# Online reports and narrations of flood events worldwide

L. Vaudor https://orcid.org/0000-0003-1844-6425, L. Bajemon https://orcid.org/0009-0003-4131-8752, M. Genouel https://orcid.org/0009-0007-8046-6300, H. Piégay https://orcid.org/0000-0002-3864-2119, D. Montagne https://orcid.org/0009-0003-6427-8386

# Abstract

The consequences of flooding on Humans is difficult to document on a large scale, especially beyond readily quantifiable damage such as deathtoll or number of people displaced by the event. This study explores the use of Wikidata and Wikipedia as viable sources for the analysis of flooding occurrence and consequences. We evaluate the reliability and scope of these platforms despite inherent challenges such as geographical biases and inconsistencies in data. By juxtaposing this dataset with an established flood database, we investigate the conditions and extent of the contents generated on these platforms. Our findings indicate that Wikidata and Wikipedia are particularly valuable for providing not only detailed textual descriptions and social context but also insights into the public’s interest and engagement with flood events. Our analysis highlights that the creation, content, and diffusion of Wikidata and Wikipedia entries are influenced by factors including geographical location, social context, and event’s severity. This underscores the potential of these platforms as complementary tools in disaster management and research.

Key Points:

* The study compares flood events-related entries on Wikidata and Wikipedia with a traditional flood database, examining the factors that influence the creation and detail of these entries.
* Wikidata and Wikipedia offer credible and extensive information about floods, despite challenges in data consistency and geographical representation.
* These platforms provide valuable context and measure public interest in flood events, offering a complementary perspective to better understand their stakes and possibly enhancing the richness of traditional data sources

# Keywords

Wikidata; Wikipedia; Crowd-sourced data; Social context; Public interest; Flood; Disaster management;

# Introduction

A flood is an overflow of water onto normally dry land, which can result from various factors, including heavy rainfall, snowmelt, storm surges, or the failure of dams or levees. It can occur gradually over an extended period or suddenly, and might cause damage to humans, infrastructure and ecosystems with varying degrees of severity, ranging from minor flooding to catastrophic events causing widespread destruction and deathtoll. Understanding the dynamics of floods globally, including their causes, patterns, and impacts, is essential for effective anticipation, mitigation, and response to minimize their effects on communities and enhance resilience to these events (Doocy et al. 2013).

Although flooding is a normal part of the functioning of a hydrosystem and is beneficial to the environment, global flood damage has increased in recent decades, primarily because of the steady growth of urban populations and economic activities in flood-prone areas (Winsemius et al., 2016). Indeed, it is estimated that 70% of the world’s population will live in urban areas vulnerable to flooding by 2050 (da Silva et al., 2012), and this risk is further exacerbated by global environmental change, in addition to anthropogenic pressures (IPCC, 2022). There has been a shift in flood risk management these last few years from risk-based approaches to resilience-based (Laidlaw and Percival, 2024), which no longer aim at controlling flooding, but at reducing its impact, particularly with the idea to take into account the different forms that this risk takes on a diversity of types of vulnerabilities. Indeed, there are many different types of flooding: by river overflow (including open-water flooding, i.e. without ice, and flooding caused by ice jams), by runoff, by marine submersion, by rising groundwater, due to infrastructure accidents… A global approach can allow us to understand this diversity, by pointing out global pressures in the Anthropocene, while showing how local representations and discourses of these events also provide global narratives.

A lot of country-based databases about floods exist -see for instance Hilker, Badoux, and Hegg (2009) for Switzerland, Jak and Kok (2000) for Netherlands, Saharia et al. (2021) for India-, but few global ones do, although by studying floods occurring in a large number of contexts, researchers are more likely to identify patterns and understand underlying causes such as climate change and land-use practices, and develop efficient strategies for governments, policymakers, urban planners, and disaster management agencies. Furthermore, understanding global flood dynamics can facilitate international collaboration and resource allocation, fostering a collective response to mitigate their consequences.

This is one of the rationale behind the Dartmouth Flood Observatory (DFO) (Brakenridge n.d.), which provides a global database of flood events. Combining remote sensing and news reports’ analysis, it produces detailed information on the extent, frequency, and impacts of floods worldwide from 1985 to 2021 (Kundzewicz, Pińskwar, and Brakenridge 2013) . It constitutes an invaluable source of information on the subject, used in many scientific studies on flood events (Doocy et al. 2013; Winsemius et al. 2013; Hu et al. 2018; Najibi and Devineni 2018; Tellman et al. 2021). Events detected through remote sensing are recorded in the database when experts assess they are significant enough in terms of damage or occurrence frequency, based on news or institutional sources (Doocy et al. 2013).

Producing a global flood event dataset poses numerous challenges due to the complex nature of data semantics, collection, standardization, and accessibility. A primary obstacle is the lack of uniform reporting standards leading to inconsistencies in definitions of flood events and in data collection methods. Varying technological capabilities (material access and practice’s acculturation) across geographical regions of the world also contribute to disparities in data availability. Disasters like floods might occur in areas where communication and infrastructure are limited, and/or not used for real-time data collection. Political, social, and economic factors also play a role, as some regions may under-report or document flood occurrences differently, due to reasons such as insufficient or different resources and infrastructures, or differences in water governance (Wehn et al. 2015). Additionally, there are inequalities in the geographies of data production that reflect and reproduce global economic core-peripheries (Graham 2014; Graham, De Sabbata, and Zook 2015; Haklay 2016; Noucher 2023). Ash, Kitchin, and Leszczynski (2018) refers to some “informational magnetism” resulting in the over-representation of the global North (the “core”). As a result, and despite advancements in satellite technology and remote sensing, ensuring comprehensive, accurate, and up-to-date global flood data remains difficult. Overcoming this challenge requires better data curation, global efforts to improve information-sharing mechanisms, and the development of a critical eye on the reasons for data production, both in terms of contexts and uses.

## Participatory approaches

Participatory data collection can significantly enhance the understanding and management of flood events. This approach involves engaging local communities, citizens, and grassroots organizations in the collection, analysis, and sharing of data related to flood occurrences. Examples of such initiatives show the potential of participatory data to complement institutional data about floods at a local scale (Dixon, Johns, and Fernandez 2021; Sekajugo et al. 2022). By incorporating local knowledge, experiences, and observations, participatory methods complement traditional data sources, offering valuable insights into flood-prone areas that might be overlooked by centralized or remote monitoring systems, controlled by certain systems in power (Nation-states, heterogeneous coalitions of digitally literate individuals, private technology companies). More broadly, volunteered geographical data at a local scale have been shown to be a valuable source of data regarding hazards (Herfort et al. 2021).

Communities living in flood-prone regions possess unique, context-specific, even vernacular knowledge about local environmental changes, historical flood memory, vulnerable areas, and coping mechanisms. Engaging these communities in data collection through citizen science initiatives, mobile applications, community mapping, or participatory workshops allows for the collection of granular, real-time information that supplements existing datasets. Ideally, this bottom-up approach not only enhances the accuracy and granularity of flood data, but also fosters community empowerment in terms of resilience, by engaging all communities to participate, including those excluded from decision-making or digital access. Moreover, participatory data can bridge gaps in official reporting by capturing small-scale or localized flood events that are not publicized and thus recognized by institutions. Integrating participatory data with conventional datasets enables a more comprehensive understanding of flood dynamics, aiding in the development of anticipation policies, effective early warning systems, and disaster response plans.

