TRAVELWARN-Crawler: Constructing longitudinal datasets of government-issued travel warnings for political and social science research

Laura Braun
Center for Crisis Early Warning
University of the Bundeswehr Munich
Neubiberg, Germany
laura.braun@unibw.de

Christian Oswald
Center for Crisis Early Warning
University of the Bundeswehr Munich
Neubiberg, Germany
christian.oswald@unibw.de

Abstract

Historical travel warnings and advisories offer a record of how governments perceive and communicate country-specific risks over time, yet large-scale quantitative analyses have been rare due to missing, fragmented, and frequently overwritten web data. We present *TRAVELWARN-Crawler*, an open pipeline that collects, recovers, cleans, stores, and standardizes advisories and warnings from archived web sources. Using the Internet Archive's Wayback Machine and issuer-specific Scrapy spiders, we reconstruct up to three decades of timelines for the United States (US), the United Kingdom (UK), and Australia. The data can be used to support comparative research in political communication, international relations, and tourism studies. Descriptively, the UK issues more country pages and updates than the US and Australia. Cross-issuer agreement about severity is modest with the highest pairwise level agreement for US–Australia ($\sim$49%). These results indicate substantial heterogeneity in how close allies communicate travel risk and underscore the value of reproducible, textual data beyond numeric severity levels alone.

1 Introduction and Motivation

The U.S. Department of State issued what is widely considered the first modern travel warning when it cautioned its nationals at the outbreak of World War I in 1914 (Löwenheim, 2007). Since the late 1990s, such warnings have increasingly become accessible online and on a regular basis, shaping individual travel decisions and redirecting tourism revenue. Beyond their immediate practical value, they also signal a government's assessment of specific countries. Although they are potentially important, researchers lack comprehensive longitudinal data. Each foreign ministry maintains its own website and frequently overwrite pages in place. There is no common official API for retrieving archived warning texts. The paper demonstrates how to navigate this terrain. We reconstruct advisory timelines for three issuers (US, UK, Australia) by requesting archived captures via the Wayback Machine API, extracting text content and metadata, and processing and standardizing the results to store in a database, to provide a ready-made and easily downloadable dataset.

Government-issued travel advisories represent a unique combination of time-connected threat assessment and official foreign policy signaling. Each text is a judgment of how the issuing state evaluates security, health, and political conditions abroad at a given moment. Since ministries publish these assessments under their own seal, language and severity levels may encode far more than objective risk. Having such textual data allows to investigate whether they also reflect strategic calculations about alliance obligations, latent disputes, or reputational costs (Chu et al., 2021; Babey, 2019; Kebede, 2018; Sharpley et al., 1996). We harvest these text data at scale and reconstruct their full issuing history to create the first longitudinal warning corpus that scholars can reuse and provide a step-by-step guide for extending the dataset to additional issuers. We focus on the US, UK, and Australia as a proof of concept and a foundation for comparative analysis. The countries are three founding members of the Five Eyes intelligence alliance, and we explore how their informal cooperation is reflected in travel warning decisions. Although we start with English sources, recent advances in automated translation may enable the inclusion of other languages.

2 Implementation

2.1 Data access

We use the Internet Archive's Wayback Machine (IAWM) API to retrieve historical digital artifacts of travel warnings and advisories from archived snapshots of provider webpages. The IAWM pre-

