

Combining Structured and Unstructured Information Sources for a Study of Data Quality: A Case Study of Zillow.com

Irit Askira Gelman
DQIQ, USA
askira@cox.net

Ningning Wu
University of Arkansas, USA
nxwu@ualr.edu

Abstract

Zillow is a web-based, leading real-estate information service in the US. We studied user-contributed facts in a sample of Zillow records. User-contributed information seems to improve the completeness and the level of detail of the information on Zillow.com. However, the accuracy of user-contributed facts may not be high. An investigation of the sources of error revealed several weaknesses, including conceptual challenges, information integration failures, and design deficiencies. A lack of shared, user-friendly, conceptual foundation has been found to be a significant drawback. In part, errors are a product of Zillow's wide geographic coverage and highly networked operation. In addition, important peculiarities of a property are often unknown to the public. Information about such peculiarities is typically shared by a small group of people, whose levels of expertise and stakes in that property, and in real estate in general, may differ. This environment poses a challenge for harnessing the collective intelligence.

The results demonstrate the success of our unique evaluation strategy, which utilizes a systematic review of a rich set of online sources. A similar strategy may also be useful for large-scale error detection and correction, if an efficient automated equivalent is developed to implement it.

1. Introduction

In a preliminary study that compared the records of real estate properties on MLS.com and Zillow.com—two popular, web-based, US real-estate information services—the results showed surprisingly high inconsistency rates [35]. These rates were as high as 40%. However, in the absence of additional information sources, the contribution of that study to an understanding of the accuracy of any of the two databases, causes of inaccuracy, and potential directions for improvement, is limited. A major goal of this work, which focuses on Zillow.com, is to extend

Wu et al.'s study through an evaluation of critical facts that Zillow exhibits about individual properties.

In response to home owners' complaints about the quality of the data that Zillow extracts from public archives across the US, Zillow added tools that enable home owners to edit facts and add information about their property. Zillow also offers listing services for home owners and real estate agents, which enable these users to edit and add information, both manually and through automated data feeds. These tools are becoming increasingly popular. At present, nearly 20% of the records in this store have been edited through such tools.¹ By default, Zillow shows the facts that are supplied by the owner or agent, and these facts are supplemented by public data. Nowadays Zillow also uses the user-contributed facts when computing its home value estimate, the Zestimate. Zillow's website declares: "we've made it easier for our users to help us improve accuracy by incorporating edited home facts into our Zestimate calculations."² Naturally, Zillow's judgment of user-contributed facts raises the hope that a growing utilization of its fact editing option will produce highly accurate data, which, ultimately, can also improve its home value estimates. A key objective of this work is to examine that assertion. This study centers on the quality, mainly the accuracy, of user-contributed facts.

Our paper contributes to a growing literature on the quality and credibility of user contributed information on the Web. Subsequent to an increasing pervasiveness of Web 2.0 [26] applications, the quality and credibility of user contributed information are now attracting substantial research interest. Recent literature has centered on assessments of the credibility of user contributed information, development of an

¹ "Have a Question about your Home or Zestimate? Check out Zillow FAQ" retrieved from <http://www.zillow.com/blog/have-a-question-about-your-home-or-zestimate-check-out-zillow-faqs/2010/02/26/>, April 30th 2010

² "Data Coverage and Zestimate® Accuracy" retrieved from <http://www.zillow.com/howto/DataCoverageZestimateAccuracy.htm>, April 30th 2010

understanding of its strengths and weaknesses compared to alternative, more traditional information sources, identification of factors that can sway the credibility of the information, and, ultimately, convergence on designs that elicit highly credible information. Studies have considered various Web 2.0 instances, including Wikipedia [10,31], question and answer (Q&A) sites [9,27], news blogs [11], health information communities [8], and other applications. Our paper contributes to this literature by presenting an exploratory study of user contributed information in a distinct setting.

Today's reality unlocks new opportunities for information quality research and practice. Our study, in particular, utilizes a diverse set of information sources, including home owners' and real estate agents' volunteered textual descriptions and uploaded photos, public facts, property parcel maps, sale history data, and information about neighboring properties. The availability of varied information sources enables us to triangulate different viewpoints on attributes of the real estate properties. This data assessment strategy has not been reported in the literature and its usefulness for the purpose at hand can not be taken for granted. Hence, an additional contribution of this work is that it serves as an initial study of our chosen assessment strategy.

Broadly, our results suggest the following:

1. The editing tools that Zillow offers do not result in highly accurate data.
2. An analysis of the causes of inaccuracy hints to several weaknesses including fundamental conceptual challenges, information integration failures, and design deficiencies. A lack of shared, user-friendly, conceptual foundation has been found to be a significant drawback. In addition, important peculiarities of a property are often unknown to the public. Information about such peculiarities is typically shared by a small group of people, whose levels of expertise and stakes in that property, and in real estate in general, may differ. This environment poses a challenge for harnessing the collective intelligence [26].
3. Zillow's editing and listing tools enhance the *completeness* and *level of detail* of the data. Agents and home owners supply information about properties that are outside the scope of Zillow's public sources, and the information that they supply often includes details that are missing from public sources.

Our evaluation strategy has yielded a number of potentially useful findings. Clearly, from a data quality perspective, the availability of a rich set of information sources is a significant potential strength. Users' textual descriptions have proved to be a particularly useful source for evaluating data accuracy, identifying

causes of errors, and predicting the correct values. In principle, a similar strategy can also be useful for detecting and correcting errors in millions of property records on Zillow. However, given the enormous number of records, such a strategy can only succeed if an efficient automated equivalent is constructed to implement it.

A principal drawback of this inquiry is the small size of the sample that it applies. While this limitation lowers the reliability of quantitative estimates, we believe that our qualitative insights regarding sources of errors are generally well founded. Future extension should be based on a larger, representative sample.