However, challenges such as ensuring data accuracy, reliability, and standardization persist in participatory approaches. While establishing protocols for data validation and quality control between participatory and formal datasets is important to maximize the potential of participatory data and to maintain data integrity (Senaratne et al. 2017), the participatory data can be standardized. In other words, the subtleties perceived by local communities can be smoothed out (i) in interfaces designed by situated engineers and (ii) in data collection via protocols designed by situated scientists. However, the stakes around data validation and homogenization are particularly high in the prospect of studying floods at a global scale.

Overall, leveraging participatory data in conjunction with conventional methods can enrich global flood databases, improve resilience, and empower communities to better respond to and mitigate the impacts of flood events. This collaborative approach can contribute significantly to more holistic and inclusive strategies for managing floods worldwide, to question (i) the data (how, by whom and for what purpose they are collected) and (ii) the collection process (highlighting cultural biases by comparison).

## Wikidata

In parallel to research-related, expert database such as the DFO dataset on the subject, the Wikimedia project stands as a remarkable source of structured or semi-structured participatory data with its vast array of user-generated content across platforms like Wikipedia, Wikidata, and Wikimedia Commons. Wikipedia, as one of the largest collaborative encyclopedias, and one of the most used and visited websites globally (Graham, De Sabbata, and Zook 2015; Ballatore, Graham, and Sen 2017) harnesses the collective knowledge of volunteers worldwide who contribute, edit, and curate articles on diverse subjects, including geographical features, history, and environmental events such as floods. It can be viewed both as a source of breaking news and as an encyclopedia, providing information and narrative regarding ongoing and past events, in particular crises and controversies (Moats 2019). Wikidata, an open database, provides structured data that can be used to categorize and link information, potentially cataloging flood events, affected regions, and relevant details. Wikimedia Commons serves as a repository for multimedia files, housing images, maps, and other visual resources related to floods, contributing to a more comprehensive understanding of such events. The open nature of Wikimedia projects allows for continual updates and contributions, making it a valuable resource for researchers, policymakers, and the public seeking information on various topics, including flood events, across the globe. The Wikimedia projects are an example of user-generated content on web 2.0 (O’Reilly 2020). Moreover, flood events are generally associated with a location or toponymic information: as such, the produced data can also be included in what Goodchild (2007) calls “volunteered geographic information”.

Wikipedia articles have been studied in the context of managing crises due to its accessibility, comprehensiveness, and real-time updates, either focusing on content creation or consultation. This has been the case for e.g. epidemiologic crises (Van Velsen et al. 2012; Colavizza 2020), natural disasters, technological accidents and violent conflict (Keegan, Gergle, and Contractor 2013). These uses of Wikipedia articles could also be contemplated for flood events. For instance, people could seek immediate information regarding safety measures, evacuation procedures, flood-prone areas, and emergency contacts, in a multilingual framework accessible to a diverse, local audience. Wikipedia’s open editing structure indeed allows for rapid dissemination of accurate information, enabling users to access critical details swiftly.

Like many volunteered mapping platforms, the Wikimedia projects are receiving growing interest from the research community (Farda-Sarbas and Müller-Birn 2019). As such, they are used in various projects: on the one hand, the content is employed directly as a data source, e.g. in history, literacy studies, archeology, philosophy (Zhao 2023), linguistics (Turki et al. 2017), medical and biological sciences (Farda-Sarbas and Müller-Birn 2019; Burgstaller-Muehlbacher et al. 2016), etc. or as an indicator of public interest (Mittermeier et al. 2021). On the other hand, studies consider the Wikimedia projects as a research object, e.g. analyzing the contributors’ motivations (Antin 2011), comparing their data to traditional databases (Lorini et al. 2020), evaluating completeness (Royal and Kapila 2009), accuracy (Giles 2005), topic coverage (Halavais and Lackaff 2008), local (Ballatore, Graham, and Sen 2017; Ballatore and De Sabbata 2020) and global distribution (Graham, De Sabbata, and Zook 2015; Graham et al. 2018), analyzing knowledge construction (Oeberst et al. 2014) and social network dynamics (Iba et al. 2010), etc.

However, using a worldwide, crowdsourced, open database implies characterizing its contributors. Firstly, studies show that user-generated content is unequally distributed on a global scale: it is generally skewed towards countries from the Global North, whether it be in terms of participation, representation or access [Graham, De Sabbata, and Zook (2015); Ballatore, Graham, and Sen (2017); Graham et al. (2018); ]. Research conducted on Wikipedia suggests that the platform follows the same pattern. For example, Graham (2014) shows that Europe and North America get the greatest number of articles per person. Secondly, participation tends to be skewed towards “a more wealthy, more educated, more Western, more white and more male demographic” (Crampton et al. 2013, 132). For example, research on OpenStreetMap – an open and crowdsourced mapping platform – suggests that it reproduces the socio-economical structures (Bittner 2017). The risk is that these crowdsourced platforms reproduce the inequalities, leaving economically and socially deprived areas also digitally deprived (Mashhadi, Quattrone, and Capra 2015). There is also a strong gender asymmetry on the web, contributing to gender-biased data (Stephens 2013; Gardner et al. 2020).

Identifying the data distribution is capital for interpretation and analysis, given that the Wikimedia projects are user-generated and are therefore subjects to biases. The question “who produces the data?” must then be addressed. Lorini et al. (2020) studied the report of floods on the English Wikipedia portal through a thorough comparison to several other databases, questioning the conditions in which the articles were created. It offers valuable insight into the biases this dataset might be subject to, which we will keep in mind in our own questioning of the contents and reach of the articles using quantitative and textometric methods. To our knowledge, no such study has been conducted on the subject of floods using extensive Wikidata and multi-language Wikipedia articles.

## Objectives

This study’s objectives are to

* collect and assess the information provided by Wikidata and Wikipedia on flood events.
* compare this data, which is crowdsourced, to an institutionally curated database on the same topic to estimate its validity and interest.
* use this data to characterize and understand (in terms of human impact especially) the flood events documented.

# Material and methods

## Primary census of flood events

We collected Wikidata about floods, using the glitter R package (an API-client to the Wikidata SPARQL endpoint API) (Vaudor and Salmon 2023). Wikidata is a free and collaborative knowledge base, structured as a graph of interconnected items representing physical entities, events, concepts or topics, associated to a unique identifier. These items are linked through properties forming statements describing the relationships between items and/or attributes. Each query we run on Wikidata can hence be viewed as a graph query, where we retrieve items and attributes and the relationships between them.

|  |
| --- |
| Figure 1: Fig. 1:“Wikidata queries represented as graphs. When properties were part of the Wikidata (wd) namespace, labels explaining their nature were displayed, instead of their codified property identifiers (e.g. ‘instance of’ instead of ‘wdt:P31’)” |

We carried out a first Wikidata query (Fig. 1.a) to get all instances of floods or subclasses of floods as well as a few descriptors of these. The events can be located through coords (geographical coordinates) loc (a location -that can be any type of geographical item documented in Wikidata, e.g. a city, region or country- and which in turn might provide coordinates), or simply through the country variable (which might also in turn provide coordinates). We used all these variables to try and provide coordinates for as many events as possible, with the greatest precision possible: we prioritized coords (when available) over coordinates corresponding to location, and coordinates corresponding to location (when available) to coordinates corresponding to country.