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 29–35, 2025

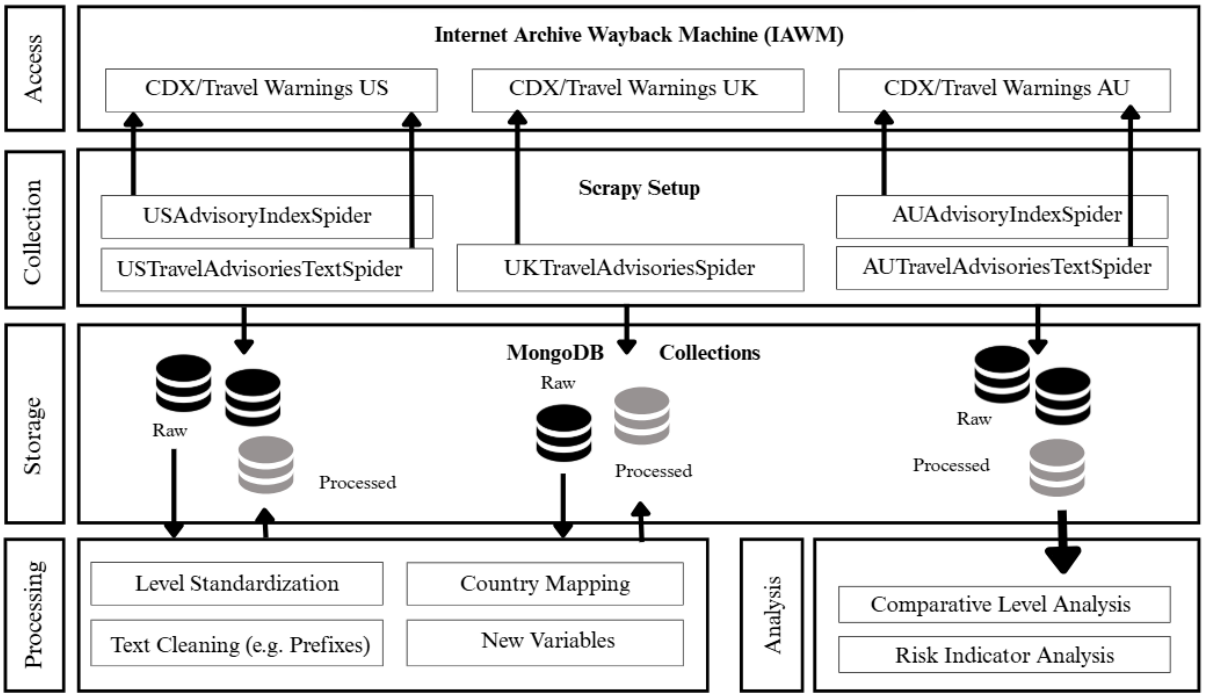

Figure 1: Implementation of dataset generation — Access, Collection, Storage

serves time-stamped URLs since 1996 and indexes more than 800 billion pages.[1] Its archival record turns otherwise ephemeral content into a longitudinal resource for scholars. A validation study demonstrates that the IAWM website age and update count measures are reliable, as the authors find convergent validity with domain-registry dates, confirming the archive's suitability as a research resource (Murphy et al., 2007). Since the foreign ministry websites we analyze went online in 1996 or 1997 (Löwenheim, 2007), the IAWM offers a long archive history with sparse coverage until the mid-2000s and increasingly dense coverage thereafter.

Previous work has outlined a complete social science workflow for sampling, crawl design, variable extraction, and transformation into structured data (Arora et al., 2016). Our approach differs in retrieving the complete archive for each issuer, prioritizing depth over breadth. This introduces practical challenges such as changing URLs, layouts, and HTML structures. We apply custom crawler logic that uses issuer- and time-specific URL prefixes to query the IAWM's CDX endpoint. Other archives are paid services or lack comparable coverage. Some issuers provide partial official archives (e.g., OSAC.gov for U.S. alerts since 2004, the UK Government archive since 2013), but these are either restricted for scientific use due to licensing or offer fewer snapshots. The IAWM with its CDX API is thus the best choice for standardized, near-complete timelines from most issuers, without copyright restrictions.[2]

2.2 Data collection and storage

We implement a modular, automated pipeline in `Scrapy`. Scrapy's abstractions (Spiders, Items, Pipelines) separate crawling from extraction and storage, while built-in concurrency, retry logic, and scheduling enable large-scale, polite crawling of mostly static content, lighter than headless-browser approaches. For the US and Australia, a two-step spider design first queries the CDX API for *one* index-page snapshot per day, extracting country links, update dates, and any listed metadata (e.g. level). For each new (country, update-date) pair, we then fetch the first available archived copy of the corresponding country page to extract full text and metadata. This design (i) detects missed updates even when country pages are sparsely archived, and (ii) avoids unnecessary downloads. For the UK, whose index pages lack explicit update dates, we instead check country pages daily and re-scrape only when their on-page update date changes. We

[1] https://archive.org/about/

[2] IAWM licensing is generally permissive, but site-specific terms should always be verified, although government pages are typically open-licensed.

observed that UK country pages also contain rich, multi-section content (safety/security, health, local laws, etc.), which can trigger additional updates and thus increase revision frequency.

