The Secret Life of Wikipedia Tables

Tobias Bleifuß
Hasso Plattner Institute,
University of Potsdam, Germany
tobias.bleifuss@hpi.de

Leon Bornemann
Hasso Plattner Institute,
University of Potsdam, Germany
leon.bornemann@hpi.de

Dmitri V. Kalashnikov*
dmitri.vk@acm.org

Felix Naumann
Hasso Plattner Institute,
University of Potsdam, Germany
felix.naumann@hpi.de

Divesh Srivastava
AT&T Chief Data Office, United States
divesh@att.com

ABSTRACT

Tables on the web, such as those on Wikipedia, are not the static grid of values that they seem to be. Rather, they have a life of their own: they are created under certain circumstances and in certain webpage locations, they change their shape, they move, they grow, they shrink, their data changes, they vanish, and they re-appear. When users look at web tables or when scientists extract data from them, they are most likely not aware that behind each table lies a rich history.

For this empirical paper, we have extracted, matched and analyzed the entire history of all 3.5 M tables on the English Wikipedia for a total of 53.8 M table versions. Based on this enormous dataset of public table histories, we provide various analysis results, such as statistics about lineage sizes, table positions, volatility, change intervals, schema changes, and their editors. Apart from satisfying curiosity, analyzing and understanding the change-behavior of web tables serves various use cases, such as identifying out-of-date values, recognizing systematic changes across tables, and discovering change dependencies.

Reference Format:
Tobias Bleifuß, Leon Bornemann, Dmitri V. Kalashnikov, Felix Naumann, and Divesh Srivastava. The Secret Life of Wikipedia Tables. In the 2nd Workshop on Search, Exploration, and Analysis in Heterogeneous Datastores (SEA Data 2021).

1 EMPIRICAL RESEARCH ON THE WEB

Traditionally, empiricism plays a minor role in the theory- and engineering-oriented field of our research community, while it has played a significant role in other disciplines of computer science (e.g., [6, 7]). Hardly ever do we pause to analyze and reflect on the observable, "natural" behavior of data and systems. Among the notable exceptions is the area of data quality research [22].

One example of observable behavior is that of web tables, in particular those that are collaboratively edited. As such, an example of a very heterogeneous and semi-structured data lake is the set of tables on Wikipedia. Such web tables are used for a variety of purposes, as was recently surveyed in [10], including entity extraction and fact generation [16], improving web search [21], and entity linking [19]. Other work seeks to enhance web tables themselves, such as generating their title [14], generating column headers [24], or finding subject columns [5, 25]. Again, all of these approaches make use of table content, headers, and surrounding text and data. Providing more such data, and in particular different versions of such data, gives these machine learning approaches a richer input set.

Party	Total seats (change)		Seat percentage
Republican Party	302	+62	69.4%
Democratic Party	131	-61	30.1%
Socialist Party	1	-	0.2%
Independent	1	+1	0.2%
Totals	**435**	**±0**	**100.0%**

Parties		Seats				Popular Vote		
		1918	1920	+/-	Strength	Vote	%	Change
Republican Party	240	302	▲62	69.43%	14,827,891	58.50%	-	
Democratic Party	192	131	▼61	30.11%	9,079,985	35.82%	-	
Socialist Party	1	1	–	0.2%	678,944	2.68%	-	
Farmer–Labor Party	1	0	▼1	0%	243,800	0.96%	-	
Prohibition Party	1	0	▼1	0%	114,057	0.45%	-	
Others	0	1	▲1	0.2%	401,289	1.59%	-	
Total		**435**	**435**	**0**	**100.0%**	**25,435,966**	**100.0%**	**-**

Figure 1: An example of an evolving table in Wikipedia (from https://en.wikipedia.org/?diff=prev&oldid=541341520).

In the context of our Janus project [3], we have been extracting and working with the histories of various structured datasets, including DBLP, IMDB, open government data, and in particular Wikipedia, for which a detailed history of every edit is available. In this empirical paper, we focus on tables as they appear on Wikipedia pages and report on our various observations across their lifetime, including their creation (Section 4), their evolution over time (Section 5), and ultimately their deletion (Section 6). We report on such varied dimensions as table-counts, users, duration, edits, table similarity, table position, and of course time itself, highlighting expected and some surprising behavior. Figure 1 shows one exemplary evolutionary step (in schema and data) for one of millions of Wikipedia tables.

Our results can help researchers better understand the volatility of web tables: a given table or corpus snapshot is not a stable basis but rather just that: a *snapshot* with a history of changes leading up to it and a future with many further changes. In fact, at the

*Work done while at AT&T Research.