We carried out a second Wikidata query (Fig. 1.b) for each flood listed by this first query, to gather information about its time of occurrence (wdt:P585), start time (wdt:P580) and end time (wdt:P582). We hence tried to provide a date for each of the events using either one of these previous items, prioritizing the date provided by time of occurrence (if available) over the average date between start time and end time (if both were available), average date between start time and end time over start time or end time (if only one was available), and start time or end time over four-digit words in the events’ label if present (as labels generally consist of a mention of type of event, time and place). Time-related information on Wikidata are specified with varying precision, which when regarding floods generally vary between dayly precision and yearly precision.

We carried out a third Wikidata query (Fig. 1.c) for each country listed by the first query, to gather information about their coordinates, the languages used by their population (either official or used languages), Human Development Index (HDI), population and area.

## Comparison to the Dartmouth Flood Observatory data

We wanted to confront the data about floods documented through Wikidata (WD) to the data gathered by the Dartmouth Flood Observatory (DFO) (Brakenridge n.d.). We downloaded the DFO dataset and curated it to make it more easily comparable to our WD dataset, especially regarding country labels. We then tried to match each event in WD to one event in DFO through the following algorithm

For each WD event flood:

* calculate the time difference between DFO events and flood
* calculate the distance between DFO events and flood
* check for all DFO events whether the time difference to flood is less than 400 days and the distance to flood is less than 400 kms (condition 1)
* check for all DFO events whether the time difference to flood is less than 400 days and country matches country registered for flood (condition 2)
* filter all DFO events to only keep those verifying condition 1 or condition 2
* if several DFO events verify those conditions, retain only the one corresponding to lowest time difference to flood, and consider that this is the DFO event that matches flood

The rather high threshold of 400 days has been chosen to account both for the possibly low precision -sometimes, yearly precision- in dates, and the fact that flood events can be quite long, leading to heterogeneities in the way the date of occurrence is defined and recorded. The also rather high threshold of 400 kms has been chosen to account for the fact that Wikidata events are not always precisely located. The flood event might indeed be located through a location corresponding to the region centroid, or even country centroid. Using as an alternative to the distance the country of occurrence recorded in both datasets aimed at alleviating this problem of events located with very low spatial precision.

## Statistical analyses

We ran all statistical analyses and produced graphics with the R software (R Core Team 2024). When comparing distributions through categories (associating this comparison with boxplots) we tested the difference of each category’s median with the other categories’ median using the Wilcoxon test, a non-parametric test based on ranks, in order not to rely on a normal distribution hypothesis or on asymptotic conditions. When we had more than two categories to compare, we compared one category to all other categories, pooling them. We generally considered models with two levels, the response depending on two categorical variables X\_1 and X\_2. We tested the significance of the difference in median of each level of X\_1 against all other levels of X\_1, and for each level of X\_1, the difference in median of each level of X\_2 against all other levels of X\_2. We displayed on the graphics the difference in median for each group associated with a code for the p-value: .:p<0.1, \*:p<0.05, \*\*:p<0.01, \*\*\*:p<0.001.

## Text analysis of Wikipedia pages

We scraped the text (all the textual content in headers from level one (h1) to level 6 (h6) as well as paragraphs) from all listed Wikipedia pages, using the rvest R package (Wickham 2024). We then translated them to English (if necessary) through Google Translate using the polyglotr R package (Iwan 2024).

We tokenized the texts with the tidytext R package (Silge and Robinson 2016), lemmatized words and filtered out stop words based on the Iramuteq English dictionnary (Ratinaud 2009). We could then identify textual specificities based on various partitions of the corpus (Lafon, Pierre 1980; Loiseau et al. 2022).

We used Reinert’s method (Reinert 1990) to classify segments of our corpus (consisting in successive parts of the Wikipedia articles, segmented in such a way that each contains 10 lemmatized, content words). To implement that method, we used the rainette package (Barnier 2023) and used its exploration tools to choose a meaningful and tractable number of classes. Each class is characterized as a collection of over-represented words, and we used them to label each class in a one-worded, synthetic way. We then used these labelled classes to examine the relationship between the flood characteristics and the topic contents of articles.

## Multivariate analysis

We considered descriptors of the events’ characteristics and of the related Wikipedia articles to examine Hypotheses 2 and 3. To account for the multiple colinearities in our dataset, we carried out a principal component analysis (PCA) of variables (rank-transformed to avoid the influence of outliers) using the FactoMineR R package (Lê, Josse, and Husson 2008). We also carried out some bivariate tests (based on simple linear regression) on these rank-transformed variables to further discuss the significance of the relationship between variables.

## Output datasets

The main output dataset generated during this study has been deposited in the Nakala repository to ensure long-term preservation and accessibility. It is available at the following DOI: https://doi.org/10.xxxx/nakala.xxxxx. All code used to process the data and produce the results presented in this article is openly available on GitHub: https://github.com/lvaudor/glourb\_floods, facilitating full reproducibility of the analysis.

## Hypotheses

We will explore the datasets with some hypotheses in mind, in particular:

* **Hypothesis 1** (regarding recorded events). We expect the WD dataset to include less events than the DFO dataset. Indeed the DFO dataset records flood events primarily based on physical criteria, before considering social criteria to decide on inclusion. This initial screening is likely to capture some floods that, while physically significant, do not attract enough attention to be recorded in WD. We also expect the WD dataset to be somewhat biased towards events occurring in countries with a higher HDI, as most crowdsourced web datasets.
* **Hypothesis 2** (regarding curation). The iterative, sometimes back-and-forth nature of the Wikipedia pages writing process (which we later refer to as curation) can be used as a measure of a particular effort having been devoted to the establishment of the current text and could thus be interpreted as a sign of possible controversy or debate in its writing process. As such, we expect curation to be linked to the characteristics of the articles and the events they describe. We hypothesize that curation will depend on the articles’ and corresponding events’ general characteristics (hypothesis 2a), in particular that it will be higher for articles that are longer (editing efforts focused on longer articles), written in a language local to the event, and dealing with events with a high deathtoll (more public scrutiny). We also consider other explanatory variables such as time (as standards for Wikipedia pages redaction might have evolved through time), HDI (as a proxy to the proportion of internet users in the population), affected country’s population and population density (as more public scrutiny and more people potentially physically affected by the flood could influence the curation process). We also hypothesize that the level of curation of an article will be correlated to the topics addressed in the article (hypothesis 2b). We hypothesize in particular that higher-quality articles will correspond to a higher level of curation and address topics which necessitate a deeper level of analysis (such as anticipation, governance, and hydrology) in higher proportions. We also expect higher-impacts events (in terms of deathtoll for instance) to cause this proportion to be higher, with complex issues around responsibility, causes, and adequacy of preventive measures only gaining attention when the damage is significant.
* **Hypothesis 3** (regarding number of views). We expect the short-term number of views of an article to be correlated to some of the articles’ characteristics such as the article’s being written in a language local to the event, article’s length, and curation. We expect that the number of views will be positively correlated to the level of curation of the article, as articles attracting more views are more likely to attract more scrutiny from editors. We expect the short-term number of views of pages to be higher in the local language, due to Wikipedia being used as an information source regarding the crisis unfolding and management. We also expect the number of views to reflect some characteristics of the events. For instance, we expect the number of views to be higher for events with a higher deathtoll, and/or events occurring in areas where the density of population is particularly high, as these events are more likely to be of interest to a larger audience. We have access to the number of views since July 2015 hence to explore this hypothesis we first filtered out all older events: we are thus able to quantify the number of views of various articles at the same stage in time from the moment the event occurred.