All advisories are stored in MongoDB, one collection per issuer. Each document includes at least `country`, `warning_date`, the IAWM `timestamp_snapshot`, the `source_url`, and the extracted `advisory_text`. We upsert on a compound key (country, update date) to prevent duplicates and preserve a stable pointer back to the archived source for full traceability.

2.3 Data processing

After collection, we standardize the records to allow cross-issuer comparisons. First, we normalize country names to a canonical mapping to accommodate spelling variants and historical names. Second, we harmonize issuer-specific severity descriptions to a unified four-tier scale using regular expressions, explicitly accounting for policy changes over time (e.g. pre-2018 U.S. materials without formal levels; UK advisory phrasing without Level 1/2 labels). We retain the original text for full transparency. Although regex-based mapping is transparent and reproducible, it can be brittle in the presence of negation or nuanced phrasing (e.g. "no longer advised not to travel…"). In practice, a small residual remains: for Australia, n=158 updates could not be mapped; for the United States, n=23 entries are labeled *Other*; and for the United Kingdom, because Level 3 and Level 4 are explicitly indicated, cases ambiguous between Level 1 and Level 2 default to Level 2. In future work, we will evaluate supervised models (e.g. ConfliBERT; Brandt et al., 2024) and LLM-assisted level assessment to reduce regex-based classification noise.

3 Results

We reconstruct longitudinal advisory timelines for the United States, the United Kingdom, and Australia and harmonize them to a common four–level scale. Coverage is complete for the US and Australia from 1997/1998 onward and extends back to 2013 for the UK. Descriptively, issuers differ in both scope and frequency.

We restrict descriptive comparisons to the overlapping ten–year window (2014–2024) in which all three issuers are covered. Table 1 summarizes this period. The UK issues the most updates (15,207) across the largest destination set (225 unique countries), followed by Australia (6,530 with 181 unique countries) and the US (3,101 with 210 unique countries). Consistent with the editorial breadth of UK pages, only ≈4% of UK updates coincide with a level change (611/15,207), compared to ≈8.4% for Australia (547/6,530) and ≈31.9% for the US (990/3,101). In other words, the typical US update is more likely to reflect a substantive change in severity, whereas UK updates more often revise text within informational sections. UK pages also yield a larger set of destinations that remain at Level 4 throughout the period ("constant L4"): 33 for the UK versus 11 for Australia and 5 for the US, reflecting both coverage scope and differing issuer thresholds to maintain a sustained "Do not travel" classification.

Figure 2 tracks the monthly average advisory level for each issuer, smoothing with a 3-month moving average to reduce month-to-month noise. The dashed line marks the US shift to the four-tier system in 2018; the gray band marks the global COVID shock. Two patterns stand out. First, during COVID all three issuers converge upward, indicating a sharp, broad-based tightening that later relaxes at different speeds. Second, outside the COVID window, UK and Australia trend closely together on average, while the US is modestly higher in the pre-2018 period and again during the 2021–2022 normalization phase.

Figure 3 provides exemplary trajectories. For **Ukraine**, Australian levels move from Level 1–2 in the early 2000s to pronounced step-ups around 2014 and again from 2022 onward, while the US remains more persistently at Level 3–4. For **Lebanon**, both issuers spend extended periods at Level 3–4, punctuated by frequent revisions (black lines) that do not always change the posted level. For **South Korea**, Australia stays mostly at Level 1 with occasional upticks, whereas the US displays the system-wide pandemic jump to Level 4 and subsequent relaxation. The dense vertical markers illustrate that issuers revise advice frequently, even when the numeric level remains, thereby underscoring the value of archiving text alongside levels.

