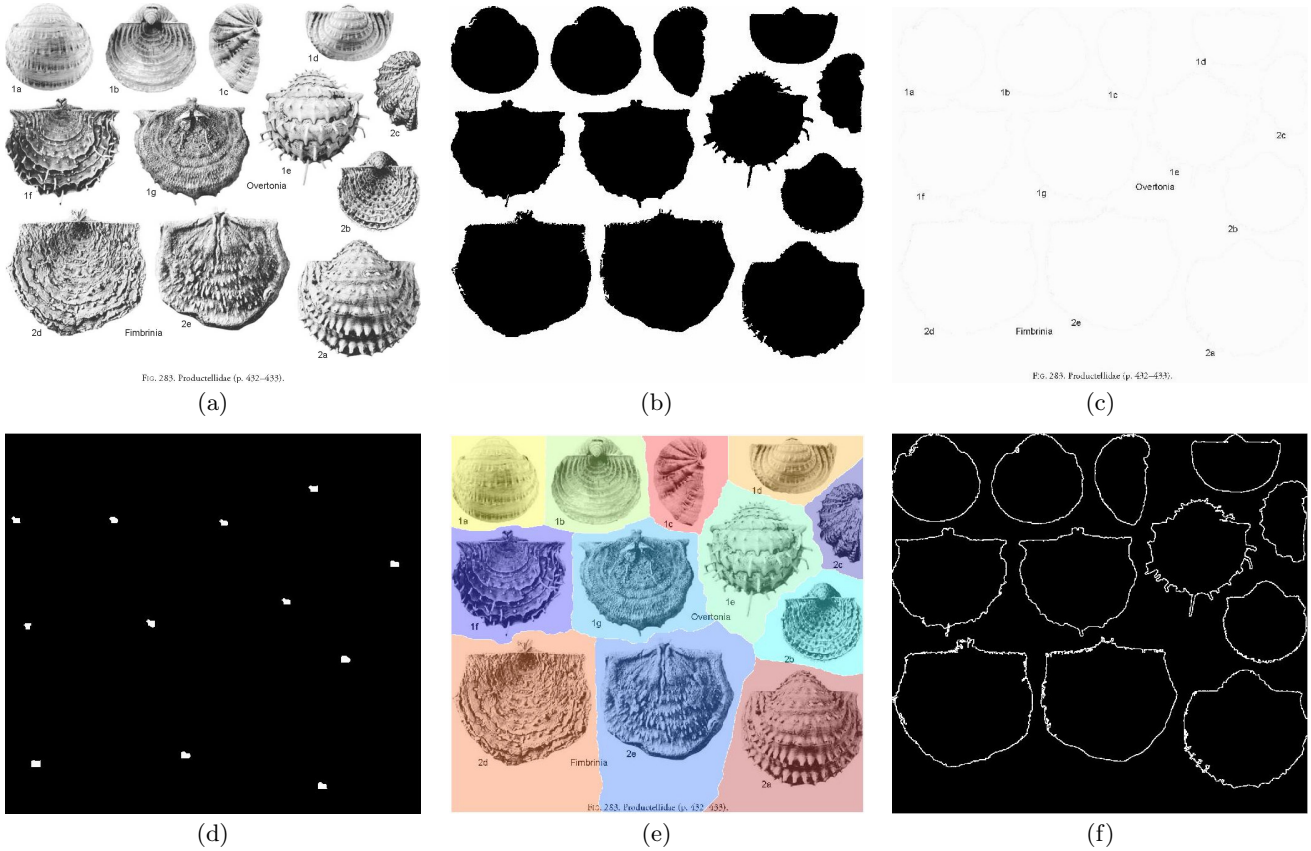

Figure 3: Image processing in IPKB: (a). the original figure extracted from from the Treatise; Images with the same numeric index belong to the same genus. (b). Opening-by-construction: fossils are masked. (c). Removing fossils from the figure, leaving only labels and captions. (d). Identified label blocks. (e). Using watershed method to split fossils. (f). Detected fossil contours.

the barrier between individual fossil images (i.e. basins), and hence split them from the figure. Figure 3 (e) shows the result of the watershed approach. After segmentation, each fossil image is padded to a square canvas and stored in separate JPG files.

The mathematic details are beyond the scope of our discussion, but interested readers may refer to [27].