# Results

## Wikidata

### Space-time framework

The initial query provided 904 Wikidata flood events. In terms of spatial precision, few events (16.1%) come with direct coordinates. Most coordinates for floods are inferred from their locations (places where they occur such as cities, regions, and mainly -52.2%- countries). Some events (10.7%) are not localized at all.

The date of most events is documented either directly or through the label of the event. Many events’ dates are provided with day precision (35.9%) but also an important proportion (38.4%) is provided with only annual precision.

|  |
| --- |
| Figure 2: Distribution of flood events through time A) all events dated with a year-accuracy B) events dated with a year-accuracy starting in 2000 |

Fig. 2 shows that the distribution of flood events through time is quite heterogeneous, starting from medieval times to current events, with a large proportion of recent events (63.8% of events occurred after 1985, 53.4% occurred after 2005). Nevertheless it implies that some older events are documented, which is actually uncommon compared to other databases (263 events documented in the WD dataset occurred before 1985).

### Comparison to DFO data

This section aims at addressing Hypothesis 1 regarding the conditions in which Wikidata records are created and how they compare to institutionally-curated records of floods.

|  |
| --- |
| Figure 3: a) Number of flood events documented in Wikidata (WD), DFO, or both datasets according to our matching algorithm. The number of events is displayed according to the Human Development Index (HDI) class of the country where the event took place (countries are pooled in the first, second, third or fourth quartile). b) Deathtoll per event according to HDI class and occurrence in one or both datasets. |

To address Hypothesis 1 we compared the number of events in both datasets (Wikidata and DFO) according to the Human Development Index of the country where the event took place (events spanning multiple countries were hence re-formatted as multiple, country-specific records). There is a total of 826 records only in Wikidata and 4883 only in DFO (499 in both). Fig. 3.a shows that, compared to DFO, a large share of floods recorded in Wikidata are associated with countries with higher HDI levels. In contrast, the distribution of events in the DFO dataset is more concentrated in the first and especially the second HDI quartiles. The DFO dataset primarily relies on physical criteria to define flood events, but also incorporates social indicators such as damage and deathtoll to assess the event’s significance. Here, we consider DFO’s distribution of flood events to be a relatively unbiased approximation of actual flood occurrence : at least, more so than Wikidata’s distribution which is also affected by biases of representation linked to digital access and visibility. This highlights the multiple criteria required to recognize a flood, but also the difficulty of reporting floods in a consistent way across all projects.

Fig. 3.b shows that compared to DFO-documented events, events included in the Wikidata dataset are generally associated with a higher deathtoll. The estimates considered for deathtoll come either from the DFO, or by default from Wikidata (some events are not associated to any deathtoll estimate, in which case they are excluded from this figure). Some events happened in places where the country HDI is missing (these countries might have undergone name changes or experienced political instability over time, causing gaps in statistical reporting): they are represented by an unknown HDI class. For DFO-documented events, the lower the HDI class, the higher the deathtoll, which could be interpreted by impoverished countries having less resources and infrastructure to mitigate and respond to such disasters effectively, as well as more inadequate housing and settlement in vulnerable areas exacerbating the impact of floods on communities. In cases with high deathtoll, the census can also be less accurate, and possibly rounded to greater values, causing a certain bias in the count of victims. As for Wikidata-documented events, the few ones which occurred in lowest HDI countries also correspond to (very) high deathtolls indeed, but so do the ones which have occurred in the richest countries and even more so the ones for which we do not have a value for HDI. In this last case, the events documented in Wikidata only might be historical ones like the 1931 China floods (wd:Q1150973). These floods occurred in the Republic of China (1912-1949), for which we failed to obtain an HDI value, and which were associated with very high and somewhat discussed deathtolls. Overall, results shown in Fig. 3.b highlight the fact that the Wikidata dataset is oriented towards events with high societal impact (in particular deathtoll), which is not surprising given the nature of the project, and consistent with Hypothesis 1.

|  |
| --- |
| Figure 4: a) Distribution of deathtolls estimates in Wikidata (WD) and DFO according to the events being paired across datasets or not. b) Comparison of deathtoll paired estimates in Wikidata and DFO for the same events. The symbol varies according to the deathtoll estimate in Wikidata being sourced. |

For some events we were able to find a match between the Wikidata dataset and the DFO dataset, so that we could compare the deathtoll paired observations. Fig. 4.a shows that the deathtoll values are generally higher in the Wikidata dataset than in the DFO dataset, even when observations are paired (i.e. the events are recorded in both datasets). There are 499 paired events, but only 293 come with a deathtoll estimate in Wikidata. For paired events, the median deathtoll is 18 points higher for Wikidata than DFO, which suggests that deathtolls could be overestimated in Wikidata and/or that deathtoll is recorded in Wikidata only when it is rather high. For unpaired events, median deathtoll is about 60 points higher in Wikidata than in DFO. Overall, Fig.4.b is consistent with the fact that the WD dataset is oriented towards events with high societal impact (and in particular, high deathtoll). However, Fig. 4.b shows that for paired observations and when deathtoll is provided by both datasets, the agreement between the deathtoll estimates from both datasets is quite high. This suggests that, although Wikidata is biased towards including events with high deathtolls, deathtolls estimates themselves are actually quite reliable. In particular, there is a portion of the matched events for which the recorded deathtoll is exactly the same (237 events i.e. 47.49% of the paired events) or of the same order of magnitude (322 events i.e. 64.53% of the paired events). This also suggests that the pairing of events is mainly correct, despite the difficulty of joining completely independently designed datasets, which probably account for discrepancies. Indeed, Wikidata deathtoll estimates’ being sourced or not does not seem to account for differences in their discrepancy to DFO estimates. This is consistent with these discrepancies being caused by a mismatch of events rather than an error in the estimates.

## Wikipedia

### Description

There are 2983 Wikipedia pages associated to the Wikidata flood events (i.e. on average 2.52 Wikipedia pages per event). 805 flood events are associated to at least one Wikipedia page. 540 Wikipedia pages are in English. Out of the 2443 remaining pages, we could translate most pages to English, except 100 pages which were in languages that Google Translate did not support. We based all subsequent textometric analyses on the texts in English, either natively so or translated.

The average length of articles (in English or translated to English) is 3478 characters (median: 1624. The 10% shortest articles represent less than 311 characters and the 10% longest ones are more than 7664 characters long.

There are 225336 revisions that produced the Wikipedia pages as they appear to this date, which represents on average 76 revisions per page. 182412 (80.95%) of these revisions appear to have been done by human editors and not bots (Wikipedia usernames do not contain the strings “bot” or “Bot”). There are 44642 distinct editors, 40288 (90.25%) of which appear to be humans. Each of the human editors in our corpus is responsible for 4.53 edits on average. Most human editors (69.03%) only edited one article, 27.39% edited up to 5 articles and 3.58% edited more than 5 different articles about floods.

