From One Star to Three Stars: Upgrading Legacy Open Data Using Crowdsourcing

Satoshi Oyama*, Yukino Baba†, Ikki Ohmukai‡, Hiroaki Dokoshi* and Hisashi Kashima†
*Graduate School of Information Science and Technology, Hokkaido University
Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido 060–0814, Japan
Email: {oyama,dokoshi}@complex.ist.hokudai.ac.jp
†Graduate School of Informatics, Kyoto University
36-1 Yoshida-Honmachi, Sakyo-ku, Kyoto 606–8501, Japan
Email: {baba,kashima}@i.kyoto-u.ac.jp
‡National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
Email: i2k@nii.ac.jp

***Abstract*—Despite recent open data initiatives in many countries, a significant percentage of the data provided is in non-machine-readable formats like image format rather than in a machine-readable electronic format, thereby restricting their usability. This paper describes the first unified framework for converting legacy open data in image format into a machine-readable and reusable format by using crowdsourcing. Crowd workers are asked not only to extract data from an image of a chart but also to reproduce the chart objects in spreadsheets. The properties of the reconstructed chart objects give their data structures including series names and values, which are useful for automatic processing of data by computer. Since results produced by crowdsourcing inherently contain errors, a quality control mechanism was developed that improves the accuracy of extracted tables by aggregating tables created by different workers for the same chart image and by utilizing the data structures obtained from the reproduced chart objects. Experimental results demonstrated that the proposed framework and mechanism are effective.**

I. Introduction

The most prominent of the recent open data initiatives to publish various kinds of data in electronic format are ones for statistical data gathered by governmental agencies [1]. Publishing such data is expected to improve government transparency, facilitate citizen participation, and create new business opportunities. These recent initiatives have led to the creation of data catalog sites by many countries, including the U.S.[1], the U.K.[2], and Japan[3], that provide data under an open reuse license.

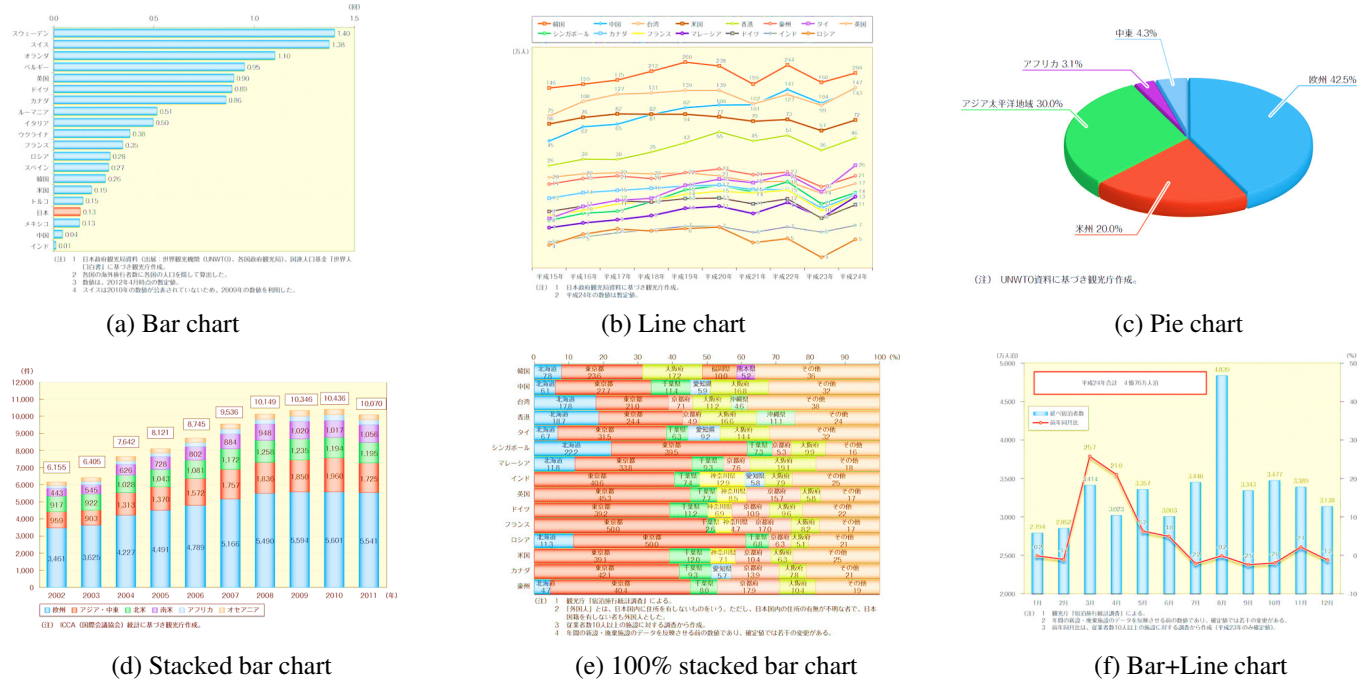
Tim Berners-Lee, the creator of the Web, developed a star rating system to encourage the publishing of data[4]:

- ⋆ The data are available on the Web (in whatever format) with an open license.
- ⋆⋆ The data are available in a machine-readable structured format (e.g., Microsoft Excel) instead of an image format.
- ⋆⋆⋆ The first two plus the data are in a non-proprietary format (e.g., CSV).
- ⋆⋆⋆⋆ The first three plus open standards from the World Wide Web Consortium (W3C), the Resource Description Framework (RDF) along with the SPARQL Protocol and RDF Query Language (SPARQL), are used to identify things.
- ⋆⋆⋆⋆⋆ The first four plus the data are linked to other people's data to provide context.

However, a significant percentage of published statistical data is published as charts or graphs in image or PDF files. For example, of the 10,410 datasets provided by the Japanese government data site, 5,452 are provided as PDF files. In the U.S. data catalog site, 4,838 of the 104,793 datasets are provided as PDF files. Such datasets earn only one star in Berners-Lee's rating scheme and are not readily reusable because extracting data from figures and tables in PDF files is not easy even if they are provided with open licenses. One of the major reasons for such hasty data publishing was limited budgets and human resources in governmental agencies. They cannot afford to convert such data into machine readable formats by themselves. The percentage of data published in machine-readable formats, such as CSV and RDF, will increase, but a certain amount of data will continue to be published in PDF or image files for a while. Furthermore, legacy data are generally published in such formats.