4.2 Fossil contour detection

In the original figure, fossil labels are manually placed "near" the corresponding fossil image. They are supposed to be placed to the lower left corner of the fossil images, however, this rule is frequently violated. The amount of the images makes it impossible to manually fix such errors. Therefore, we have designed an automatic method to associated fossils and their labels.

Although fossils are segmented and extracted from the figures, we still need an accurate identification of the fossil contours to compute the minimum distance between labels and fossil images. Edge detection has been well studied in the image processing literature, among various methods [24, 10], we have employed the *sobel* edge detection approach. A sobel operator is a 3×3 kernel, which is applied to the image using 2-D convolution, to compute the (approximate) gradient of pixel intensity. For a 2D image, two kernels are used: one for the horizontal dimension and the other for the vertical dimension:

$$\mathbf{G}_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * \mathbf{I} \text{ and } \mathbf{G}_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * \mathbf{I} \quad (1)$$

The magnitude of the gradient can be obtained in the following way: one for the horizontal dimension and the other for the vertical dimension:

$$\mathbf{G} = \sqrt{{\mathbf{G}_x}^2 + {\mathbf{G}_y}^2} \quad (2)$$

The edge pixels are those with high values of $\mathbf{G}$. If the image is masked, as in our algorithm, $\mathbf{G}$ is simply zero for all the non-edge pixels. Figure 3 (f) demonstrates the detected contours of the fossil photos.

4.3 Fossil label association

We link labels with fossils based on the distance between them. At high level, each label is associated with the nearest fossil. However, there could be confusion in some cases. For instance, in Figure 3, label `1d` is placed between two fossils, with similar distance to both of them. In this case, if we can first associate the lower fossil with label `1e`, we can confidently link `1d` to the upper fossil. Theoretically, this becomes a *bipartite graph matching* problem, i.e. to find a *perfect matching* with minimum distance.

Labels \ Fossils	1	2	3	4	5	6	7	8	9	10	11	12
1	17.493	47.043	100	100	100	100	100	100	100	100	100	100
2	100	19.235	64.498	100	100	100	100	100	100	100	100	100
3	100	100	21.19	100	100	100	100	100	100	100	100	100
4	76.217	45.706	100	18.028	68.622	100	100	100	100	100	100	100
5	100	100	76.968	100	21.587	74.465	100	100	100	100	100	100
6	100	100	100	100	100	17.464	100	100	100	100	100	100
7	100	100	100	88.482	100	100	16.971	98.732	100	100	100	100
8	100	100	100	100	100	100	100	23.77	100	100	100	100
9	100	100	100	100	100	100	100	25	100	16.971	100	100
10	100	100	100	100	100	100	100	100	17.493	100	100	100
11	100	100	100	100	100	100	100	100	31.385	100	18.385	100
12	100	100	100	100	100	100	100	69.318	100	100	50.537	16.031

Figure 5: The distance matrix. Each column represents a photo and each row represents a label. The value 100 is a dummy distance for those too far away.

First, we compute the Euclidean distance between the center pixel of the label block and all pixels of a fossil image contour. The minimum value is saved as the "distance" between the label and the fossil. After all labels loop over all fossils, we obtain a distance matrix. We have extracted the distance matrix for labels and fossils from Figure 3, and demonstrate the matrix in Figure 5. All distances greater than 100 are set to 100 (and ignored), while distances smaller than 100 are kept as candidates. We scan through the columns to identify the minimum distance between labels and images. We first identify columns with only one candidate (e.g. columns 7, 10, 12 in the figure), and eliminate the corresponding rows from the matrix. For instance, we link fossil 10 with label 9 (column 10) without confusion. Therefore, although label 9 is also close to fossil 8 (column 8), they cannot be associated. We repeat this process until all labels and fossils are linked.

4.4 Label image recognition

Since the labels have a limited scope of inventory and are small in size, it is adequate to use templates for recognition. We select some label images, binarize them and save them as templates.