Published in the Proceedings of the 2nd Workshop on Search, Exploration, and Analysis in Heterogeneous Datastores, co-located with VLDB 2021 (August 16-20, 2021, Copenhagen, Denmark) on CEUR-WS.org.

time of writing any given Wikipedia table was changed twice in the past year, on average, but with a standard deviation of 9.1, and some tables changed multiple times per day. But not only the content of tables change, also their schema evolves over time. This information about evolving schema can serve, for example, to identify synonymous attributes [23].

In the following, we highlight selected analyses in this paper and outline their possible implications for researchers:

Need for timeliness. Figure 8 shows how quickly a snapshot becomes outdated. As a consequence, all models that are trained on static snapshots run the risk of quickly becoming obsolete. Efficient methods for updating these are therefore desirable.

Help with maintenance and updating. A large portion of tables is created and maintained by power-users, as can be seen in Figures 5 and 10. Knowledge about update patterns could be used to notify these editors about (potentially) outdated values and the need for updates.

Suggestions for cleaning and improvement. Figure 16 illustrates an example of the inconsistencies that can emerge in tables. Based on the knowledge of how similar tables evolve, one can make concrete suggestions to improve their (in this case) schemata, such as to rename or add certain columns.

2 RELATED WORK

We discuss two types of related work: structured datasets with history and Wikipedia change analysis. There is a lot of research on web tables, so we can only provide a high-level overview in this short paper and refer to surveys and research papers for more details.

Related corpora. Wikipedia provides access to its entire version history, allowing us to track very fine-grained changes. A variety of datasets that also deal with (semi-) structured content have been extracted from the web and Wikipedia before. Multiple corpora of web tables [12, 18] provide extracts of static versions of tables on the web and have since been subject to extensive research [10]. For example, Lautert et al. establish a taxonomy of web tables and thus give an insight into the general structure of static web tables in [17], whereas we focus on the temporal evolution of web tables. The infobox history dataset WHAD [2] comprises structured information on Wikipedia, namely the changes of infoboxes. This dataset is orthogonal to our dataset, since it does not cover general tables, which are more diverse and also more complex in comparison.

Analyzing changes in Wikipedia. The content of Wikipedia has been the subject of much research [20]. While large parts of that research were conducted on static snapshots of Wikipedia, a variety of works analyzes changes on Wikipedia. Both the evolution of content [11] and the evolution of the page link graph [8] have been studied. Specifically, the study of content evolution can help detect conflicts [15] or controversy that may result in edit-wars [9]. The edit histories serve as input to event-extraction [13] and are also valuable for trust assignment [1]. Our approach can provide a better understanding of what has really changed, from which many of these studies should benefit.

3 TABLE CORPUS

To explore tables over time, we need to be able to *track* tables over time, which is a non-trivial task as tables and their context can change over time. We consider tables as objects with an identity, which in contrast to its concrete shape and content, stays constant over time. A table can have multiple versions, where each version is an edit of the previous version. However, tables on the web usually lack a stable identifier, which is why we have to infer that identifier.

We proposed a solution for this identity inference through a table matching procedure. The details of this work are described in [4], where we also evaluate the matching and show that it works much better than related work. Our input for the table change extraction process is a dump of web page versions – either snapshots that have been crawled, and specifically for Wikipedia the complete edit history as a set of XML files[1]. These XML files contain the actual version content encoded as Wikitext, a markup language including table markup, as well as additional metadata for each of the revisions, such as its timestamp, author, and comments.

For every page version, we extract a (possibly empty) list of parsed table nodes. This list of table nodes for every page revision constitutes the input for our table matching. For each page revision, it is necessary to decide whether the tables therein are versions of previously identified tables or entirely new tables. It is not sufficient to consider only tables of the directly preceding version, because tables can be deleted and be restored several revisions (and sometimes several years) later. To determine the quality of our matching, we have manually created a gold standard of table matchings comprising 1,445 tables, with a total of 16,919 distinct table versions selected from 90 pages. We show that the matching works well, matching *all* versions that belong to a given table correctly for 88.58% of all tables in the gold standard. For a majority of the remaining tables, we misclassify only a small number of versions of individual table histories (1 mistake: 6.09%; 2 mistakes: 2.98%). Our matching decisions for *individual* table versions reach >99% F_1-measure.

We next present various statistics based on the 3,471,609 Wikipedia table objects we have collected that have been linked using our matching process. These statistics are based on the Wikidump of September 1, 2019. Both our gold standard and output dataset are available at our project website www.IANVS.org. The different statistics and findings are grouped by the three phases of a table's existence: creation, evolution and deletion. Already these three phases of existence show that establishing a table identity is essential for the following statistics, because without it, it would not be possible to determine statistics that aggregate on a per-table basis (but only on a per-version basis).

