The following is a pre-print of the paper:

Hazeline U. Asuncion. SourceTrac: Tracing Data Sources within Spreadsheets. IPAW 2012, Springer-Verlag, June 2012 (to appear).

SourceTrac: Tracing Data Sources within Spreadsheets

Hazeline U. Asuncion

Computing and Software Systems
University of Washington, Bothell
Bothell, WA USA

hazeline@u.washington.edu

Abstract. Analyzing data from multiple sources is a common task in scientific research. In particular, spreadsheet data is often aggregated from a variety of sources to identify patterns and synthesize reports. Yet, techniques are lacking for automatically capturing the provenance of such data within spreadsheet environments like Excel. We present a novel approach for fine-grained tracing of tabular data that may have been obtained from files, databases, or the Web. Our approach provides relevant provenance information at both the micro-level (per cell) and the macro-level (per sheet). Initial results suggest that our approach is scalable and beneficial to data analysts.

Keywords: data provenance, spreadsheets, multiple sources

1 Introduction

Aggregating data from multiple sources is a fundamental operation in data-intensive domains, including the natural sciences, social sciences, and business. In the eScience domain, raw scientific data may be stored by different instruments in different file locations or may be obtained from different research organizations. As an example, the Jaffe Atmospheric Research Group at UW Bothell analyzes tabular data from their own field sites as well as published data from the EPA, NOAA, and NASA on a weekly basis to understand global and regional sources of air pollution.

While methods exist for tracking multiple sources of data in the context of scientific workflows, databases, and grid computing [5, 6, 7, 11, 14], provenance techniques are lacking in spreadsheet environments (like Microsoft Excel) which are ubiquitously used by data analysts and researchers. Because spreadsheet environments are highly interactive, multiple data sets are often merged together into one spreadsheet at multiple time points. Thus, without

provenance it becomes difficult for researchers to recover the original source(s) of a record. In addition, it is difficult to determine updates to downstream files if a source file has changed or contains an error.

In response to these challenges, we present SourceTrac, a novel approach for tracing multiple sources of data within a spreadsheet. Our approach captures the source as the users are obtaining data (e.g., from the Web), annotates the immediate source of each record at the granularity of individual cells, calculates source ancestors of formulas, and provides mechanisms for connecting a file to parent files as well as visualizing this provenance data. In this paper, our approach is tailored towards Microsoft Excel but can be generalized to other environments.

In previous work, we focused on capturing the operations that users perform on a spreadsheet while analyzing data [2]. This paper, meanwhile, is complimentary since the focus is on capturing operations that pertain to fine-grained tracking of data sources.

The rest of the paper is organized as follows. The next section discusses techniques we use to track data sources. Section 3 describes the implementation details of the SourceTrac tool. Section 4 presents use cases, initial user feedback, and scalability measures. Section 5 covers related work. Finally, we conclude with a discussion of future avenues of research.

2 Provenance Technique

In this section, we delve into the core aspects of our provenance technique for tracking sources within spreadsheets.

2.1 Tracking the source as the user obtains the data

To support the tracking of heterogeneous data sources, we capture the data source *in situ*, at the time a user obtains the data. Otherwise, retrospectively determining the source may be difficult, if not impossible. Users also explicitly specify when they wish to track data sources. This avoids the accidental capturing of sources that are unnecessary to researchers. Users may obtain data from the Web, a database, or other tabular files.