The recognition task is to compute the distances between target images and templates, and the one with the minimum distance is the match. Since the label images are relatively small, we used the most intuitive method: computing the normalized pixel-wise sum of square error (SSE) between images (Equation 3). The pairs with the minimum SSE is a match. In reality, the target image and the template image do not often have the same size, so we need to try different positions. Again we used a pixel-by-pixel loop.

$$SSE = \frac{\sum_{(x,y)}(\mathbf{I}_{temp}(x,y) - \mathbf{I}_{targ}(x,y))^2}{\sum_{(x,y)}\mathbf{I}_{temp}(x,y)^2 + \sum_{(x,y)}\mathbf{I}_{targ}(x,y)^2} \qquad (3)$$

Once we know what the label is, we name the associated fossil image with that label (and the figure index). Then the fossil images are properly named.

5. WEB ACCESS

An important goal of our project is to digitally deliver the Treatise to the research community and the general public. Therefore, we built a web interface to facilitate browsing and advanced searching functions. We have adopted the three-tier architecture, where the database server is powered by MySQL, the Web server is powered by Apache+PHP, and JavaScript is used on the client side. An schema shredding is enforced to convert the XML data generated in Section 3 into relational model to be handled in RDBMS, for its maturity and performance. Figure 6 shows the homepage of IPKB website. Since the Treatise is not an open-source product, subscribers need to login to view full volumes.

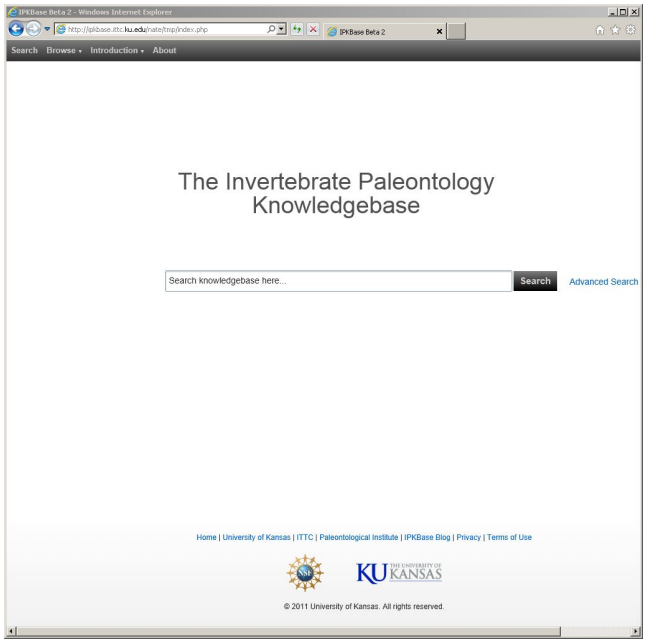

Figure 6: The homepage of the web interface.

5.1 Searching and browsing

A major function of the web access module is search. At present, we provide two search modes: quick search, which takes free text queries, and advanced search, which takes more complex predicates.

5.1.1 Quick search

Quick search accepts a free text query, which could be the name of a genus, or a family, or an order etc., and returns the records that match the name. Nomenclature in paleontology is often confusing even to professionals, hence, we support approximate matching. The algorithm runs int he following way:

(1) search for exact match of genus name;

(2) if no results are found, try a natural language full-text search over taxon names;

(3) if no results are found, try approximate search of genus name;

(4) if no results are found, ask the user to try another request.

The exact match function only returns entries which contain the exact keyword(s) in the query. The approximate search will allow partial matching. For instance, query "admi" will yield the genus records of "Adminiculoria" and "Adminixtella", as shown in Figure 8. When full-text search is implemented, the records are ranked according to relevance. For more details about relevance, please refer to the discussion on advanced search in next section.

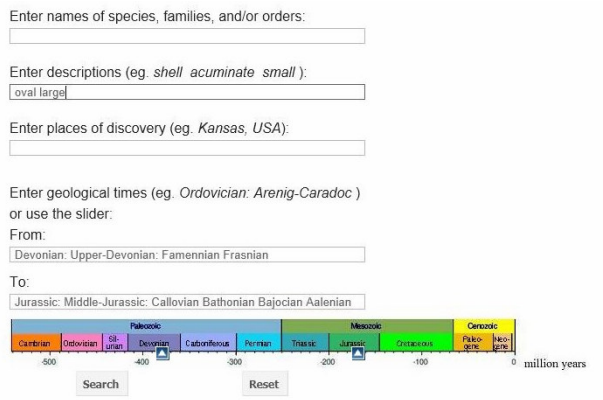

Figure 7: The webpage for advanced search.

