Socializing Data with Google Fusion Tables

Hector Gonzalez Alon Halevy Anno Langen Jayant Madhavan Rod McChesney
Rebecca Shapley Warren Shen Jonathan Goldberg-Kidon

Abstract

We describe the social features of Google Fusion Tables, a cloud-based data management service whose goal is to facilitate collaboration around data sets. The social features include the ability to specify attribution of data sets, a mechanism for conducting discussions on data (at fine granularity, such as row, column or cell), the ability to merge tables that belong to different owners, and the ability to share specific queries and visualizations and embed them in other properties on the Web. We describe the rationale for designing these features and our experiences after our first year of interacting with users.

1 Introduction

Google Fusion Tables [7, 8] is a cloud-based service for data management and integration. The over-arching goal of Fusion Tables is to facilitate new types of collaboration around private and public data sets and to enable data management by a broader audience of users. Though we have witnessed a wide range of applications of Fusion Tables, we have seen considerable usage by organizations that are struggling with making their data available internally and externally, and communities of users that need to collaborate on data management across multiple enterprises.

Fusion Tables enables users to upload tabular data files (spreadsheets or CSV) of up to 100MB. The system provides several ways of visualizing the data (charts, maps, time-lines), and the ability to query by filtering and aggregating the data. We support integrating data from multiple sources by performing joins across tables that may belong to different users. Derived tables can be created by joining multiple tables or by projecting a subset of the columns from a single table. Users can keep the data private, share it with a select set of collaborators, or make it public. Fusion Tables also supports an API which many have used to build interesting applications.

As we gathered requirements for Fusion Tables it was clear that provenance and lineage need to be attached to tables in such an open environment. However, we quickly realized that there is a much broader set of features that are needed to support rich social interactions. This paper focuses on the social features of Fusion Tables:

- Attribution: when a user uploads a table into Fusion Tables, she is given the option to attach an attribution to the data. The attribution is the mechanism for specifying the provenance of the data. The attribution is attached to any visualization or table that is derived from the table. Attribution is crucial for two reasons. First, it is important for users to know the source of the data they are looking before they can trust it or use combine it in interesting ways. Second, many owners of data are reluctant to share data unless the attribution is attached. Hence, attribution plays a key role to entice organizations to share their data.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

- Discussion: Fusion Tables enables users to discuss the data at a fine granularity: individual rows, columns and cells. Discussions are important in order to enable users to improve the quality of the data, clarify the assumptions underlying it, or opine on its meaning.

- Data integration: as users find tables produced by others, they want to combine them to create new tables. Fusion Tables enables a user to merge any pair of tables as long as the user has read permissions on it. The system also enables users to inspect the lineage of derived tables.

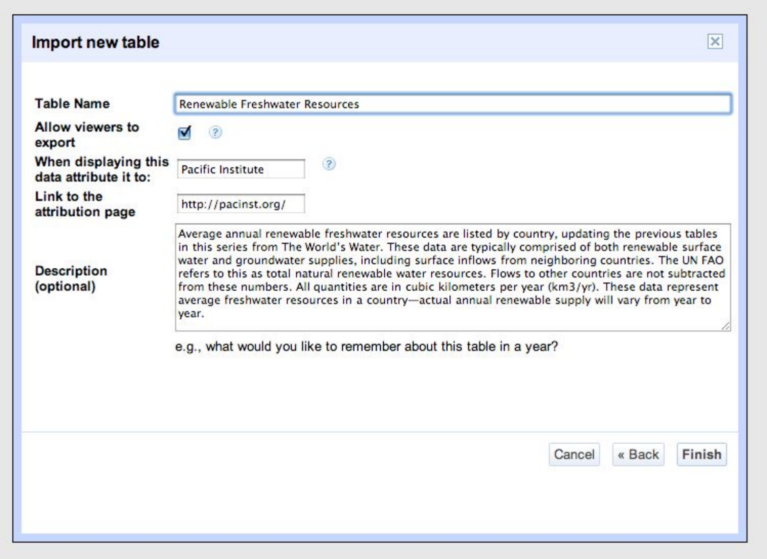
- Publishing: Fusion Tables offers several easy ways of publishing data. A user can create a saved link of a table or a visualization that can be shared via social media. Users can also embed a visualization in other web pages by cutting and pasting an HTML snippet. Publishing data effectively is an important component of enticing people to create interesting data sets.