To quantify convergence, we align updates within countries using a nearest-neighbor match inside a ±15-day window (see Appendix for details) and then compare levels. Across all matched updates, **US–Australia** aligns the most, agreeing on **48.5%** of 890 matches (432 agreements). **UK–US** agrees on **33.1%** of 2,533 matches (839), and **UK–Australia** on **29.3%** of 5,299

Table 1: Summary statistics (2014–2024) for UK, US, and Australia advisories.

Issuer	L3 count	L4 count	Unique countries	Level changes	Updates	Most changes	Constant L4
UK	2,810	3,904	225	611	15,207	China (41)	33
Australia	651	2,014	181	547	6,530	Bangladesh (8)	11
US	844	758	210	990	3,101	D.R. Congo (21)	5

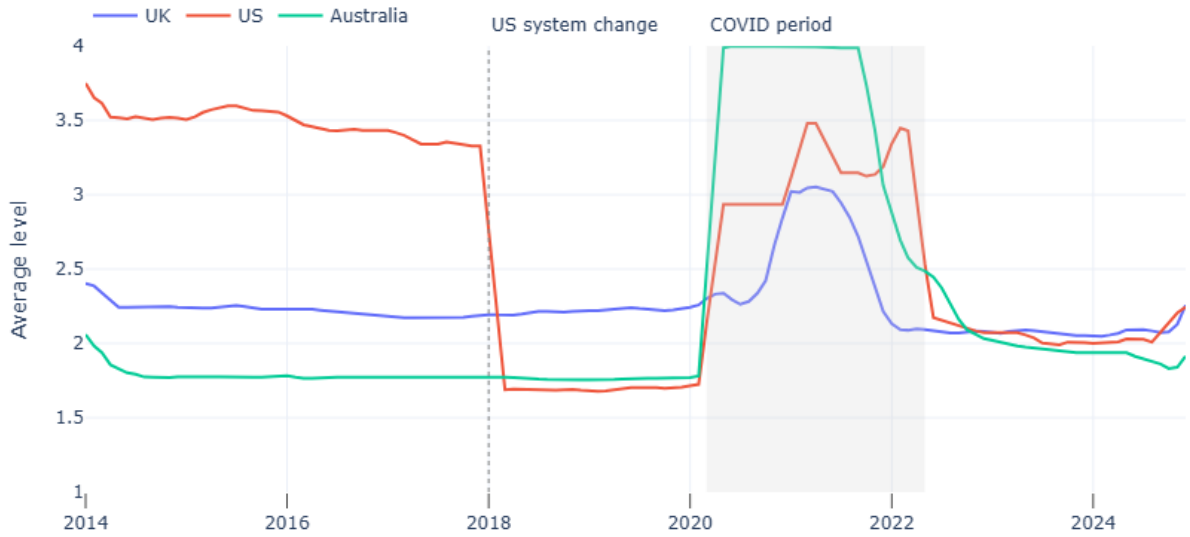

Figure 2: Monthly average advisory level across a common basket of countries for the UK, US, and Australia. Lines show 3-month moving averages; shaded bands are interquartile ranges. The dashed vertical line marks the US system change in 2018; the light gray rectangle marks the COVID period (2020–2022).

(1,550). When restricting to events where any issuer assigns Level 3/4, agreement rates drop: **37.5%** for US–Australia (238/635), **29.4%** for UK–US (584/1,987), and **23.3%** for UK–Australia (852/3,662). Requiring *all three* issuers to coincide produces very low agreement: **16.9%** across all matched triples (203/1,199) and **15.9%** for the Level 3/4 subset (165/1,039). Substantively, even among close allies, severity judgments are far from harmonized. The comparatively higher US–Australia alignment is consistent with shared terminology and, post-2018, more comparable US level definitions; by contrast, the UK's revision-intensive advice style lowers the probability of within-window convergence at the same numeric level.