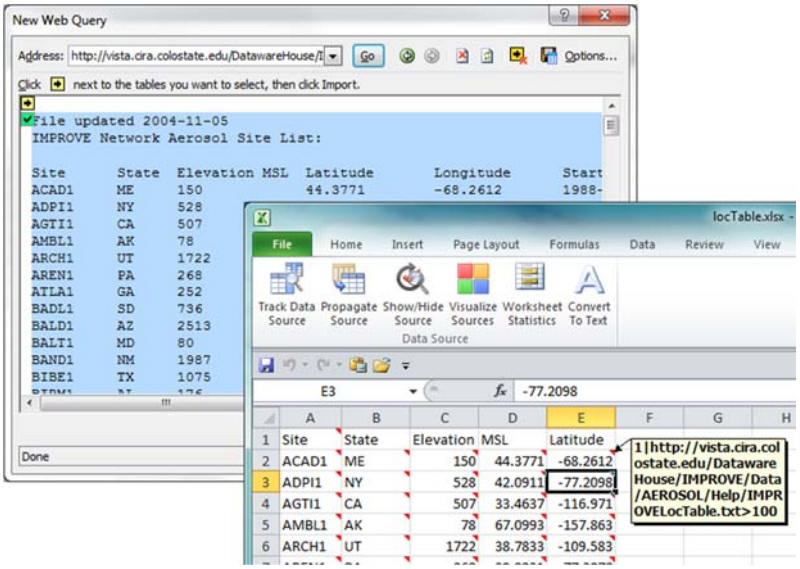

Fig. 1. Data obtained from the Web are annotated with the URL.

There are several ways that sources can be recorded when users obtain data from the internet. Within Excel, SourceTrac can detect data scraped from the web through Excel's "Get External Data From Web" interface. If data is manually copied from a Web page or if a file is downloaded from a web site, it is also possible to determine the source by automatically inspecting the immediate history of the user's web browser (e.g., Firefox) and obtaining the visited URL (using a traceability technique [3]). Once the URL is captured, the cells are annotated with this URL.

Figure 1 shows data scraped using the Excel interface (background) which is then pasted onto a spreadsheet (foreground). After the extracted data has been cleaned up and formatted, we see that each cell is automatically annotated with the source, shown as a small red triangle at the top-right of each cell. In this example, the annotation reveals that the data comes from one source, indicated by "1", followed by the URL of the source. The "100" at the end indicates that the data in the cell is 100% derived from the specified URL.

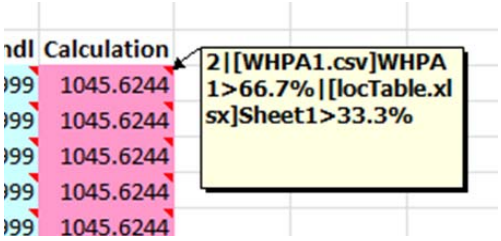

Fig. 2. Data sources are annotated at the cell level.

Another source of data is external files, such as other spreadsheet files or text files. In this case, we can record the name of the file (including the file path) when the user copies the data. When the data is pasted onto the spreadsheet, the cells are then annotated with the file name of the source. In principle, it is also feasible to perform provenance tracking as users query a database, using Excel's facilities. In this case, the SQL query would be the recorded source. It is worth noting that once a cell is annotated with a data source, the annotation will remain with the cell even if the user copies the cell to another sheet.

2.2 Querying data sources at multiple levels of granularity

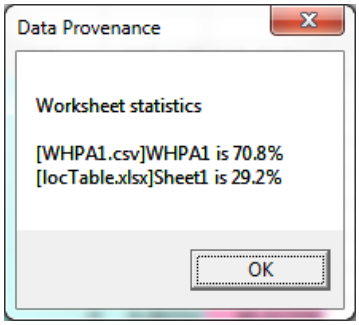

Fig. 3. Source statistics can be calculated at the worksheet level.

Since sources, or provenance metadata, are recorded at the cell level, SourceTrac allows for provenance reporting at multiple levels of granularity: the cell level, the spreadsheet level, and the file level. At the cell level, individual cells are annotated with the source. If the data contained within a cell is a function or a calculated formula, the sources of the dependency cells are indicated. For instance, Figure 2 shows that the calculated cell is derived from two sources (as indicated by the number 2 at the beginning of the annotation): two-thirds of the cells which the calculated cell depends on is derived from the file "WHPA1.csv", while one-third comes from worksheet "Sheet1" of the file "locTable.xlsx". These ratios are based on the

number of cell references specific to a source and the total number of cell references. The vertical bar "|" is used to delimit the sources.

One may also determine the source composition for a given spreadsheet. The distribution of data sources can be determined by scanning through the data cells (i.e., non-empty cells below the row headings) and summing up the individual source distributions for each cell. Figure 3 shows example worksheet statistics which list the data sources as well as the percentage of the cells that belong to each source.

Finally, one can obtain the source dependencies at a file level. This can be achieved by iterating through the different worksheets in the file and compiling the source distributions of each worksheet. This information can then be saved as the metadata of the spreadsheet file. As an example, Figure 4 shows the source of the current file ("WHPA_process.xlsx") in the Comments field of the file properties. Such metadata is useful when researchers need to trace the file to the parent spreadsheet file (e.g., to check whether the formulas used were correct).