There have been certain demands for extracting values from statistical charts among the scientific community, typically for reusing data published in old papers. To meet such demands, various types of chart digitizing software such as WebPlotDig-

978-1-4673-8273-1/15/$31.00 © 2015 IEEE
[1]http://catalog.data.gov/
[2]http://data.gov.uk/
[3]http://www.data.go.jp/
[4]http://www.w3.org/DesignIssues/LinkedData.html



(a) Bar chart (b) Line chart (c) Pie chart

(d) Stacked bar chart (e) 100% stacked bar chart (f) Bar+Line chart

Fig. 1. Examples of charts used in the 2013 White Paper on Tourism published by the Japan Tourism Agency

itizer[5] and DataThief[6] have been developed. However, such software is designed for manual use and thus requires human intervention, such as in calibrating the chart axes, making it unsuitable for automatically extracting data from a large number of data charts.

Figure 1 shows examples of charts used in the 2013 White Paper on Tourism[7] published by the Japan Tourism Agency. Some charts are very complicated or in non-standard formats. For example, in the line chart (b), lines representing different data series cross each other. The pie chart (c) is represented as a cylinder in a 3D space rather than a simple circle. In chart (f), both bars and lines are used represent the data. Such variety and complexity make it difficult to automatically extract data using chart-digitizing software.

Human computation [2] has been attracting attention as a way to solve problems that are difficult to solve solely with a computer but are solvable with some human help. The increasing amount of workforce available online provided by crowdsourcing services such as Amazon Mechanical Turk[8] has been pushing widespread applications of human computation to data processing tasks in various domains. This will help overcome the bottleneck in promoting open data due to limited human resources. Open data are especially suitable for being processed by human computation because the data confidentiality/privacy issue, which is sometimes a barrier to crowdsourced data processing, does not have to be considered.

Data charts are designed to help people better understand data, and people are better at understanding them than computers. We have thus taken a human computation approach to the *datafication* of legacy data: use crowdsourcing to extract structured data from charts in legacy file formats such as image and PDF files. Doing this will improve the ranking of such data from one star in Berners-Lee's scheme to two or three stars. While Berners-Lee ranks CSV above Microsoft Excel since the former is an open format while the latter uses a proprietary one, in practice, the distinction is not substantial because recent versions of Excel use the Office Open XML format, and data in this format are readable by other software. Moreover, Excel spreadsheets are more expressive than the plain text of CSV and can represent the data structure, making it easier to automatically process the data by computer and upgrade it to four-star data. Thus, we use Excel spreadsheets as the format to save data extracted from chart images.

To the best of our knowledge, this paper presents the first unified framework for converting legacy open data in image format into a machine-readable and reusable format by using crowdsourcing. Also presented is a quality control mechanism that improves the accuracy of extracted tables by aggregating tables created by different workers for the same chart image and by utilizing the data structures obtained from the reproduced chart objects. Testing showed that the proposed framework and mechanism are effective.

II. Related Work

Studies to promote open data in accordance with the road map put forth by Berners-Lee have been conducted by researchers investigating the Semantic Web. However, their research interests have been mainly focused on pulling three-star data up to the fourth or fifth levels and building services on top of them—few have focused on dealing with one-star data.

[5]http://arohatgi.info/WebPlotDigitizer/
[6]http://datathief.org/
[7]http://www.mlit.go.jp/statistics/file000008.html
[8]http://aws.amazon.com/mturk/

Han *et al.* [3] developed an open-source tool for converting spreadsheet data into RDF format. Users define the relationships between the columns of a spreadsheet table in a *map graph* by using a graphical interface. A Web service then takes the spreadsheet or CSV file and the map graph and provides an RDF file as output. Mulwad *et al.* [4] presented techniques for automatically inferring the semantics of column headers and cell values and the relationships between columns. Their techniques are based on graphical models and probabilistic reasoning augmented with background knowledge from the Linked Open Data cloud (Web of Linked Data [5]).

The RDF Data Cube Vocabulary[9] is a W3C recommendation for publishing and sharing statistical data on the Web. In the Data Cube model, data observations (values in table cells) are characterized by dimensions, attributes, and measures. Meroño-Peñuela *et al.* [6] converted historical census data into RDF format by using the Data Cube vocabulary and a semi-automatic process: First, an expert manually annotated tables in Microsoft Excel workbooks, and then software called TabLinker[10] was used to convert them into RDF data cubes.

Government data made public with open licenses are called *open government data (OGD)* and are considered important for enhancing the transparency of governance and improving public services by promoting participatory decision making. Shadbolt *et al.* [1] described a project for integrating CSV data and spreadsheet data published on the U.K. data catalog site (data.gov.uk) in the Web of Linked Data. Kalampokis *et al.* [7] asserted that the real value of OGD comes from performing data analytics on top of combined datasets from different sources. As a test case, they used various published datasets including the unemployment rate dataset published in spreadsheet form on data.gov.uk and datasets regarding the outcome of UK general elections published in spreadsheet form on the Guardian newspaper's web site under an open license. They converted the datasets into RDF data cubes and performed data analytics using the combined datasets. They showed that there was a correlation between the probability of one of the two main political parties winning a particular constituency and the unemployment rate for that constituency.

Crowdsourcing is a means for asking many workers to perform tasks via the Internet. There are various types of crowdsourcing, which can be classified in terms of workers, requesters, tasks, and platforms [8]. Crowd workers can be paid via a crowdsourcing market such as Amazon Mechanical Turk or be volunteers. In the latter case, games with a purpose [9] are usually used. Human computation is a computing paradigm used to solve problems that are difficult to solve solely by computer but are relatively easy to solve if human assistance is provided. It is generally achieved by asking people to complete small tasks, called *microtasks* or *human intelligence tasks*, by crowdsourcing. Human computation approaches have been successfully applied to various domains including computer vision and natural language processing, where current artificial intelligence is still no match for human intelligence. One of the most successful examples of human computation is reCAPTCHA [10], which is a system for both verifying that an on-line user is actually human and for deciphering words unrecognized by optical character recognition (OCR) software used in, for example, a book digitization project. Two image files are presented to the user, one containing a word known by the computer and used for verification, and one containing an unrecognized word. The user must type the characters shown in both files. If the known word is typed correctly, the user is recognized as human. If enough users interpret the unknown word as a certain word, that word is considered valid. This process can be considered to be digitizing characters in images using human computation. The Semantic Web is an endeavor to make the Web machine understandable, so both machine and human intelligence are naturally needed to achieve it. Human computation practices have been applied to various sub-problems for enabling the Semantic Web [11], [12], [13], [14], [15].