5.1.2 Advanced search

Advanced search allows users to send structured queries to IPKB. Users could provide predicates on names, morphological features (descriptions), geological times, and/or geographic locations.

The fossils are dated from lower Cambrian (540 million years ago) to present. In order to help users select a geological time range, we designed a scrollbar with two pointers. In our database, the geological time of fossils are encoded into integers. For a user query, the provided starting and ending times are encoded in the same way, and sent to the database as predicates on a integer field. Moreover, we also send the corresponding geological *period* and *epochs* (if any) of the starting and ending points as keywords in the search query as well. This will yield records that contain geological period and epochs in their morphological descriptions (not in the geological stratigraphy field). This is also useful when people search in a small time range, and helps to provide better ranking.

We have employed the natural language full-text search function in MySQL to support full-text search and similarity-based ranking. In the database, the relevant columns are included in a FULLTEXT index. We have assigned different weights w to different columns to reflect the "importance" of the information in the search context (e.g. genus names are more important than descriptions): 3.0 for taxnon names, 1.5 for geological and geographical distribution, and 1.0 for morphological features.

MySQL implements a model similar to TF-IDF in the full-text search function. The *significance* of a term is calculated as:

$$s = \frac{log(dtf) + 1}{\sum dtf} \times \frac{U}{1 + 0.0115U} \times log(\frac{N - nf}{nf}) \qquad (4)$$

where dtf is the number of times the word appears in a document (i.e. a database record); $\sum dtf$ is the sum of $(log(dtf) + 1)$'s for all words in the same document. U is the number of unique words in the document (see [30] for more details). N is the total number of documents. nf is the number of documents that contain the term. The combined relevance of a term and a document is:

$$R = w \times s \times qf \qquad (5)$$

where qf is the frequency of the term in the query, and in most cases it is 1. It should also be mentioned that MySQL natural language full-text search ignores words shorter than 4 letters by default. In addition, there is a stopword list which excludes words with too high frequencies.

5.1.3 Browsing

The *Browse* function is $defacto$ search by names, too. It provides a wizard for users to explore the hierarchical structure of fossil categories. Users see all the orders of brachiopods when the web page is loaded. When an order is clicked on, its subordinate suborders appear, and so on. In this manner, a user does not need to type in anything to locate the genus he wants to view. People can browse all the category names, even if they are not willing to view any specific record.

5.2 Genus record displaying

In designing the interface to show genus records, we consider the following factors: it must be easy to view and users should be able to perceive the relevant information immediately; there should be some links that allow users to check related information. The results of quick search and advanced search are essentially delivered in the same format. Up to 10 records (genera) are shown on each page. As shown in Figure 8, for each genus record, there are representative photos, if possible; and there are a few lines of textual information. On the top of the text, the name of the genus is shown. Under the genus name, there are names of higher taxa (i.e. family, order etc.) governing the genus. Users can click on a taxon name to view a general description for that taxon. Further down other information can be found, including morphological description, geological and geographic distribution.

In the results of advanced search, the keywords will be highlighted. This feature applies to keywords in geological and geographic distributions, as well as in morphological descriptions. For geological time in particular, the terms marking the beginning and the end of the searched time range will be highlighted. Many of the terms in geolog-

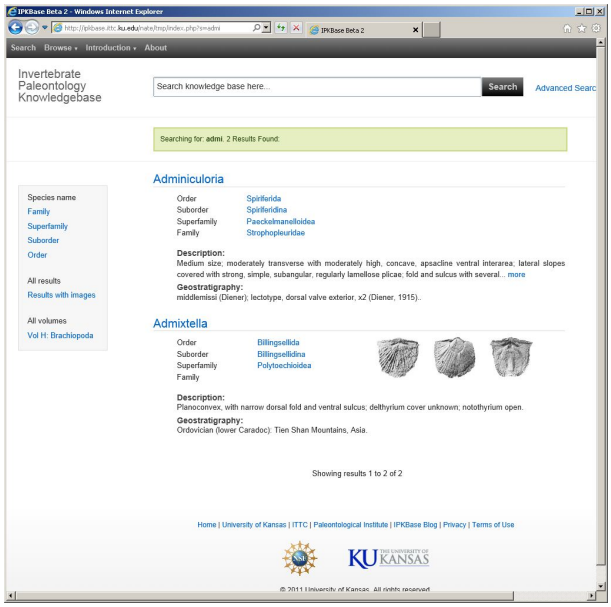

Figure 8: The result of quick search, using "admi" as keyword.