This paper discusses the social aspects of fusion tables and the rationale underlying their design (Section 2). We then discuss some of the experiences we have had with these features and the challenges that lie ahead 3).

2 Social features of Fusion Tables

We designed the social features of Fusion Tables to facilitate collaboration among multiple parties, the ability to enhance the data, create new data sets, and to provide easy discovery and dissemination of data. In what follows we describe the features of Fusion Tables that are focused on social interactions.

Attribution: In an open data environment such as the Web, it is crucial that data owners get credit for publishing their data and that users know the provenance of the data as they inspect it or use it further. As we were building Fusion Tables, the need to specify attribution came up as a key requirement.



Import new table

Table Name	Renewable Freshwater Resources
Allow viewers to export	☑
When displaying this data attribute it to:	Pacific Institute
Link to the attribution page	http://pacinst.org/
Description (optional)	Average annual renewable freshwater resources are listed by country, updating the previous tables in this series from The World's Water. These data are typically comprised of both renewable surface water and groundwater supplies, including surface inflows from neighboring countries. The UN FAO refers to this as total natural renewable water resources. Flows to other countries are not subtracted from these numbers. All quantities are in cubic kilometers per year (km3/yr). These data represent average freshwater resources in a country—actual annual renewable supply will vary from year to year.

e.g., what would you like to remember about this table in a year?

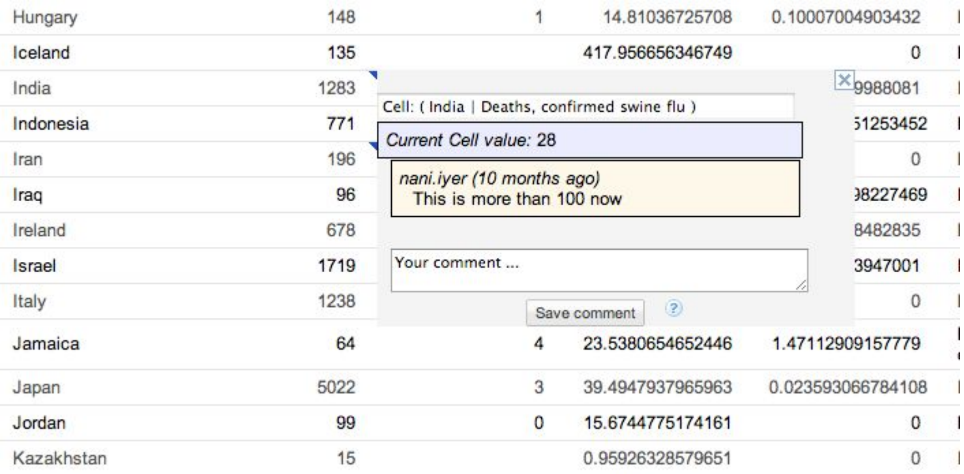
Figure 1: Specifying attribution in Fusion Tables. Users are encouraged to provide the attribution as a string and a URL. They are also encouraged to describe their data during the import process. In addition, users can specify whether the data can be exported into CSV files by other parties.

Users specify the attribution of a table in initial import flow (see Figure 1). Our import flow is designed to have as few steps as possible in order to make import simple. Even so, we felt that specifying attribution is important enough to be included. The attribution includes both a string value and a URL of the source. It is important to note that you do not need to be the originator of the data in order to give the attribution. For example, in Figure 1 a user is importing data that he found (and attributes to) the Pacific Institute. On the Web it is common for users to upload data that is attributed to other organizations.

The attribution is always shown along with the data, whether it is visualized as a table or using a chart or a map. As the data gets processed, the attribution is propagated appropriately.[1] When two tables are joined, the attributions of both tables are shown and the interface on the table view shows the columns from the different sources in slightly different colors. The attribution is also propagated to views and selections of the data. Given the current query capabilities of Fusion Tables, there are no subtle issues related to propagating attribution – the attribution always includes the attributions of all the base tables involved in a view. Users can also inspect how a table is derived from other base tables in the About menu. We note that lineage on views is shown at the level of tables. We do not currently show how each row in a derived table is derived from rows in the base tables.

Discussion: Viewing and querying data are just a first step in a real collaborative effort. In many applications where data is shared among multiple parties, there needs to be significant discussion about the data. The discussion can lead to correcting data values, expressing additional opinions about the data, and elucidating the assumptions underlying the data.