### Wikipedia articles’ history

|  |
| --- |
| Figure 5: Time delay between events and a) revisions to the Wikipedia articles documenting them b) views of the Wikipedia articles, displayed according to the period in which the event occurred. The grey areas thus correspond to unobservable delays. The color indicates the precision of the recorded event date (it can be accurate to the day, month or year). |

Fig. 5 displays the dynamics in time surrounding contents creation and consultation in Wikipedia, showing the delays between events occurrence, and article revisions (Fig. 5.a) and article views (Fig. 5.b). The delays depend on the period of occurrence all the more that, for instance, events which occurred before 2000 cannot correspond to edits in the first year after their occurrence since Wikipedia was created in 2001, and that for events which occurred before 2014 we do not know the views in the first few months after the event (since the views are only accessible starting from 2015). Also, the delay might in some cases appear negative, because the edits or views about an event began before the recorded, average date for the event, hinting to an event with long duration or event with a low-precision recorded date.

Fig. 5.a shows that most of the writing of Wikipedia articles related to flood events take place during the month of their occurrence (except of course when Wikipedia did not exist (before 2001) or when it was still moderately used to document current events. This highlights the fact that Wikipedia articles about recent flood events are written as news rather than encyclopedic reports on past events. Fig. 5.b shows that views (consultations) of the pages, on the other hand, are distributed more evenly through time. We cannot access the data relative to views before July 2015 so that early views of articles dated before 2015 cannot be known. For articles dated after 2015, there is a peak in the number of views in the 1.5 months after the fact, but the data about the events between 2015 and 2020 shows that the total number of views is far more important in the long run (at least 6 months after the fact) than this initial number of views.

### Wikipedia articles’ topics

|  |
| --- |
| Figure 6: Topics identified by the Reinert classification method. |

Fig. 6 shows the topics identified by the Reinert classification method. Based on the over-represented terms in each cluster we propose the following labelling of the classes: ‘governance’ for cluster 1, ‘relief’ for cluster 2, ‘damage’ for cluster 3, ‘hydrology’ for cluster 4, ‘anticipation’ for cluster 5, and ‘weather’ for cluster 6.

|  |
| --- |
| Figure 7: Link between total article’s length and topics identified by the Reinert classification method. |

In Fig. 7 the articles are classified into 4 categories of lengths according to their number of segments (used for the Reinert classification and identification of topics). The 25% shortest articles represent 1<k<=4 segments while the 25% longest represent 26<k<860 segments. The proportion of segments falling into each category is calculated and the distribution of these proportions is displayed through boxplots for each article length category. The plus signs correspond to the median proportions of topics throughout all articles irrespective of articles’ length. The underlying model for the tests is proportion~topic+length\_category. For each topic the difference in median due to length\_category is represented by the gap between the + sign and the mid-line of each boxplot: this difference is tested through a Wilcoxon rank test and the result is displayed in the plot.

Fig. 7 shows that the length of the article is probably a very important factor in the topics addressed (or conversely, the topics addressed determine the length of the article), with shorter articles corresponding to factual statements regarding physical features of the event (weather), and justifying reporting that event based on its consequences (damage), while longer articles might convey more extensive reports on the physical causes (hydrology) and management of the crisis (anticipation, governance, relief).

### Curation

This section aims at addressing Hypothesis 2 regarding the curation of articles and its relationship with the events’ and articles’ other characteristics.

|  |
| --- |
| Figure 8: Total length of articles as a function of number of edits, and resulting curation metric. |

Fig. 8 shows the relationship between number of edits and total length of the article (all articles having been translated to English). The articles in language ‘simple’ (for ‘simple English’) being destined to be particularly synthetic, we did not take include them in this model. Beyond a few edits (here, we consider the threshold is 10), the total length for an article depends on the total number of edits for that same article, in an approximately linear relationship with both variables being log-transformed. The more edits a page has undergone, the lower the mean edit length: there is a tendency towards lower mean edit length as the page ‘matures’, with edits in a more advanced writing stage corresponding to suppressions, corrections, or slight modifications while edits at the beginning of the writing process tend to be more frequently text additions.

We consider the difference between actual mean edit length edit and predicted mean edit length (based on number of edits) as a measure of relative edit length (it corresponds to a measure of mean edit length decorrelated from this previously described writing stage effect). We calculate the curation score as corresponding to minus the relative edit length (i.e. high level of curation corresponds to smaller-sized edits). These residuals are normalized to correspond to scores ranging from -1 (rather lowly curated articles) to 1 (very highly curated articles). These curation scores are only available for articles that have undergone at least 10 edits, so that article stubs created mainly from an automatic translation process from another Wikipedia page are not included in the analysis of curation.

## Linking Wikipedia pages characteristics and events characteristics

|  |
| --- |
| Figure 9: Crossed correlations between metrics (events and pages characteristics), displayed as a) a correlation matrix plot of metrics and b) a factorial map of metrics derived from a Principal Components Analysis |

|  |
| --- |
| Figure 10: Crossed correlations between metrics (events and pages characteristics), displayed as a) a correlation matrix plot of metrics and b) a factorial map of metrics derived from a Principal Components Analysis |

Regarding Hypothesis 2 and as expected, curation is linked to the characteristics of the articles and the events they describe. curation is heavily negatively correlated to the articles’ length and also to other article’s characteristics such as the proportion of complex topics in an article. Curation is higher for articles written in a language local to the event. On the other hand, curation is rather lowly correlated to deathtoll, population and population density, or even views, which suggests that these variable which we had identified as causing more scrutiny and.

##ICI

We also consider other explanatory variables such as time (as standards for Wikipedia pages redaction might have evolved through time), HDI (as a proxy to the proportion of internet users in the population), affected country’s population and population density (as more public scrutiny and more people potentially physically affected by the flood could influence the curation process). We also hypothesize that the level of curation of an article will be correlated to the topics addressed in the article (hypothesis 2b). We hypothesize in particular that higher-quality articles will correspond to a higher level of curation and address topics which necessitate a deeper level of analysis (such as anticipation, governance, and hydrology) in higher proportions. We also expect higher-impacts events (in terms of deathtoll for instance) to cause this proportion to be higher, with complex issues around responsibility, causes, and adequacy of preventive measures only gaining attention when the damage is significant. - **Hypothesis 3** (regarding number of views). We expect the short-term number of views of an article to be correlated to some of the articles’ characteristics such as the article’s being written in a language local to the event, article’s length, and curation. We expect that the number of views will be positively correlated to the level of curation of the article, as articles attracting more views are more likely to attract more scrutiny from editors. We expect the short-term number of views of pages to be higher in the local language, due to Wikipedia being used as an information source regarding the crisis unfolding and management. We also expect the number of views to reflect some characteristics of the events. For instance, we expect the number of views to be higher for events with a higher deathtoll, and/or events occurring in areas where the density of population is particularly high, as these events are more likely to be of interest to a larger audience. We have access to the number of views since July 2015 hence to explore this hypothesis we first filtered out all older events: we are thus able to quantify the number of views of various articles at the same stage in time from the moment the event occurred.

