Landslides (2020) 17:2205–2215
DOI 10.1007/s10346-020-01431-5
Received: 23 December 2019
Accepted: 5 May 2020
Published online: 11 June 2020
© The Author(s) 2020

Thomas M. Kreuzer · Bodo Damm

Automated digital data acquisition for landslide inventories

Abstract Landslide research relies on landslide inventories for a multitude of spatial, temporal, or process analyses. Generally, it takes high effort to populate a landslide inventory with relevant data. In this context, the present work investigated an effective way to handle vast amounts of automatically acquired digital data for landslide inventories by the use of machine learning algorithms and information filtering. Between July 2017 and February 2019, a keyword alert system provided 4381 documents that were automatically processed to detect landslide events in Germany. Of all those documents, 91% were automatically recognized as irrelevant or duplicates; thereby, the data volume was significantly reduced to contain only actual landslide documents. Moreover, it was shown that inclusion of the document's images into the automated process chain for information filtering is recommended, since otherwise unobtainable important information was found in them. Compared with manual methods, the automated process chain eliminated personal idiosyncrasies and human error and replaced it with a quantifiable machine error. The applied individual algorithms for natural language processing, information retrieval, and classification have been tried and tested in their respective fields. Furthermore, the proposed method is not restricted to a specific language or region. All languages on which these algorithms are applicable can be used with the proposed method and the training of the process chain can take any geographical restriction into account. Thus, the present work introduced a method with a quantifiable error to automatically classify and filter large amounts of data during automated digital data acquisition for landslide inventories.

Keywords Landslide inventory · Data acquisition · Machine learning · Document classification · Information filtering

Introduction

Landslide research chiefly relies on landslide inventories (here synonymous with databases) for a multitude of spatial, temporal, or process analyses (Van Den Eeckhaut and Hervás 2012; Klose et al. 2015). Generally, it takes high effort to populate a landslide inventory with relevant data. Therefore, researchers have applied different strategies that, following a similar classification as Guzzetti et al. (2012), can be differentiated into two main categories: for one, data derived from morphological examination by fieldwork, remote sensing products, or cartographic analysis and, secondly, data derived from textual sources and, if present henceforth always considered, their accompanying images (usually ground images). Such textual sources can be acquired from scientific publications, reports of varying agencies (e.g., civil protection, police, building authorities, road construction offices), newspaper articles, and unpublished documents (e.g., church records or historical archives). Overall, data acquisition from textual sources is an effective method (Wohlers et al. 2017; Rupp et al. 2018) that could further profit from digitalization. While there are many inventories in use which were created in large parts, or even exclusively, based on textual sources (Guzzetti et al. 1994; Devoli et al. 2007; Foster et al. 2012; Liu et al. 2013; Hess et al. 2014; Pereira et al. 2014; Damm and Klose 2015; Raska et al. 2015; Rosser et al. 2017; Valenzuela et al. 2017; Piacentini et al. 2018), few publications provide a methodological approach on how to acquire those sources in digital form. Battistini et al. (2013) and Taylor et al. (2015) propose a keyword search of digital news archives, where Innocenzi et al. (2017), Klimeš et al. (2017), and Voumard et al. (2018) use a keyword-based Internet monitoring service from Google, called Google Alerts. It makes all sites newly registered with Google available to the user of the service if they contain user-defined keywords. Calvello and Pecoraro (2018) propose a combination of both previously mentioned methods. In sum, all approaches start with a keyword search, which most people know from entering a search term into a search engine like Google Search. Furthermore, keywords can be logically combined with AND, OR, and NOT for these searches.