AskSheet [16] is a system that uses crowdsourcing to make spreadsheet tables. It does not ask a worker to make an entire table but instead asks workers to gather the information needed to fill in the blank cells in a spreadsheet. As with our approach, it uses a quality control mechanism to compensate for erroneous inputs by workers, but it is applied to individual values while ours is applied to the entire table. Fan *et al.* [17] proposed a framework for utilizing crowd workers to aggregate tables. They prepared a concept catalogue and asked the workers to select the concept that best represented the values in each column. The selected concepts were used to match the columns that appeared in different tables but corresponded semantically. They focused only on one-dimensional tables and assumed that there were no incorrect values in the tables. In contrast, we focused on two-dimensional tables and assumed that there were incorrect values. Ermilov *et al.* [18] proposed a formalization of tabular data as well as its mapping and transformation to RDF, which enable the crowdsourcing of large-scale semantic mapping of tabular data. They mentioned automatic header recognition in CSV files as important future work. In our approach, headers in tables can be easily recognized by referring to the properties of the chart objects in the spreadsheets.

III. Datafication of Legacy Open Data Using Crowdsourcing

A. Overview

One of the most important aspects of human computation is designing the task so that intuitive and effective instructions can be given to non-expert workers so that they can work efficiently and accurately. The objective in our study was extraction of data from charts in image format and conversion of them into a form convenient for computer processing. A possible task design is to simply ask workers to extract graph data and place them in a CSV- or Excel-formatted file; however, the output with this approach does not provide a data structure, such as a distinction between row/column headers and data values, which is inconvenient for later data processing steps like data integration. Therefore, in our method, workers are asked to instead visually reproduce a chart image as a chart object in a spreadsheet using the functions of the spreadsheet software. This enables us to obtain a table linked to a chart object representing the data in the table and obtain the structure of the data, such as row and column headers and data sequences, from the properties of the chart object. It is not a

[9]http://www.w3.org/TR/vocab-data-cube/

[10]https://github.com/Data2Semantics/TabLinker

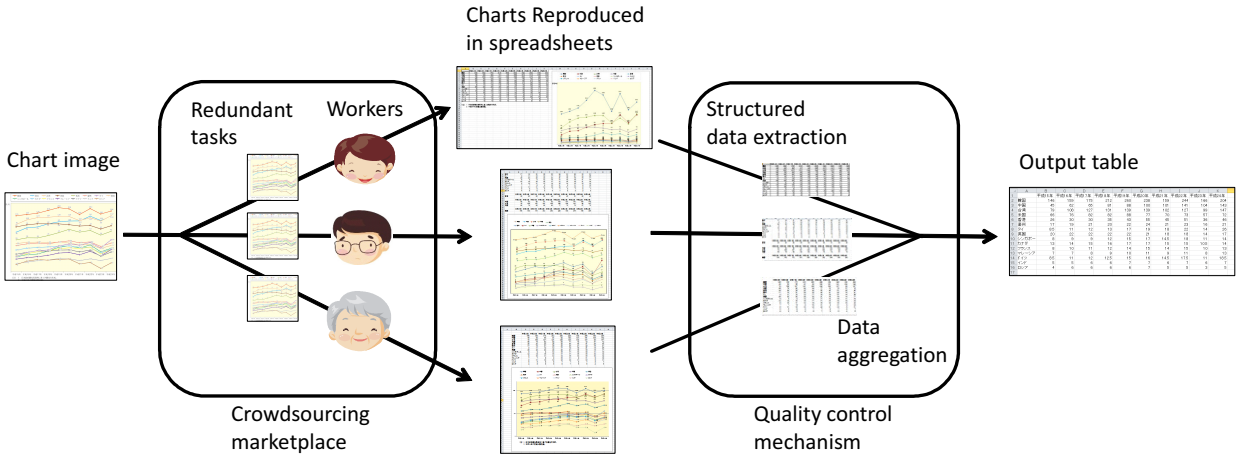


Fig. 3. Framework for digitizing chart image using crowdsourcing. Microtasks are generated and posted to a crowdsourcing marketplace. Several workers are assigned to each task. Each worker converts the image into a spreadsheet. The spreadsheets obtained from the different workers are integrated into a single, higher quality spreadsheet.

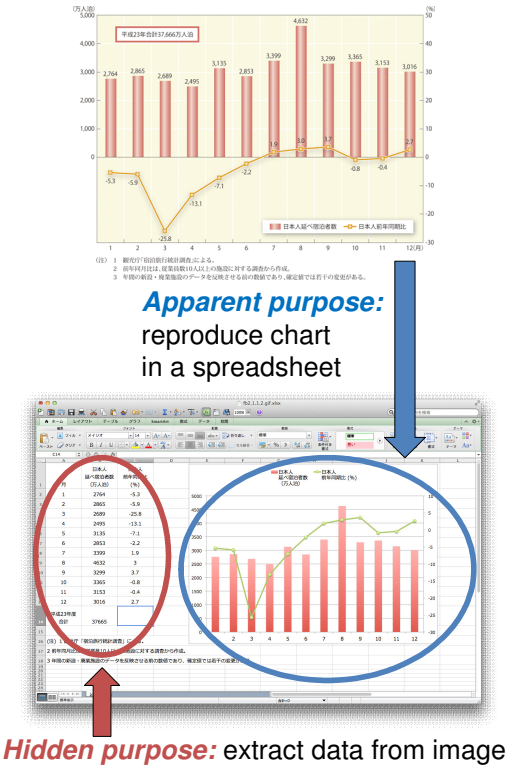


Fig. 2. Task design for extracting data from charts in image format. Workers are asked to visually reproduce a chart image as a chart object in a spreadsheet.

straightforward task to automatically identify row and column headers in a table in a CSV file or a spreadsheet without the chart, but they can be easily obtained using an application programming interface if the chart object is provided with the table. This task design (Figure 2) is an example of having a real purpose (extracting structured data) hidden behind an apparent purpose (reproducing a chart), which is common in human computation such as in the reCAPTCHA system [2].