Given that we aim to support large collections of data, it was necessary that we enable users to discuss data at varying levels of granularity – discussing 1 million rows in one block of text would be an ineffective way of keep track of conversations. Hence, Fusion Tables enables conducting discussions at different levels of granularity, including individual rows, columns and cells (see Figure 2).



Hungary	148	1	14.81036725708	0.10007004903432
Iceland	135		417.956656346749	0
India	1283			9988081
Indonesia	771			51253452
Iran	196			0
Iraq	96			98227469
Ireland	678			8482835
Israel	1719			3947001
Italy	1238			0
Jamaica	64	4	23.5380654652446	1.47112909157779
Japan	5022	3	39.4947937965963	0.023593066784108
Jordan	99	0	15.6744775174161	0
Kazakhstan	15		0.95926328579651	0

Cell: (India | Deaths, confirmed swine flu)

Current Cell value: 28

nani.iyer (10 months ago)
This is more than 100 now

Your comment ...

Figure 2: Discussion on cells in Fusion Tables. Discussions can be conducted on individual rows, columns or cells. Discussions are shown as trails, and if the value of the data is changed, then the change is shown as part of the trail to provide context.

Discussions are presented as a record of statements through time, similar to instant messaging interfaces. When a value in a cell changes, the change is shown as part of the discussion trail, and therefore the context

[1] When data is embedded in other web sites, we currently do not show the attribution, mostly because we have not yet found an appropriate place in the UI.

of the change is visible. Fusion Tables also provides a discussion panel, where the user can search through a collection of discussions, rather than having to view each one.[2]

Unlike attribution, we do *not* propagate discussions on a table to tables that are derived from it. The rationale for this design is that a view may be used for a different purpose. For example, you may want to conduct two parallel discussions on a different views of the same underlying data with two distinct set of collaborators. There are cases where users would like to propagate discussions to views, and therefore in, in principle, we could leave the propagation decision to the user (using the mechanisms developed in DBNotes [5]).

One of the emergent use cases of the discussion feature has been to annotate the data with additional explanation. For example, a column may use a convention to describe different health facilities in a country, and a comment on the column can explain the meaning of the different conventions. While the same information could have been put in the text description of the table, putting it in a comment makes it more accessible while viewing the data. However, explanations of the data are likely valuable to all users of the data, and less likely to need to be kept limited to a private set of collaborators. This is an example of where the capability to over-ride the decision not to propagate discussions to views would be useful.

Data integration: The ability to find tables produced by others raises the potential of combining multiple tables. For example, if you own a table about coffee production by different countries in the world, you may want to combine it with tables containing the GDP or population of these countries. More generally, combined data sets enable users to discover deeper correlations or trends. As shown in Figure 3, Fusion Tables currently lets users merge (i.e., join) two tables based on a shared column. When two tables are combined, the permissions on the individual columns are the same as for the base tables.

Some subtle issues arise as we propagate permissions on tables that are derived by merging. Suppose user A wants to grant permissions on table T to user B. If the table T is the result of fusion tables T_1 and T_2, then A may not have write permissions on all of T's columns. Hence, A can only grant write permissions to B on columns of T for which A himself has write permissions. We note that join column used to create a merged table is never directly editable on the merged table itself. All changes to the merge column need to be made on the base tables.