In general, the problem with a keyword search is its provision of the user with duplicates and unwanted results (false positives) thus increasing the data volume for the user. For this reason, Battistini et al. (2013) and Taylor et al. (2015) restrict their search to specific news archives to take advantage of a preselection of textual sources and further propose to adapt the used keywords based on previous results. This means, they manually identify keywords that have a high probability of occurring in either wanted or unwanted results and use these detected keywords to minimize false positives in further searches. In comparison, Innocenzi et al. (2017) use only one keyword—"frana" (landslide in Italian)—to avoid too many unwanted results; the remaining results are filtered manually. The other studies do not provide details to reduce unwanted results; however, Calvello and Pecoraro (2018) also use only "frana" as a keyword (albeit in singular and plural form), which may indicate the same reasoning behind this decision as reported by Innocenzi et al. (2017).

The overarching goal of the present work is to increase the productivity of digital data acquisition for any type of source with any number of keywords through automated reduction of irrelevant data; hence, the use of digital documents available on the Internet and related machine learning algorithms from the field of information retrieval (IR). After Manning et al. (2009), IR is defined as follows:

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Therefore, an automated process chain is facilitated that processes vast amounts of potential documents on landslides to enable its user to handle them effectively. In this context, the

questions that arise are if the method can reliably filter unwanted results and are the remaining results relevant for landslide inventories.

The present work further pursues the idea of digital data acquisition by extending an automated keyword search with automated extraction and classification methods for texts and images. Therefore, an automated process chain is established that monitors new entries on the Internet for landslide events and, at the same time, identifies and discards false positives. Moreover, multiple documents that provide information about the same event are identified automatically and hidden from the user until they are specifically requested. The present work applies these methods to German language texts in order to use the results for the "National Landslide Database for Germany" (Damm and Klose 2015; Kreuzer et al. 2017); however, these methods are meant to be applicable to other languages with minor modifications. Furthermore, in addition to an automated process chain for landslide document aggregation, this study assesses the quality and characteristics of its textual and visual results. For that reason, various information classes are introduced that are tested for their respective presence in the results, for example, location, magnitude, activity, and date of the respective landslide process. Moreover, an inventory map is compiled from the results of the process chain to present a practical example.

Methods

Automated digital data acquisition and processing

Established methods from the field of IR (Manning et al. 2009) are used to create an automated process chain for digital data acquisition and processing. A short introduction into the applied methodical principles of IR is given, since the target audience of this work might not be familiar with them.

Introduction to IR

For this study, the three main principles of information retrieval from texts are (i) tokenization, (ii) part of speech tagging (POS), and (iii) word vectoring. All three methods rely on statistically deduced insights about the respective language; for example, how is the relative occurrence of a certain word in the analyzed texts, how often does a specific word occur left or right from another specific word, or what is the likelihood for two specific words to occur together within a sentence. These results are stored in a *corpus*. The present work uses the publicly available "Leipzig Corpora Collection" for the German language (Goldhahn et al. 2012).

I. *Tokenization* means to decompose text into tokens. In the first order, these tokens are sentences, and, in the second-order, words. Sentence tokenization relies on statistical information for punctuation that means a sentence like "A landslide occurred near St. Louis." is not split into two sentences because the cut-off point is mistaken for a full stop. Word tokenization further removes tokens that represent either punctuations or stopwords (words that occur very frequently in language texts, e.g., in English "a" and "is" are stopwords). That is because punctuation and stopwords are usually irrelevant for text analysis (Aggarwal and Zhai 2012). The purpose of tokenization is to compare two sets of tokens and for this reason, the tokenized words need to be reduced to their word stem, since languages use flexion to alter words in a grammatical context. Thus, two sentences should be very similar if they carry the same meaning and words but are, for example, written in different tenses. In this work, the word stemmer from Weissweiler and Fraser (2018) is used to acquire a set of stemmed tokens.
II. *Part of speech tagging* uses probabilities of word positions to determine the sentence unit they represent. For example, "slide" can be either a noun or a verb in the English language. In this case, a probabilistic decision has to be made based on the position of the word in the sentence. The present work implements a decision tree–based method with a reported accuracy of 96.36% (Schmid 1999).
III. *Word vectors* describe the semantic distance, i.e., similarity, of a word to all other words in the corpus. For example, the semantic distance from "politician" to "president" is very close; thus, when assessing word similarity, they could be considered equal. Moreover, word vectors allow for word calculations, e.g., statistically the occurrence of "mother" is strongly correlated, i.e., has a short distance, to the occurrences of "woman" and "child"; thus, "woman" + "child" ≈ "mother."