Additionally, the structure of the data is essential for controlling the quality of digitizing work; it provides an efficient way to aggregate tables made by different workers and enables use of the common practice of asking multiple workers to complete the same task and then aggregating the results. Figure 3 shows our framework for digitizing chart images using crowdsourcing. The inputs are charts in an image file format such as JPEG. Microtasks asking crowd workers to reproduce the images in spreadsheets are generated and posted to a crowdsourcing marketplace. Several workers are assigned to each image. Each worker converts the image into a spreadsheet (in Microsoft Excel format) with an embedded graph. The axis labels, legends, and sequence values are extracted from the submitted file, resulting in pairs of attribute names and values. Finally, the spreadsheets obtained from the different workers are integrated into a single, higher quality spreadsheet.

B. *Structured Data Extraction through Visualization*

During the process of visually reproducing a chart image, a worker has to specify the properties of the chart object in the spreadsheet to reflect the structure of the data represented in the chart. Such properties can be accessed by using a computer program and an application programming interface. Although there are various kinds of charts including bar charts and line charts, most spreadsheets use a common format for their internal representations; for example, Microsoft Excel uses a three-item format.[11]

- A chart (`Chart`) has several data series (`Series`).
- Each data series (`Series`) has a name (`Name`).
- Each data series (`Series`) has x-axis values (`XValues`) and values (`Values`)

Figure 4 shows the relationships between a table and the properties of a `Chart` object. Although a two-dimensional table has several possible chart representations, the column

[11] http://msdn.microsoft.com/en-us/library/office/ff194068%28v=office.14%29.aspx

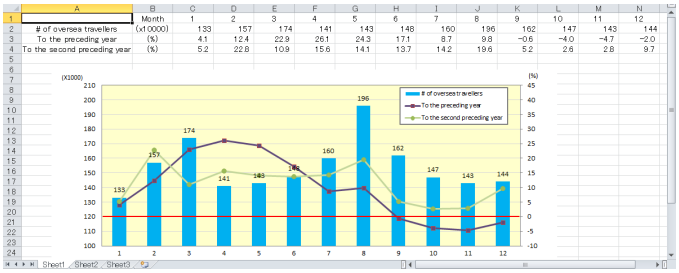


Fig. 5. Example of table in which data categories correspond to rows and months correspond to columns

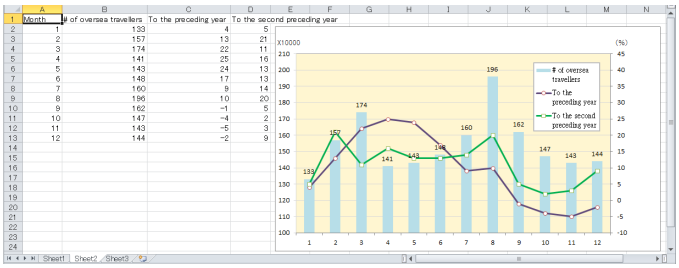


Fig. 6. Example of table in which data categories correspond to columns and months correspond to rows

labels and row labels correspond to the labels of the x-axis and the legends; they are extracted as `XValues` and `Name`. The data values are extracted as `Values` in `Series` objects. In tables, the choice of columns and rows is arbitrary; for example, with Figures 5 and 6, the data categories can correspond to the rows and the months can correspond to the columns, and vice versa. In either case, `Name` corresponds to categories, and the `XValues` property corresponds to months; this is essential information for integrating multiple tables since the choice of rows and columns does not need to be made. Moreover, such information is also beneficial when converting tables into RDF format using the RDF Data Cube Vocabulary, which is the next step toward 5-star open data.

IV. Quality Control Mechanism

Human-created tables may contain errors such as typos and missing values. This is especially true for crowdsourcing as the workers are usually anonymous and not trained well. The requester thus has to take into account errors and introduce a quality control mechanism that makes the results error-tolerant. A common approach to quality control in crowdsourcing is introducing redundancy by asking multiple workers to do the same task. For example, in classification tasks such as image classification, using the majority criterion generally improves the accuracy of the final result.

In our case, however, the worker outputs are tables, which are complex objects composed of several headers and many values, so a simple majority criterion cannot be applied. We thus developed a two-stage algorithm for integrating multiple tables made by different workers for the same chart image. First, the rows and columns in the different tables are aligned, and then the values in corresponding cells are aggregated. Figure 7 shows an example of aggregating the tables created by three workers.

A. Alignment of Rows and Columns

In general, the order of rows and columns in a table is arbitrary (except obvious cases in which the order is naturally defined such as for dates), so different workers may order the rows and columns differently. For example, different workers may arrange the row labels ("London," "New York," "Tokyo") in the chart in Figure 4 differently. Therefore, when aggregating the tables created by different workers, we first have to align the rows and columns among the different tables.

The names of rows (or columns) are the most important clue for judging whether two rows (columns) are identical; however, the names may contain errors or are sometimes missing in tables created by crowd workers. In that case, if the values in the rows (columns) are the same, the rows (columns) can be judged to be the same. Therefore, we introduce the similarity of two rows (columns) considering both their names and values and use it to find matching between rows (columns).

The measure we introduce here for measuring the similarity between two rows made by different workers is based on the probability of disagreement between the row headers and between the row values. Assume two workers w and w' transcribe the same row in a table and produce rows $(z^w, x_1^w, \ldots, x_N^w)$ and $(z^{w'}, x_1^{w'}, \ldots, x_N^{w'})$ respectively, where z represents the header and x_n represents a row value. (We assume the two rows have the same number of values but later explain how we deal with rows with different numbers.) We assume that with probability α_z the two workers use the same label for the header and that with probability $1-\alpha_z$ they use different labels. We also assume that the probability that two corresponding values are the same is α_x. These probabilities give the probability of having the two rows transcribed by the two workers:

$$\alpha_z^{I(z^w=z^{w'})}(1-\alpha_z)^{I(z^w\neq z^{w'})}\prod_{x_n^w=x_n^{w'}}\alpha_x\prod_{x_n^w\neq x_n^{w'}}(1-\alpha_x) \quad ,$$

where I is a function that returns 1 when the condition holds and 0 otherwise. Using this probability, we define the similarity measure between rows transcribed by different workers as

$$I(z^w=z^{w'})\ln\alpha_z + I(z^w\neq z^{w'})\ln(1-\alpha_z) + \sum_{x_n^w=x_n^{w'}}\ln\alpha_x + \sum_{x_n^w\neq x_n^{w'}}\ln(1-\alpha_x) \quad .$$

We define the similarity between columns in the same way.