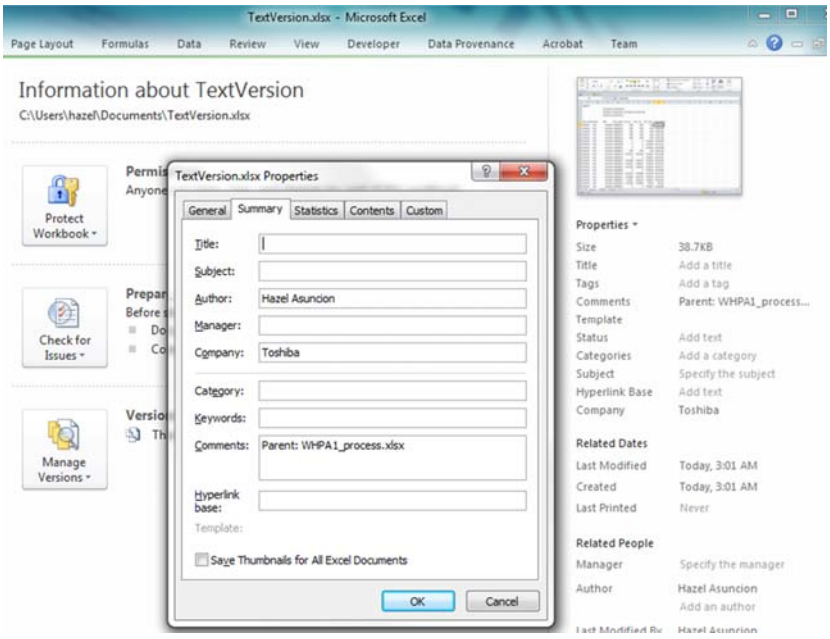

Fig. 4. Data source dependencies at the file level can be annotated within the file's metadata.

2.3 Calculating the ancestors of the data

While the previous techniques allows for tracking to the immediate source, there are situations where tracking the line of source ancestors (across files) is necessary (e.g., when an error appears in a descendant spreadsheet). There are two possible ways to track the ancestors of data within a cell. The first method is by following the backward links from the cell to the immediate source and on to its source until we arrive at an external root source, which is can be a web URL, a database connection, or another file with no source annotations.

Another method is to build source trees as the user obtains data from the different sources. Root nodes are created each time data is extracted externally. Branches are created each time the extracted data is copied to another file or is inserted via an Excel "lookup" function across files. Branch nodes contain the filename of the descendant spreadsheet, as well as specific location in the spreadsheet where the data has been copied or inserted, such as the worksheet name and the range of cells.

If the data is extracted from the web, it is also possible to further query the source of the published data on the web. This will require the existence of structured metadata or a web service that takes in a URL of a data source and outputs the source(s) of the data. In the event that the data has been obtained from a published source of another research organization, one can envision the output of the web service to be another URL, or perhaps another web service that points to yet another published data source on the web.

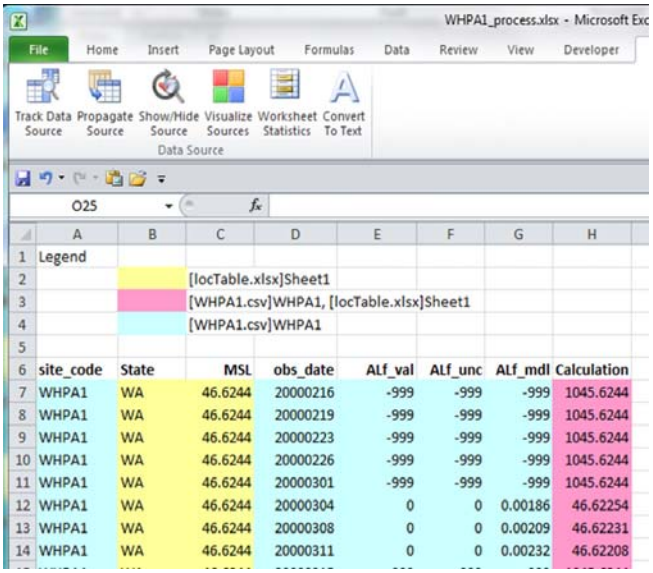

Fig. 5. Data sources can be visualized for quick analysis of provenance.