# Discussion

This study demonstrates the possibility to list and collect information on flood events on a large temporal and spatial scale through Wikidata, and Wikipedia. Although these data exhibit some heterogeneity (Lorini et al. 2020; Ruprechter, Burghardt, and Helic 2023) due to economic and digital inequities, being able to rely on a preexistent, rich global dataset spanning a large time period is still a priceless asset in studying floods, and could be a first step to better-quality curated datasets. Besides, large-scale institutional data sources themselves are not immune to spatial heterogeneities due to scarcity of research activity in some areas, linguistic barriers, varying degrees of public investment in disaster management, or other sources of environmental, political and social inequities. The overcoming of all these difficulties require an ongoing specialist effort which might not always be guaranteed or easy to sustain.

Linked Open Data such as Wikidata hold great interest for the study of natural or social events due to their inherent ability to offer contextual information to documented items. Linking semantically events when they are documented in fundamentally different datasets (such as DFO and Wikidata) might remain very difficult. Nevertheless, the interconnected nature of Wikidata’s model (as a knowledge graph) as well as their link to other data silos in the Linked Open Data framework allow users to access additional information or related resources linked to a particular event. In particular, the present study demonstrates the strength of such a data source to provide context on a global scale (social context, in particular) through the integration of Wikipedia discourse elements.

In the present study we focused on the extrinsic reasons why entries about floods existed in Wikidata and Wikipedia, discussing it as indicators of the events’ social significance. These platforms also provide guidance of their own as to which items are notable enough to warrant the existence of, primarily, Wikipedia pages (Wikipedia 2024a) and subsequently Wikidata pages (as the first condition to the existence of an item for Wikidata is that it should have a link to another Wikimedia-related project e.g. Wikipedia). The criteria for the existence of a page in Wikipedia might differ across language editions but share the same general principles: a subject typically needs to have received significant coverage from reliable, independent sources to be considered notable, information must be verifiable through reliable, independent, and trustworthy sources, and articles must be written from a neutral perspective, presenting all significant views fairly and without bias. The existence of a page in Wikipedia and Wikidata is thus a strong indicator of the event’s notability, but it might be a bit restrictive compared to other lists of events that do not need to comply with as many criteria. Incidentally, Wikipedia itself does in fact offer pages that list items in certain categories, and in particular lists of floods (e.g. Wikipedia 2024b), which might be a good source of information for events that are not as well documented.

This study hints at Wikipedia articles being generally less opinionated than one might think. Wikipedia’s content is indeed based on verifiable sources and strives to present information in a neutral and unbiased manner. While individual contributors may have their own perspectives, Wikipedia’s community guidelines emphasize the importance of maintaining a neutral point of view in articles. Editors are encouraged to cite reliable sources and provide balanced coverage of topics to ensure accuracy and verified, factual content. Here, the possibly controversial topics in the narration of floods (governance, relief, hydrology) are finally less represented in the articles with high levels of curation. This means that a study with a particular interest in debates or controversies would need to rely either on sources of information outside Wikipedia or more specifically on the discussion pages of or revisions history of Wikipedia articles (rather than final article’s content).

The difficulty we faced with the spatial information available in Wikidata regarding floods demonstrates it is not, as such and on its own, a really convenient and rich data source for spatial analysis, which is not entirely surprising as it is not primarily designed for this purpose. This limitation could be overcome by the use of other data sources, such as OpenStreetMap, which is more spatially oriented as a community-driven mapping initiative, and akin in principles to Wikipedia. However, although OpenStreetMap would be a suitable source of spatial data for Wikidata items corresponding to physical objects, it does not deal directly documents events: OpenStreetMap disaster-related projects such as the ones of the Humanitarian OpenStreetMap Team rather focus on making sure that places and people are well mapped throughout the world to help provide relief or anticipate future crises (Specht 2022). Whatever the database’s structure though, providing spatial information regarding an event (such as a flood) is actually not a trivial task, as it requires the definition (through time) of the event’s spatial extent, location(s), and/or impact area(s). Actually, using Wikipedia pages as a source of information on floods could be a good way to overcome this limitation, as the articles often provide a detailed textual description of the event’s impact area(s) and location(s). Nevertheless, extracting a map from each of the pages and inferring a standardized location from it would probably have arisen many methodological but also semantic questions similar to those surrounding narrative maps (Caquard 2013).

# Conclusion

This study demonstrates that Wikidata and Wikipedia can be used as a source of valuable information on flood events. In particular, the importance and stakes associated to floods can be inferred both from direct (Wikidata) and indirect (Wikipedia) metrics. It could support these data being used to study other kinds of events or even topics, as it demonstrates in particular their ability to provide (con)textual information on these on a wide or even global scale, despite some need for adjustement in the analysis due to possible problems of representativity or precision.

### Acknowledgements

This work is founded by the French National Research Agency (ANR) (Project GloUrb n° ANR-22-CE03-0005). It is also cofounded by the Labex IMU (ANR-10-LABX-0088) and the EUR H2O’Lyon (ANR-17-EURE-0018) of Université de Lyon, within the programme ‘Investissements d’Avenir’ operated by ANR.

Antin, Judd. 2011. “My Kind of People?: Perceptions about Wikipedia Contributors and Their Motivations.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3411–20. Vancouver BC Canada: ACM. <https://doi.org/10.1145/1978942.1979451>.

Ash, James, Rob Kitchin, and Agnieszka Leszczynski. 2018. “Digital Turn, Digital Geographies?” *Progress in Human Geography* 42 (1): 25–43. <https://doi.org/10.1177/0309132516664800>.

Ballatore, Andrea, and Stefano De Sabbata. 2020. “Los Angeles as a Digital Place: The Geographies of User‐generated Content.” *Transactions in GIS* 24 (4): 880–902. <https://doi.org/10.1111/tgis.12600>.

Ballatore, Andrea, Mark Graham, and Shilad Sen. 2017. “Digital Hegemonies: The Localness of Search Engine Results.” *Annals of the American Association of Geographers* 107 (5): 1194–1215. <https://doi.org/10.1080/24694452.2017.1308240>.

Barnier, Julien. 2023. *Rainette: The Reinert Method for Textual Data Clustering*. <https://juba.github.io/rainette/>.

Bittner, Christian. 2017. “Diversity in Volunteered Geographic Information: Comparing OpenStreetMap and Wikimapia in Jerusalem.” *GeoJournal* 82 (5): 887–906. <https://doi.org/10.1007/s10708-016-9721-3>.

Brakenridge, G. Robert. n.d. “Global Active Archive of Large Flood Events.” Dartmouth Flood Observatory, University of Colorado, USA. Accessed December 1, 2023. <https://floodobservatory.colorado.edu/Archives/index.html>.

Burgstaller-Muehlbacher, Sebastian, Andra Waagmeester, Elvira Mitraka, Julia Turner, Tim Putman, Justin Leong, Chinmay Naik, et al. 2016. “Wikidata as a Semantic Framework for the Gene Wiki Initiative.” *Database* 2016: baw015. <https://doi.org/10.1093/database/baw015>.

Caquard, Sébastien. 2013. “Cartography I: Mapping Narrative Cartography.” *Progress in Human Geography* 37 (1): 135–44. <https://doi.org/10.1177/0309132511423796>.

Colavizza, Giovanni. 2020. “COVID-19 Research in Wikipedia.” *Quantitative Science Studies* 1 (4): 1349–80. <https://doi.org/10.1162/qss_a_00080>.

