A Standardized Aerosol Repository System for Knowledge Services and Crowdsourcing

1st Xiaoyi Huang
State Key Laboratory of CAD&CG
Zhejiang University
Hangzhou, China
choyeehuang@gmail.com

2nd Jun Zheng
State Key Laboratory of CAD&CG
Zhejiang University
Hangzhou, China
zzderry@zju.edu.cn

3rd Kejun Zhang
State Key Laboratory of CAD&CG
Zhejiang University
Hangzhou, China
zhangkejun@zju.edu.cn

4th Dong Liu, Jie Chen
State Key Laboratory of Modern Optical Instrumentation
Zhejiang University
Hangzhou, China
liudongopt@zju.edu.cn, jieziwaiwai@126.com

5th Lei Bi
Department of Atmospheric Sciences
Zhejiang University
Hangzhou, China
bilei@zju.edu.cn

***Abstract*—A web-based Standardized Aerosol Repository System (SARS)[1] was developed to facilitate relevant associated with atmospheric aerosols. A new method was proposed to obtain the digitized data from literatures, followed by data processing, storage and services. Moreover, a repository was set up to support an online aerosol-literature-knowledge-service module which provides special services including "keyword query, thesis topic classification, duplicate check, paper evaluation and data visualization". Furthermore, crowdsourcing, a distributed problem solving mechanism was proposed that the users can submit formatted data to update and expand the repository. At the same time, the web page of crowdsourcing submission also optimizes the user interaction such as adding thesaurus and drop-down options to ensure the users' experience. The SARS is expected to provide a convenient platform for sharing aerosol data and maximize the data usage in the atmospheric aerosol research community.**

***Keywords*—aerosol, standardized, standardization format strategy, repository, knowledge service, data visualization, crowdsourcing**

I. Introduction

An aerosol is a colloid of fine solid particles or liquid droplets, in air or another gas. Aerosols can be natural or artificial. For example, fog, forest exudate and geyser steam are natural aerosols, artificial aerosols include haze, dust, particulate air pollutants, smoke and so on. Now, aerosols play an important role in regulating climate change [1]. In the context of global climate change, this regulation is global. In order to conduct aerosol-related research, global aerosol observation data are urgently needed. Currently, the data is still scattered in a large number of documents, open source repositories and other types of storage medias. For readers who need data, it is cumbersome to transcribe and read a large amount of literature data between various data platforms. For ordinary readers, it is necessary to evaluate literatures' authority in order to acquire such statistical knowledge. The acquired knowledge is also limited by the region and the exact year that the author wants to display. When facing a large number of intricate data, if we rely on traditional literature collection methods, then it would be difficult to extract aerosol data in massive documents quickly and efficiently.

- **Repository.** Since the concept of institutional repository was proposed in the 1990s, theoretical and practical research on repository has never stopped [2]. Theoretical research mainly focuses on the relationship between the basic concepts of repository and open access, as well as the impact on traditional academic exchanges. Practical research mainly focuses on different academic research institutions to build a repository according to their own needs and goals, and provides institutional members with service [3]–[5]. At present, many atmospheric research institutions and observatories have independently developed a repository for storing and sharing observed or simulative data of their own. It then motivates our objective which is to introduce our solution for this issue: cross-media data acquisition and a new open access method.
- **Standardization format.** At present, storage and applications for standardized data have emerged in atmospheric field and other fields such as clinical medicine and social government [6]–[9]. A standardized format with a rich metadata description, which stands for a new culture in data archiving: highly standardized, easy to use and free data available for the whole research community [6]. In doing so we hope to encourage the use of its data, but also to encourage others to adopt the methods, quality control procedures, and data description methods we developed, to expand the archive for use by future generations.
- **Crowdsourcing.** Since Jeff Howe (2006) proposed the concept of crowdsourcing, there has been a significant increase in attempts at crowd sourcing research and practice [10], [11]. Crowdsourcing is considered to be: a problem-solving

[1]SARS-Standardized Aerosol Repository System.