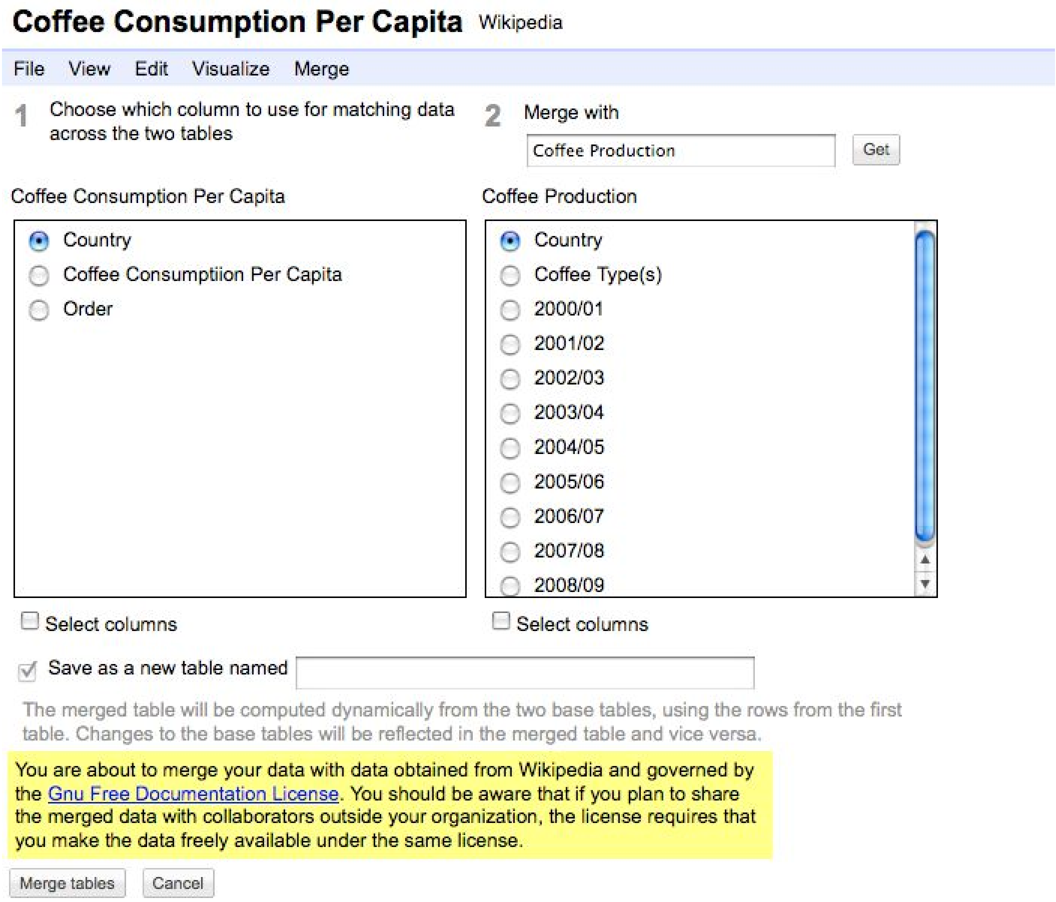
Publishing data: There are several mechanisms for sharing data in Fusion Tables. For starters, an owner of a table can share it with a set of collaborators by specifying their email addresses. The collaborators can be granted read and/or write permissions on the table. The owner can also specify *merge* permissions, where the collaborator can add new columns to the table but cannot change the data in the existing columns. Merge permissions are useful in cases where multiple parties are creating a data set together, but each one has authority over a particular set of columns.

Tables can also be made public to be viewed by anyone. The table owner can decide whether to allow search engines to crawl the table or not. If not, the data may be shared by sending around a link. In addition to showing up in general search results, Fusion Tables provides a search facility over the tables that have been made public and crawlable in the spirit of [6].

We quickly found out that users want share visualizations of the data in addition to the data itself. The reason is that creating a visualization (which may apply to a carefully chosen subset of the table) can convey a particular point and involve more intimate knowledge of the data than can be expected by the viewers. Hence, Fusion Tables enables users to create *saved links* of the data. A saved link is a visualization (e.g., map, bar chart) applied to some query over the data. The link can be shared using any social media (e.g., Twitter or Facebook).

Visualizations of the data can be embedded in other Web properties. For example, as illustrated in Figure 4, a common use case of Fusion Tables is by journalists who want to back up their article by relevant data. Such

[2] At the storage level, comments are stored in a separate table than the data itself. We store all the comments for all user tables in a single Bigtable table. The key of the comments table is the subject of the com- ment, which is the triple: (table, row, column). It uniquely identies the element a comment applies to. The value of a row is the text of the comment, the author, and the date the comment was posted.



1 Choose which column to use for matching data across the two tables

2 Merge with

Coffee Production

Coffee Consumption Per Capita

● Country
○ Coffee Consumptiion Per Capita
○ Order

□ Select columns

Coffee Production

● Country
○ Coffee Type(s)
○ 2000/01
○ 2001/02
○ 2002/03
○ 2003/04
○ 2004/05
○ 2005/06
○ 2006/07
○ 2007/08
○ 2008/09

□ Select columns

☑ Save as a new table named

The merged table will be computed dynamically from the two base tables, using the rows from the first table. Changes to the base tables will be reflected in the merged table and vice versa.