Automated process chain

The automated digital data acquisition works in a timely manner. Textual sources that potentially describe landslides are processed as soon as possible after they are published. For evaluation of the process chain, digital textual sources have been collected between July 2017 and February 2019. The major steps of the automated process chain are outlined in Fig. 1 and described in the following:

A. *Define keywords*. Keywords are the most language-dependent item in the automation process. Their definition commonly stems from empiricism and knowledge about language ambiguity. In this case, the keywords come from science—mostly inspired by Cruden and Varnes (1996)—and media through previously conducted manual searches since 1998 (Damm and Klose 2015). Here, they are single words and logically "or"-combined; i.e., a result is found if any of the keywords occur in the document. The keywords have to be adjusted as often as the respective language changes. Since this is not expected to happen frequently, the adjustment is the only manual item in the process chain.
B. *Monitor the internet for newly published textual sources containing any of the keywords from step A*. Google LLC runs an alert service called "Google Alerts." It provides a machine-readable interface, a so-called RSS-Feed, for documents freshly registered with Google Search. These documents also contain a logical "or"-combination of predefined keywords, here taken from step A. Google Alerts also uses word stemming to find documents with derivatives of the supplied keywords. The RSS-Feed always contains entries about the last 16 documents that were registered with Google. If a new document is registered with Google, the oldest document entry in the feed is deleted and the new document is inserted. Therefore, a custom-written program queries the RSS-Feed hourly.

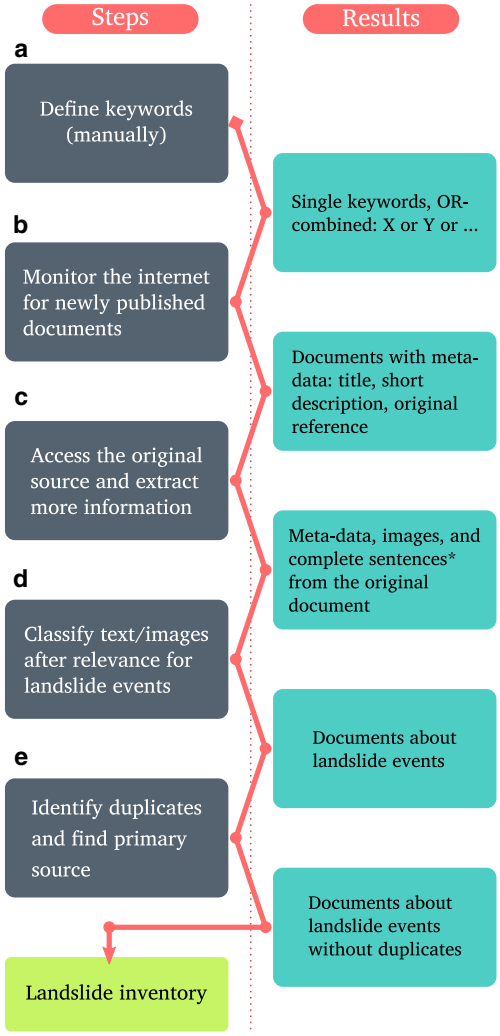


Fig. 1 Outline of the automated process chain for digital data acquisition. A is the only manual process step and it is not continuously repeated. Steps B–E are programmatically executed every hour

Furthermore, the RSS-Feed contains only document meta-data, i.e., title, short content description, publishing date, and the reference to the original source.

