What the Corpus Can No Longer Find

Legacy YouTube comment IDs in a 2017 research deposit no longer resolve, although most of the comments are still there: an undocumented identifier discontinuity and what it means for ID-only data sharing

Max Sedlmair · ORCID 0009-0007-2471-5595 · Independent Research, Germany · research@woidwide.net

Version 1.1 · 30 October 2026 · Research note, not peer reviewed. Refutations, review, or comments are much appreciated.
Data collected 25 October 2026 (YouTube Data API v3).
This note (all versions): https://doi.org/10.5281/zenodo.23642703 · Verification package (scripts, outputs, design documents): https://doi.org/10.5281/zenodo.23642957
Part of a series on YouTube comment corpora: What the Corpus Does Not Say, https://doi.org/10.5281/zenodo.23339873 · What the Corpus Still Holds, https://doi.org/10.5281/zenodo.23805575

Abstract. Many YouTube comment datasets are shared as lists of comment IDs. Anyone who wants the comments back has to retrieve them through the YouTube Data API, which only works if the IDs still work. I tested this on a real case: the 703,339 comment IDs that Veletsianos et al. (2018) collected in 2017 and deposited with their published study.

The deposit itself is sound. Its published frequency table is reproduced exactly (22 of 22 cells), and an independent reanalysis (Motnikar, 2019) is reproduced to within 0.06 on every mean.

The IDs are not. On 25 October 2026, not one of them could be found: 0 of 3,941 IDs from three pilot videos, and 0 of a random sample of 500 from the whole deposit (95% upper bound 0.74%). The videos are still there (660 of 665), and so, for the most part, are the comments. On ten random videos, the number of comments visible today from the same period equals 87.5% of the deposited number, but these comments carry IDs in a different format. Most comments from that period appear to remain visible, under a different identifier format; the deposited keys no longer lead to them. YouTube has documented no such change, although it announced a comparable one for caption IDs in 2022.

In practical terms, an ID-based survival test of this deposit would have reported a loss of 100% while most comments are still visible. Anyone who reuses or shares YouTube comment data by ID should therefore test a sample of IDs one by one before relying on them, and should deposit the analytical variables alongside the IDs. This deposit is still reproducible only because it contains them.

The note is the second part of a small research programme on YouTube comment corpora:

- *What the Corpus Does Not Say* (Sedlmair, 2026a) asks what a corpus leaves out when it is collected;
- this note asks whether a deposited corpus can still be traced back to the platform years later;
- *What the Corpus Still Holds* (Sedlmair, 2026b) measures, on three younger deposits whose IDs still work, how much of a corpus survives and whether the loss changes its results.

The three papers complement each other: each looks at one stage in the life of a corpus.

Keywords: YouTube; comment IDs; rehydration; data sharing; reproducibility; persistent identifiers; data decay; computational social science

1. Introduction

Platform rules often allow researchers to share identifiers but not content. For Twitter, this produced "dehydrated" datasets: lists of tweet IDs that others rehydrate through the API. For YouTube, the same practice is common. Recent deposits publish comment IDs together with labels and explicitly invite others to retrieve the original comments through the YouTube Data API (e.g. Chan et al., 2026).

The weakness of this practice is known for content. Deleted or protected posts cannot be rehydrated, and the loss is not random:

- Zubiaga (2018) showed that older Twitter datasets lose a growing share of their tweets;
- Bastos (2021) showed that deletions in the Brexit debate were concentrated in particular accounts;
- Küpfer (2024) showed that non-random tweet mortality can change the results of published studies when only IDs may be shared;
- for YouTube, Wang and Weller (2025) tracked the deletion and editing of comments over repeated snapshots.

This note is the second part of a small research programme on what YouTube comment corpora do and do not contain.

The first part, *What the Corpus Does Not Say* (Sedlmair, 2026a), looks at the moment of collection. The API flattens reply structure, interface extraction caps reply depth and reconstructs timestamps from labels such as "1 year ago", and runs silently lose comments. A corpus can look complete and still not say what researchers read into it. That report also commits to re-checking its premises every six months, because tools and platforms change.

The present note looks at what happens after collection, once a corpus has been deposited: can it still be traced back to the platform years later? A third part, *What the Corpus Still Holds* (Sedlmair, 2026b), takes three younger deposits whose IDs still work and measures how much of them survives and whether the loss changes their results (Section 7).