You are about to merge your data with data obtained from Wikipedia and governed by the Gnu Free Documentation License. You should be aware that if you plan to share the merged data with collaborators outside your organization, the license requires that you make the data freely available under the same license.

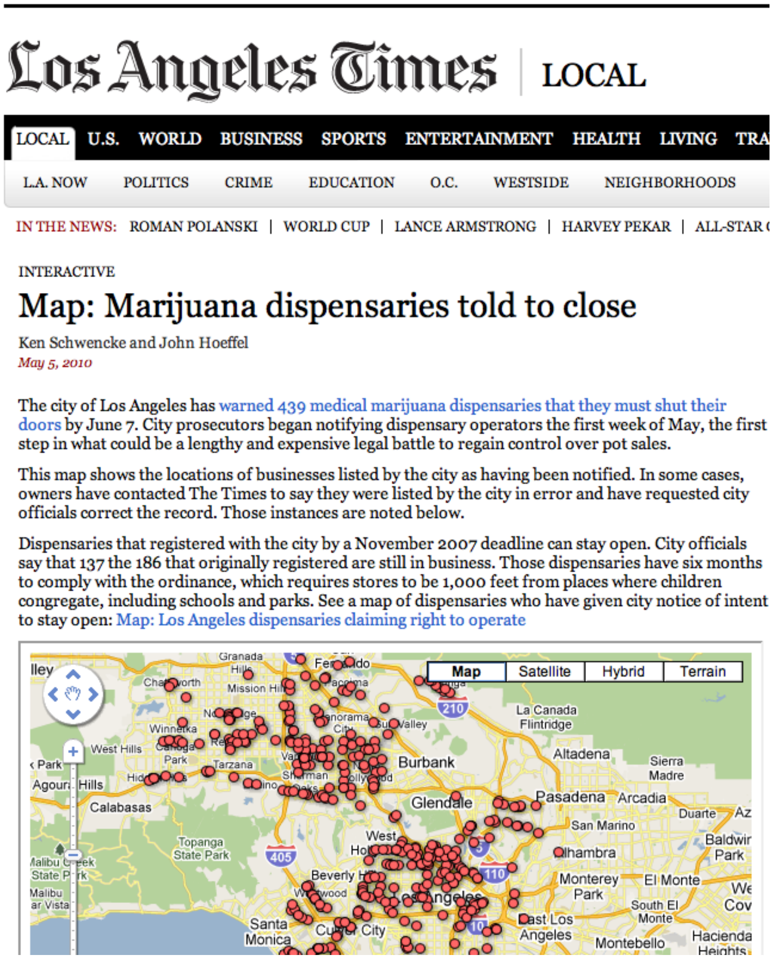
Figure 3: Integrating two tables via a merge operation. The user can search for tables to merge with, and then specifies the join column in both tables. In the figure, the user is merging the table Coffee Consumption per Capita with the table Coffee Production on the Country' column.

embeddings are a significant source of our traffic, because the data is served to users who are browsing other popular Web pages. The owner of the data can see how many views a data set received, which is another mechanism to create an incentive for owners to share their data.

3 Experiences and Challenges

We learned several lessons from our experience with Fusion Tables and from users' requests. These lessons have given rise to some challenges in our system and lead to some interesting future work. We summarize the main ideas below.

Attribution and provenance: We observed that 12% of the tables in the system provide some attribution, and among tables that have been made public 34% have attribution. One of the issues that was raised by our users is the desire to distinguish between the case in which the data is uploaded by the actual owner/producer of the data versus the case in which someone uploaded data they found and want to provide the appropriate attribution. Of course, in the current state attribution can be abused, but we have not witnessed that happening yet. Currently, the only way to check whether the data was uploaded by the owner is if their email address matches the appropriate organization.



Los Angeles Times | LOCAL
LOCAL U.S. WORLD BUSINESS SPORTS ENTERTAINMENT HEALTH LIVING TRA
L.A. NOW POLITICS CRIME EDUCATION O.C. WESTSIDE NEIGHBORHOODS
IN THE NEWS: ROMAN POLANSKI | WORLD CUP | LANCE ARMSTRONG | HARVEY PEKAR | ALL-STAR

INTERACTIVE

Map: Marijuana dispensaries told to close

Ken Schwencke and John Hoeffel

May 5, 2010

The city of Los Angeles has warned 439 medical marijuana dispensaries that they must shut their doors by June 7. City prosecutors began notifying dispensary operators the first week of May, the first step in what could be a lengthy and expensive legal battle to regain control over pot sales.