978-1-7281-4528-0/19/$31.00 ©2019 IEEE
DOI 10.1109/DSC.2019.00043

TABLE I
THE GAP

	Professor	Teacher	Student	Laypeople	The Gap
Requirements	Access to domain information and data				Data is distributed across a variety of platforms
	Timely access to the latest literature and data			NA	There is a delay in the publication of papers and articles
	Promote academic influence		NA	NA	1) The low citation rate of the paper makes the data value difficult to be discovered, 2) The problems of data diversification exist in the personal or institution homepage
	NA	NA	Promote academic ability	NA	Lack of comparison with excellent papers
	Grasp the direction of the field	NA	NA	NA	Lack of all-sided rofessional data
	NA	NA	NA	Learn about the basics	1) Lack of visualization to help learning, 2) Lack of knowledge assistance and service

tool [12], an online distributed problem-solving model and production model [13], a learning paradigm that forms an open collaboration [14], a resource for new product development and so on [15], [16]. Our work is to solve the problem of repository expansion through crowdsourcing-distributed problem solving mechanism.

Inspired by these researches, we have built a standardized repository and developed a comprehensive literature knowledge website system dedicated to the aerosol field. In addition, we use the crowd-sourcing method to assist in the update and expansion of the repository. We hope that the research methods of this project will be able to enlighten the relevant scholars all over the world to solve such review questions, and have important significance for promoting the progress and maturity of relevant methods.

The rest of this paper is organized as follows: Section II formulates the problems and proposes a standardization format strategy(SFS) [2] to build aerosol repository. In Section III, Web-system functions are illustrated for the knowledge services. Section IV proposes a format of crowd-sourcing submission. At last, conclusions are stated in Section V.

II. AEROSOL REPOSITORY

The construction of the institutional repository requires good management, continuous planning, and coordination of different stakeholders [17].

In this work, we planed and coordinated different stakeholders in advance, including academic experts, scientific research personnel, college students and the general public. We analyzed the gaps between stakeholders' realities and requirements. See Table 1 for details.

A. Standardization Format Strategy(SFS)

We proposed a standardization format strategy(SFS): the system stores standardized core data so that the users can find key information. Therefore, we determine the data information and properties through user studies.

[2]SFS: Standardization format strategy.

This work was cooperated with the State Key Laboratory of Modern Optical Instrumentation. After discussing with the graduate students of the cooperative lab, we designed the user experiments.

- **Questionnaire.** We invited 30 (10 females and 20 males) college students of aerosol field from the age of 20 to 25 to participate in the questionnaire experiment. The questionnaire consisted of five parts: literature publication information, aerosol basic information, optical parameter information, space-time information, particle microphysical properties, It is necessary to give these five parts an important attribute that they think may contain a piece of literature related to aerosol data. We compiled and merged the attributes of the 30 questionnaires collected, and obtained preliminary results. The questionnaire results contain preliminary 33 attributes that will roughly cover the important information needed.
- **User interviews.** In the follow-up, we interviewed experts and scholars in the field of aerosol, taking into account the importance of the data and the clarity of the records. We found that in the literature, there is numerical observation or simulation of aerosol with low probability. The detail data values are usually represented in the form of images(a picture may have hundreds of data), while the text is used to describe the statistic data(mean value, deviation) and the external information(aerosol type, methods, etc.). At last, the data attribute is determined to be 15 types, including 5 literature attributes, 5 text data attributes and 5 image data attributes. See Table 2 for details.

B. Data Acquisition

- **Literature download and abstract extraction.** Our team members manually downloaded 10,966 aerosol documents and crawled 37,302 abstracts legally.
- **Data extraction based on image recognition.** We proposed a program with four steps: 1 Extract the literature

TABLE II
DATA ATTRIBUTE.