This paper is organized as follows: The next section adds details about Zillow and its user-contributed data functions. Section 3 offers a literature review. Later, Section 4 describes the research method, and Section 5 reports the results. We end the paper with an account of our conclusions.

2. Zillow

Launched in February 2006, less than five years ago, today Zillow is a leading online real estate marketplace. Zillow aims to help homeowners, buyers, sellers, renters, real estate agents, mortgage professionals, and other stake holders find and share vital information about homes, real estate, and mortgages.

As of April 2010, Zillow's property database covers 96,893,696 US homes, i.e., the majority of the homes in the US.³ A major component of the information that Zillow offers about real estate properties is the "Zestimate," its estimate of the value of a property. Zestimates are computed by proprietary formulas.

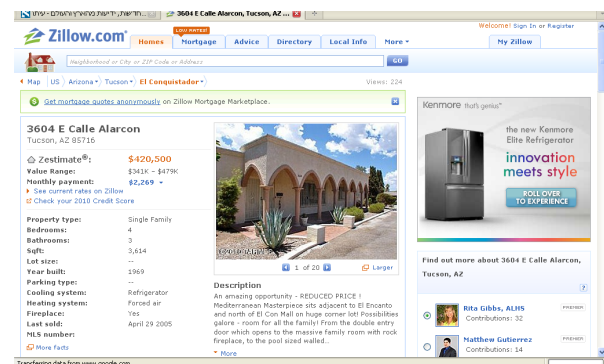


Figure 1. A property page on Zillow.com

³ Retrieved from WWW.Zillow.com ; May 27th 2010

In addition to information on nearly all homes in the U.S., visitors can search homes for sale, homes for rent and recently sold homes, find mortgage solutions on Zillow mortgage marketplace, and more. The business model that Zillow has developed is based on selling advertising and connecting homeowners, renters, buyers, and sellers with professionals. Since its inception Zillow has also steadily expanded its reach through a growing number of business collaboration agreements, and has entered the mobile arena as well. As a private enterprise, Zillow is not required to disclose financial details regarding revenues or expenses. However, Zillow's ongoing reports on its web site indicate that real estate agents account for a substantial portion of its revenues. For instance, at the time this paper was written there were over four million properties for sale by agents on Zillow.

User-contributed data: A considerable part of the information is extracted from countless local public archives. However, many home owners have complained about the accuracy of these data sources. In response, Zillow added tools that enable users to edit the data and create their personalized value estimate on any home, free of charge. For the most part, these updates remain private, part of the personal account of the user. However, home owners can publish information about their own properties. As mentioned in the introduction, Zillow also offers a listing service to home owners and agents. A user can enter the data manually, or, if they use the listing service extensively they can channel the data to an automated data feed. Zillow's automated data feeds draw content from a variety of brokers, agents, multiple listing services (MLSs), and syndication partners. When sources overlap, Zillow essentially looks for the best available source. Specifically, Zillow implements "trumping rules" which favor providers that are most directly connected to the listing source.⁴ Accordingly, Zillow assigns higher priority to home owner and agent-contributed facts over public data. Whenever the former are available, Zillow, by default, displays the edited facts. Data from public archives supplement the user-contributed data, and are shown separately upon request.

Evidently, home owners and agents have been using the editing and listing tools extensively. A Zillow content manager, Diane Tuman, noted on February 26, 2010, that: "nearly 18 million homes on Zillow have had their facts updated by owners or agents to date."⁵

Zillow's President and co-founder Lloyd Frink had commented earlier that: "I have been monitoring the way owners use this feature and must say, I'm pleasantly surprised that the vast majority are not inflating the facts about their homes."⁶ In return, Zillow is demonstrating its trust of owners' contributed facts in an important way—it is currently using the edited data in its computations of the Zestimate.⁷

Zillow's judgment of user-contributed data raises the hope that a growing utilization of the fact editing tools will yield highly accurate data. A key goal of this work is to examine that assumption.

3. Related work

A number of case studies have empirically explored data accuracy in several organizational settings [17,23, 12,18,30,6]. Morey [23] studied the quality of the marine corps manpower management data. Laudon [18] examined the quality of the criminal-record system in the US. DeHoratius and Raman [6] studied inventory errors at a large retail organization.

While our work may be viewed as an empirical study of data quality in an organizational setting, it is also strongly related to recent trends on the Web, especially Web 2.0. Notably, "Web 2.0 is all about harnessing collective intelligence" [26]. However, utilization of the crowds as information sources implies a dramatic shift from traditional information flows, which requires a fresh inquiry of information quality and information quality management. In recent years we have witnessed a surge of studies on these topics. Several studies aim to assess the quality of information in selected Web 2.0 settings (e.g., [10,8,11]); the findings of such inquiries vary. Our paper contributes to this literature through an assessment of a previously unexplored Web 2.0 scenario. Past inquiries of the quality of information of Zillow [13,16,35] ignored that aspect. In addition, a key contribution of this paper is its account of sources of errors, which provides a concrete (though preliminary) basis for a study of potential ramifications and solutions. Our findings suggest that the sources of error are varied such that they are only partly directly related to Web 2.0. One research stream in the Web 2.0 literature that is potentially relevant from this perspective examines the link of "organizational" facets of the information

⁴ Zillownews, "How Zillow Handles Receiving A Listing From Multiple Sources - Trumping Order" retrieved from <http://www.youtube.com/watch?v=zLgLVd9rQlg> , May 20th 2010.

⁵ "Have a Question about your Home or Zestimate? Check out Zillow FAQ" retrieved from <http://www.zillow.com/blog/have-a->

[question-about-your-home-or-zestimate-check-out-zillow-faqs/2010/02/26/](http://www.zillow.com/blog/question-about-your-home-or-zestimate-check-out-zillow-faqs/2010/02/26/) , April 30th 2010.