All of this assumes that the identifier itself stays valid, so that a missing item means a missing post. This note reports a case where that assumption fails: the identifiers stopped working as keys, regardless of what happened to the comments. I came across it while preparing a survival test on the deposit of Veletsianos et al. (2018). That test could not be run, and the reason is the finding.

The claim is deliberately narrow:

> Google+-era YouTube comment identifiers preserved in a 2017 research deposit are no longer usable as rehydration keys in the current YouTube Data API, and I found no public documentation of this identifier discontinuity.

2. The deposit

Veletsianos et al. (2018) studied comments on YouTube TED and TED-Ed videos and reported that replies on videos with female presenters were more polar than replies on videos with male presenters. The comments were collected through the YouTube Data API in February 2017. The deposit at BYU ScholarsArchive (Kimmons, 2018; CC BY 4.0) contains:

- per-comment SentiStrength scores and grouping variables;
- a cross-reference from internal comment numbers to YouTube comment IDs;
- video data and video IDs.

It contains no comment texts.

The deposit is intact. Before any API query, I checked that the deposit reproduces what was published:

- **Frequency table.** All 22 cells of the paper's frequency table (Table 3; 354,539 comments and 348,800 replies) are reproduced exactly from the deposited scores.
- **Independent reanalysis.** A reanalysis without replies reproduces all six means reported by Motnikar (2019) to within 0.06, in the correct order.
- **Model.** The paper's multilevel model is reproduced closely (Appendix A).

Every valid row maps to a unique comment ID. IDs without a dot belong to top-level comments, and IDs with a dot to replies, in all cases.

The identifiers. All 703,339 IDs share the same format:

- top-level comments: `z12…` or `z13…`, 32–37 characters;
- replies: `<parent ID>.<number>`, e.g. `z13wen1gxuipflevg22bjbrobyyjhzzzb04.1401760611033308`.

This is the format of the comment system that YouTube introduced with its Google+ integration in November 2013 (Google, 2013). Contemporary examples confirm that the v3 API returned such IDs in 2017 (Truelogic, 2017a, 2017b). The deposit itself is further evidence, since it was collected through the API in that year.

3. Method

All queries used the YouTube Data API v3 with the author's API key on 25 October 2026. Only counts, status codes and ID prefixes were recorded. No comment text, author or other content was stored. In the date count (F), the API response includes the comment text; it was held in memory only and never written.

An API detail that matters. `comments.list` with the `id` parameter accepts up to 50 IDs. If any one of them is not found, the API returns HTTP 404 (`commentNotFound`) for the whole request rather than omitting the missing ID (Google, n.d.-a). A batch failure therefore says nothing about the other IDs. All conclusions below rest on IDs looked up one at a time.

The checks, each fixed in writing before it was run (Section 8):

1. **Pilot (validation gate).** Three videos were drawn with a fixed seed from those with 300–2,000 deposited rows and at least 50 replies. For each:
   - today's top-level threads were listed (`commentThreads.list`, `part=id`);
   - every deposited ID was looked up through `comments.list`; failed batches were split down to single IDs.
2. **Diagnostic, one video.**
   - format of the 20 newest current threads;
   - lookup of one current ID as a positive control;
   - lookup of 1 and of 50 old IDs;
   - lookup of 50 old thread IDs through `commentThreads.list`.
3. **Confirmation run.**
   - **E0:** availability of all 665 videos (`videos.list`).

- **E1:** a random sample of 500 deposited IDs (250 top-level comments, 250 replies; seed 20261024), each looked up individually.
- **E2:** one page of current threads (100, default order) on each of 20 randomly drawn available videos, counting ID prefixes.

4. **Date count (F).** On 10 videos drawn with a fixed seed (seed 20261025; public, 200–1,500 deposited top-level comments), today's top-level threads were paged through from newest to oldest (`commentThreads.list`, `part=snippet`, `order=time`) until a thread older than 7 November 2013 appeared or the list ended. Counted: threads dated 7 November 2013 to 31 January 2017, their ID prefixes, and whether the window was reached. Only counts were stored; comment text was not written.
5. **Documentation search.**
   - the YouTube Data API revision history back to 2014 and the current API reference;
   - developer communities (Stack Overflow; GitHub issue trackers of common comment tools, via web search);
   - the literature on data decay and rehydration.