C. *Access the original source from newly generated content through step B and extract more information.* The same program from step B accesses all original sources from the newly acquired content and searches them for sentences that are complete; i.e., the sentence has at least a subject, predicate, and object; has punctuation; and contains any keyword of step A. If a document does not contain a complete sentence, it is marked as irrelevant; otherwise, these sentences are added to the respective database entry for further analysis.

D. *Classify text and images.* For text classification, the multinomial naive Bayes algorithm is applied (Zhang 2005). The naive Bayes classifier is a supervised learning algorithm; i.e., before it can be applied, it is trained and verified with a pre-classified data set. Thus, for the first inception of the classifier for continuous use, the data collected from steps B and C has to be manually classified as invalid and valid data. "Valid" means the textual source is about a landslide event of interest (here any landslide event within Germany), "invalid" that it is not. After the manual classification, the aforementioned program randomly shuffles and divides the classified textual sources into two equally sized sets of documents and utilizes one set to train and the other in order to verify the classifier. The textual information comes from the title, the short description, and the additionally extracted sentences described in steps B and C. The basic principle behind the classifier is to count occurrences of tokenized words in classes of the training set and thus deduce the probability of specific words to indicate a specific class. The complementary probability of all assessed words indicates the resulting class. Subsequently, the classifier predicts classes for textual sources from the verification set. These predictions are then compared with the manual classifications. In addition to text classification, all images from textually valid documents are also automatically classified as valid and invalid through the method of logistic regression (León et al. 2007). In this case, "valid" means the image depicts an actual landslide and "invalid" that it does not. A prerequisite for image classification is that all images have the same size. In this case, all images are automatically resized to 640 × 480 pixels and converted to greyscale mode. The latter is important to reduce features for the classifier. This means, only the brightness of a pixel is taken into account, not the original color value. For classification, the pixel occurrences of an image are fitted against a logistic model. Based on the model parameters, the validity class of the image is deduced. Logistic regression classification is also a supervised classification like the naive Bayes classifier and, as such, was trained and verified analogous to text classification.

Both classification results are stored in the database entry of the document.

E. *Identify duplicates.* The automated duplicate identification compares the tokenized sentences from steps B and C against each other for similarity. If the similarity metric of the sentences from two different documents X and Y falls below a certain threshold, these documents are considered duplicates. Furthermore, the duplicate detection conjoins multiple results; thus, for documents X, Y, and Z, it follows that

$$X \equiv Y \quad \wedge \quad Z \equiv Y \quad \Rightarrow \quad Z \equiv X$$

even though document Z was not marked as a duplicate of document X at first. Here, the "Fast Word Mover's Distance" (WMD) (Kusner et al. 2015) is used to calculate the similarity metric through the sum of the word vector distances, see (iii), of the respective sentences. It also utilizes word calculations, so that multiple words can be compared with a single word and vice versa. Two completely equal sentences have a distance of 0. If a sentence is included in another sentence word for word but also with additional text, the WMD is not 0, even though it implies equality. Therefore, word tokens are removed from the longer sentence until both sentences have an equal number of tokens. Here, tokens that would increase

the overall distance the most are removed first. The utilized similarity threshold is determined by manually identifying duplicates in the collected data set and testing their automated identification by systematically increasing the threshold value for as long as no false duplicates are detected. The start value for this procedure is the theoretical minimum distance of θ. So far, the logic of the algorithm leads to the declaration of the document that was published first as the primary source, of which all subsequently identified duplicates are later published documents. This is not practical, since the primary source should be the document with the highest information content. For this reason, the duplicate with the highest token count coming from complete landslide sentences extracted in step C and the most amount of valid landslide images detected from step D are marked as the primary source in the database. The duplicates are hidden but stay retrievable at any time and are not discarded.

Together, steps A–E provide the methodical approach for the automated process chain. Step A, as well as the training and verification of step D, and the threshold detection of step E do not have to be repeated for the continuous use of the process chain. However, since steps D and E rely on statistical results, periodic re-training might increase their performance, and thus, this is incorporated into the software design for the process chain.