Attribute Type	Attribute	Content
Document	Journal	75 journal options
	DOI	Unique Number of article
	Paper Name	Full name of the document
	Author	Corresponding author
	Unit	Company or school
Text	Key Physics Parameters	Lidar ratio, depolarization ratio, backscatter coefficient, extinction coefficient, optical depth, depolarization spectral ratio, color ratio, angstrom exponents
	Aerosol Type	Dust, smoke, clean continental, polluted continental, polluted dust, clean marine
	Method	Observation, retrial
	Mean value	Optical parameter average, decimal form
	Standard deviation	Optical parameter standard deviation, decimal form
Image	value	Concrete numerical value
	Wavelength	355nm, 440nm, 532nm, 645nm, 780nm, 870nm, 1020nm, 1540nm
	Height	Kilometer
	Time	Year, month, day, hour, minute, second
	Location	Country, province, city, the country specific to the school, latitude and longitude and other information

publishing data and the text statistical information by reading literature manually. 2 Find and intercept the data analysis images corresponding to the text statistical information. 3 Perform data extraction. 4 Save the text statistical information and the image data into the normalized database.

During the review and analysis of the aerosol literature, we found that aerosol data is usually stored in scatter plots, column plots, and line graphs. We need to process the image first and then use different image recognition methods for different image types to extract the aerosol data stored in.

In order to accurately identify the data in the process of image recognition, it is first necessary to perform preprocessing of the image. This section mainly performs the steps of image size adjustment and image graying. We first intercepted the image coordinate axis region in the aerosol literature, and used the method of scaling the image in TesnsorFlow[3], and then grayed out the image to reduce the influence of irrelevant factors on image recognition.

The detectability of effective information in aerosol images is enhanced by the image preprocessing, and then it can be used to improve the accuracy of image recognition and to extract the radar ratio and related data.

Scatter diagram. In the procedure of extraction, we first applied the Hough Transform [18] algorithm to identify the data points. The Hough Transform recognizes circular data points mainly based on the following principles: (a) detecting the edges of the input image; (b) for an edge line, assuming it is a circle and traversing all points on the circumference, taking any three points as a combination. The intersection of the lines determines the candidate center. After the center position is obtained, its pixel coordinates can be outputted.

In the aerosol literature image, in addition to the detection of circular data points, it may involve the detection of other types of data points, such as squares, triangles, and the like. At this time, all the straight lines in the coordinate axis can be identified by using the Hough Transform to detect, and a closed area is formed among the straight lines, which is the data point area.

At last, it is also necessary to determine the specific pixel coordinates of the data point by the center point detection. Usually, the center point of a square or a triangle is its center of gravity. Therefore, the starting and ending coordinates of the straight line can be detected by the Hough Transform, and the coordinates of each vertex in the area can be determined. At last, we can receive the result of the corresponding data point pixel coordinates.

Histogram. The data extraction of the histogram is relatively simple. The basic idea is to traverse all the pixels and use the edge detection method to identify the horizontal line. For the identified horizontal line, when there is a large blank area above it, it is the effective horizontal line, which the middle point coordinate is the pixel coordinate of the required data.

Among them, the edge detection algorithm is used in the process of horizontal line recognition. Edge detection is a way of imitating the human visual system in computer graphics, the principle is to identify the intersection between the elements in the image and the background of the image by an algorithm. In edge detection, the edge point is the point where the gray level jumps in the image is more intense. By the calculation of the gradient image, the brighter part in the gradient image is the simple edge part. In general, edge detection requires the steps of eliminating noise, enhancing boundary contours, and detecting edges, then using the Canny algorithm [19] in OpenCV[4] for edge detection. The procedure is shown in Algorithm 1.

Line chart. The data extraction method of the line graph is similar to the histogram. By using the image recognition technology, traversing all the pixels on the line graph,

[3]https://tensorflow.google.com

[4]https://opencv.org/

Algorithm 1 Abstract data from Histogram

$img_{gray} \leftarrow$ grayscale the image
$img_{gray} \leftarrow$ reduce the noise by the 3×3 core
$lines \leftarrow$ compute based on the Canny

identifying the abscissa of all valid data points, and finding the median, we can receive the result of all the pixel coordinates in the line chart.