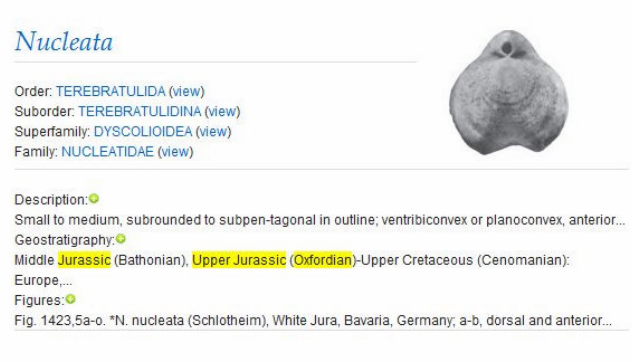

Figure 9: The result of advanced search.

ical time contain modifiers such as "upper", "middle" and "lower". When users are searching for "upper Jurassic", they certainly do not want to see records with a lot of "upper"s but no "Jurassic". At the same time, nevertheless, they do want the system to return and highlight records with "Jurassic" even though the modifier is absent. We took cautious treatments for this situation. As we can see from the result, when "Upper Jurassic" is searched for, the strings of "Upper Jurassic" or "Jurassic" are highlighted, but "Upper Cretaceous" and alike are not.

Once a user clicks on the genus name to view the details, a new page containing information of that specific genus is loaded. All the descriptions and photos of this genus display. Moreover, geographical distribution is crucial information in paleontology. Although it is nontrivial to visualize geological stratigraphy, we have employed Google Earth API [2] to visualize the sites of discovery of the fossils. The Google Earth API only accepts exact longitudes and latitudes to mark the locations. Therefore we first use the Google Map

[2]http://code.google.com/apis/earth/

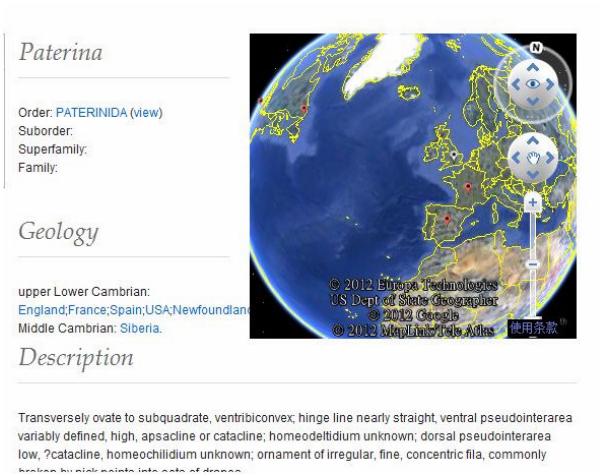

Figure 10: View information of a genus. The photos are on the left side. Google Earth displays the geographical distribution.

API to convert the name of locations to coordinates, and then pass it to the Google Earth API. If a certain genus is found in more than one location, multiple markers will show up on Google Earth as well. We also employ the Google Image Search API [3] to embed results of Google Image Search in our system. Such plug-ins not only help with data interpretation, but also significantly improves user experiences.

6. RELATED WORK

In 1990s, digital libraries have been introduced to provide fast and flexible online access to digitalized information repositories (e.g. [7, 14, 6]). Traditional libraries, which house and provide access to collections of books, have made metadata (e.g. catalogs) available and searchable online (e.g. The Library of Congress online catalogs [4]). However, access to full content is usually not available over the web since most traditional media are yet to be digitized. Meanwhile, these projects only propose to provide digital access to traditional media, but do not make further explorations on the digitized contents.

We also have digital libraries that focus on a specialized area. For example, bibliographic databases (e.g. PsycINFO, PubMed, arXiv) are repositories for academic publications. On the other hand, Digital Himalaya [29] is an anthropology library to digitize, organize and publicize multimedia ethnographic materials from the Himalaya region. The goal is to preserve valuable information in digital format, and provide searchable access to the research community as well as general public.