⁶ Lloyd Frink, "Zestimate Accuracy" retrieved from <http://www.zillow.com/blog/zestimate-accuracy/2006/10/> , April 30th 2010.

⁷ "What is a Zestimate?" Retrieved from <http://www.zillow.com/wikipages/What-is-a-Zestimate/> , April 30th 2010.

contribution process to information quality. Studies in that stream have considered variables such as the number of editors [19,14], contributor incentive schemes [1,14], roles [14,20], and collaboration patterns [20].

As for the research method in use, our method adds to new developments in the area of sample-based data quality assessments [28,3] and potentially also to the growing literature on data quality mining [15,21,29,5, 4]. A distinctive characteristic of our work is its utilization of computer-based information sources that became readily available over the recent years. Today one can often find rich sets of information sources—including structured and semi-structured data sources, texts, maps, images, and more—that enable multiple, complementary viewpoints of the attributes of an object of interest. This new wealth creates new opportunities for accuracy assessment, error correction, and data quality management in general.

4. Research method

If Zillow's fact editing functions actually produce highly accurate data, then, arguably, properties that were recently listed for sale, such that the respective Zillow records have been recently edited by users, would show highly accurate data. Given this hypothesis, our study re-examines a sample of 4,677 MLS properties for sale and their matching Zillow records, which were collected as part of the research by Wu et al. [35].

In conformance with the accepted convention, we say that a recorded value is accurate if it agrees with the true value.

There are two fundamental strategies for investigating errors [25]: *data event analysis* and *error cluster analysis*. Data event analysis involves a review of all the processes that capture or manipulate the data: data capture processes, time durations of data decay, data movement and restructuring, and conversion to information products [25]. Alternatively, error cluster analysis centers on a data subset that contains data suspected of being incorrect. This strategy carefully scrutinizes all the available data in the selected subset in order to identify sources of errors. Error cluster analysis is considered to be a relatively quick and simple strategy.

Since Zillow's information processing internals are not publicly available, data event analysis is not feasible. Subsequently, the method that this inquiry employs largely conforms to error cluster analysis. In particular, given a sample of 4,677 properties as above, we focus on properties whose Zillow records were found to be in conflict with the respective MLS records

on one or more out of four key characteristics: *size of house* (square footage), *number of bedrooms*, *number of bathrooms*, and *year built*. Data that display conflicts with the related records are natural candidates for error cluster analysis, since such a conflict implies that one of the sources is incorrect (possibly both). We found 1,461 properties in this class (over 31%), i.e., properties where the data exhibited at least one conflict between two non-null data values. We selected an arbitrary subset of 70 properties and re-examined their Zillow records in depth. We evaluated the edited facts in this subset using public archive data, descriptive texts, sale history information, photos, maps (mainly maps that show parcel boundaries), and information about neighboring properties. All of these information sets are accessible on Zillow.com. Each inconsistency between Zillow and MLS in any of the former attributes was studied with the objectives of clarifying its causes and identifying the correct value. Specifically, we looked for patterns that would point to the correct data and/or source of error. Whenever this process uncovered a new inconsistency among any of the sources in any of the attributes, the new inconsistency was marked and studied in a similar manner. We have also added a fifth attribute, the *lot size* of a property, as part of this study. Edited lot size facts have been compared against public facts and other Zillow sources.

	This study	Properties that showed inconsistencies	Entire property sample
# bedrooms	< 4: 31% 4-5: 54% >5: 15% based on all 70 properties (no missing values)	<4: 38% 4-5: 51% >5: 11% based on 1,461 properties that showed inconsistencies (no missing values)	<4: 45% 4-5: 48% >5: 7% based on 4,652 properties (25 missing values)
sqft	Average: 3,868 sqft based on 59 properties (11 missing values)	Average: 3,588 sqft based on 1,160 properties (301 missing values)	Average: 3,588 sqft based on 4,351 properties (301 missing values)
year built	>1990: 43% (11 missing values)	>1990: 63% (307 missing values)	> 1990: 61% (1,105 missing values)

Table 1. Property profile

Validity: Given that error cluster analysis is correctly applied, the data subset that describes the 70 chosen properties should demonstrate a *significantly higher* error rate than the error rate of the original sample of 4,677 properties. If sample representativeness had been guaranteed for both sets we would have been able to derive from the observed error rates estimates (lower boundaries) of the error rates in Zillow's property database. However, a study of the data indicates that the samples may not be representative. Table 1 contains profiles of the entire property sample and subsets that served this investigation. This table implies that the data are not representative of the populations from which they have been taken (e.g., in 2007 the average size of a new single-family home in the US peaked at 2,479 sqft⁸). Evidently, extensive quantitative conclusions from this preliminary study are not encouraged.

When a data sample is collected at one time and a related data sample is collected at a later time, there is a risk that an analysis of the data will be affected by changes due to the passage of time (the real estate property may have changed). However, in our research the time period between Wu et al.'s study and this investigation has been relatively short, approximately one year. Therefore, the passage of time has probably not played a major role in this research.

Finally, the validity of our judgments regarding the correct data values and sources of error is not ensured; further tests are needed in order to substantiate these conclusions.

5. Results

87 Zillow records matched (by address) the 70 properties that have been selected for this study. Almost half of the records (40) were *created* by users, mainly by agents, either manually or through an automated feed (Table 2). User-created records contain no public facts, only edited facts, about basic characteristics of the house and land. In most cases (30 properties), user-created records described properties that would not have been covered otherwise by Zillow. Furthermore, when public data were incomplete in records from public archives, e.g., when the public facts about a house were limited to its size, or lot size, edited facts often filled the gaps. As will be illustrated below, user-contributed information (texts, photos) is also characterized by rich detail which is not available through public sources. Therefore, taken as a whole, user-created records, and user-contributed information

in general, appear to enhance both the completeness (see also Table 3) and the amount of detail of the information on Zillow.com.