The data extraction method of the line chart is similar to the histogram. In the line chart of the aerosol literature, the radar ratio is usually compared under two different parameters. In this case, the data extraction process of the line chart involves the extraction of two polyline data, which needs to be distinguished based on the characteristics of the two lines. Usually, polylines that represent different meanings are distinguished by color or virtual solid line. When distinguishing by color, data extraction is needed based on different RGB color values from the original image before image graying. When the pixel points meet the color value requirements, they output the effective pixel points. When the polyline is distinguished by a virtual solid line, the data of the solid line portion may be first extracted. When the polyline is satisfied, the pixel points are extracted, and the pixel points in the dashed line are extracted. The procedure is shown in Algorithm 2.

Algorithm 2 Abstract data from Line diagram

step $\leftarrow \left\lceil \frac{image\ height}{600} \right\rceil$
index $\leftarrow 0$
while index $<$ image height **do**
  $(x, y) \leftarrow image[index]$
  if match the RGB **then**
    $v_1 \leftarrow$ compute
    if $v_1 > min_scale$ **then**
      $v \leftarrow$ save v_1
    end if
  end if
  $v_{medium} \leftarrow$ get medium from v
  if v_{medium} *is* 0 *or* $v_{medium} > \max(image)$ **then**
    $v \leftarrow$ save v_{medium}
  end if
end while

Data conversion . The data we extracted is the unconverted pixel point pixel coordinates. If we want to get really accurate data, we need to convert the pixel coordinates into coordinate coordinates. The pixel coordinates of the image represent the position of the pixel in the image [43]. In the aerosol-related literature, the quadrant usually starts from the lower left corner. In the image coordinate system, the Cartesian coordinate system in pixels is usually established with the upper left corner of the image as the origin. Therefore, the pixel coordinate to coordinate coordinate transformation is performed. When the object limit conversion is required, the calculation is performed. Assume that the pixel coordinates of the pixel are represented by (a, b), X_1, X_2, Y_1, and Y_2 represent the maximum and minimum values of the horizontal and vertical axis, respectively. Pixel coordinates can be derived by (1) and (2).

$$x = \frac{X_1 + (X_2 - X_1) \times b}{width} \tag{1}$$

$$y = \frac{Y_1 + (Y_2 - Y_1) \times (height - a + 1)}{height} \tag{2}$$

After the data conversion, most of the image data has been normally stored into the database. However, some images may have a specified implication. For example, the abscissa of the column chart may be a parameter that represents a specific meaning(time, date, etc.). It is difficult to extract data by image recognition, so it is necessary to supplement the manual process to ensure the accuracy of the data.

C. Database

The SARS is constructed using free, open-source tools.The database is a standard MySQL database, managed using Python and Django[5].

The current literature data of the aerosol repository cover 75 core aerosol-related journals, with 10966 aerosol-related literature data, 37302 abstracts and 10623 format data. It is worth mentioning that 10623 format data are recorded from 32 literatures, 55 images and contain two aerosol parameters-Lidar Ratio and Depolarization Ratio. The literatures cover China, South Korea, Japan, India, Africa, the United States and other countries and continents. In China, the data cover Beijing, Chongqing, Xinjiang, Jiangsu Province, Anhui Province, Taiwan Province, Guangdong Province, Hubei Province, Guizhou Province, Gansu Province, and Shanxi Province, see Figure 1.

In addition, the system backstage will automatically detect the recent literature update of each academic platform, which promptly notifies administrators to update the data(download full-text, select images, etc.). Meanwhile, the system will download and extract abstracts from web legally.

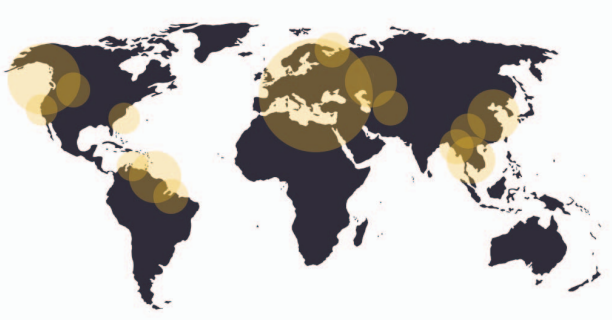


Fig. 1. Data distribution on world map.

[5]https://www.djangoproject.com/

III. Knowledge Services

When targeting the user groups, we need to understand the specific usage requirement and usage preferences of the users for determining the system service functions.