Using this similarity measure and the following procedure, we align the rows and columns in the tables created by the two workers.

1) Calculate the similarities between all rows produced by the two workers.
2) Assume that row pairs with high similarity contain identical data, and pair them up.
3) Pair any remaining unmatched rows with dummy rows.

A similar procedure is applied to columns. The alignment can begin with either rows or columns. In many cases, a `Series` represents a time series and, in such cases, the `XValues` represent time or date. For such cases, there is usually less disagreement on the order of the `XValues` among

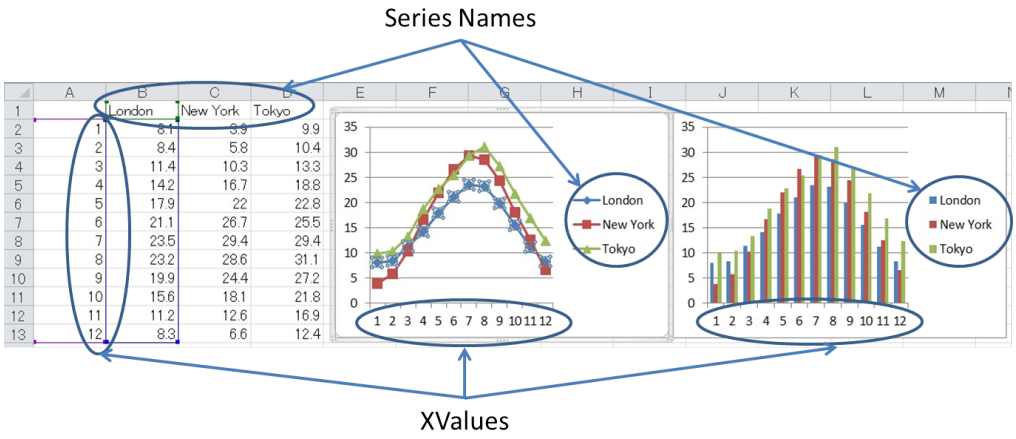
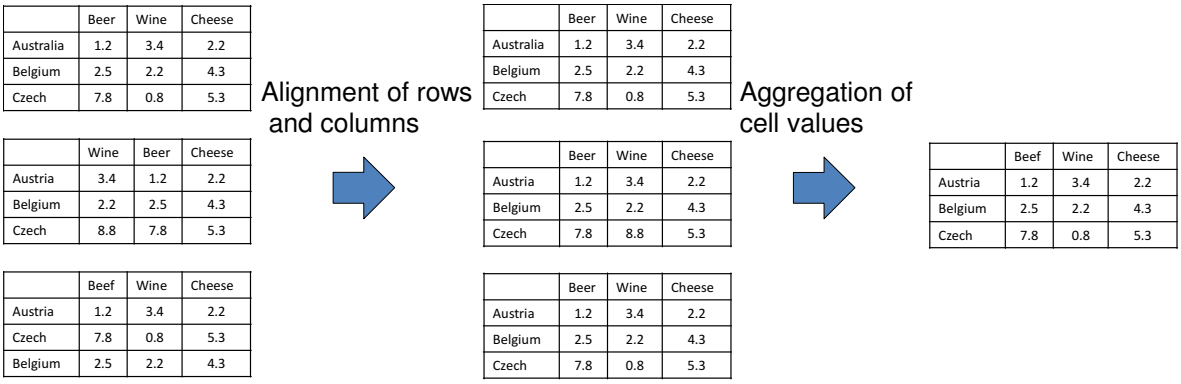


Fig. 4. Relationships between table and properties of `Chart` object. Although a two-dimensional table has several possible chart representations, the column labels and row labels correspond to the labels of the x-axis and the legends.



	Beer	Wine	Cheese
Australia	1.2	3.4	2.2
Belgium	2.5	2.2	4.3
Czech	7.8	0.8	5.3

	Wine	Beer	Cheese
Austria	3.4	1.2	2.2
Belgium	2.2	2.5	4.3
Czech	8.8	7.8	5.3

	Beef	Wine	Cheese
Austria	1.2	3.4	2.2
Czech	7.8	0.8	5.3
Belgium	2.5	2.2	4.3

	Beer	Wine	Cheese
Australia	1.2	3.4	2.2
Belgium	2.5	2.2	4.3
Czech	7.8	0.8	5.3

	Beer	Wine	Cheese
Austria	1.2	3.4	2.2
Belgium	2.5	2.2	4.3
Czech	7.8	8.8	5.3

	Beer	Wine	Cheese
Austria	1.2	3.4	2.2
Belgium	2.5	2.2	4.3
Czech	7.8	0.8	5.3

	Beef	Wine	Cheese
Austria	1.2	3.4	2.2
Belgium	2.5	2.2	4.3
Czech	7.8	0.8	5.3

Fig. 7. Example of aggregating tables created by three workers. After the rows and columns are aligned, the corresponding values from the three tables are integrated to obtain a single table.

workers' tables than there is on the orders of the `Series` objects. Therefore, starting with row alignment is a useful convention if the `Series` are arranged in a row (that is, the `Name` of a `Series` is used as the row header, and the `XValues` are used as the column headers). Note that changing the order of columns changes the similarity between rows, and vice versa; therefore, we should order rows and columns simultaneously. However, for simplicity, we assume they are independent and order the rows first and then order the columns. If three or more tables are to be aligned, one is chosen as a reference table, and the remaining tables are aligned with respect to the reference table.

B. Aggregating Cell Values

Since the results produced by crowdsourcing may contain errors, after the rows and columns of the tables are matched, the corresponding values from the tables are integrated to obtain a single table. The majority criterion is used to determine the final values for nominal values such as names. The median value is used for numerical values since the median is more robust to outliers than the average. Considering human errors as outliers rather than noise (as they are in instrument measurements) is appropriate because crowd workers can make errors of great magnitude. For example, consider a case in which the inputs from three workers are 1982, 1892, and 1982 and the actual value is 1982; the median matches the actual value while the average value greatly differs.