However, the record set that we have extracted reveals a negative side of user-created records, too. A subset of this record set, including over one third of the records (30), was such that two, three, or even four records described a single property (Table 2). Specifically, 13 properties (19%) had multiple records. A property in this category was described by two or more user-created records (observed in four properties), or by one or more user-created records and a record that was generated from public sources (four properties), or by duplicate records from public sources (five properties).

Our findings indicate that the scale of this duplication problem is significant. Moreover, as the ensuing sections demonstrate, duplicate records are accompanied by increasing inconsistency and outdated information.

	# of properties	# of Zillow records	# of records originally created from public archives	# of user-created records
Properties with 4 Zillow records	1	4	1	3
Properties with 3 Zillow records	2	6	1	5
Properties with 2 Zillow records	10	20	12	8
Totals:	13	30	14	16
Totals for the entire sample set:	70	87	47	40

Table 2. Information source distribution

Accuracy of Size of House (Square Feet): Edited size facts appeared in 66 of the 70 properties (Table 3). Wu et al.'s data [35] showed 13 inconsistencies between Zillow and MLS (Table 4), and the data that we collected uncovered nine more inconsistencies.

The square footage of a house can be important to the degree that appraisers have been sued over square footage calculations. Price per square foot is often used as a basis for comparisons with other properties, and square footage is a significant factor in property tax assessments and lending decisions.

⁸ National Association of Home Builders (NAHB), retrieved from http://www.nahb.org/fileUpload_details.aspx?contentID=80051, April 30th 2010

The numbers that Zillow draws from public sources, which are also frequently quoted by real estate agents, are considered unreliable. “Those records are a convenient source, but public records were never intended to be used by the real estate industry as a source of square footage. These estimates were created by and for a mass appraisal system” [34]. Unfortunately, if both Zillow and MLS draw their numbers from a single, inaccurate source then consistency between them does not necessarily imply accuracy.

Attribute	# of properties with edited data	# of properties with public data
<i>sqft</i>	66 (94%)	30 (43%)
<i>lot size</i>	53 (78%)	35 (50%)
<i># bathrooms</i>	70 (100%)	24 (34%)
<i># bedrooms</i>	70 (100%)	22 (31%)
<i>year built</i>	49 (70%)	30 (43%)

Table 3. Distributions of edited and public data

Nonetheless, regardless of the source of the data, public or individual user, differences between appraisers are common. The measured square footage can vary depending on the preferred standard, tools, and techniques, and its calculation also involves personal judgment. Interestingly, until 1996 there had been no national standard in the US for measuring the square footage of residential buildings. This standard, now called ANSI Z765-2003, is, however, voluntary, and does not cover apartment multifamily buildings. A central element of this standard is the concept of “finished area.” A finished area is “an enclosed area in a house that is suitable for year-round use, embodying walls, floors, and ceilings that are similar to the rest of

the house” [24]. While the ANSI standard has been gaining popularity, it is not universally used.

A study of the records of three of the properties showed that, in each of them, one address matched two Zillow records and the root of the inconsistency was confusion in the public records—two distinct physical properties were assigned the same address. Naturally, these data exhibited other inconsistencies as well.

Ten other properties showed minor inconsistencies, up to 5% of the higher value, which can be easily explained by variations between appraisers. Five properties exhibited substantial inconsistencies which were accompanied by evidence of changes in the properties themselves. In two of them we point to the more plausible figure (Table 4). The public figure of one such property, in Pennsylvania, was 1,068 sqft, while the edited and MLS data showed 1,352 sqft. The seller introduced the property as “beautifully remodeled” and elaborated on its remodeling. The photos showed a small attic turned into a bedroom and a beautifully finished basement. If these additions satisfy the ANSI standard then the higher figure could be closer to reality as per ANSI Z765-2003.

In a second instance there is evidence to suggest that the figure contributed by the user violates ANSI Z765-2003. The public figure of that Arizona property was 1,514 sqft, while the edited and MLS data showed the higher number of 1,930 sqft. The text explained that the owner was a contractor who renovated this house and, among the improvements, added heating and cooling to a relatively large garage. This description can explain the difference of over 400 sqft between the two numbers above. However, if we choose to obey ANSI Z765-2003 then a garage cannot be included in the calculation of the finished square footage regardless of year-round climate control or any other improvement. Obviously, in order for a user to

Attribute	Inconsistencies in the original records from Wu et al. 2009	Newly discovered inconsistencies	Possible cause identified	Correct value predicted	User's fact incorrect
<i>sqft</i>	13	9	18	2	1
<i>lot size</i>	NA	17	12	9	9
<i># bathrooms</i>	48	3	30	27	23
<i># bedrooms</i>	22	8	11	14	5
<i>year built</i>	4	2	5	1	2

Table 4. Distribution of inconsistencies

update the data accurately they need to be knowledgeable about the relevant criteria. Familiarity with one's own property does not guarantee accuracy in this sense.

Accuracy of Lot Size: Lot size values are typically derived from land surveys, carried out by professional surveyors. Land survey outputs have diverse uses in public decision making and elsewhere; they are used by assessors, planners, engineers, and others. New technologies (e.g., GPS) can yield higher accuracy and/or data collection speed, but the accuracy of the outcome depends on a myriad of factors [33]. The preparation, output, and accuracy of a land survey are governed by various local and national standards that aim to guarantee a known level of data quality.