60 participants (45 from aerosol field, 15 from non-professional fields) are invited to join in user interview. Interviews contained 1. What approaches are available to determine the direction of the paper before writing the paper; 2. How to obtain the data; 3. What form do you wish to evaluate your paper after completion? etc. In this section, we initially acquired functional points. Then, we proposed different aerosol-related tasks and record the requirements through the problem-solving discussion between the participants, and let the participants score the importance of the function points we selected before. At last, 6 functions of the system are determined: Keyword Query, Topic Classification, Duplicate Check, Paper Evaluation, Visualization, Crowdsourcing.

A. Functional Service

- **Keyword Query.** After entering the keyword, the system will search all the documents in the database, and output all the document information containing the keywords in descending order, click to view the documents, and automatically locate the fields of the searched keywords.
- **Topic Classification.** Users can upload documents, the system will accurately locate the document level and category, and inform the document subject classification. According to the similarity in the topic, the relevant literature is recommended for the user. The user can click to view the recommended documents.
- **Duplicate Check.** When users upload their own document, the system will perform data preprocessing such as word segmentation and deactivation, and convert it into a text file, compare the similarity with other documents in the library, and feedback the user and upload the document. Similar published literature, users can click to view the literature, the system will automatically locate similar fields in the literature, which is convenient for users to view.
- **Paper Evaluation.** In the same way, the uploaded PDF document will be considered as the target files, then the system will compare them with the published papers or the more influential papers in the library. At last, the system will give the feedback on the document and recommend the corresponding excellent papers.

Currently, SARS has been developed to offer 5 functions except "Paper Evaluation" to explore new data or getting all necessary information about them. Besides, SARS is accessible through one data portal:http://39.108.69.117/WDCLab/hazeInfoExtract.

B. Visualization

The aerosol data visualization is mainly served by the aerosol database. Through visual display, information can be obtained more intuitively, and it can better serve the research work of scholars in the field of atmospheric aerosols.

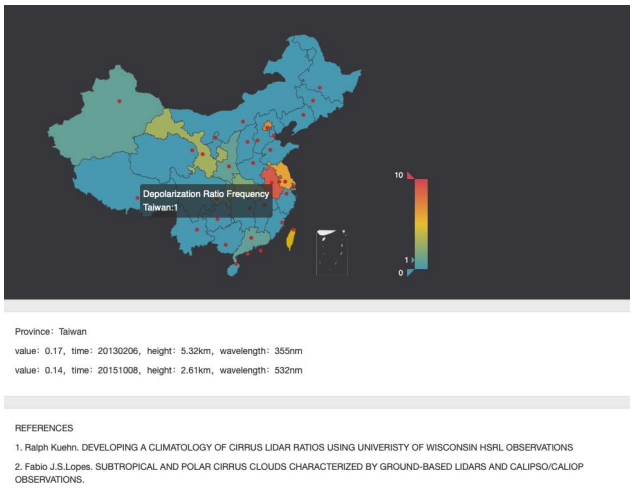


Fig. 2. The visualization of Depolarization Ratio in Taiwan, China.

In figure 2, taking the map of China as an example, after choosing an aerosol parameter, the map shows the distribution of the data collected from the paper, with a red color representing more observations. When clicking on the province area of the map, the number of observations of the aerosol parameters corresponding to the province will be displayed, and the specific values, time, wavelength, height and source of the data will be displayed at the bottom of the page.

In the visual analysis function, it is necessary to select appropriate charts for visual analysis for different data. Based on this, we analyzed the advantages and disadvantages of various types of charts in the demand analysis stage, and summarized the applicable scenarios. It aims to provide a theoretical basis of selecting the appropriate charts for data analysis in the visualization stage. Among the common chart types, there are scatter plots, column charts, line graphs, pie charts, etc., which can be selected for visual analysis. Each type of chart has its own advantages and applicable scenarios. Therefore, this section has carried out summary of chart type characteristics and applicable scenarios:

- **Scatter plot**: In the case of a large amount of data, the scatter plot works well in dealing with distribution of values and data clustering. Therefore, the scatter plot is suitable for displaying the relationship between the values in several data series, which it is easy to observe whether there is an association between the variables.
- **Column chart**: The advantage of the column chart is that it can use the height of the column to visually reflect the difference of the data, but it is only suitable for small and medium-sized datasets. Therefore, column chart is usually used to display data changes over time or show comparisons between items.
- **Line chart**: The advantage of the line chart is that it easily reflects the trend of data, and is suitable for comparison of two-dimensional big data sets and multiple two-dimensional data sets.