V. Case Study

A. Data set and software

We evaluated our proposed framework and quality control mechanism experimentally by using chart images from the 2013 White Paper on Tourism[12] published by the Japan Tourism Agency. The white paper is published under the Creative Commons CC-BY license, but most of the statistical data are provided as figures embedded in HTML documents or PDF files, i.e., as one-star open data in Berners-Lee's scheme. Among the 104 images used in the white paper, 61 explicitly show values as data labels, and we used them as the gold standard for evaluating the correctness of the extracted values.

The dataset contains various types of charts. We categorized them into six categories: (a) bar chart, (b) line chart, (c) pie chart, (d) stacked bar chart, (e) 100% stacked bar chart,

[12]http://www.mlit.go.jp/statistics/file000008.html

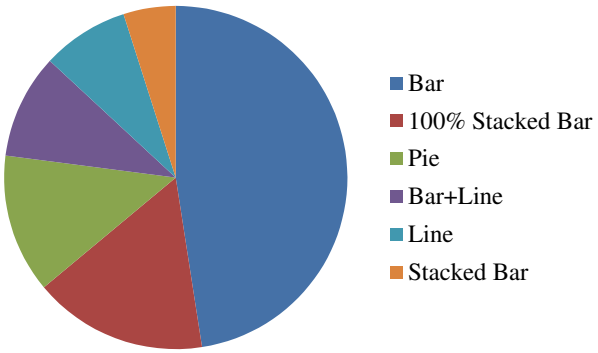


Fig. 8. Percentages for different types of charts in the dataset

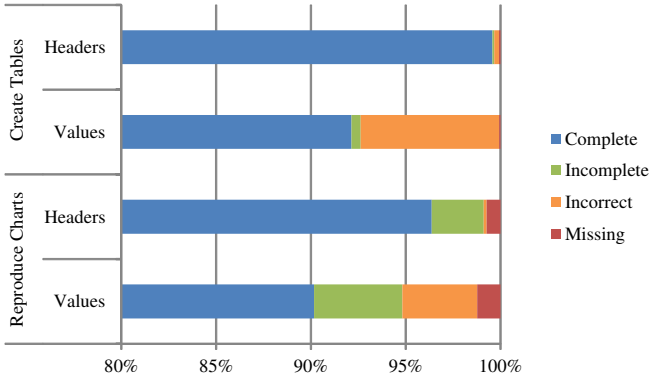


Fig. 9. Percentages for different types of errors in worker tables

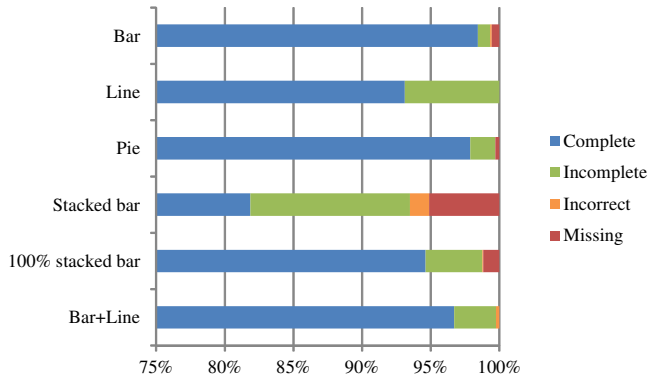


Fig. 10. Errors in table headers in worker tables for different types of charts

and (f) bar+line chart. The percentages for these chart types in the dataset are shown in Figure 8. Although the bar charts account for about half the dataset, the other chart types are also prevalent.

We compared the results of two different crowdsourcing tasks. One simply asked workers to extract data from charts and put them in a spreadsheet ("Create Table" tasks), and the other asked workers to reproduce charts in a spreadsheet ("Reproduce Chart" tasks). We asked five workers to complete each task. We used the Lancers crowdsourcing service[13] and paid 200 JPY (approximately 2 dollars) for each completed task. A worker was not required to do all tasks—23 different workers did at least one Create Table task, and 30 different workers did at least one Reproduce Chart task.

We implemented the table aggregation software using Microsoft Excel VBA (Visual Basic for Applications). This software took the Excel files created by the workers as input and searched each file for a `Chart` object. If multiple `Chart` objects were found, it used only the one with the first index number. For each `Series` object in the `Chart` object, it extracted the values of the `Series.Name`, `Series.XValues`, and `Series.Values` properties from the corresponding cells in the worksheet as the row headers, column headers, and cell values, respectively. The table aggregation algorithm was then applied to the set of worker tables (with $\alpha_x = \alpha_z = 0.9$), and the aggregated table was stored as a worksheet in a new Excel file.

B. Accuracy of Worker Tables

We manually made gold standards for the transcribed tables and evaluated the accuracy of the tables created by the crowd workers. We separately evaluated the accuracies of the row and column headers and the cell values. Figure 9 shows the percentages of different types of error cells for both tasks. "Incomplete" means some data values were not exactly the same as the gold standard, such as different spelling or values without appropriate units. "Incorrect" means that the values were simply incorrect mainly due to mistyping or another mistake, and "Missing" means some data values were missing from the table. Although the dataset contained complex charts such as the ones shown in Figure 1, both tasks resulted in accuracies higher than 90% for both table headers and cell values, which indicates that our approach using crowdsourcing is promising. The Reproduce Chart task resulted in fewer incorrect header and cell values than the Create Table task. This might be because the reproduced charts made it easier for the workers to spot errors. On the other hand, the Reproduce Chart task resulted in more incomplete and missing values.

Figures 10 and 11 show the percentages of errors in table headers and cell values generated by the Reproduce Chart task for different types of charts. Most "incomplete" headers were missing units of measure. The pie charts had many missing values. Pie charts usually display percentages as well as numeric values, but many workers did not transcribe them into their tables but instead calculated them from the numeric values using a function of Excel. For example, the pie chart in Figure 12 show percentages, but they are not found in the table. They are displayed by specifying data label formats. The totals for the stacked bar chart are also missing for the same reason. Although we counted them as "missing values" in our evaluation, they can be recovered from the numeric values in the table and thus should not cause major problems in practice.