Edited lot size facts were observed on the Zillow records of 53 properties. Our study uncovered 17 addresses where the edited lot size disagreed with a respective value by another source (Table 3 and Table 4). In five instances, the difference in lot size was relatively small (up to 0.1 acre and 5% of the higher value, e.g., 0.22 acre vs. 0.21 acre). Interestingly, in four of these five instances the edited facts showed the lower number. Our analysis focuses on the remaining 12 properties. The evidence at hand supports the conclusion that the edited facts were incorrect in nine instances (see below). No evidence exists to support either number in three other inconsistencies that have apparently been created when two distinct physical properties were assigned the same address (see earlier).

Inaccuracies due to incompatible uses of the data:

Up to six instances of inconsistent lot size values displayed a pattern that is illustrated by the following two examples. The first example is that of a Vermont property that corresponded to two Zillow records, both created from public sources and one of them flagged for sale. Its sale history suggested that the property had been up for sale for about two years (but never sold), and its sale price had been declining steadily. The property was being offered for sale at half (\$1,495,000) its price tag in 2008 (\$2,995,000). Public data on both records said that the lot had 164 acres. However, an edited number on the record that was flagged for sale, 16 acres, was far below that figure. The accompanying description of this property said "Additional acreage available including large acreage."

A second example refers to a Washington State property that was similarly flagged for sale. The address of that property matched four different Zillow records. According to its sale history, the property has been on and off the market a few times in the past three years (again, never sold). The sale price went down during that time from \$875,000 to \$650,000. One of the records contained the public lot size figure, which was 1.3 acres. The text that accompanied that number

supported it. However, the agent explained that the seller was actually offering "5.3 acres... The property is separated into two separate tax parcels. The second four acre parcel can be built on." A study of the boundaries of this property and neighboring properties verified this description. However, since that record was not flagged for sale, the former offer was outdated. The edited facts and matching agents' descriptions in two other records, neither of which flagged for sale, claimed that the property had 1.5 acres. Finally, the edited lot size in the record that was flagged for sale was 1 acre only. Apparently, buyers would assume a lot size of 1 acre. Nonetheless, given three (or four) conflicting numbers, we argue that the public figure of 1.3 acres, which was supported by a detailed, coherent explanation, was the accurate number. In other words, the figure of 1.3 matched the legal state of the land. According to this interpretation, the viewpoint which was expressed in the first record, which focused on the unit for sale rather than the legal unit (the sale encompassed two parcels rather than one) is reflected also in the other records. All the numbers other than 1.3 describe the "product" that the agent was offering—in this case, the product deviated from the legal state of the property. A similar explanation may hold true for the first example. While the property was legally 164 acres in size, only part of it was being offered at the specified asking price.

The examples hint to a sad reality in which the values of properties have decreased dramatically over a relatively short time period. In the face of this dramatic depreciation, people turned to creative solutions in order to get the most out of their assets. For the purpose of this study, however, these examples reveal the possibility of a significant tension between two different uses of the data. While millions consider Zillow as a source of information about the present state of the world (of properties), others, including sellers and real estate agents, approach the information on Zillow from a sales perspective. In essence, the lot that a seller is offering does not necessarily agree with the legal lot—the seller may prefer to combine multiple lots, or, alternatively, sell lot fractions.

Unfortunately, Zillow's design scheme fails to support these distinct perspectives, and that neglect appears to have a negative effect on accuracy. In particular, the edited facts deviate from legal reality.

Inaccuracy due to poorly designed data entry interface: Three instances of inconsistent lot size values are attributed to data entry errors. In one case, lot size units were chosen incorrectly. As a result, a lot that had 0.72 acres according to public data (31,363.20, sqft), was a sprawling 31,363.20 acre property according to the edited facts. A study of the parcel map of this property in comparison to neighboring

properties verified that similarly sized parcels were all less than 1 acre in size.

In two other cases, the edited numbers showed the unreasonably small numbers of 158 sqft and 206 sqft, and the records contained no public facts. A study of the parcel map, mainly this property and neighboring properties of a similar lot size, concluded that the lot size values were probably missing their last two digits (zeros?). In summary, all three inconsistencies appear to have been caused by errors in the edited data. These errors hint to potential design flaws in Zillow's data entry interface.

Accuracy of Number of Bathrooms: All 70 properties had records that contained edited bathroom facts. The data collected by Wu et al. [35] showed 48 properties with inconsistencies between MLS and Zillow. In addition, the new data revealed three more properties whose bathroom data suffered from inconsistencies among the sources. However, in 23 properties the inconsistencies were minor (up to 0.75 bathrooms and 25% of the highest value). Our investigation produced possible explanations for 30 of the cases, and in 27 of them we also predict the correct values; in 23 instances there is reason to believe that an edited fact was inaccurate (Table 4).

In the US, the term "bathroom" is refined to convey additional information about the features that the bathroom contains. Four fixtures are considered: sink, toilet, shower, and bathtub. The US National Association of Realtors⁹ (NAR) proposes the following definition: "A 'full bathroom' is a room with a toilet, a sink and a bathtub. A 'three-quarter bathroom' has a toilet, a sink and a shower. A 'half bathroom' or powder room has only a toilet and a sink."¹⁰ There is, however, variation across markets in the US with regard to the interpretation of these terms. For instance, a bathroom that contains a sink, a toilet, and either a shower or a bathtub but not both (i.e., three fixtures out of four) is sometimes categorized as a "three-quarter bathroom," while a full bathroom requires all four fixtures. The term "one-quarter bath" occasionally designates an area with one fixture, e.g., a sink and vanity area. Nowadays, as homeowners are adding new types of fixtures to their bathrooms, the traditional categories are viewed by some as unsatisfactory.