Document quality assessment

The valid documents collected by the process chain from the "Automated digital data acquisition and processing" are brought into a qualitative context to motivate their usage in landslide inventories. Therefore, in case a document's text or image is determined to relate to landslides, the next step consists of the identification of its precise content. Thus, thirteen attribute classes commonly used in landslide inventories are used to manually check the document (text or images) whether it contains any corresponding information. For example, a document could contain information about the attribute class "corrective measure" through an image that depicts a road cleanup or the text mentions a specific slope stabilization method like the construction of a safety fence. Table 1 shows the list of the utilized attribute classes and expected content that needs to be found in the document's text or images to be counted as fulfilling the respective information need.

Additionally to the manual check for information in images as described above, images are also automatically scanned for information from their metadata. The most widely used standard for metadata in images is the "Exchangeable image file format" (Exif) developed by Japan Electronics and Information Technology Industries Association (2019). Specifically, Exif is a machine-readable file header attached to image files from digital cameras, including smartphones. For example, such metadata may contain any of the following: the geographic coordinates where the image was taken, the name of the location depicted in the image, the date the image was taken, or a short description of the image.

Last, an inventory map is compiled from the results of the process chain. Therefore, the location information found during the aforementioned document review is manually cross-referenced with topographic maps in order to acquire coordinates and assess their respective accuracy. In this context, the determined coordinates are the midpoint of a circle encompassing all possible locations for the landslide event and its radius is used as an accuracy metric. This means, in case a document provides coordinates or the exact position is reliably inferred (e.g., with the help of photos or morphological constraints), the radius equals θ (exact location). Moreover, for comparison purposes, the accuracy is specified by three "spatial confidence descriptors" modified after Calvello and Pecoraro (2018): exact (Sd1), less or equal than 500 (Sd2), and greater than 500 (Sd3).

Results

Automated process chain

Ten keywords (cp. Fig. 2) were defined that are nouns and are used in scientific or colloquial language; thus, the meaning of some keywords may be redundant. During the course of 87 weeks, from July 2017 to February 2019, the use of these keywords produced 4381 documents that were automatically collected from the Google Alert RSS-Feed. Out of these documents, 480 (10.95%) were manually classified as *valid* landslide sources. In Fig. 2, the performance of the ten keywords with respect to *valid* and *invalid* results is presented; particularly, five keywords are responsible for ≈96% of the *valid* documents.

For the given period of time, the weekly average for the number of the RSS-Feed provided documents lies at 5.51 for *valid* and at 44.84 for *invalid* documents; i.e., *invalid* documents are ≈8 times more frequent on average. The range of the number of provided documents lies between 7 and 134 documents per week. In Fig. 3, the time series of collected documents is shown for the year 2018.

With respect to the automated content extraction, complete sentences were automatically extracted from 2768 (63.18%) of the 4381 overall documents. Furthermore, the algorithm retrieved landslide images from the majority (54%) of the 480 *valid* documents (Fig. 4).

The classification results from the naive Bayes classifier for texts and the logistic regression for images are presented in two-class (*valid*/*invalid*) "confusion matrices" (Ting 2017). In these confusion matrices, relative agreement of correct and incorrect class predictions of the respective classifier with the actual manual classifications is shown. Specifically, the sum of the matrix's main diagonal, and thus the overall agreement of *valid* and *invalid* classifications from automated and manual classification, corresponds to the accuracy of the classifier. Since the process chain algorithm discards exclusively *invalid* documents, only documents erroneously classified as *invalid* are lost for the process. Falsely classified *valid* documents are retained for the process; however, they increase the data volume unnecessarily. Table 2 shows the confusion matrix for the naive Bayes classifier and Table 3 for the results of the logistic regression on images.