4 CREATION

In this section, we focus on the first insert of every table – its creation – even when during its lifetime a table might be deleted and recreated multiple times.

4.1 Where and when are tables created?

Figure 2 shows that Wikipedia pages in their initial years (2001–2003) had almost no tables. Using tables in Wikipedia became more

[1] https://dumps.wikimedia.org/

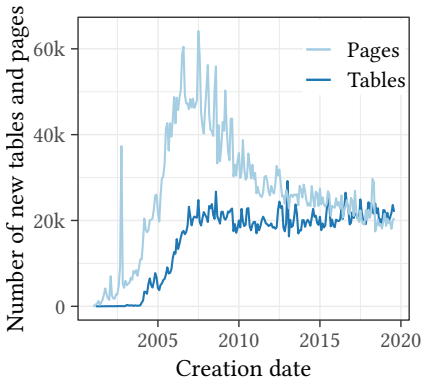

Figure 2: Number of tables and pages created per month.

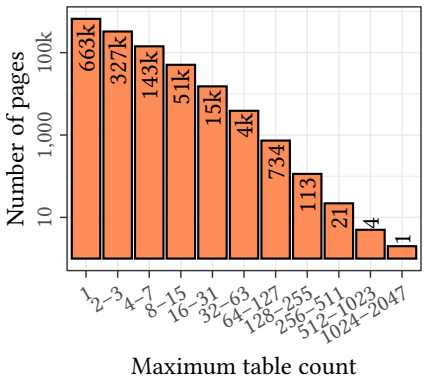

Figure 3: Histogram of the maximum table count per page.

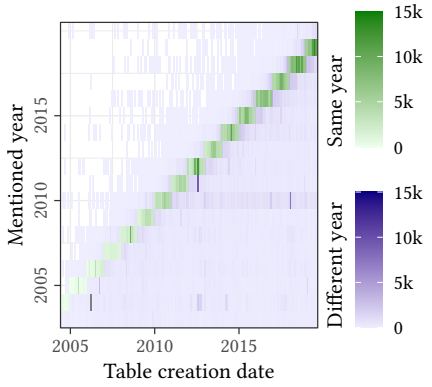

Figure 4: Correlation of years mentioned in categories and time of creation.

popular only around 2004 and tables were fully adopted by end of 2006. Since then, every month around 20,000 new tables are created (about one every two minutes). The hypothesis that insertion frequency would decrease once tables are inserted at all relevant locations seems false: While the number of new pages created per month drops since 2007[2], the insertion-rate of new tables remains constant. This relative increase in tables per page shows that more and more data is stored in a structured fashion, raising the relevance of methods to extract knowledge from said tables.

We observed separately that most tables are created at the same time or soon after the page containing them is created. Only for pages that were created at the beginning of Wikipedia, when tables were not so popular, larger gaps between the page creation and the creation of the first table on the page are common.

Figure 3 shows a histogram of the maximum number of tables that ever existed simultaneously on a Wikipedia article. The vast majority of Wikipedia articles contain only a few tables (we omitted the even larger number of pages that do not contain any tables at all). On the other hand, most tables appear on pages together with other tables. Only 19.1% of all tables appear alone on a Wikipedia article. The many tables that exist in the vicinity of each other can be assumed to be related in terms of content.

On Wikipedia, every article can link to categories, which are used to group related articles to a topic and can themselves be organized in categories. We investigate how the creation dates of tables correlate with any year mentioned in these page categories (such as 2020 for "2020 United States presidential election"), which we assume to be the relevant years for that table. Figure 4 shows that the extracted years and the creation year match for most tables. For every mention of a year in the page categories, a table is counted in a cell that represents the month of creation (on the x-axis) and the mentioned year (y-axis). If those two dimensions would perfectly align, we would only see marks close to the diagonal of the plot. There is a tendency that tables are rather created in the second half of the mentioned year or in the beginning of the following year, which shows as a small shift to the right in the plot. For those years that are covered by our dataset (2004–2019), in 50.8% of all cases the mentioned year and the year of creation align, for 5.6% of the cases the tables are created before the mentioned years and in 43.5% of the cases, the tables are created after the mentioned year.