4. Results

Table 1. Lookup of deposited 2017 comment IDs on 25 October 2026.

Check	IDs tested	Returned	Outcome
Pilot, three videos, all deposited IDs, individually	3,941	0	all "not found"
Diagnostic, one video	1 + 50 (comments), 50 (threads)	0	HTTP 404 in all three requests
E1, random sample, individually	500 (250 top-level, 250 replies)	0	HTTP 404 in all 500

The upper bound of the 95% Clopper–Pearson interval for the share of resolvable IDs is 0.74% for E1 alone and 0.09% for the pilot alone, although the pilot covers three videos and is not a random sample.

Table 2. The videos and their current comments on 25 October 2026.

Check	Result
E0: videos still returned by `videos.list`	660 of 665, all public
E2: current threads listed on 20 random public videos	2,000 (100 per video)
E2: prefixes of current thread IDs	`Ugw` 486, `Ugx` 444, `Ugy` 465, `Ugz` 489, `Ugg` 33, `Ugh` 33, `Ugi` 31, `Ugj` 19
E2: current threads with an old-format (`z1…`) ID	0
Pilot, video with complete listing: deposited top-level IDs found in today's listing	0
Diagnostic: current ID looked up by `comments.list`	found (1 of 1)

In short:

- the videos are there;
- the comment sections are active;
- current IDs resolve;
- the old IDs resolve neither by direct lookup nor appear in current listings.

Table 3. Today's top-level threads dated within the deposit window (7 November 2013 – 31 January 2017), 10 random videos, all listed completely.

Video	Deposited top-level comments	Threads in window today	Ratio
1	295	277	0.94
2	1,179	1,068	0.91
3	283	236	0.83
4	298	303	1.02
5	1,181	1,034	0.88
6	546	478	0.88
7	212	192	0.91
8	1,033	814	0.79
9	227	196	0.86
10	378	329	0.87
Total	**5,632**	**4,927**	**0.875**

The 4,927 threads in the window carry the prefixes `Ugg` (1,212), `Ugh` (1,274), `Ugi` (1,237), `Ugj` (1,203) and `Ugw` (1). The newest threads sampled in E2 mostly carry `Ugw`–`Ugz`. No thread in either run carried an old-format ID.

The ratio is an aggregate, video-level indicator, not a survival rate:

- the counts are not linked comment by comment;
- the window ends before the February 2017 collection, so deposited comments from early 2017 are not matched;
- one video exceeds 1.0, which shows that dates and inclusion rules do not align exactly.

The reply counts of these threads today exceed the deposited replies, because replies can be added to old threads after 2017. They are not interpreted.

Documentation.

- **Revision history.** The YouTube Data API revision history contains no entry on a change, migration or regeneration of comment IDs (Google, n.d.-b). The Google+-related comment entries it does contain concern the removal of a sharing parameter (2015) and of a Google+ profile field (2016).
- **A precedent.** On 5 August 2022 the same history announced that YouTube was “assigning new caption IDs to all caption tracks” and that “on or after December 1, 2022, YouTube will stop supporting the old caption track IDs”, after which old IDs return `captionNotFound` (Google, n.d.-b). YouTube therefore documents identifier regeneration when it treats it as a breaking change. For comments, no such notice exists.
- **The current reference.** `comments.list` still lists an error stating that “the id filter is only compatible with comments based on Google+” (Google, n.d.-a). This wording predates the change and contradicts the behaviour observed here.
- **Developer communities and literature.** I found no Stack Overflow question, GitHub issue or paper reporting that `z…` comment IDs no longer resolve. A widely used R package still shows a `z13…` ID as its example (tuber documentation, 2026).

5. Interpretation

The evidence supports one statement: **for this deposit, the 2017 comment IDs no longer work as keys in the current API.** The breadth of the result rests on two facts:

- the failure is complete: none of the 3,941 pilot IDs and none of the 500 IDs sampled from the whole deposit resolved;
- the videos and their comment sections are still there.