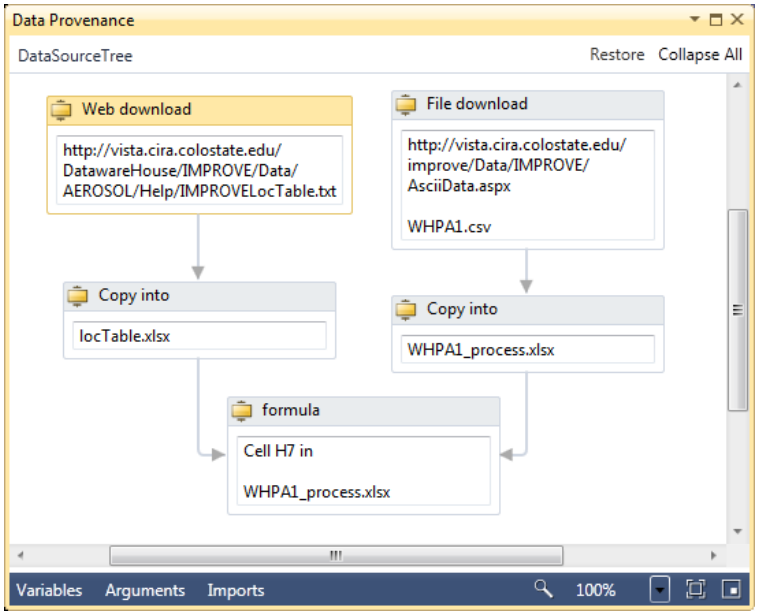

Fig. 6. The source ancestry dependency graph can also be visualized.

2.4 Visualizing sources at different levels of granularity

With the ability to query data sources at different levels of granularity, SourceTrac can also visualize the source information. One approach is to visualize the entire spreadsheet to allow a user to gain a high level-view of which regions of the spreadsheet are obtained from which sources. Figure 5 shows that columns B and C, color-coded in yellow, were obtained from "Sheet1" of the "locTable.xlsx file", while column H, color-coded in pink, was obtained from both "WHPA1.csv" and "locTable.xlsx" and the rest of the columns, color-coded in light blue, were obtained from "WHPA1.csv".

One may also be interested in visualizing the ancestors of a cell location. Figure 6 depicts an example where the cell H7 has two source lines due to the fact that H7 contains a formula that depends on two cells which have two separate sources. The first source is the file, "locTable.xlsx", which was obtained via the "Get External Data" interface in Excel. The other source is "WHPA1_process.xlsx" which originated from a download. Thus, this view allows a user to quickly identify the various sources of a calculated field.

3 SourceTrac Tool Support

The following sections describe the tool's design and implementation as well as use cases of interest to data analysts.

3.1 Tool design and implementation

We designed the SourceTrac tool to be easy to use and accessible, requiring minimal setup and configuration. To this end, we provide a user interface within the Excel environment (specifically Excel 2010). The functionalities mentioned in the previous section are accessible via the Excel ribbon (seen in Figure 1). To set up the tool, users would simply run an installation file.

SourceTrac also leverages Excel's interfaces for obtaining data from the Web and from a database. For example, we use Excel's XLQueryType property of the QueryTable object to obtain the URL of the source data. SourceTrac automatically determines which sections of the spreadsheet received the data and annotates the cells with the sources.

We use various means of storing the source annotations. At the cell level, sources are stored as comments to individual cells. At the file level, sources are stored as a file property (within the "Comments" metadata field). Users also have the option of viewing or hiding the cell comment indicator (i.e., the red triangle at the top-right corner).

The current tool implementation includes most of the capabilities mentioned in the previous section. Dependency tracking is only partially implemented. We also plan to integrate the capture of the source URL from the Firefox browser's history into our tool.

3.2 Use cases

We envision the following use cases for SourceTrac. A researcher obtains data published on the Web through Excel's "Get External Data" interface. The researcher then obtains another data set online by downloading a spreadsheet from a different URL. The researcher proceeds to combine the two data sets by using an Excel "vlookup" function in the downloaded spreadsheet to get the data from the first file. After the data has been combined, the researcher adds formulas to the spreadsheet.