Primary documents and their respective duplicates were manually and automatically identified. In Fig. 5, the respective results are presented. In the total sum of 480 detected landslide documents, a percentage of 43.51% are actual duplicates. In comparison, the algorithm identified 35.83% of the 480 landslide documents as duplicates, thus 86.12% of the manual identification. In this case, the value of 86.12% directly corresponds to the overall accuracy, since the algorithm does not produce any falsely identified duplicates with a document similarity threshold set to 0.89

Table 1 Common landslide attribute classes whose presence is examined in texts and images from the automated process chain to assess the respective document's eligibility for landslide inventories

Name	Document content
Activity	Concerns/facts about an ongoing process
Corrective measure	Type of corrective measures
Cost	Actual/estimated costs for damages or corrective measures
Damage	Damaged objects/persons
Date	Date of occurrence
Lithology	Lithology involved
Location	Coordinates/geographic names
Magnitude	Actual/estimated landslide volume
Morphometry	Extent, depth, or gradient
Movement speed	Estimated speed/timespan of the event
Preparatory factor	Previous destabilizing conditions/incidents
Trigger	Directly preceding event
Type	Process or block size

(cp. "Automated digital data acquisition and processing"). It is emphasized that 93% of the primary sources have 3 duplicates or less.

All in all, during the test period of 87 weeks, the automated process chain yielded 385 potential landslide documents and 214 duplicates (error 8.24%). Furthermore, the automated process chain provided 297 images of landslide events (error 4.09%) out of these documents (including duplicates).

Quality assessment

In Fig. 6, the results of the quality assessment after 2 are presented for the manually identified *valid* results: 480 landslide documents

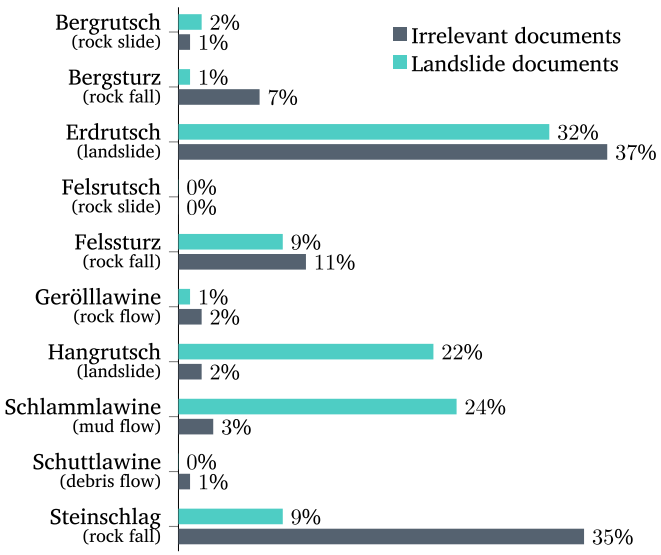


Fig. 2 Share of results from the RSS-Feed of the respective German keyword for valid ($n = 480$) and invalid documents ($n = 3901$) respectively. English terms after Cruden and Varnes (1996), the redundancy comes from doubling of colloquial and scientific terms, as well as a different classification system in German

(271 primary sources and 209 duplicates) and their respective 259 images.

Almost all texts contained information on location and date of the landslide event (94% and 96% respectively). Date and location information for images (Fig. 6) are provided by the corresponding image metadata. Overall, the metadata of 18 (6.56%) images provided a date when the image was taken. Metadata for the location of the subject of the image was present in 11 (4.35%) images; in 1 of these cases, the metadata provided geographic coordinates, and the other 10 cases provided geographic names.

During the quality assessment, the locations of 201 landslide events were discovered and cataloged in an inventory map (Fig. 7). In general, the discovered location information consisted of geographic names, which had to be cross-referenced with topographic maps. Only four documents provided geographic coordinates. The accuracies of the detected locations range from 05.5 and are divided into Sd1 =22.40%, Sd2 =51.37%, and Sd3 =26.23%. An example of how the respective accuracies come about is presented in Fig. 8.