The summary counts also reveal distinct issuer profiles. The UK posts many more absolute Level 3 and Level 4 observations (2,810 and 3,904), but as a *share* of all UK updates, these are smaller than Australia's, which has fewer updates overall, yet a higher fraction of high-severity postings. The US sits between the UK and Australia in scope, but shows the highest ratio of level-changing to total updates, reflecting more tightly scoped, advisory-centric editing pre-2018. Importantly, the aggregate lines in Figure 2 are computed on the *common basket* of countries to minimize composition bias.

Low pairwise and triple alignment can arise from (i) different thresholds for moving between Level 3 and Level 4, (ii) non-synchronous response timing around the same underlying event, and (iii) composition effects (issuers emphasizing different destination portfolios and sub-national guidance). Figure 3 shows that many micro-revisions do not alter levels, while the COVID panel in Figure 2 illustrates the opposite case, globally synchronized shocks that temporarily compress issuer differences and then re-diverge as governments normalize at different speeds.

A large share of updates, especially for the UK and Australia, revise or expand guidance without changing the numeric level. Most are minor editorial adjustments (wording, formatting, refreshed links), but many introduce substantively useful text: added/removed specific risk indicators, tighter regional exclusions, updated entry/exit rules, or clarifications of recent incidents. These "text-only" re-

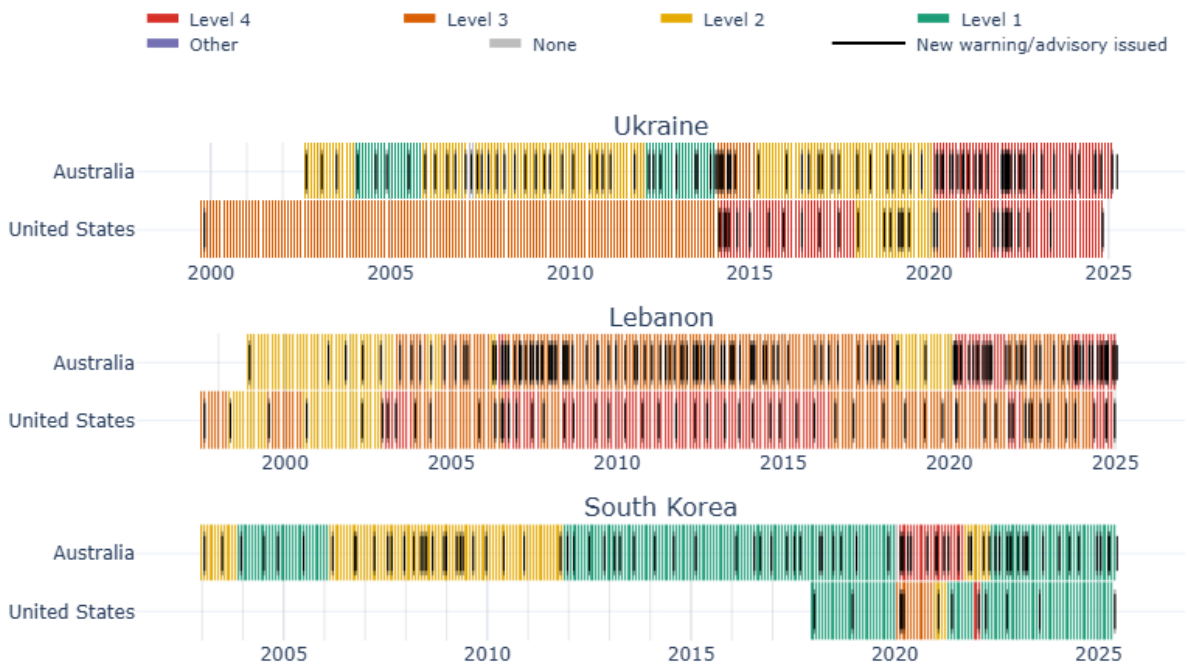

Figure 3: Monthly travel-advisory levels for three illustrative cases (Ukraine, Lebanon, South Korea). Rows show issuer (Australia, United States). Black marks indicate publication days of new warnings/advisories.

visions still signal issuer attention and policy stance even when the severity is unchanged.