[2]Source: https://stats.wikimedia.org/#/en.wikipedia.org/contributing/new-pages/normal|line|2001-01-01~2020-10-01|page_type~content|monthly

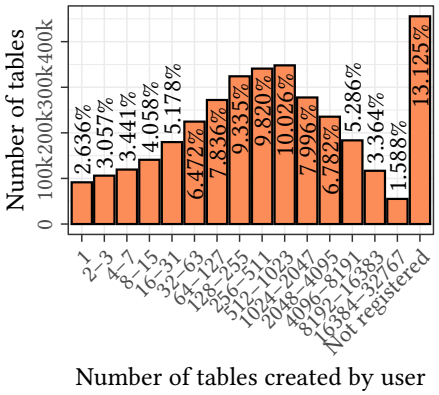

Figure 5: Histogram of tables bucketed by the total number of tables an author created.

4.2 Who creates tables?

The distribution of the number of tables that a user creates is shown in Figure 5. Only 13.1% of the tables are created by non-registered users. The figure also clearly shows that tables are more likely to be created by power-users: More than half of the tables are created by users who each have also created at least 128 other tables. The record for the highest number of tables created by a single user is 20,194 (in this case on a variety of sports events).

A possible explanation for this behavior could be that the effort and skill it takes to create a new table is too high for many users. On the other hand, there are very dedicated users who must have acquired the necessary skills and possibly tools to create thousands of tables. One insight that we can take from this observation is that any random sample of tables is likely to be influenced by those power users.

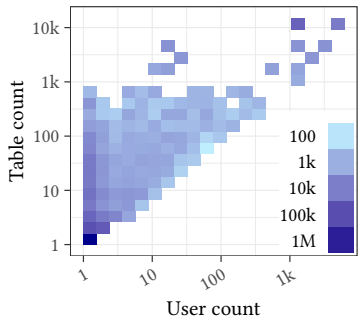

Figure 6: Numbers of users that use the same template for different numbers of tables.

Header 1	Header 2	Header 3
row 1, cell 1	row 1, cell 2	row 1, cell 3
row 2, cell 1	row 2, cell 2	row 2, cell 3
row 3, cell 1	row 3, cell 2	row 3, cell 3

(a) A general table template

Player	G	AB	H	Avg.	HR	RBI

(b) A sport specific table template

Figure 7: Two concrete examples of table templates.

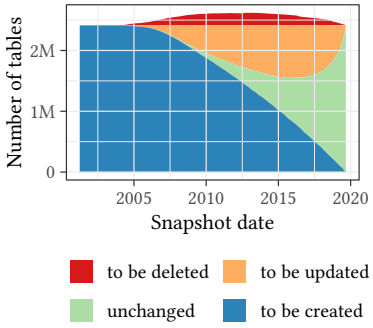

Figure 8: Missed updates in relation to snapshot date.

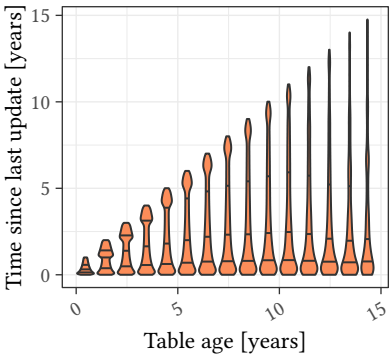

Figure 9: Table freshness over time.

4.3 How are tables created?

Creating tables is a tedious job, especially for inexperienced users, who might not be familiar with the syntax of tables. An obvious hypothesis is therefore that users copy & paste similar tables (created by other users or themselves) and adapt them according to their needs. To investigate this hypothesis, we studied the frequency with which the same content appears in the first version of different tables. For a more accurate picture, we also analyzed how many users chose to use exactly the same table markup code, presumably as table templates.

The first observation we made is that 3,004,883 of the 3,471,609 tables (86.6%) in our corpus appear to have unique first versions. This does not imply that they are not copied from somewhere else, but were modified prior to the first save of the page. On the other end of the spectrum, there are templates that are used more than 15,000 times to create tables.

As can be seen in Figure 6, the ratio of tables and number of users that use the same template greatly varies. Some templates are used for thousands of tables, but also by thousands of different users (top right in the plot). This is usually the case for example tables that contain only dummy values (an example can be seen in Figure 7a). However, there are other templates that are also used for thousands of tables, but only by a few dozen users (top left in the plot). These are usually domain-specific templates, such as tables for sports results or statistics (see Figure 7b for an example).

5 EVOLUTION

The second phase in the lifetime of a table is its evolution. This phase encompasses all changes that happen between the initial creation and the possible final deletion, including changes to data, to schema, and to shape.