Crampton, Jeremy W., Mark Graham, Ate Poorthuis, Taylor Shelton, Monica Stephens, Matthew W. Wilson, and Matthew Zook. 2013. “Beyond the Geotag: Situating ‘Big Data’ and Leveraging the Potential of the Geoweb.” *Cartography and Geographic Information Science* 40 (2): 130–39. <https://doi.org/10.1080/15230406.2013.777137>.

Dixon, Barnali, RebeccaA. Johns, and Amada Fernandez. 2021. “The Role of Crowdsourced Data, Participatory Decision-Making and Mapping of Flood Related Events.” *Applied Geography* 128 (March): 102393. <https://doi.org/10.1016/j.apgeog.2021.102393>.

Doocy, Shannon, Amy Daniels, Sarah Murray, and Thomas D. Kirsch. 2013. “The Human Impact of Floods: A Historical Review of Events 1980-2009 and Systematic Literature Review.” *PLoS Currents*. <https://doi.org/10.1371/currents.dis.f4deb457904936b07c09daa98ee8171a>.

Farda-Sarbas, Mariam, and Claudia Müller-Birn. 2019. “Wikidata from a Research Perspective – A Systematic Mapping Study of Wikidata.” arXiv. <http://arxiv.org/abs/1908.11153>.

Gardner, Z., P. Mooney, S. De Sabbata, and L. Dowthwaite. 2020. “Quantifying Gendered Participation in OpenStreetMap: Responding to Theories of Female (Under) Representation in Crowdsourced Mapping.” *GeoJournal* 85 (6): 1603–20. <https://doi.org/10.1007/s10708-019-10035-z>.

Giles, Jim. 2005. “Internet Encyclopaedias Go Head to Head.” *Nature* 438 (7070): 900–901. <https://doi.org/10.1038/438900a>.

Goodchild, Michael F. 2007. “Citizens as Sensors: The World of Volunteered Geography.” *GeoJournal* 69 (4): 211–21. <https://doi.org/10.1007/s10708-007-9111-y>.

Graham, Mark. 2014. “Inequitable Distributions in Internet Geographies: The Global South Is Gaining Access, but Lags in Local Content.” *Innovations: Technology, Governance, Globalization* 9 (3-4): 3–19. <https://doi.org/10.1162/inov_a_00212>.

Graham, Mark, Stefano De Sabbata, Ralph Straumann, and Sanna Ojanperä. 2018. “Uneven Digital Geographies … and Why They Matter.” In *Sozial- Und Kulturgeographie*, edited by kollektiv orangotango+, 2nd ed., 26:308–15. Bielefeld, Germany: transcript Verlag. <https://doi.org/10.14361/9783839445198-041>.

Graham, Mark, Stefano De Sabbata, and Matthew A. Zook. 2015. “Towards a Study of Information Geographies: (Im)mutable Augmentations and a Mapping of the Geographies of Information.” *Geo: Geography and Environment* 2 (1): 88–105. <https://doi.org/10.1002/geo2.8>.

Haklay, Mordechai (Muki). 2016. “Why Is Participation Inequality Important?” In *European Handbook of Crowdsourced Geographic Information*, Cristina Capineri, Muki Haklay, H Huang, Vyron Antoniou, J Kettunen, F Ostermann, and R Purve, 35–44. London: Ubiquity Press.

Halavais, Alexander, and Derek Lackaff. 2008. “An Analysis of Topical Coverage of Wikipedia.” *Journal of Computer-Mediated Communication* 13 (2): 429–40. <https://doi.org/10.1111/j.1083-6101.2008.00403.x>.

Herfort, Benjamin, Sven Lautenbach, João Porto De Albuquerque, Jennings Anderson, and Alexander Zipf. 2021. “The Evolution of Humanitarian Mapping Within the OpenStreetMap Community.” *Scientific Reports* 11 (1): 3037. <https://doi.org/10.1038/s41598-021-82404-z>.

Hilker, N., A. Badoux, and C. Hegg. 2009. “The Swiss Flood and Landslide Damage Database 1972–2007.” *Natural Hazards and Earth System Sciences* 9 (3): 913–25. <https://doi.org/10.5194/nhess-9-913-2009>.

Hu, Pan, Qiang Zhang, Peijun Shi, Bo Chen, and Jiayi Fang. 2018. “Flood-Induced Mortality Across the Globe: Spatiotemporal Pattern and Influencing Factors.” *Science of The Total Environment* 643 (December): 171–82. <https://doi.org/10.1016/j.scitotenv.2018.06.197>.

Iba, Takashi, Keiichi Nemoto, Bernd Peters, and Peter A. Gloor. 2010. “Analyzing the Creative Editing Behavior of Wikipedia Editors.” *Procedia - Social and Behavioral Sciences* 2 (4): 6441–56. <https://doi.org/10.1016/j.sbspro.2010.04.054>.

Iwan, Tomer. 2024. *Polyglotr: Translate Text*. <https://github.com/Tomeriko96/polyglotr/>.

Jak, Martine, and Matthijs Kok. 2000. “A Database of Historical Flood Events in the Netherlands.” In *Flood Issues in Contemporary Water Management*, edited by Jiri Marsalek, W. Ed Watt, Evzen Zeman, and Friedhelm Sieker, 139–46. Dordrecht: Springer Netherlands. <https://doi.org/10.1007/978-94-011-4140-6_15>.

Keegan, Brian, Darren Gergle, and Noshir Contractor. 2013. “Hot Off the Wiki: Structures and Dynamics of Wikipedia’s Coverage of Breaking News Events.” *American Behavioral Scientist* 57 (5): 595–622. <https://doi.org/10.1177/0002764212469367>.

Kundzewicz, Zbigniew W., Iwona Pińskwar, and G. Robert Brakenridge. 2013. “Large Floods in Europe, 1985–2009.” *Hydrological Sciences Journal* 58 (1): 1–7. <https://doi.org/10.1080/02626667.2012.745082>.

Lafon, Pierre. 1980. “Sur La Variabilité de La Fréquence Des Formes Dans Un Corpus.” *Mots* 1 (1): 127–65. <https://doi.org/10.3406/mots.1980.1008>.

Lê, Sébastien, Julie Josse, and François Husson. 2008. “FactoMineR: An r Package for Multivariate Analysis.” *Journal of Statistical Software* 25 (1): 1–18. <https://doi.org/10.18637/jss.v025.i01>.

Loiseau, Sylvain, Lise Vaudor, Matthieu Decorde, and Serge Heiden. 2022. *Textometry: Textual Data Analysis Package Used by the TXM Software*. <https://cran.r-project.org/web/packages/textometry/>.

Lorini, Valerio, Javier Rando, Diego Saez-Trumper, and Carlos Castillo. 2020. “Uneven Coverage of Natural Disasters in Wikipedia: The Case of Flood.” <https://doi.org/10.48550/ARXIV.2001.08810>.

Mashhadi, Afra, Giovanni Quattrone, and Licia Capra. 2015. “The Impact of Society on Volunteered Geographic Information: The Case of OpenStreetMap.” In *OpenStreetMap in GIScience*, edited by Jamal Jokar Arsanjani, Alexander Zipf, Peter Mooney, and Marco Helbich, 125–41. Cham: Springer International Publishing. <https://doi.org/10.1007/978-3-319-14280-7_7>.