- **Pie chart**: It is easy to display the proportion of data. The disadvantage is that the naked eye is far from sensitive to the size of the area. It is suitable for displaying the ratio of the size of each item to the sum of the items without requiring accurate data.
- **Radar chart**: It is applicable to multi-dimensional data, and each dimension must be sorted. It is mainly used to analyze the change of data and its good and bad trend.

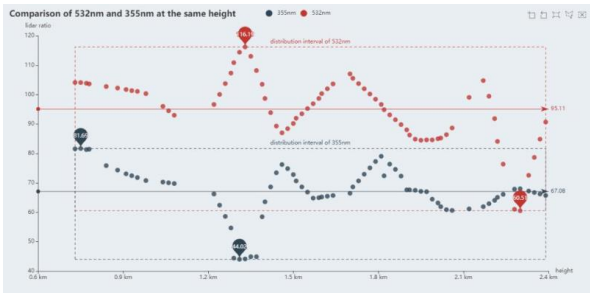


Fig. 3. Radar ratio and observation height distribution at different wavelengths in 2005.

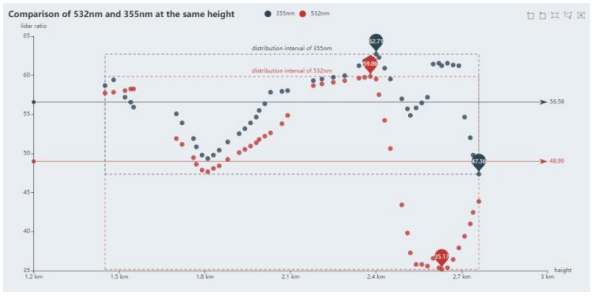


Fig. 4. Radar ratio and observation height distribution at different wavelengths in 2004.

Figure 3 and Figure 4 show the visualization results based on observations in the same season between 2004 and 2005 in Korea. Generally, at the same height, the lidar ratio, which is obtained from different observation wavelengths at same height in 2004 and 2005, shows different change trends. Specifically, in 2004, the lidar ratio at the 355nm observation wavelength is generally greater than the lidar ratio at the 532 nm observation wavelength, while in 2005 the data trend shows the opposite. Considering the reason of such result, besides the different accuracy of lidar observation. it is also related to the more diversity of aerosol types and human activity at 532 wavelength.

In Figure 3, the curve is relatively consistent between 0.75km and 1.2km, while the trend of the curve shows the opposite between 1.2km and 2.4km. In Figure 4, the curve of lidar ratio at 532nm and 355nm observation shows the same trend between 1.5km and 2.5km, while the lidar ratio appears opposite situation between 2.5km-2.8km. The same change trend of lidar ratio represents the aerosol species are generally consistent under different wavelength observations while the opposite change trend indicates that the relevant parameters of aerosol particles obtained by different observation wavelengths are different at this height.

Based on the analysis of the above chart types, in the visual analysis stage, the appropriate chart type can be used for visual analysis, which provides a theoretical basis for the visualization of aerosol data.

IV. Crowdsourcing

Since each discipline has its own data sharing practices, data sharing differs across disciplines. Furthermore, even in the same discipline, data sharing behaviors can vary because of their technological infrastructures, institutional set-ups, and individual expectations [20]. Therefore, it is very important to understand how diverse data sharing factors.

Extensibility is one of the major strengths of SARS. In this work, we integrated three factors including web system, aerosol information attributes (Table II) and benefits for repository members (Table I), and proposed crowd sourcing to solve data sharing and data extension.