A common coding scheme aggregates the numbers that match the bathrooms in a house in order to determine the total number of bathrooms. For instance, a house that has one full bathroom and two three-

quarter bathrooms, has, according to this coding scheme, 2.5 bathrooms. Obviously, the information that this coding scheme provides is ambiguous, e.g., the number 2.5 can also describe a house that has two full bathrooms and one half bathroom, as well as other configurations. A second, newer, coding scheme specifies the number of full bathrooms to the left of the decimal point and the number of half bathrooms to the right of the decimal point (e.g., the number 2.1 means two full baths and one half bath).¹¹ The expressive power of this coding scheme is somewhat limited as well.

As this discussion implies, a service provider that reaches beyond local markets (here, a national provider) would face substantial challenges with regard to the core definitions, as well as the coding of the number of bathrooms. Zillow has not addressed these challenges—it obeys the simple aggregate coding scheme despite its significant ambiguity, and has not engaged in the semantic challenges.

The majority of the inconsistencies (up to 30 properties) can be linked to these issues. In contrast to Zillow, MLS enables a detailed account of the bathrooms by category. For instance, MLS data about the Washington state property that we described earlier showed two full bathrooms, one half bathroom, and one three-quarter bathroom. This specification was verified by an agent's textual description on Zillow. In terms of Zillow's aggregate code scheme, the total is 3.25. Arguably, some users may not have been fully familiar with NAR's definitions or with the former aggregate scheme, while others might have found the previous number confusing and looked for a better representation. Accordingly, the edited facts were somewhat, though not drastically, different from the value based on such conventions. The property matched four Zillow records such that on two records the edited facts showed the number 3.5, and a third record showed the number 4 (a fourth record had no bathroom data). We predict that the data by MLS, which were supported by the text, were valid.

Another example: a Virginia property which was described on three Zillow records had, according to MLS, five full bathrooms and two half bathrooms. A conversion of this specification to Zillow's code yields the number 6. The edited facts on two Zillow records agreed with this number. However, on a third record the edited facts showed the number 5.5, and, moreover, the textual description mentioned seven bathrooms. We propose that the specification by MLS, which was backed by the edited facts in two of the three existing records, is valid.

⁹ North America's largest trade association, whose members, including NAR's institutes, societies, and councils, are involved in all aspects of the residential and commercial real estate industries.

¹⁰ Retrieved from

<http://www.realtor.com/basics/buy/looking/lingo.asp>, April 30th 2010.

¹¹ Bathroom, Wikipedia, retrieved from

<http://en.wikipedia.org/wiki/Bathroom>, April 30th 2010.

An interesting point is that the descriptive texts of several properties mention various features that go beyond the expressive capacity of the structured data, such as double vanities, Jacuzzi tub, sauna, soaking tub, and custom steam shower.

Accuracy of Number of Bedroom: All 70 properties had records that contained edited bedroom data. Wu et al.'s data [35] displayed 22 properties with an inconsistent account of the bedrooms, and the data that we collected revealed eight more properties whose data were inconsistent. Our investigation has produced possible explanations for 11 of these cases, and in 14 instances we predict the correct value (Table 4).

According to the US National Association of Realtors (NAR), “‘Bedroom’ usually means a sleeping area with a window and a closet, but the definition varies in different places.”¹²

A few (up to 11) of the inconsistencies hint that people struggle with the prevailing interpretations of the term “bedroom”. For example, one property in South Carolina was described by MLS as having three bedrooms, while the respective edited fact on Zillow said that it had two bedrooms. The textual description explained: “.. 3 bedrooms. Two with queen beds and one with bunk beds for the kids but no closet..” Plainly, the third room does not satisfy a common requirement and the edited number conforms to that requirement (i.e., the edited fact is correct). In another instance the edited fact reflects a compromise of the definition, such that it is incorrect. The edited fact shows that the property has four bedrooms, although an explanation makes clear that the fourth room is, in fact, a “bonus room over garage that is used as 4th bedroom..” In other words, the fourth bedroom does not meet the requirements on a bedroom.

Old houses, in general, exacerbate the challenge that the common definitions pose, since bedrooms did not include closets in the past—instead, people used *armoires* that could be moved around.

In conclusion, the high rate of properties that present symptoms similar to these examples implies that the common definitions are barely serving their purpose when one describes the number of sleeping areas. Variations across different markets add to the confusion.

Accuracy of Year Built: 49 properties had records that contained edited facts about this attribute. The inconsistency rate is lower than the corresponding number in other attributes. Four inconsistencies have been identified in the original data of Wu et al. [35] and two more inconsistencies surfaced through the new

data. Our investigation has yielded possible explanations for four of the inconsistencies, and in one of them we predict the correct values (Table 4).

As mentioned earlier, three of the inconsistencies that Wu et al. [35] found are explained by confusion in the public records where a single address pointed to two distinct legal entities.

The remaining inconsistencies raise concern about the value of allowing users to update this attribute. One substantial disparity has been found in a California property where the edited data said that the house was built in 2000, while the public figure is 1965. The sale history of the property supported the earlier date, and so did a study of the neighboring properties. Although we cannot explain this inconsistency based on the evidence, such evidence implies that the public figure is correct. In a second, Vermont property, which was designated by two Zillow records (both of which were user-created, such that the records did not contain public facts), MLS data showed the year 2007 while the edited numbers on Zillow varied between 2004 and 2005. A third property, also in Vermont, which, again, matched two Zillow records (user-created), was built in 2001 according to both MLS and the edited facts in one of the records. A second record had no edited value for this attribute. However, a textual description in the second record suggested that the house was not built in 2001, but rather, was *remodeled* in 2001. Photos portrayed an older architectural style. If this house had indeed been built years earlier than 2001, its seller could benefit from falsely claiming that it was built in 2001.