Two readings were possible before the date count:

1. YouTube re-keyed older comments under new IDs, and the old IDs were retired.
2. YouTube removed the old comments, and the comments now visible are later ones.

The date count (Table 3) supports the first. Most of the deposit period is still visible on the platform: 87.5% of the deposited top-level count on ten randomly drawn videos. These comments carry IDs from a distinct `Ugg…`–`Ugj…` range, which is not used for the newest comments. In short, as far as aggregate counts can show, the comments were not lost; the keys to them were.

This is still an inference at the aggregate level. Without matching on text or author, which this study deliberately avoids, it cannot be shown that a particular deposited ID corresponds to a particular current thread.

On the timing. Google shut down consumer Google+ on 2 April 2019 and deleted Google+ comments on other sites. That is a plausible point in time, but no source links it to YouTube comment IDs, and I do not attribute the change to it.

Consequences.

- **For deposits of this era.** An ID-only YouTube comment deposit from the Google+ period cannot be rehydrated through `comments.list`. The deposit of Veletsianos et al. remains fully reproducible only because it contains the analytical variables. Had it contained IDs alone, nothing of the study could be rebuilt today.
- **For current deposits.** Datasets published now with `Ug…` IDs and an explicit invitation to rehydrate (e.g. Chan et al., 2026) rest on the assumption that failed here once. Nothing guarantees it will hold. A follow-up study has since tested this assumption on the Chan et al. deposit and two further deposits (Sedlmair, 2026b). It first checked that the deposited IDs are still valid keys; they are. It then measured how many comments remain and whether the loss changes the deposited results (Section 7).
- **For survival studies.** An ID-based survival test must first establish that the IDs are still valid keys. Otherwise identifier failure will be counted as deletion. Here, a naive ID-based test would have reported a loss of 100%, while most of the comments are still visible. The validation gate used here (Section 3) is a simple way to do this.
- **For the API.** The batch behaviour (one missing ID fails the whole request) makes ID-based rehydration of decayed datasets expensive. Every failed batch must be split down to single IDs.

Practical recommendations.

1. Deposit the analytical variables alongside the IDs wherever the platform rules permit.
2. Record the ID format and the collection date in the codebook.
3. Before any rehydration or survival analysis, test a sample of IDs individually and check that the videos are still available.
4. Report identifier failure separately from content loss.

6. Limitations

- **One deposit, one platform.** Other Google+-era deposits were not tested.
- **Scope of the sample.** The random sample covers 500 IDs, and the pilot covers three videos (plus one diagnostic video). The 95% upper bound for resolvable IDs is below 1%, but rare exceptions cannot be excluded.

- **Lookup methods.** Only `comments.list` and `commentThreads.list` were tested. Other endpoints or the YouTube interface were not.
- **Timing and cause.** The change cannot be dated or attributed.
- **Identity of current comments.** The date count shows that comments from the deposit period are visible under new IDs at the aggregate level. It does not link individual comments, and it cannot separate re-keying from later edits of the date field.
- **The documentation search** may have missed closed issues, non-English forums or internal notices that search engines do not index.
- **Collection setup.** The API calls were made from an Android phone (Termux) over a mobile network. The API result does not depend on the client, and current IDs resolved normally in the same session.

7. Conclusion and outlook: a small research programme

A comment ID is only as durable as the platform's willingness to keep it. In this case, the 2017 IDs of a published and otherwise fully reproducible deposit no longer work, although most of the comments they pointed to are still visible, and no public notice records the change. The general lesson goes beyond this case: a deposited platform identifier cannot be assumed to remain a valid key, and identifier failure must be separated from content deletion before rehydration or attrition is measured.

This note and *What the Corpus Does Not Say* (Sedlmair, 2026a) ask the same question at different points in a corpus's life: what the data can and cannot support.

Table 4. Three stages of a YouTube comment corpus and the question asked at each.

Stage	Question	Study
Collection	What does the extraction deliver, and what does it leave out or reconstruct?	*What the Corpus Does Not Say* (Sedlmair, 2026a)
Deposit, years later	Do the deposited keys still lead back to the platform?	This note
Survival	With valid keys, how much of a corpus remains, and does the loss change its results?	*What the Corpus Still Holds* (Sedlmair, 2026b)

All three share one rule: test a premise before relying on it, fix the test in writing before running it, and report a failed premise as a result.