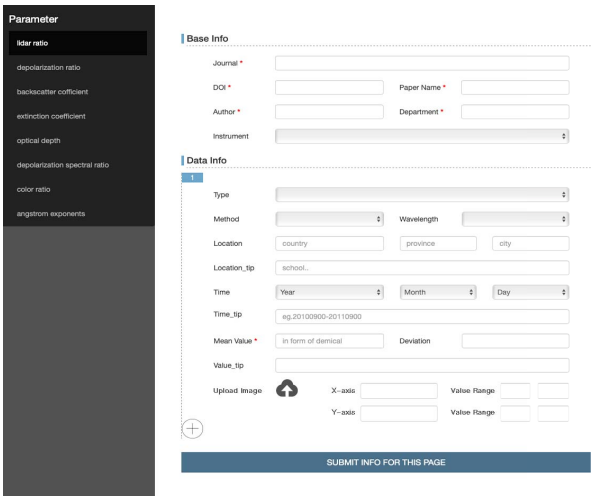


Fig. 5. Main Interface of crowd sourcing.

A. Organizational method of crowd sourcing

There are 2 organizational methods of data submission have been proposed: invitation and voluntariness.

By sending e-mails, we invite experts and scholars who publish literature in the aerosol field conferences and journals to submit data on the crowd sourcing submission page of the system, which will also help to improve the accuracy of the system's "paper evaluation" function.

In addition, we encourage contributions of individual data. Individuals or labs interested in contributing should consult with the authors for advice about data formatting and upload.

B. Interactive mode of submission

As is shown in Figure 5, there are 8 aerosol parameters(lidar ratio, depolarization ratio, backscatter coefficient, extinction coefficient, optical depth, depolarization spectral ratio, color

ratio, angstrom exponents) in the navigation bar on the crowdsourcing page, which guide the users to fill the corresponding information(15 data attributes in section II) in the boxes.

In addition, users can contribute new image data to the site and connect to it directly using a function button for image uploading. Currently, the upload only supports image data extraction of scatter plots and line graphs. After uploading the image, enter the attributes corresponding to the horizontal and vertical coordinates of the image and the range of values. Finally, click the "Submit" button to import the relevant data into the aerosol database.

In terms of user experience, in order to save user data filling time, we set the thesaurus function, and the drop-down options in multiple options such as Journal, Instrument, Type, and province cities, users can click to complete. In terms of data security, the system will automatically save the data to avoid data loss caused by web page error shutdown, network disconnection and other reasons.

V. Conclusions

In this paper, we have presented SARS, an open repository for aerosol data from kinds of data platforms and literatures. Extensive atmospheric measurement data are gathered every day by various devices and a large amount of papers are published by institutions. Thus it will become even more challenging and important in the future to standardize these data and make them accessible for the whole research community. In addition, thermodynamic diagram is also a common expressions in aerosol field. Therefore, further observational datasets will be integrated into the steadily growing SARS archive.

Future collaborations, data interoperability and meta data exchange with other worldwide existing earth system databases will further increase the outreach of the available SARS database, but at the same time, the SARS needs to extend itself through its advantages and unique features compared to the numerous other new and existing repositories. Also, knowledge services need to be improved such as perfecting the system's functions. The image recognition can extract some data from the diagrams but it can not get the all specific meanings. Some manual work are still needed as the supplement. So we need to replace manual work with automatic service(such as NLP) to reduce the processing time and the cost of system.

One of the biggest challenges in the future will be to make the rapidly increasing data easily searchable and accessible to the users to gain new insights.

Acknowledgment

The authors would like to thank the National Key Research and Development Project of China (2016YFC0200700), the Key Project of National Science Foundation of Zhejiang Province(No. LZ19F020002) and the State Key Laboratory of CAD&CG. Especially, this work was partially supported by the State Key Laboratory of Modern Optical Instrumentation Innovation Program, the Department of Atmospheric Sciences of Zhejiang University, the Key Laboratory of Design Intelligence and Digital Creativity of Zhejiang Province and Alibaba-Zhejiang University Joint Institute of Frontier Technologies.

References

[1] W. C. Hinds, "Aerosol Technology: Properties, Behaviour and Measurement of Airborne Particles," Elsevier Ltd, 1999.