5.1 How often are tables changed?

The average table in our corpus is re-inserted 0.62 times, deleted 0.93 times, and updated 13.89 times. Of the 0.62 re-inserts, 0.10 are fresh table versions, i.e., the table's content is different from any previous version, which means 0.52 of the inserts restore previously existing table versions that were deleted at some prior point in time. For the 13.89 updates, the ratio of fresh and old versions is 11.97 fresh versus 1.92 updates that restore previously existing versions. While these average numbers seem quite low, there is a large skew: there is a table on social networking websites that was updated more than 10,000 times during its lifetime. At least 1,310 tables were each updated more than 1,000 times during their lifetimes.

Figure 8 shows the number of tables that would have been created/updated/deleted by the date of our analyzed snapshot (September 1, 2019), if the snapshot were taken at a previous point in time (shown on the x-axis). In a one-month-old snapshot, already 4.4% of tables are outdated. If the snapshot were taken a year earlier, 26.6% of the tables would no longer represent the current state. In a 5 years time range, this number rises to 60.6%.

The violin plot in Figure 9 shows how the change frequency behaves with increasing table age. The shape of a violin plot follows the distribution of the values: the wider the line, the more probable the value. Within the violin plot, there are marks at the 0.25/0.5/0.75 quantiles. In particular, it shows how the time since the last update is distributed for tables at different ages. The median rises until a certain point, after which it stays constant or slightly decreases again. However, the distribution is skewed towards the two ends of the spectrum: tables either are very frequently updated or are hardly ever changed. For example, considering the quantiles for the 5-year-old tables, more than 25% of these tables were updated in

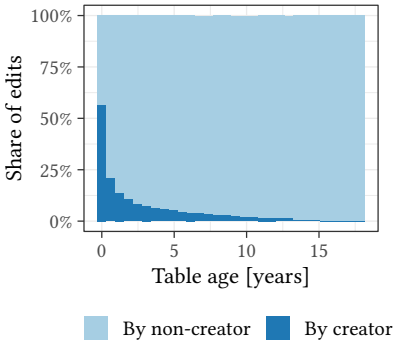

Figure 10: Creator update activity for tables created by registered users.

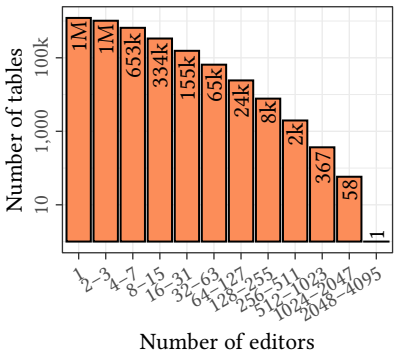

Figure 11: Number of editors per table.

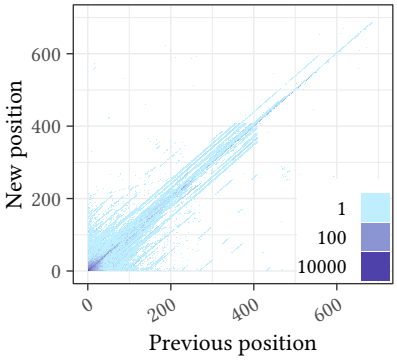

Figure 12: A density plot of table position differences between two consecutive table versions.

the last year and for another 25% the last update was almost more than four years ago.

5.2 Who changes tables?

Figure 10 shows how long the original creator is active as an updater of a table. We distinguish between registered and unregistered creators, because for unregistered creators we have only the IP address as an identifier, which might change from time to time. Therefore, it is not too surprising that the share of edits that is done by an unregistered creator quickly drops and, hence, we exclude tables created by unregistered users from this plot. On the other hand, for registered users, there are tables that are still updated by the original creators years after they have been created. In reality, the influence of the original author on a table could be even higher than what this plot suggests: the number of edits for a table decreases over time, so the first buckets contain more edits.

When we look at the number of editors that change individual tables in Figure 11, we see that a large number of tables (35%) are updated by the creator of the table only. In this analysis, we only consider users as editors of a table who create a new version (a simple revert to a previous version is not counted). While most of the tables are updated by only a few users, there are some exceptions where thousands of users contribute to the table. Again, the previously mentioned table on social networking websites holds the record with contributions by 4235 distinct users.

5.3 Are tables moved?

Figure 12 shows how much tables move in relation to other tables on their page. While for most page revisions, the tables do not move or move only slightly, there are page revisions for which tables move by up to 1,574 positions for a single page (we removed this one extreme case as an outlier). We observe that if tables move, this is often due to the insertion or deletion of tables and that tables rather move down on the page (64.09%) than up (35.91%). One obvious reason for this imbalance is the fact that a table inserted in the middle of the page causes other tables to move down, and insertions are more common than deletions. On average, a table's position changes 1.66 times during its lifespan.