A number of digital paleontological databases have sprung up over the last decade or so, the best known being the *Paleobiology Database* (PBDB) [5]. This NSF-funded database is the foremost online paleontological database. It differs from the *Treatise* and IPKB in a number of important ways. First, it is populated in a wiki-style process: large numbers of contributors, from a variety of backgrounds, input

[3]http://code.google.com/apis/imagesearch/
[4]http://catalog.loc.gov/
[5]PBDB: http://paleodb.org

data with varying degrees of accuracy. In many cases, it is student labor that inputs the data. As a consequence, the quality is highly variable. Second, the data is predominantly text-based, and thus similar to any other queryable dataset; there are few fossil images, and these are random additions by particular workers. In contrast, *Treatise* data is collected, peer-reviewed and verified by the experts in the field, and images of type specimens are an integral feature.

Meanwhile, numerous other paleontological databases can be found on the internet, with each one set up by generally a single worker to reflect their own interests. These vary greatly in scope and utility. For example, the Florissant Fossil Database [6] is essentially a collection of high-quality images of fossils from this National Park.

7. CONCLUSION AND FUTURE WORK

In this paper, we report our initial accomplishments in IPKB, a project aiming at digitalization, utilizing and sharing of the rich information from the *Treatise on Invertebrate Paleontology*, which is the most authoritative compilation of data on invertebrate fossils. In the project, we first extract raw data from the Treatise, and processes textual and image data objects separately. We have designed an ontology to capture fossil information, and explored ways to segment images, recognize labels, and link fossil (genera) descriptions with images. Finally, a web interface is designed to present the information, and provide easy browsing and searching functions. Google Earth and Google Image Search APIs are employed to help with presentation and improve user experiences.

IPKB provides a solid foundation for future discoveries based on the rich information repository of invertebrate paleontology data. Our next step is to extend the IPKB framework to provide the scientists with a knowledge discovery platform, which hosts complex data analysis, content-based retrieval, modeling, data mining, image processing and understanding, animation and visualization functions. We are also implementing a more illustrative and information-rich web portal for the general public. Meanwhile, to improve the accessibility and usability of IPKB, we are implementing a web services interface, and designing apps for mobile devices. For instance, paleontologists working in the field will be able to browse and search IPKB on hand-held devices, and submit their fossil images for identification and search.

8. ACKNOWLEDGMENTS

This work is supported by the US National Science Foundation Award CDI-1028098.

The authors would like to thank reviewers for their feedback, and thank Jill Hardesty from the Department of Geology, University of Kansas for her advice and help.

9. REFERENCES

[1] L. J. Barker. Science teachers' use of online resources and the digital library for earth system education. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '09, pages 1–10, 2009.

[2] L. Bolelli, X. Lu, Y. Liu, A. Jaiswal, K. Bai, I. Councill, P. Mitra, J. Wang, K. Mueller, J. Kubicki, B. Garrison, B. J., and C. Giles. Chemxseer: A chemistry web portal for scientific literature and datasets. In *Open Repositories Conference*, 2007.

[3] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986.

[4] C.-c. Chen. Global memory net: New collaboration, new activities and new potentials. In Z. Chen, H. Chen, Q. Miao, Y. Fu, E. Fox, and E.-p. Lim, editors, *Digital Libraries: International Collaboration and Cross-Fertilization*, volume 3334 of *Lecture Notes in Computer Science*, pages 543–575. Springer Berlin / Heidelberg, 2005.

[5] H.-H. Chen, L. Gou, X. Zhang, and C. L. Giles. Collabseer: a search engine for collaboration discovery. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11, pages 231–240, 2011.

[6] Y. Choi and E. Rasmussen. What do digital librarians do. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '06, pages 187–188, 2006.

[7] E. A. Fox, R. M. Akscyn, R. K. Furuta, and J. J. Leggett. Digital libraries. *Commun. ACM*, 38(4):22–28, 1995.

[8] C. L. Giles, K. D. Bollacker, and S. Lawrence. Citeseer: an automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, DL '98, pages 89–98, 1998.

[9] C. L. Giles and I. G. Councill. Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51), 2004.

[10] R. Gonzalez, R. E. Woods, and S. L. Eddins. *Digital Image Processing Using MATLAB*. Gatesmark Publishing, 2009.

[11] H. Hsieh, B. Draxler, N. Dudley, J. Cremer, L. Haldeman, D. Nguyen, P. Likarish, and J. Winet. Facilitating content creation and content research in building the city of lit digital library. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11, pages 145–148, New York, NY, USA, 2011. ACM.