Unlike the number of bathrooms, number of bedrooms, square footage, or lot size, the year in which a house was built does not change throughout the life of a house. Therefore, data currency is hardly ever an issue. The accuracy of the data on such an attribute is expected to increase over time [2]. However, the value of allowing users to update this attribute when public data are available is not obvious. If users are not careful enough then the data that they contribute may have lower accuracy than the public figures.

6. Discussion and conclusions

The results of this work indicate that edited facts improve the completeness of the information that Zillow has in store, but the accuracy of Zillow's edited facts is not high. Zillow's proclaimed optimism about the accuracy of edited facts may have been motivated by the assumption that home owners are familiar with their properties. However, an investigation of the causes of inaccuracy portrays a far more complex reality. Numerous factors affect the accuracy of the data aside from home owners' familiarity with their

¹² <http://www.realtor.com/basics/buy/looking/lingo.asp> ; accessed on April 30th 2010.

properties. We have identified several weaknesses including design deficiencies, fundamental conceptual challenges (bedroom, bathroom), information integration failures, and comparable conceptual deficiencies as well as measurement limitations and lack of uniformity among the information sources that supply the data.

The lack of shared, user-friendly, conceptual foundation seems to be a significant drawback. Conceptual inconsistency is a common problem which frequently arises as a result of an expansion or, generally, a change in the organization or industry [25 p.5]. Web presence often stimulates geographic expansion; hence, conceptual gaps such as we have uncovered are a natural outcome in this environment. A popular approach to the lack of uniformity in the definitions of a bedroom and a bathroom, as well as the weakness of the bathroom aggregation schemes would call for a nation-wide concerted standardization effort. This solution has the advantage that it may offer an opportunity to re-assess the existing definitions. However, such a solution would be costly, and may take many years to materialize. An alternative partial solution would be to clarify the conceptual differences and highlight them.

The idea to let the crowds update the data has apparently worked well as far as the completeness and the level of detail of the information are concerned. However, there is evidence to suggest that home owners' lack of expertise and conflicts between users' goals and providers' goals have a negative effect on data accuracy. A distinct characteristic of real estate property information is that, apart from publicly known attributes that a property may share with similar (e.g., neighboring) properties, important peculiarities of a given property may be unknown to the public. Typically, such details are shared by a small group of people, whose levels of expertise and stakes in that property, and in real estate in general, may differ. A question of interest from a research perspective is how to design a Web 2.0 application for optimal information accuracy under these conditions. Web 2.0 literature on a related topic has indicated that a higher number of editors can improve the quality of the information (e.g., [19,14]), and, furthermore, expertise (e.g., [7]), as well as the choice of incentive scheme (e.g., [1,27,14]) can have a significant impact. However, it is not clear if and how these findings apply to our scenario (e.g., can an educational component be part of such a solution such that it actually improves the outcome?). As explained in the introduction, Zillow's opted to restrict the right to edit the information and, in essence, rely on a single editor.

Regarding the evaluation strategy that this research has adopted, mainly its use of a rich set of

information sources, the results demonstrate the success of this approach. We have identified potential sources or causes of inaccuracy in a substantial percentage of the inconsistencies; we have often been able to derive a prediction of the correct value, and the explanations that we offer have practical implications and can guide future research. Obviously, however, an empirical evaluation of the predictions that this work proposes could add to the value of this research.

User-contributed texts have proven to be a particularly useful information source about all the attributes that we have investigated. The usefulness of other information sources has typically been limited to specific attributes (e.g., parcel maps helped with lot sizes). In principle, a largely similar strategy can be useful for detecting and correcting errors in millions of records in Zillow's database. An error cluster analysis of "suspected" data (identified through inconsistency with a chosen source), which employs a variety of information sources, can work as one more remedy to accuracy deficiencies. However, given the enormous size of Zillow's database, this strategy can only work if an efficient automated equivalent is constructed to implement it. Mainly, an automation effort should probably start with textual descriptions, due to their apparent superior informativeness. A few of the textual descriptions that we investigated contained direct specifications, comparable to structured data. Such instances should not be too hard to automate. Broadly, however, the abundance of textual information in the recent years has motivated a new interest in the area of information extraction and the older field of natural language processing [32,22]. Solutions may be found in these and other directions.

A major limitation of this study, which should be addressed through future research, is its utilization of a relatively small sample of properties, whose profile (Table 1) is probably not representative of Zillow's property population. This limitation lowers the reliability of quantitative estimates. Future extension should be based on a larger, representative sample.

7. References

- [1] Anthony, D., S. Smith, and T. Williamson, "Explaining Quality in Internet Collective Goods: Zealots and Good Samaritans in the Case of Wikipedia", *Innovation & Entrepreneurship Seminar*, MIT, Fall 2005.
- [2] Ballou, D.P., and H. Pazer, "Designing Information Systems to Optimize the Accuracy-Timeliness Tradeoff", *Information Systems Research* Vol. 6, No. 1, 1995, pp. 51-72.
- [3] Ballou, D.P., I.N. Chengalur-Smith, and R.Y. Wang, "Sample-Based Quality Estimation of Query Results in