5.4 How much do tables change?

Figure 13 shows how the content of tables develops over time. More precisely, it shows a similarity score of each table compared to the first version of that table (calculated on a random subset of 1,000 tables). We use a similarity metric that is based on a word vector representation of both table versions: $\text{sim}(\vec{v}, \vec{w}) = \frac{\sum_i \min(v_i, w_i)}{\sum_i \max(v_i, w_i)}$, where $\vec{v}$ and $\vec{w}$ are word vectors of the two table versions that should be compared. In general, the similarity is expected to decrease over time, but it can also rise if the table content becomes more similar to its original version. While there are some tables that stay almost unchanged throughout their lifetime, there are other tables that rapidly change within the first few days of their existence. One reason for this could be that people copy & paste other tables as templates and then adjust the content, as explained in Section 4.3.

During their lifetime, 23.6% of all tables either grow or shrink in the number of columns, 37.0% grow or shrink in the number of rows. However, 57.3% of all tables retain their original size throughout their lifetime.

5.5 How much do schemata change?

About half of all tables never change their schema, as can be seen in Figure 14 (note that this is a log-log plot). On the other side, there are tables that change their schema hundreds of times, up to 443 changes. On average, each table has 1.86 schema versions. The types of schema change can be manifold. For example, columns are renamed, columns are added or removed.

Figure 16 shows a vivid example of how schemata of web tables evolve over time. To create this plot, we created a clustering of schemata based on tables that evolve from one schema to another. This particular plot shows a cluster of schemata that all contain information about league results of football teams. There are almost 500 tables for which at least one of the snapshots had one of the Schemata 2–7. More than half of those tables followed Schema 6 at the beginning of 2011, while the other half mostly did not yet exist (Group 1). The splines show how this schema evolved to many different specializations until 2018. While in some cases these specializations make sense (such as a clarification about the league system), in other cases they are due to inconsistent changes (such as

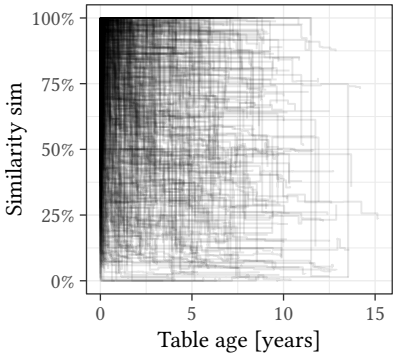

Figure 13: Similarity of table versions and the table's first version. Each line represents one table.

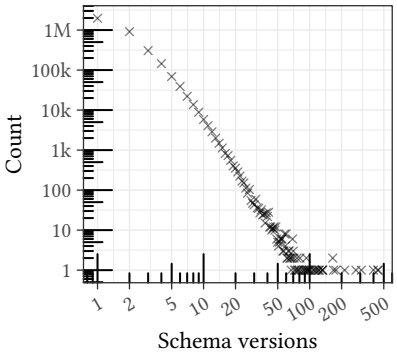

Figure 14: Number of schema versions per table.

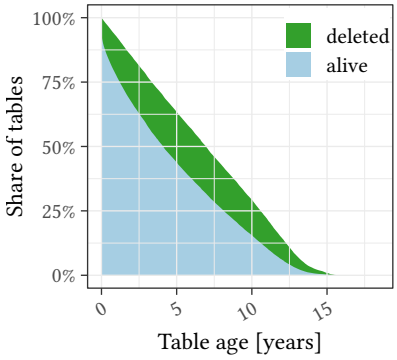

Figure 15: Time until deletion.