This map shows the locations of businesses listed by the city as having been notified. In some cases, owners have contacted The Times to say they were listed by the city in error and have requested city officials correct the record. Those instances are noted below.

Dispensaries that registered with the city by a November 2007 deadline can stay open. City officials say that 137 the 186 that originally registered are still in business. Those dispensaries have six months to comply with the ordinance, which requires stores to be 1,000 feet from places where children congregate, including schools and parks. See a map of dispensaries who have given city notice of intent to stay open: Map: Los Angeles dispensaries claiming right to operate

Figure 4: Embedding a map visualization from Fusion Tables in a news article. Updates to the underlying table will be automatically propagated to the map visualization.

Given our current experience, we view the attribution feature of Fusion Tables as reasonably effective. We note that comments in discussions also contain the email address of the person commenting. Currently, we do not let users query for comments by owner, though that could be a useful feature.

We see attribution and provenance playing a significant role in socializing data. In particular, there are domains in which the source of the data and the exact method that derived data was computed are crucial. For example, in climate science or bio-diversity many of the discussions among scientists and the messaging of science to the general public is based on derived data sets. But given the diversity of opinions in these fields and the fact that the science is constantly evolving, it must be possible to specify where the data came from and when and how it was derived from other data. Such provenance is necessary in order to ensure that the assumptions underlying the data are appropriate for the implied conclusion, and in order for other scientists to be able to recreate the same results.

Discussion: The experience with the discussion feature has been much more mixed and we believe there are many more challenges there. Our main observation from talking to users is that the role of the discussion feature in collaboration is not yet clear. Specifically, users do not always understand the purpose of leaving a comment and what effect it will have. For example, they do not know whether someone will actually reply or act on a comment that was left on a cell or row. As a result, they resort to the usual ways of discussing data (e.g., email and chat).

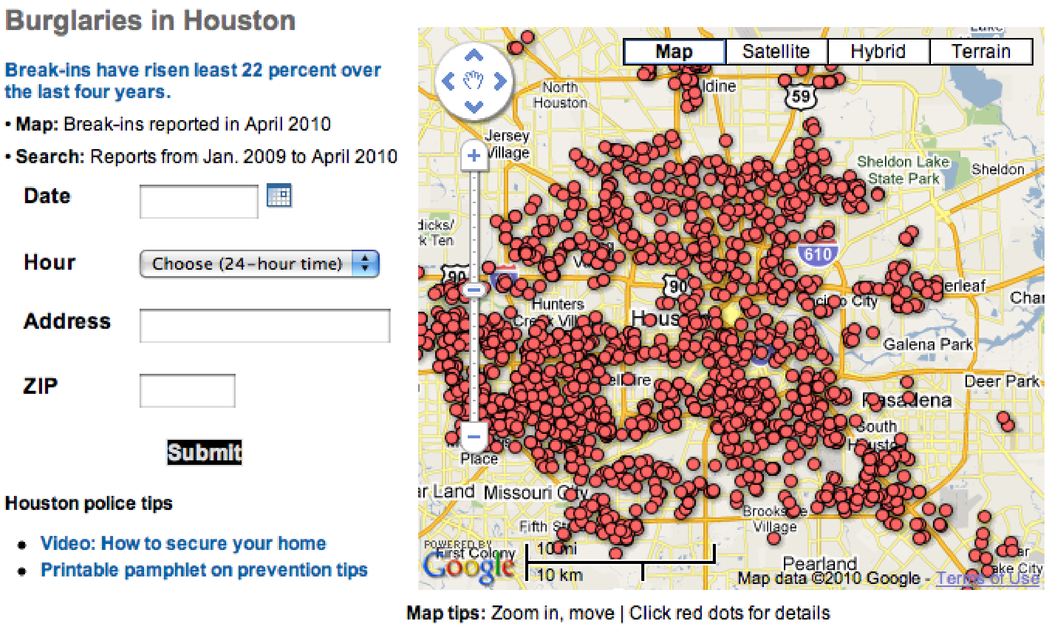
One strategy to address the confusion issue is to provide discussion mechanisms with more specific semantics. For example, we can enable a data owner to set up a comment where she solicits ratings for particular rows. As another example, we can specify that a comment is an explicit proposal to change or correct a value in a cell. The challenge here would be to select a relatively small set of discussion types that would satisfy users' most common needs.