[2] Lilu, Zhu, S. Shijun, W. Lin, "The Development Status of Worldwide Institutional Repositories and Some Corresponding Measures," Journal of Modern Information, 2017, 37(3), pp. 109-115.

[3] Borchert, Martin, P. Callan, "Strategies for gaining and maintaining academic support for the institutional open access repository," Organisation of Information&Knowledge Resources (2013).

[4] L. Yuan , M. Billings, "Strategies for Developing an Institutional Repository: A Case Study of ScholarWorks@UMass Amherst," Journal of Library&Information Science, 2011, 37(1), pp. 81-98.

[5] Michael C. Frank, Mika Braginsky, Daniel Yurovsky, Virginia A. Marchman, "Wordbank: an open repository for developmental vocabulary data*," Journal of Child Language, 2017, 44(3), pp. 1-18.

[6] A. Lammert, A. Hansen, F. Ament, S. Crewell, G. Dick, V. Gru tzun, H. Klein- Baltink, V. Lehmann, A. Macke, B. Pospichal, W. Schubotz, P. Seifert, E. Stamnas, and B. Stevens, "A standardized atmospheric measurement data (SAMD) archive for distributed cloud and precipitation process-oriented observations in Central Europe," Bull. Amer. Meteor. Soc. doi:10.1175/BAMS-D- 18-0174.1, in press.

[7] Jingling Wu, Long Ge, Junhua Zhang, "Comparation of accuracy of different diagnostic tests: an introduction of network meta-analysis methods," Chinese Journal of Evidence-Based Medicine, 2017, pp. 125-130.

[8] Joseph, C. Rhoda, "A structured analysis of e-government studies: Trends and opportunities," Government Information Quarterly, 2013, 30(4), pp. 435-440.

[9] Manninen, Lauren, "Describing Data: A Review of Metadata for Datasets in the Digital Commons Institutional Repository Platform: Problems and Recommendations," Journal of Library Metadata, 2018, pp. 1-11.

[10] Jie Yan, et al, "A Literature Review of Domestic and Foreign Crowdsourcing Research," Forum on Science and Technology in China, 2017, 8, pp. 59-68.

[11] Jianhong Feng, Guoliang Li, Jianhua Feng, "A Survey on Crowdsourcing," Chinese Journal of Computers, 2015, 38(9), pp. 1713-1726.

[12] Chiu, Chao Min, T. P. Liang, E. Turban, "What can crowdsourcing do for decision support?," Decision Support Systems, 2014, 65, pp. 40-49.

[13] Brabham, D. C., "Crowdsourcing the public participation process for planning projects," Planning Theory, 2009, 8(3), pp. 242 -262.

[14] A. Ståhlbröst, N. Bjørn-Andersen, "Organizational learning with crowdsourcing:The revelatory case of LEGO," Journal of the Association for Information Systems, 2014, 15(11), pp. 754-778.

[15] Chris Zhao, Q. Zhu, "Effects of extrinsic and intrinsic motivation on participation in crowdsourcing contest," Online Information Review, 2014, 38(7), pp. 896-917.

[16] Marion K. Poetz, Martin Schreier, "The value of crowdsourcing:can users really compete with professionals in generating new product ideas?," Journal of Product Innovation Management, 2012, 29(2), pp. 245 -256.

[17] Palmer, L. Carole, L. C. Teffeau , M. P. Newton, "Strategies for Institutional Repository Development: A Case Study of Three Evolving Initiatives," Library Trends, 2008, 57, pp. 143-167.

[18] Lowe, G. David, "Distinctive Image Features from Scale-Invariant Keypoints." International Journal of Computer Vision, 2004, 60(2), pp. 91-110.

[19] Danos, J. Rebecca, Andrew R. Frey, Yi Wang, "Canny Algorithm: A New Estimator for Primordial Non-Gaussianities." Physical Review D, 2012, 86(4).

[20] Youngseek Kim, "Fostering scientists' data sharing behaviors via data repositories, journal supplements, and personal communication methods," Information Processing&Management, 2017, 53(4), pp. 871-885