C. Accuracies of Aggregated Tables

We next measured the accuracies of the aggregated tables. At least three tables are required so that using the majority criterion or the median works well. We compared two different

[13] http://www.lancers.jp

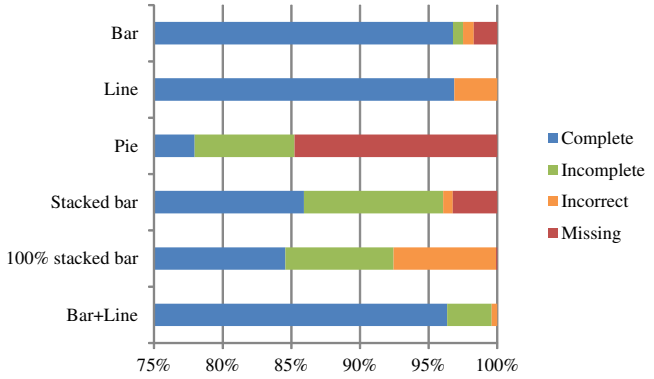


Fig. 11. Errors in cell values in worker tables for different types of charts

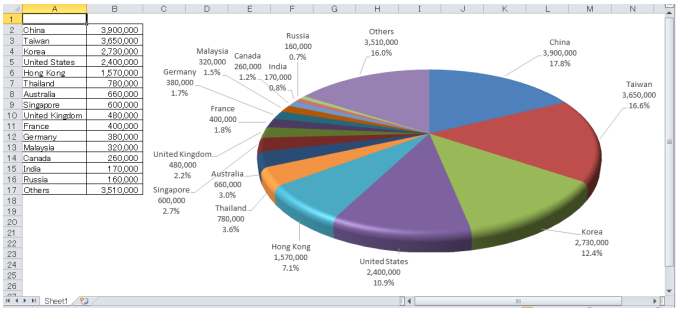


Fig. 12. Example of pie chart and corresponding table. Although pie chart displays percentages, they are not shown in table.

settings. In one, all five worker-generated tables were used for each chart image; in the other, three randomly selected tables were used for each image. Figure 13 shows the percentages for different types of errors after aggregation. Aggregation greatly improved the accuracy for cell values. It also eliminated most of the incorrect and missing headers while it was not very effective for reducing the incomplete headers.

Figures 14 and 15 show the percentages of errors in the aggregated tables from five workers for different types of charts. Most of the incomplete headers were due to lack of appropriate units. Many workers did not write them in the cells, so the majority criterion did not work well. Although we could recover some missing "percentages" by retrieving cell style information, a more general handling of missing units is part of our future work. Most missing cell values in the stacked bar charts were the total of stacked values, which can be recovered, as explained in the previous subsection.

VI. Summary and Future Work

Converting legacy open data in, for example, image files into machine-readable format is an important step for realizing the potential of Linked Open Data, but it is labor intensive, and there have been few related research efforts. We proposed using crowdsourcing to digitize chart images and introduced a task design for crowd workers. We asked them to reproduce chart images as embedded chart objects in spreadsheets, which enabled us to automatically identify the data structures of tables from the properties of the chart

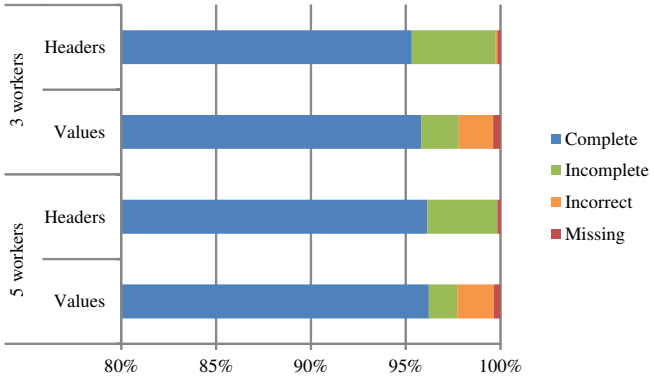


Fig. 13. Percentages for different types of errors in aggregated tables

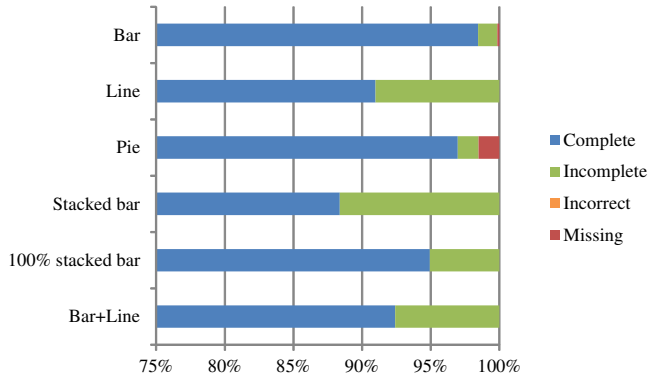


Fig. 14. Errors in table headers in aggregated tables for different types of charts

objects by using a computer program. To reduce the number of errors inherent in crowdsourced results, we developed a quality control mechanism. Multiple workers were asked to digitize the same chart image, and the tables they created were aggregated. The data structure identified from the chart objects played an important role in the automatic processing of data by computer. The experimental results demonstrated that our approach was effective, but many tasks remain for future work.

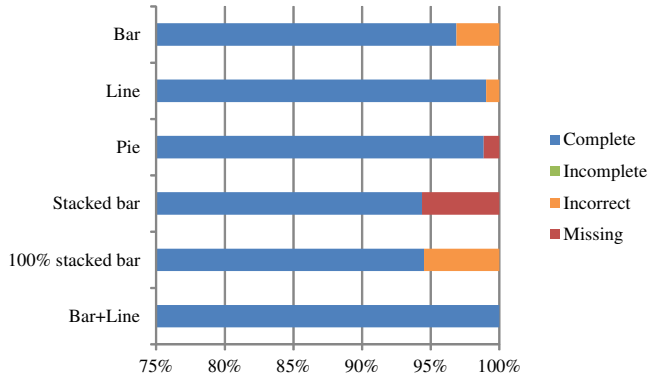


Fig. 15. Errors in cell values in aggregated tables for different types of charts

They can be grouped into four main areas, as described below.