At a later point in time, the researcher discovers that the first data set has been corrected by the publishers of the data. The researcher opens the aggregated file, visualizes the sources, and immediately identifies which regions of the spreadsheet to update using the SourceTrac tool.

When the researcher decides to publish his results, he proceeds to duplicate his spreadsheet by pasting the data values and removing the formulas. Sometime later, he wishes to double-check whether he used the correct statistical formula in his published result. Looking at the "Comment" metadata field in the file properties of the published file, he finds the path to the parent file and verifies that he has indeed used the correct formula.

4 Evaluation

In this section, we discuss the provenance queries that can be answered by the tool as well as initial user feedback and a discussion of the tool's overhead.

4.1 Provenance queries

The SourceTrac tool can answer the following important provenance queries.

1. Where did the data from this spreadsheet file come from? This query can be answered in multiple ways. One may view the individual annotations at the cell-level (Figure 2), or one may choose to look at the worksheet source distribution statistics (Figure 3). Furthermore, file-level annotation is saved in the "Comment" field in the file properties (Figure 4).

2. How was the data in this cell derived? Again, one may view the cell annotation, which would show the sources from which that cell was derived. This query can also be answered by viewing the ancestor source tree (Figure 6).

3. If a source file changes, which parts of the worksheet need to be updated? This query can be answered by visualizing sources (Figure 5). If the first source file, "locTable.xlsx", has been changed, then one can find (by color) the regions of the spreadsheet to change.

4. Does my final accuracy in cell H18 depend at all on faulty data source X? One can look at the cell annotation or the ancestor tree. Moreover, SourceTrac returns the degree (percentage) to which H18 depends on data source X.

4.2 User feedback

We have solicited feedback regarding our tool from a senior scientist in the Jaffe Atmospheric Research Group at the University of Washington, Bothell. This scientist analyzes data from multiple sources frequently, at least on a weekly basis. A typical size of his resulting dataset is 30,000 records with 10 to 50 columns. He performs environmental analyses of data from multiple data sources.

His current approach to tracking sources is naming the spreadsheet based on the source and manually entering the data sources in another worksheet within the same file. A drawback to this approach is that there are times where he may forget to document the data source. In addition, since he transforms his spreadsheet data into another spreadsheet by pasting by value, the formulas he used in the original spreadsheet are not readily available. In order to find this information, he needs to search through his file system to find the parent file (or the source file) of the text file.

According to the scientist, the tool will allow him to improve his documentation of the sources. It will also save him a substantial amount of time by avoiding the duplication of his analysis. Without the tool, if he is unsure of the data source or the version of the data source, he would have to re-analyze the source data again to verify his results. He comments that he is looking forward to using the tool in day-to-day analysis operations.

4.3 Scalability of tool

Since we provide a source annotation for each data cell in the spreadsheet, a potential concern is that the spreadsheet files would become too large in size due to the necessity to store this additional metadata. However, we find that the space overheads are reasonable. For an Excel spreadsheet containing 5,000 cells with formulas and annotations, the file size is only 3.25 times larger than the equivalent spreadsheet without annotations. With 10,000 cells, the file size is only 3.69 times larger. With 100,000 cells, the file size is only 4.63 times larger than the corresponding spreadsheet without annotations. In all these cases, the file size is less than a few megabytes. These numbers suggest that the overhead cost of adding annotations at a cell-level is reasonable.

5 Related work

Tracking multiples sources of data has been addressed in different contexts: databases, scientific workflows, grid computing, and web pages. In databases, one technique analyzes the database query issued to obtain the source of data [6]. Other techniques include propagating source annotation along with the data [5] or using provenance polynomials [14]. In the context of scientific workflows, one can show the derivation path of information [7] and a dataset derivation graph [15] or one can use the notion of a strong link to connect a workflow instance with the input and derived data [11]. In the context of grid computing, one may track files and processes [12] or use a web crawler [4]. In the context of the web, multiple data sources may be presented in a web page. To determine the data sources, one can extract provenance metadata embedded on the page [1, 8].