Another feature requested by our users is to be able to provide a comment on a *set* of cells (which cannot be described as a query in a simple fashion). For example, such a comment would specify that a set of values seem inconsistent, or that the cells together exhibit an interesting phenomenon that deserves attention.

Users have also noted the desire to discuss specific visualizations or parts of a visualization. For example, when viewing a scatter plot of a data set you may notice a set of outliers, and it is much more effective to mark them on the visualization itself. Annotating on visualizations raises the challenge of propagating the annotation to the underlying data (and from there to other visualizations, when relevant). Basic functionality for annotating visualizations is supported in ManyEyes [1].

Discussions often refer to specific versions of the data. Fusion Tables does not currently support creating snapshots of the data and referring to them later. Several users have noted that adding snapshots would increase the variety of cases in which our discussion features could be used.

Data integration: Our data integration features are quite basic at this point, and this was an explicit decision when we designed Fusion Tables. The goal was simply to let users see different data sets side by side, without having to perform any schema-level integration. One of the common requests is to allow more flexibility in joining two tables. In particular, the rules for matching pairs of cells from the two join columns should be more flexible. Typically, users merge a pair of tables, but we have seen cases of up to up to eleven base tables being merged[3]



Burglaries in Houston

Break-ins have risen least 22 percent over the last four years.

- **Map:** Break-ins reported in April 2010
- **Search:** Reports from Jan. 2009 to April 2010

Date

Hour Choose (24-hour time)

Address

ZIP

Submit

Houston police tips

- Video: How to secure your home
- Printable pamphlet on prevention tips

Map tips: Zoom in, move | Click red dots for details

Figure 5: A page on the Houston Chronicle that crowd-sources data about burglaries in the Houston area. Readers are asked to report known burglaries on the accompanying web form.

[3] See http://tables.googlelabs.com/DataSource?dsrcid=183101.

Crowd sourcing: One of the interesting applications enabled by Fusion Tables is the ability to crowd-source data that would otherwise be hard to obtain. For example, Figure 5 shows how the Houston Chronicle asks their readers to report burglaries in the Houston area. We have seen several cases where Fusion Tables has been used for this purpose. Crowd-sourcing raises several issues that we will need to address in Fusion Tables, such as being able to determine which data is reliable, which is accurate and up to date and which is redundant. Currently all of these issues need to be addressed in the code that implements the crowd-sourcing.

4 Conclusion

Recently, there have been many calls for an *open data movement*, where data is made public by organizations, curated and enhanced by people who care about it, and then disseminated to a much wider audience. Such open data projects have the potential of affecting decision making throughout many organizations, and making high-quality data more readily available to those who need it.

Fusion Tables and related products such as Factual [2], Socrata [3] and Swivel [4] aim to support the interactions needed for open data using a set of social mechanisms to encourage collaboration. This article discussed the social features we currently support in Fusion Tables, but we note that the area of supporting social aspects of data management is still very much in its infancy. We described some of the challenges we have gleaned from our users, but we are sure more challenges will soon become evident.

References

[1] www.manyeyes.com.

[2] www.factual.com.

[3] www.socrata.com.

[4] www.swivel.com.

[5] D. Bhagwat, L. Chiticariu, W. Tan, and G. Vijayvargiya. An Annotation Management System for Relational Databases, In *VLDB Journal*, 2005.

[6] M. J. Cafarella, A. Halevy, Y. Zhang, D. Z. Wang, and E. Wu. WebTables: Exploring the Power of Tables on the Web. In *VLDB*, 2008.

[7] H. Gonzalez, A. Halevy, C. Jensen, A. Langen, J. Madhavan, R. Shapley, and W. Shen. Google Fusion Tables: Data Management, Integration and Collaboration in the Cloud. In *SOCC*, 2010.

[8] H. Gonzalez, A. Halevy, C. Jensen, A. Langen, J. Madhavan, R. Shapley, W. Shen, and J. Goldberg-Kidon. Google Fusion Tables: Web-Centered Data Management and Collaboration. In *SIGMOD*, 2010.