- Relational Database Environments”, IEEE Trans. Knowl. Data Eng. 18(5), 2006, pp. 639-650.
- [4] Bugajski, J., R. Grossman, E. Sumner, and T. Zhang, “A Methodology for Establishing Information Quality Baselines for Complex, Distributed Systems”, 10th International Conference on Information Quality (ICIQ), 2005.
- [5] Dasu, T., T. Johnson, S. Muthukrishnan, and V. Shkapenyuk, “Mining Database Structure; or, how to Build a Data Quality Browser”, Proceedings of the 2002 ACM SIGMOD international conference on Management of data, June 03-06, 2002, Madison, Wisconsin
- [6] DeHoratius, N., A. Raman “Inventory Record Inaccuracy: an Empirical Analysis”, Management Science, Vol. 54, No. 4, 2008, pp. 627-641.
- [7] Edelman, B. “Earnings and Ratings at Google Answers”, Unpublished Manuscript, 2004.
- [8] Esquivel, A., F. Meric-Bernstam, and E.V. Bernstam, “Accuracy and Self Correction of Information Received from an Internet Breast Cancer List: Content Analysis”, British Medical Journal, Vol. 332, No. 7547, 2006, pp. 939-942.
- [9] Gazan, R. “Specialists and Synthesists in a Question Answering Community”, Proc. American Society for Information Science and Technology, 43, 2006.
- [10] Giles, J., “Internet Encyclopaedias Go Head to Head”, Special Report. *Nature*. 438, 2005, pp. 900-901.
- [11] Gunter, B., V. Campbell, M. Touri, and R. Gibson, “Blogs, News and Credibility”, Aslib Proceedings, Vol. 61 Iss: 2, 2009, pp.185 – 204.
- [12] Ham, J., D. Losell, and W. Smieliauskas, “An Empirical Study of Error Characteristics in Accounting Populations”, The Accounting Review, 60, 1985, pp.387-406.
- [13] Hagerty J.R. “How Good are Zillow’s Estimates?”, Wall Street Journal February 14, 2007.
- [14] Harper, M. F., D.R. Raban, J.K. Konstan, and S. Rafaeli, (2008). “Predictors of Answer Quality in Online Q&A sites”, CHI 2008.
- [15] Hipp, J., U. Guntzer, and G. Grimmer, “Data Quality Mining: Making a Virtue of Necessity”, Proceedings of the 6th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2001.
- [16] Hollas, D.R., R.C. Rutherford, and T.A. Thomson “Zillow’s Estimates of Single-Family Housing Values”, The Appraisal Journal, Winter 2010.
- [17] Johnson, J. R., R.A. Leitch, and J. Neter, “Characteristics of Errors in Account Receivable and Inventory Audits”, The Accounting Review, 56, 1981, pp. 270-293.
- [18] Laudon, K. C. “Data Quality and Due Process in Large Interorganizational Record Systems”, Communications of the ACM, 29, 1986, pp. 4-11.
- [19] Lih, A. (2004). “Wikipedia as Participatory Journalism: Reliable Sources? Metrics for Evaluating Collaborative Media as a News Resource”, 5th International Symposium on Online Journalism:16-17
- [20] Liu, J., and S., Ram, “Who Does What: Collaboration Patterns in the Wikipedia and Their Impact on Data Quality”, 19th Workshop on Information Technologies and Systems, pp. 175-180, 2009.
- [21] Luebbers, D., U. Grimmer, and M. Jarke. “Systematic Development of Data Mining-Based Data Quality Tools”, Proceedings of the 29th international conference on Very Large Data Bases Volume 29 (VLDB 2003), pp. 548-559.
- [22] Mooney, R. J., and R. Bunescu, “Mining Knowledge from Text Using Information Extraction”, SIGKDD Explorations (special issue on Text Mining and Natural Language Processing), 7, 1, 2005, pp. 3-10.
- [23] Morey, R. C., “Estimating and Improving the Quality of Information in a MIS”, Communications of the ACM, 25, 1982, pp. 337-342.
- [24] NAHB (National Association of Home Builders) Research Center: American National Standard for Single-Family Residential Buildings: Square Footage–Method For Calculating: ANSI Z765-2003, 2003.
- [25] Olson, J, Data Quality – The Accuracy Dimension. Morgan Kaufmann, San Francisco, 2003.
- [26] O’Reilly, T., and J. Battelle, “Web Squared: Web 2.0 Five Years On”, Special Report, Web 2.0 Summit, San Francisco, 2009.
- [27] Rafaeli, S., D. Raban, and G. Ravid, “How Social Motivation Enhances Economic Activity and Incentives in the Google Answers Knowledge Sharing Market”, Int. Journal of Knowledge and Learning 3, 1, 2007, pp. 1-11
- [28] Rakov, I., “Data Quality and its Use for Reconciling Inconsistencies in Multidatabase Environment”, unpublished PHD, dissertation, 1998.
- [29] Soler, S. V., and D. Yankelevich “Quality Mining: A Data Mining Based Method for Data Quality Evaluation”, 6th International Conference on Information Quality (ICIQ), 2001.

- [30] Sheppard, G.M., and K.A. Brown, "Predicting Inventory Record Keeping Errors with Discriminant Analysis: A Field Experiment", *International Journal of Production Economics* Vol. 32, 1993, pp. 39-51.
- [31] Stvilia, B., M. B. Twidale, L. C. Smith, and L. Gasser. "Information Quality Work Organization in Wikipedia", *Journal of the American Society for Information Science and Technology (JASIST)*, Vol. 59, No. 6, 2008, pp. 983–1001
- [32] Sarawagi, S., "Information Extraction", In: *Foundations and Trends in Databases*, Vol.1, No. 3, 2008.
- [33] Taupier, R., N. Beliaeva, L. Gnadinger, S. Kovalenko, and M. Wheeler, "Parcel Mapping Using GIS: A guide to Digital Parcel Map Development for Massachusetts Local Governments", 1999
- [34] Thomas D. H. "Yes, Size Matters: Inaccurate Square Footage Invites Trouble", *Realtor® Mag: The Business tool for Real Estate Professionals*, September 2008, retrieved from: http://www.realtor.org/rmonews_and_commentary/opinion/0809commentarysquarefootage
- [35] Wu, N., I. Askira Gelman, and I.O. Osesina, "How Consistent is Web Information - A Case Study on Online Real Estate Databases", 15th Americas Conference on Information Systems (AMCIS), 2009.