[12] M. Jones, E. Thom, D. Bainbridge, and D. Frohlich. Mobility, digital libraries and a rural indian village. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '09, pages 309–312, 2009.

[13] D. J. Kennard, W. B. Lund, and B. S. Morse. Improving historical research by linking digital library information to a global genealogical database. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '09, pages 255–258, 2009.

[14] D. M. Levy and C. C. Marshall. Going digital: a look at assumptions underlying digital libraries. *Commun. ACM*, 38:77–84, April 1995.

[15] H. Li, I. Councill, W.-C. Lee, and C. L. Giles. Citeseerx: an architecture and web service design for an academic document search engine. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 883–884, 2006.

[16] N. Li, L. Zhu, P. Mitra, K. Mueller, E. Poweleit, and

[6] http://planning.nps.gov/flfo

C. L. Giles. orechem chemxseer: a semantic digital library for chemistry. In *Proceedings of the 10th annual joint conference on Digital libraries*, JCDL '10, pages 245–254, 2010.

[17] F. Meyer. Topographic distance and watershed lines. *Signal Processing*, pages 113–125, July 1994.

[18] A. Mikeal, J. Creel, A. Maslov, S. Phillips, J. Leggett, and M. McFarland. Large-scale etd repositories: a case study of a digital library application. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '09, pages 135–144, New York, NY, USA, 2009. ACM.

[19] P. Mitra, C. L. Giles, B. Sun, and Y. Liu. Chemxseer: a digital library and data repository for chemical kinetics. In *Proceedings of the ACM first workshop on CyberInfrastructure: information management in eScience*, CIMS '07, pages 7–10, 2007.

[20] R. C. Moore and other editors. *Treatise on Invertebrate Paleontology: Part H.* Geological Society of America and the University of Kansas Press, http://www.ku.edu/ paleo/treatise.html, 1953-2006.

[21] F. Morchen, M. Dejori, D. Fradkin, J. Etienne, B. Wachmann, and M. Bundschus. Anticipating annotations and emerging trends in biomedical literature. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 954–962, 2008.

[22] T. Nagatsuka and C.-c. Chen. Global memory offers new innovative access to tsurumi's old japanese waka poems and tales, and maps. In E. Fox, E. Neuhold, P. Premsmit, and V. Wuwongse, editors, *Digital Libraries: Implementing Strategies and Sharing Experiences*, volume 3815 of *Lecture Notes in Computer Science*, pages 149–157. Springer Berlin / Heidelberg, 2005.

[23] C. Neuhaus and H. D. Daniel. Data sources for performing citation analysis: an overview. In *Journal of Documentation*, volume 64, 2008.

[24] J. R. Parker. *Algorithms for Image Processing and Computer Vision*. John Wiley & Sons, Inc., New York, 1997.

[25] Y. Petinot, P. B. Teregowda, H. Han, C. L. Giles, S. Lawrence, A. Rangaswamy, and N. Pal. ebizsearch: an oai-compliant digital library for ebusiness. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, JCDL '03, pages 199–209, Washington, DC, USA, 2003. IEEE Computer Society.

[26] R. Sanderson, B. Albritton, R. Schwemmer, and H. Van de Sompel. Sharedcanvas: a collaborative model for medieval manuscript layout dissemination. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11, pages 175–184, 2011.

[27] J. Serra. *Analysis and Mathematical Morphology*. Academic Press, London, 1982.

[28] B. Shaparenko and T. Joachims. Information genealogy: uncovering the flow of ideas in non-hyperlinked document databases. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 619–628, 2007.

[29] S. Shneiderman, M. Turin, and the Digital Himalaya Project Team. Digital himalaya: an ethnographic archive in the digital age. *European Bulletin of Himalayan Research*, 20(1):136–141, 2002.

[30] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 21–29. ACM, 1996.

[31] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 306–315, 2004.

[32] L. Vincent. Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms. *IEEE Transactions on Image Processing*, 2(2):176–201, April 1993.

[33] J. Wang, J. Li, and G. Wiederhold. Simplicity: semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947 –963, sep 2001.

[34] C. Yang, B. Wei, J. Wu, Y. Zhang, and L. Zhang. Cares: a ranking-oriented cadal recommender system. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '09, pages 203–212, 2009.