The follow-up study. *What the Corpus Still Holds* (Sedlmair, 2026b) took the deposit of Chan et al. (2026) and two further deposits with IDs in the current format, collected between 2020 and 2024. The design followed from this note:

1. It first checked that the deposited IDs are still valid keys, by testing a sample individually and confirming that the videos are available.
2. Only then did it measure survival: overall, by comment age, for top-level comments and replies, and by original label.
3. Finally, it recomputed the deposited results on the surviving comments only.

As in this note, every step was fixed in writing before any query was made. In all three deposits the keys still work.

Keeping the premises current. *What the Corpus Does Not Say* re-checks its premises every six months. The next check (May 2027) is planned to include a lookup of a fixed sample of current comment IDs, so that a future change of identifiers is noticed when it happens rather than years later.

8. Provenance

A note from the author. My interest is primarily linguistic: I want to know what YouTube comment data can tell us about language and conversation. That question kept leading back to the data themselves, and so this series has become, I hope, useful measurement work as well. I have worked as carefully as I could and fixed every step in advance. Even so, I may have made mistakes or overlooked work that should have been cited, and I apologise to any author whose study is missing here. I would be grateful to anyone who checks the figures against the deposit and the verification package, and who tells me where I am wrong.

The work began as a planned eight-year survival test of the Veletsianos et al. deposit. Each step was fixed in writing before it was run:

Document	Content
Freeze v0.1	Target study, inputs with checksums, reproduction gates, decision lines
Addenda A1–A3	Reproduction (exact, close) and positive control against Motnikar (2019)
Freeze v0.2	Design of the survival test
Addendum V1	Validation gate, added before any API query because the deposited IDs use an older format
Addendum V2	Gate failure (the pilot)
Addendum V3	Diagnostic run
Freeze v0.3	This note's claim and the confirmation run E, fixed before E was run
Addendum F	The date count, fixed before it was run

The survival test itself was not carried out, because the gate failed. No further analysis was run after the date count.

AI assistance. The author designed the study, set all analytic decisions and decision rules, and revised all text. Claude (Anthropic) was used under the author's direction to draft text and analysis code, which the author then edited. All numbers come from scripts run on the deposited data. The author is responsible for the content.

9. Data and code availability

Original data: BYU ScholarsArchive data/3 (CC BY 4.0).

Scripts:

- `gate_a1.py`, `gate_a2.py`, `gate_a2_std.py`, `step_b.py`, `step_c.py`: reproduction;
- `d_survival.py` (SHA-256 prefix e3b4c0b1f49b6549): the pilot and gate;
- `diag.py`: the diagnostic;
- `e_confirm.py` (SHA-256 prefix e2f3ee3597fbb119): the confirmation run;
- `f_dates.py` (SHA-256 prefix 8b25ce6ad23ad9f5): the date count.

Outputs: the text outputs of all reproduction scripts, and `pilot_2026-10-25.json`, `vel_e_confirm_2026-10-25.json` and `vel_f_dates_2026-10-25.json`, which contain counts and prefixes only.

Design documents: all freezes and addenda, unchanged.

ID bundle: `bundle.txt.gz`, the deposited video and comment IDs, without any other field.

All of these are deposited together as a verification package, with a README that maps every table of this note to its script and output: https://doi.org/10.5281/zenodo.23642957. No comment text, author name or author ID is contained in it.

Appendix A. Further reproduction checks

Model. The original model was a multilevel SEM in Mplus with FIML; its syntax is not published. A linear mixed model with a random intercept per video reproduces the unstandardised presenter coefficients to the second decimal (male → positivity −0.066 against −0.07; male → negativity +0.103 against +0.10).

Standardised coefficients were first computed under STDYX; this failed the pre-set tolerance of 0.10. Under STDY, chosen afterwards and recommended for binary covariates, all four published values were reproduced within 0.10:

- male: −0.46 against −0.52, and +0.54 against +0.62;
- animation: −1.63 against −1.70, and +0.87 against +0.87.

The animation coefficients were not part of the pass rule. This is a close, not an exact, reproduction, and the choice of convention was made post hoc.

Open points. The deposit contains 665 videos with valid rows, while the paper's abstract reports 655. One entry of the paper's Table 5 (51,807) appears to be a typographical inconsistency with the 58,107 parent comments.