4 Discussion

Our results show that even among close allies, advisory severity is far from harmonized and that issuers differ markedly in update cadence. A natural next step is to exploit the text itself rather than levels alone. Information extraction and named-entity recognition (NER) can recover subnational geography (regions, corridors, border zones) and relevant actors, enabling granular, region-level timelines layered onto the national series. Topic discovery and topic classification can map risk drivers (crime, terrorism, civil unrest, health, disasters, kidnapping, wrongful detention) and track how their composition shifts across issuers and time; change-point detection on topic proportions or keyphrase series can sharpen the timing of substantive updates that occur without a level change. Cross-issuer harmonization can be improved with supervised classifiers and LLM-assisted labeling to normalize historical phrasing into a shared taxonomy of risks and severities. Together, these textual transformations open the door to event-aware models of advisory dynamics that incorporate both global shocks and country-specific triggers, and to theory tests on alliance behavior, responsiveness, and the strategic use of language in official risk communication.

Limitations

Archive coverage. Our timelines are bounded by what the Internet Archive preserved. Gaps exist, particularly in the late 1990s and for low-traffic URLs. Missing snapshots may selectively underrepresent short-lived advisories.

Site drift and parsing. Ministry redesigns and evolving HTML lead to fragile extraction rules. Although our Scrapy spiders handle many variants, long-term maintenance is required. Edge cases (e.g. split country pages, temporary microsites) can still slip through.

Harmonization choices. Mapping issuer-specific phrasing to a four-level scale inevitably introduces noise (e.g. negations in "no longer advised to…"). Our regex strategy is transparent and reproducible but imperfect; supervised or LLM-assisted classifiers are a promising replacement once labeled data exist.

Structural changes. The US system changed in 2018, and pandemic-era global notices created atypical level spikes. Such policy regime shifts should be modeled explicitly when drawing causal inferences from the time series.

Generalizability. We focused on three English-language issuers for proof of concept. Extending to

non-English issuers will require additional country-name normalization and machine translation, and coverage may vary by local archival practices.

Ethical considerations

The data processed and analysed through this pipeline consists solely of publicly available travel advisories from official government sources (United States, United Kingdom, and Australia). The dataset contains no personally identifiable information (PII) or other sensitive data.

Acknowledgments

We thank David Bencek, André Bluhm, Vanessa Gottwick, Marius Hofmann and Daniel Racek, as well as the editors and three anonymous reviewers, for their helpful comments and feedback. The Center for Crisis Early Warning (Kompetenzzentrum Krisenfrüherkennung) is funded by the German Federal Ministry of Defense and the German Federal Foreign Office. The views and opinions expressed in this article are those of the author(s) and do not necessarily reflect the official policy or position of any agency of the German government.

Data and code availability

Upon publication, Scrapy spiders, processing scripts, and analysis notebooks will be released, together with issuer-specific schemas and harmonization code, under an open license. In subsequent work, we will publish the full dataset accompanied by an in-depth analysis paper that documents construction choices, validation checks, and known limitations.

References

Sanjay K. Arora, Yin Li, Jan Youtie, and Philip Shapira. 2016. Using the wayback machine to mine websites in the social sciences: A methodological resource. *Journal of the Association for Information Science and Technology*, 67(8):1904–1915.

Nicholas George Babey. 2019. The Politics of Travel Advisories: Foreign Policy and Error in Canada's Traveller Information Program. *The Journal of Intelligence, Conflict, and Warfare*, 2(1):15–33. Number: 1.

Patrick T. Brandt, Sultan Alsarra, Vito J. D'Orazio, Dagmar Heintze, Latifur Khan, Shreyas Meher, Javier Osorio, and Marcus Sianan. 2024. ConfliBERT: A Language Model for Political Conflict. *arXiv preprint*. ArXiv:2412.15060 [cs].

Yinxiao Chu, Xiaoyu Huang, and Tao Jin. 2021. Political relations and tourism: evidence from China. *Applied Economics*, 53(45):5281–5302. Publisher: Routledge _eprint: https://doi.org/10.1080/00036846.2021.1922591.