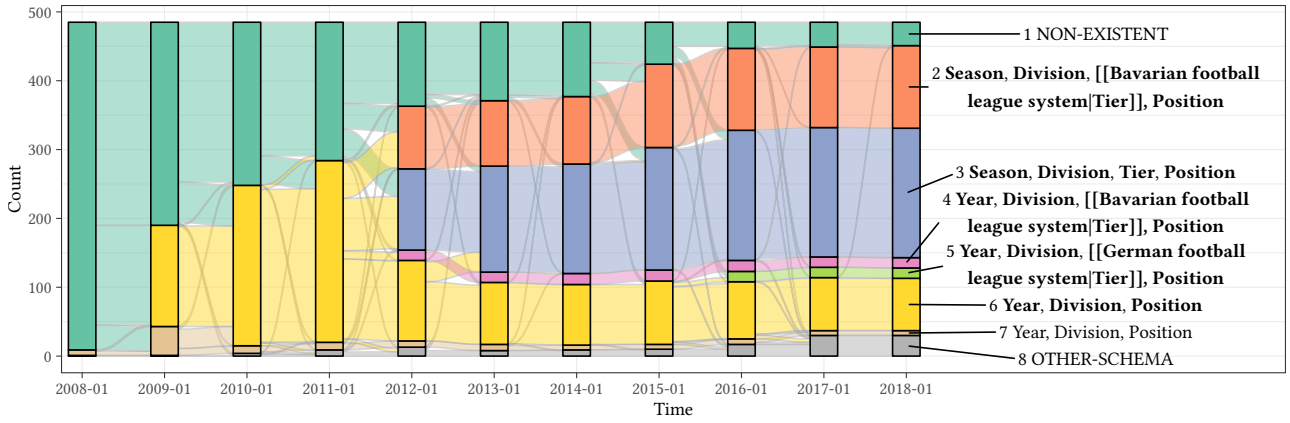

Figure 16: Example of schemata evolving over time.

the header "Year", which after manual inspection should actually be "Season", a range spanning two consecutive years, in most cases). As these tables are webtables, the header can also be formatted differently and we can see that for most tables of Schema 7, the header was changed to bold-type (Schema 6) in between 2009 and 2010. Still, there is a small number of tables that even after almost a decade still did not make this transition.

6 DELETION

Figure 15 shows how long tables survive, counting the days from their creation. The blue part shows the percentage of tables that reached the respective age without being deleted. The green part represents those tables that have been created long enough ago such that they could have reached the respective age, but were deleted before reaching that age. 69.5% of all tables ever created have survived until the end-date of our dataset. If a table is deleted, then this usually happens at the beginning of its lifetime. The longer a table exists, the less likely it becomes that it will be deleted.

From the time a table is created until it is deleted (or until the end-date of the dataset), the average in our table corpus is 4.93 years. In 97.7% of that time, the table is truly part of the page, while in the remaining 40.50 days the table is (temporarily) deleted.

While the vast majority of tables is never deleted (57.2%) or deleted only once (29.9%), there is a larger skew in the distribution of deletes. One table that explains the Wiki syntax was deleted 620 times during its lifetime, mostly from vandalism.

7 CONCLUSIONS

In summary, we have seen how fast tables on Wikipedia change and how fast they come and go. When working with this corpus, it is important to keep this additional temporal dimension in mind and leverage it when possible. The history also makes other dimensions of the corpus accessible, such as the creators, editors, or templates, which together provide a perspective on the tables that is more holistic than single snapshots of individual tables or a table corpus.

As future work, we plan to explore whether other structured corpora, such as Wikipedia infoboxes or lists, for which we also provide histories, behave similarly in terms of their dynamics. Furthermore, we want to use the gained insights to assign trust to values and improve data quality. We encourage researchers to explore their datasets in a similar manner to uncover hidden information in a dataset's history.

REFERENCES

[1] B. Thomas Adler, Krishnendu Chatterjee, Luca De Alfaro, Marco Faella, Ian Pye, and Vishwanath Raman. 2008. Assigning trust to Wikipedia content. In *Proceedings of the International Symposium on Wikis (WikiSym)*. 26:1–26:12.

[2] Enrique Alfonseca, Guillermo Garrido, Jean-Yves Delort, and Anselmo Peñas. 2013. WHAD: Wikipedia historical attributes data - Historical structured data extraction and vandalism detection from the Wikipedia edit history. *Language Resources and Evaluation* 47, 4 (2013), 1163–1190.

[3] Tobias Bleifuß, Leon Bornemann, Theodore Johnson, Dmitri V. Kalashnikov, Felix Naumann, and Divesh Srivastava. 2018. Exploring Change – A New Dimension of Data Analytics. *PVLDB* 12, 2 (2018), 85–98.

[4] Tobias Bleifuß, Leon Bornemann, Dmitri V Kalashnikov, Felix Naumann, and Divesh Srivastava. 2021. Structured Object Matching across Web Page Revisions. In *Proceedings of the International Conference on Data Engineering (ICDE)*.

[5] Leon Bornemann, Tobias Bleifuß, Dmitri V Kalashnikov, Felix Naumann, and Divesh Srivastava. 2020. Natural Key Discovery in Wikipedia Tables. In *Proceedings of The Web Conference 2020*. 2789–2795.

[6] Andrei Z. Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet L. Wiener. 2000. Graph structure in the Web. *Comput. Networks* 33, 1-6 (2000), 309–320.