Collector cap (sensitivity only). Truncating every video to 2,500 rows (random order or by like count) leaves the published reply contrast robust:

- negativity: +6.4 and +6.7 percentage points against +9.2;
- positivity: +4.4 and +4.1 against +3.9.

Changes in version 1.1

The first page gives a contact address and links to the two companion papers. The abstract, Sections 1, 5 and 7 and the references now cite the follow-up study, *What the Corpus Still Holds* (Sedlmair, 2026b), which has been published; Section 7 describes what it did instead of what it planned. Section 8 adds a note from the author. Unchanged: every figure and every finding.

References

Bastos, M. T. (2021). This account doesn't exist: Tweet decay and the politics of deletion in the Brexit debate. *American Behavioral Scientist, 65*(5), 757–773. https://doi.org/10.1177/0002764221989772

Chan, J., Wang, Y.-L., Kok-Shun, B. V., & Woo, M.-W. (2026). *A dual-annotated YouTube comment corpus on precision fermentation technology (2019–2024)* (Version 2) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.21146055

Google. (2013). *Changes to comments in the legacy YouTube Data API v2* (September 2013, updated 2 December 2013). https://developers.google.com/youtube/articles/changes_to_comments

Google. (n.d.-a). *Comments: list*. YouTube Data API reference. Retrieved 25 October 2026, from https://developers.google.com/youtube/v3/docs/comments/list

Google. (n.d.-b). *YouTube Data API revision history*. Retrieved 25 October 2026, from https://developers.google.com/youtube/v3/revision_history

Kimmons, R. (2018). *YouTube TED talk comment sentiment data* (ScholarsArchive Data 3) [Data set]. BYU ScholarsArchive. https://scholarsarchive.byu.edu/data/3

Küpfer, A. (2024). Nonrandom tweet mortality and data access restrictions: Compromising the replication of sensitive Twitter studies. *Political Analysis, 32*(4), 493–506. https://doi.org/10.1017/pan.2024.7

Motnikar, M. (2019). Comment sentiment associations with linguistic features of educational video content. In *Proceedings of the 22nd International Multiconference Information Society (IS 2019)*. Jožef Stefan Institute. https://is.ijs.si/wp-content/uploads/2019/10/10-Motnikar-Comment-sentiment-associations-with-linguistic-features-of-educational-video-content.pdf

Sedlmair, M. (2026a). *What the corpus does not say: Reply structure, time and coverage in YouTube comment data* [Preprint]. Zenodo. https://doi.org/10.5281/zenodo.23339873

Sedlmair, M. (2026b). *What the corpus still holds: Testing keys, survival and result stability in deposited YouTube comment corpora* [Preprint]. Zenodo. https://doi.org/10.5281/zenodo.23805575

Truelogic. (2017a, 10 August). *YouTube Data API – comments list function*. https://truelogic.org/wordpress/2017/08/10/25-youtube-data-api-comments-list-function/

Truelogic. (2017b, 16 August). *YouTube Data API – comments insert function*. https://truelogic.org/wordpress/2017/08/16/26-youtube-data-api-comments-insert-function/

tuber documentation. (2026). *get_comments*. https://rdrr.io/cran/tuber/man/get_comments.html

Veletsianos, G., Kimmons, R., Larsen, R., Dousay, T. A., & Lowenthal, P. R. (2018). Public comment sentiment on educational videos: Understanding the effects of presenter gender, video format, threading, and moderation on YouTube TED talk comments. *PLOS ONE, 13*(6), e0197331. https://doi.org/10.1371/journal.pone.0197331

Wang, Y., & Weller, K. (2025). Tracking ephemerality in YouTube comments: Towards methods for building dynamic CMC corpora. In A. Fábián & I. Trost (Eds.), *Impulses and approaches to computer-mediated communication: Proceedings of the 12th International Conference on Computer Mediated Communication and Social Media Corpora for the Humanities* (pp. 36–39). University of Bayreuth.

Zubiaga, A. (2018). A longitudinal assessment of the persistence of Twitter datasets. *Journal of the Association for Information Science and Technology, 69*(8), 974–984. https://doi.org/10.1002/asi.24026