Nafbek Solomon Kebede. 2018. The Fate of Tourism during and in the Aftermath of Political Instability: Ethiopia Tourism in Focus. *Journal of Tourism & Hospitality*, 7(1):1–7. Publisher: Longdom Publishing S.L.

Oded Löwenheim. 2007. The Responsibility to Responsibilize: Foreign Offices and the Issuing of Travel Warnings. *International Political Sociology*, 1(3):203–221.

Jamie Murphy, Noor Hazarina Hashim, and Peter O'Connor. 2007. Take Me Back: Validating the Wayback Machine. *Journal of Computer-Mediated Communication*, 13(1):60–75.

Richard Sharpley, Julia Sharpley, and John Adams. 1996. Travel advice or trade embargo? The impacts and implications of official travel advice. *Tourism Management*, 17(1):1–7.

A Appendix

Collections, document examples and structure: For the US and Australia, each advisory is represented by two related documents:
(1) an "index" record (e.g. `us_advisories_index`) containing update metadata, and
(2) a "full-text" record (e.g. `us_advisories`) containing the parsed text and additional metadata (See examples below). For the UK, the index collection is not available; only full-text records exist.

Australia example (Lebanon, 1998):


```
{
  "country": "Lebanon",
  "warning_date": "1998-12-08",
  "advisory_text": "Australians
      travelling or resident in Lebanon
      should keep [...],
  "date_updated": "1998-12-08",
  "source_url": "https://web.archive.org
      /web/19990203024850/http:/www.dfat
      .gov.au/consular/advice/lebanon.
      html",
  "timestamp_snapshot": "19990203024850"
}
```


UK example (Lebanon, 2013):


```
{
  "country": "Lebanon",
  "level": "against all travel",
  "warning_date": "2013-03-28",
  "advisory_text": "Summary Still
      current at: 7 April 2013 [...]",
```

```
  "source_url": "https://web.archive.org
      /web/20130407005707/https://www.
      gov.uk/foreign-travel-advice/
      lebanon",
  "timestamp_snapshot": "20130407005707"
}
```


US example (Lebanon, 2024):


```
{
  "country": "Lebanon",
  "warning_date": "2024-12-27",
  "level": "Level 4: Do Not Travel",
  "advisory_text": "Updated to reflect
      the lifting of ordered departure
      [...]",
  "tooltips": [
    "Other: There are potential risks
        not covered [...]",
    "Kidnapping/Hostage Taking: [...]",
    "Civil Unrest: [...]",
    "Terrorism: [...]",
    "Crime: [...]"
  ],
  "source_url": "https://web.archive.org
      /web/20250103135018/https://travel
      .state.gov/content/travel/en/
      traveladvisories/traveladvisories/
      lebanon-travel-advisory.html",
  "timestamp_snapshot": "20250103135018"
}
```


Appendix A: Cross-issuer matching and agreement

We quantify cross-issuer alignment by per-country nearest-neighbor matching. For a given issuer A with update dates $t^A_{c,i}$ for country c, and issuer B with dates $t^B_{c,j}$, we match each $t^A_{c,i}$ to the chronologically nearest $t^B_{c,j}$ within a symmetric tolerance window of ± 15 days. Ties are broken by absolute time distance; updates without a counterpart inside the window are dropped. Agreement is computed over matched pairs (or triples) by comparing harmonized levels.

Table 2: Pairwise and triple agreement of advisory *levels*. Rows matched by country and nearest update within ± 15 days. "All" uses all matched updates; "L3/L4" restricts to matches where any issuer assigns Level 3 or Level 4.

Pair	Rows (all)	Agree % (all)	Agree n (all)	Rows (L3/L4)	Agree % (L3/L4)	Agree n (L3/L4)
UK–Australia	5,299	29.25	1,550	3,662	23.27	852
UK–US	2,533	33.12	839	1,987	29.39	584
US–Australia	890	48.54	432	635	37.48	238
UK–US–Australia	1,199	16.93	203	1,039	15.88	165