Mittermeier, John C., Ricardo Correia, Rich Grenyer, Tuuli Toivonen, and Uri Roll. 2021. “Using Wikipedia to Measure Public Interest in Biodiversity and Conservation.” *Conservation Biology* 35 (2): 412–23. <https://doi.org/10.1111/cobi.13702>.

Moats, David. 2019. “Following the Fukushima Disaster on (and Against) Wikipedia: A Methodological Note about STS Research and Online Platforms.” *Science, Technology, & Human Values* 44 (6): 938–64. <https://doi.org/10.1177/0162243918815234>.

Najibi, Nasser, and Naresh Devineni. 2018. “Recent Trends in the Frequency and Duration of Global Floods.” *Earth System Dynamics* 9 (2): 757–83. <https://doi.org/10.5194/esd-9-757-2018>.

Noucher, Matthieu. 2023. “Blancs Des Cartes Et Boîtes Noires Algorithmique.” Paris: CNRS Éditions.

O’Reilly, Tim. 2020. “4. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software.” In *The Social Media Reader*, edited by Michael Mandiberg, 32–52. New York University Press. <https://doi.org/10.18574/nyu/9780814763025.003.0008>.

Oeberst, Aileen, Iassen Halatchliyski, Joachim Kimmerle, and Ulrike Cress. 2014. “Knowledge Construction in Wikipedia: A Systemic-Constructivist Analysis.” *Journal of the Learning Sciences* 23 (2): 149–76. <https://doi.org/10.1080/10508406.2014.888352>.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

Ratinaud, Pierre. 2009. “Iramuteq: Interface de R Pour Les Analyses Multidimensionnelles de Textes Et de Questionnaires.” <http://www.iramuteq.org>.

Reinert, Max. 1990. “Alceste Une Méthodologie d’analyse Des Données Textuelles Et Une Application: Aurelia De Gerard De Nerval.” *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 26 (1): 24–54. <https://doi.org/10.1177/075910639002600103>.

Royal, Cindy, and Deepina Kapila. 2009. “What’s on Wikipedia, and What’s Not . . . ?: Assessing Completeness of Information.” *Social Science Computer Review* 27 (1): 138–48. <https://doi.org/10.1177/0894439308321890>.

Ruprechter, Thorsten, Keith Burghardt, and Denis Helic. 2023. “Poor Attention: The Wealth and Regional Gaps in Event Attention and Coverage on Wikipedia.” Edited by Robin Haunschild. *PLOS ONE* 18 (11): e0289325. <https://doi.org/10.1371/journal.pone.0289325>.

Saharia, Manabendra, Avish Jain, Ronit Raj Baishya, Saagar Haobam, O. P. Sreejith, D. S. Pai, and Arezoo Rafieeinasab. 2021. “India Flood Inventory: Creation of a Multi-Source National Geospatial Database to Facilitate Comprehensive Flood Research.” *Natural Hazards* 108 (1): 619–33. <https://doi.org/10.1007/s11069-021-04698-6>.

Sekajugo, John, Grace Kagoro-Rugunda, Rodgers Mutyebere, Clovis Kabaseke, Esther Namara, Olivier Dewitte, Matthieu Kervyn, and Liesbet Jacobs. 2022. “Can Citizen Scientists Provide a Reliable Geo-Hydrological Hazard Inventory? An Analysis of Biases, Sensitivity and Precision for the Rwenzori Mountains, Uganda.” *Environmental Research Letters* 17 (4): 045011. <https://doi.org/10.1088/1748-9326/ac5bb5>.

Senaratne, Hansi, Amin Mobasheri, Ahmed Loai Ali, Cristina Capineri, and Mordechai (Muki) Haklay. 2017. “A Review of Volunteered Geographic Information Quality Assessment Methods.” *International Journal of Geographical Information Science* 31 (1): 139–67. <https://doi.org/10.1080/13658816.2016.1189556>.

Silge, Julia, and David Robinson. 2016. “Tidytext: Text Mining and Analysis Using Tidy Data Principles in R.” *JOSS* 1 (3). <https://doi.org/10.21105/joss.00037>.

Specht, Doug. 2022. “The Codification of Local Knowledges Through Digital Cartographic Artefacts: A Case Study of the Humanitarian OpenStreetMap Team in Dar Es Salaam, Tanzania.” <https://doi.org/10.34737/W0607>.

Stephens, Monica. 2013. “Gender and the GeoWeb: Divisions in the Production of User-Generated Cartographic Information.” *GeoJournal* 78 (6): 981–96. <https://doi.org/10.1007/s10708-013-9492-z>.

Tellman, B., J. A. Sullivan, C. Kuhn, A. J. Kettner, C. S. Doyle, G. R. Brakenridge, T. A. Erickson, and D. A. Slayback. 2021. “Satellite Imaging Reveals Increased Proportion of Population Exposed to Floods.” *Nature* 596 (7870): 80–86. <https://doi.org/10.1038/s41586-021-03695-w>.

Turki, Houcemeddine, Denny Vrandecic, Helmi Hamdi, and Imed Adel. 2017. “Using WikiData as a Multi-Lingual Multi-Dialectal Dictionary for Arabic Dialects.” In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, 437–42. Hammamet: IEEE. <https://doi.org/10.1109/AICCSA.2017.115>.

Van Velsen, Lex, Julia E. W. C Van Gemert-Pijnen, Desirée J. M. A Beaujean, Jobke Wentzel, and Jim E Van Steenbergen. 2012. “Should Health Organizations Use Web 2.0 Media in Times of an Infectious Disease Crisis? An In-Depth Qualitative Study of Citizens’ Information Behavior During an EHEC Outbreak.” *Journal of Medical Internet Research* 14 (6): e181. <https://doi.org/10.2196/jmir.2123>.

Vaudor, Lise, and Maëlle Salmon. 2023. *Glitter: Glitter Makes SPARQL*. <https://lvaudor.github.io/glitter/>.

Wehn, Uta, Maria Rusca, Jaap Evers, and Vitavesca Lanfranchi. 2015. “Participation in Flood Risk Management and the Potential of Citizen Observatories: A Governance Analysis.” *Environmental Science & Policy* 48 (April): 225–36. <https://doi.org/10.1016/j.envsci.2014.12.017>.

Wickham, Hadley. 2024. *Rvest: Easily Harvest (Scrape) Web Pages*. <https://rvest.tidyverse.org/>.

Wikipedia, Contributors. 2024a. “Citing Wikipedia.” <https://en.wikipedia.org/wiki/Wikipedia:Notability>.

———. 2024b. “List of Floods.” <https://en.wikipedia.org/wiki/List_of_floods>.

Winsemius, H. C., L. P. H. Van Beek, B. Jongman, P. J. Ward, and A. Bouwman. 2013. “A Framework for Global River Flood Risk Assessments.” *Hydrology and Earth System Sciences* 17 (5): 1871–92. <https://doi.org/10.5194/hess-17-1871-2013>.

Zhao, Fudie. 2023. “A Systematic Review of Wikidata in Digital Humanities Projects.” *Digital Scholarship in the Humanities* 38 (2): 852–74. <https://doi.org/10.1093/llc/fqac083>.