A. Task Design

Several lessons were drawn from the results of our experiments. The inconsistency in the use of units could be reduced by defining units as cell styles in advance and asking the workers to use them. We used chart images that explicitly showed values as data labels to enable us to objectively evaluate the accuracy of the generated tables, but in practice many charts are published without data labels. Since eye-based measurement for such charts would not be accurate, the workers should be asked to use chart digitizing software as it would be more effective. In the experiments, we asked the workers to upload their Excel files to a file server, and we downloaded them manually to a local machine, on which we ran the table aggregation program. This procedure is prone to problems, such as lost or misidentified files. Performing all the processes in a single cloud environment would greatly improve the throughput and accuracy of the chart digitizing tasks.

B. Table Aggregation

In this study, we assumed that the workers had consistent abilities, and we used simple aggregation methods for table integration; however, in a crowdsourcing setting, the workers generally have various degrees of ability. Application of more sophisticated statistical quality control methods that consider worker-dependent abilities (e.g., [19], [20]) is a possible future direction for improving integration accuracy.

C. Converting Tables into RDF Format

The next step according to the roadmap for Linked Open Data is converting tables into RDF format. The dimensions, measures, and attributes in the RDF Data Cube Vocabulary, a framework for representing statistical tables, generally correspond to headers, values, and units of values in statistical tables. After the headers and values are successfully extracted using our chart digitizing approach, we have to relate them to a formally defined vocabulary. This process is also difficult to perform solely by computer; it requires human intelligence. Therefore, the use of crowdsourcing to convert tables into RDF format is important future work. Collecting publicly available spreadsheets with charts and extracting names from them would help in constructing a vocabulary for describing statistical data.

D. Structured Data Extraction through Visualization

In this study, we reproduced image-formatted charts in spreadsheets to enable us to extract table-formatted data from them. However, there are many other data types that are not provided in visualized formats such as CSV-like texts and spreadsheets without charts. Producing charts from these non-visualized data would make the data easier to understand; moreover, such processes would help in extracting the structures of the data as a byproduct. This is referred to as *unsupervised* visualization of the data, while chart reproduction from images is referred to as *supervised* visualization.

REFERENCES

[1] N. Shadbolt, K. O'Hara, T. Berners-Lee, N. Gibbins, H. Glaser, W. Hall, and m. c. schraefel, "Linked Open Government Data: Lessons from Data.gov.uk," *IEEE Intelligent Systems*, vol. 27, no. 3, pp. 16–24, 2012.

[2] E. Law and L. von Ahn, *Human Compuation*. Morgan & Claypool Publishers, 2011.

[3] L. Han, T. Finin, C. Parr, J. Sachs, and A. Joshi, "RDF123: from Spreadsheets to RDF," in *Proceedings of the 7th International Semantic Web Conference (ISWC 2008)*, 2008, pp. 451–466.

[4] V. Mulwad, T. Finin, and A. Joshi, "Automatically Generating Government Linked Data from Tables," in *Working Notes of AAAI Fall Symposium on Open Government Knowledge: AI Opportunities and Challenges*, 2011.

[5] C. Bizer, "The Emerging Web of Linked Data," *IEEE Intelligent Systems*, vol. 24, no. 5, pp. 87–92, 2009.

[6] A. Meroño-Peñuela, R. Hoekstra, A. Scharnhorst, C. Guéret, and A. Ashkpour, "Longitudinal Queries over Linked Census Data," in *Proceedings of the 10th Extended Semantic Web Conference (ESWC 2013)*, 2013, pp. 306–307.

[7] E. Kalampokis, E. Tambouris, and K. Tarabanis, "Linked Open Government Data Analytics," in *Proceedings of the IFIP Electronic Government Conference (EGOV 2013)*, 2013, pp. 99–110.

[8] M. Hosseini, K. Phalp, J. Taylor, and R. Ali, "The Four Pillars of Crowdsourcing: A Reference Model," in *Proceedings of the IEEE Eighth International Conference on Research Challenges in Information Science (RCIS 2014)*, 2014, pp. 1–12.

[9] L. von Ahn and L. Dabbish, "Designing Games with a Purpose," *Communications of the ACM*, vol. 51, no. 8, pp. 58–67, 2008.

[10] L. von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, "reCAPTCHA: Human-Based Character Recognition via Web Security Measures," *Science*, vol. 321, no. 5895, pp. 1465–1468, 2008.

[11] E. Simperl, B. Norton, and D. Vrandečić, "Crowdsourcing Tasks in Linked Data Management," in *Proceedings of the 2nd Workshop on Consuming Linked Data (COLD 2011)*, 2011.

[12] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux, "ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-scale Entity Linking," in *Proceedings of the 21st International Conference on World Wide Web (WWW 2012)*, 2012, pp. 469–478.

[13] N. F. Noy, J. Mortensen, M. A. Musen, and P. R. Alexander, "Mechanical Turk as an Ontology Engineer?: Using Microtasks as a Component of an Ontology-Engineering Workflow," in *Proceedings of the 5th Annual ACM Web Science Conference (WebSci 2013)*, 2013, pp. 262–271.

[14] M. Acosta, A. Zaveri, E. Simperl, D. Kontokostas, S. Auer, and J. Lehmann, "Crowdsourcing Linked Data Quality Assessment," in *Proceedings of the 12th International Semantic Web Conference (ISWC 2013)*, 2013, pp. 260–276.

[15] D. DiFranzo and J. Hendler, "The Semantic Web and the Next Generation of Human Computation," in *Handbook of Human Computation*. Springer, 2013, pp. 523–530.

[16] A. J. Quinn and B. B. Bederson, "AskSheet: Efficient Human Computation for Decision Making with Spreadsheets," in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW 2014)*, 2014, pp. 1456–1466.

[17] J. Fan, M. Lu, B. C. Ooi, W.-C. Tan, and M. Zhang, "A Hybrid Machine-Crowdsourcing System for Matching Web Tables," in *Proceedings of the 30th IEEE 30th International Conference on Data Engineering (ICDE 2014)*, 2014, pp. 976–987.

[18] I. Ermilov, S. Auer, and C. Stadler, "User-driven Semantic Mapping of Tabular Data," in *Proceedings of the 9th International Conference on Semantic Systems (I-SEMANTICS 2013)*, 2013, pp. 105–112.

[19] A. P. Dawid and A. M. Skene, "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statics)*, vol. 28, no. 1, pp. 20–28, 1979.

[20] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan, "Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise," in *Advances in Neural Information Processing Systems 22*, 2009.