[7] A.D. Broido and A. Clauset. 2019. Scale-free networks are rare. *Nature Communications* 10, 1017 (2019).

[8] Luciana S. Buriol, Carlos Castillo, Debora Donato, Stefano Leonardi, and Stefano Millozzi. 2006. Temporal analysis of the wikigraph. In *International Conference on Web Intelligence (WI)*. 45–51.

[9] Siarhei Bykau, Flip Korn, Divesh Srivastava, and Yannis Velegrakis. 2015. Fine-grained controversy detection in Wikipedia. In *Proceedings of the International Conference on Data Engineering (ICDE)*. 1573–1584.

[10] Michael Cafarella, Alon Halevy, Hongrae Lee, Jayant Madhavan, Cong Yu, Daisy Zhe Wang, and Eugene Wu. 2018. Ten years of webtables. *PVLDB* 11, 12 (2018), 2140–2149.

[11] Andrea Ceroni, Mihai Georgescu, Ujwal Gadiraju, Kaweh Djafari Naini, and Marco Fisichella. 2014. Information evolution in Wikipedia. In *Proceedings of the International Symposium on Open Collaboration (OpenSym)*. 24:1–24:10.

[12] Julian Eberius, Maik Thiele, Katrin Braunschweig, and Wolfgang Lehner. 2015. Top-k Entity Augmentation Using Consistent Set Covering. In *Proceedings of the International Conference on Scientific and Statistical Database Management (SSDBM)*. 8:1–8:12.

[13] Mihai Georgescu, Nattiya Kanhabua, Daniel Krause, Wolfgang Nejdl, and Stefan Siersdorfer. 2013. Extracting event-related information from article updates in Wikipedia. In *Advances in Information Retrieval (ECIR)*. Springer, 254–266.

[14] Braden Hancock, Hongrae Lee, and Cong Yu. 2019. Generating Titles for Web Tables. In *Proceedings of the International World Wide Web Conference (WWW)*. ACM, 638–647.

[15] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He says, she says: conflict and coordination in Wikipedia. In *Proceedings of the International Conference on Human Factors in Computing Systems (SIGCHI)*. 453–462.

[16] Flip Korn, Xuezhi Wang, You Wu, and Cong Yu. 2019. Automatically Generating Interesting Facts from Wikipedia Tables. In *Proceedings of the International Conference on Management of Data (SIGMOD)*. 349–361.

[17] Larissa R. Lautert, Marcelo M. Scheidt, and Carina F. Dorneles. 2013. Web Table Taxonomy and Formalization. *SIGMOD Record* 42, 3 (2013), 28–33.

[18] Oliver Lehmberg, Dominique Ritze, Robert Meusel, and Christian Bizer. 2016. A large public corpus of web tables containing time and context metadata. In *Proceedings of the International Conference Companion on World Wide Web*. 75–76.

[19] Pei Li, Xin Luna Dong, Andrea Maurino, and Divesh Srivastava. 2011. Linking temporal records. *PVLDB* 4, 11 (2011), 956–967.

[20] Mostafa Mesgari, Chitu Okoli, Mohamad Mehdi, Finn Årup Nielsen, and Arto Lanamäki. 2015. The sum of all human knowledge: A systematic review of scholarly research on the content of Wikipedia. *Journal of the Association for Information Science and Technology* 66, 2 (2015), 219–245.

[21] Rakesh Pimplikar and Sunita Sarawagi. 2012. Answering Table Queries on the Web Using Column Keywords. *PVLDB* 5, 10 (2012), 908–919.

[22] Shazia Wasim Sadiq, Tamraparni Dasu, Xin Luna Dong, Juliana Freire, Ihab F. Ilyas, Sebastian Link, Renée J. Miller, Felix Naumann, Xiaofang Zhou, and Divesh Srivastava. 2017. Data Quality: The Role of Empiricism. *SIGMOD Record* 46, 4 (2017), 35–43.

[23] Paolo Sottovia, Matteo Paganelli, Francesco Guerra, and Yannis Velegrakis. 2019. Finding Synonymous Attributes in Evolving Wikipedia Infoboxes. In *Advances in Databases and Information Systems (ADBIS)*. 169–185.

[24] Jingjing Wang, Haixun Wang, Zhongyuan Wang, and Kenny Q. Zhu. 2012. Understanding tables on the web. In *Proceedings of the International Conference on Conceptual Modeling (ER)*. Springer, 141–155.

[25] Ziqi Zhang. 2017. Effective and Efficient Semantic Table Interpretation using TableMiner+. *Semantic Web* 8, 6 (2017), 921–957.