Within Excel spreadsheets, tracking and visualizing dependent data sources within a spreadsheet can be performed with Excel's built-in "Trace Precedents" or "Trace Dependents" interface [13]. A more intuitive visualization of data sources within a spreadsheet has also been proposed [9]. Excel also has a built-in mechanism for linking to external data when a formula, chart, pivot table, or object link is created [13]. However, for the usage scenarios of importing data or copying/pasting data into a spreadsheet, tracking sources is not provided. Another technique can trace the relationships between spreadsheets, but only at the file level [10].

6 Conclusion

In this paper, we presented SourceTrac, a provenance technique for spreadsheets that supports capturing data sources in situ, querying data sources at different levels of granularity, calculating ancestors of data, and visualizing source compositions at different levels of granularity. Preliminary results suggest that this tool has minimal overhead and is beneficial to data analysts and researchers. There are many potential directions for future work. An interesting future work is tracking whether the source data has moved or has been modified. Combining source tracking with provenance techniques for analyzing data manipulations is also another promising direction to pursue in the future.

Acknowledgement

The author thanks Alex Dioso for development support and Dan Jaffe for helpful feedback. Research was supported in part by the University of Washington Royalty Research Fund No. A65951 and the UWB Collaborative Undergraduate Research Grant.

References

[1] EXIF. http://www.exif.org/.

[2] Hazeline U. Asuncion. *In Situ* data provenance capture in spreadsheets. In *Proc of the 7th International Conference on e-Science*, 2011.

[3] Hazeline U. Asuncion and Richard N. Taylor. *Software and Systems Traceability*, chapter Automated Techniques for Capturing Custom Traceability Links Across Heterogeneous Artifacts, pages 129–146. Springer-Verlag, 2012.

[4] Ammar Benabdelkader, Mark Santcroos, Souley Madougou, Antoine H.C. van Kampen, and Silvia D. Olabarriaga. A provenance approach to trace scientific experiments on a grid infrastructure. In *Proc of the 7th International Conference on e-Science*, 2011.

[5] Deepavali Bhagwat, Laura Chiticariu, Wang-Chiew Tan, and Gaurav Vijayvargiya. An annotation management system for relational databases. *VLDB Journal*, 14, 2005.

[6] Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Why and where: A characterization of data provenance. In *Proc of 8th International Conf on Database Theory*, 2001.

[7] Mark Greenwood, Carole Goble, Robert Stevens, Jun Zhao, Matthew Addis, Darren Marvin, Luc Moreau, and Tom Oinn. Provenance of e-science experiments - experience from bioinformatics. In *The UK OST e-Science Second All Hands Meeting*, 2003.

[8] Paul Groth. ProvenanceJS: revealing the provenance of web pages. In *International Provenance and Annotation Workshop*, 2010.

[9] Felienne Hermans, Martin Pinzger, and Arie van Deursen. Supporting professional spreadsheet users by generating leveled dataflow diagrams. In *Proc of ICSE*, 2011.

[10] Carlos Jensen, Heather Lonsdale, Eleanor Wynn, Jill Cao, Michael Slater, and Thomas G. Dietterich. The life and times of files and information: A study of desktop provenance. In *Proc of International Conf on Human factors in Computing Systems*, pages 767–776. ACM, 2010.

[11] David Koop, Emanuele Santos, Bela Bauer, Matthias Troyer, Juliana Freire, and Cláudio T. Silva. Bridging workflow and data provenance using strong links. In *Proc of International Conf on Scientific and Statistical Database Management*, 2010. Springer-Verlag.

[12] Tanu Malik, Ashish Gehani, Dawood Tariq, and Fareed Zaffar. Sketching distributed data provenance. In *Data Provenance and Data Management for eScience*, volume 7092 of *Lecture Notes in Computer Science*. Springer, 2012.

[13] Microsoft Corporation. MS Excel. http://office.microsoft.com/en-us/excel/.

[14] Dan Olteanu and Jakub Zavodny. On factorisation of provenance polynomials. In *USENIX Theory and Practice of Provenance*, 2011.

[15] Leon J. Osterweil, Lori A. Clarke, Aaron M. Ellison, Emery Boose, Rodion Podorozhny, and Alexander Wise. Clear and precise specification of ecological data management processes and dataset provenance. *IEEE Trans on Automation Science & Engr*, 7:189–195, 2010.