Discussion

Particularly relevant for the discussion are the works of Taylor et al. (2015) and Innocenzi et al. (2017). Both publications provide a detailed evaluation of their respective acquisition process. Other works (Battistini et al. 2013; Klimeš et al. 2017; Calvello and Pecoraro 2018; Voumard et al. 2018) focus on (automated) analysis of data rather than acquisition processes and thus do not provide details that would be needed for a comparison here. This is also the case for works from other fields of research like medicine (Weichelt et al. 2018).

In principle, Innocenzi et al. (2017) applied steps A and B from the proposed process chain, whereas the remaining steps were manually implemented. Taylor et al. (2015) applied step A in a more extensive manner; their additional search for a logical combination of keywords to reduce irrelevant results can be seen as a modification of a naive Bayes classifier, where every word has a

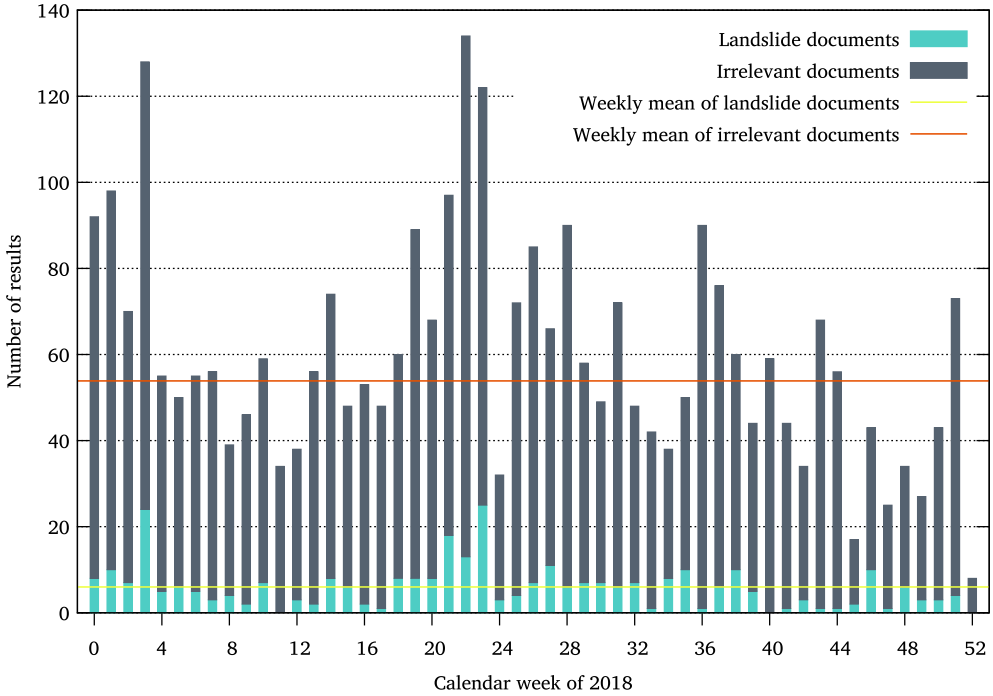


Fig. 3 Time series of document results per week from the automated process chain for the year 2018. The results are based on expert-based classification. The annual mean for landslide documents is 5.98 and for irrelevant results 53.87 documents per week

probability of 100% to refer to a specific class. Thus, Taylor et al. (2015) partly applied step D semi-automatically, whereas steps B, C, and E are applied manually. Table 4 gives an overview of the respective automated and manual steps from the process chain. The comparison is for textual sources only, since neither of Innocenzi et al. (2017) nor Taylor et al. (2015) report results for images.

In general, the keyword search is language-specific; however, common principles apply and motivate a deeper comparison with studies of different languages. The present work found 10 German keywords producing 3172 results of which 2855 (90.01%) were irrelevant for the year 2018. Innocenzi et al. (2017), who used the same monitoring service (Google Alerts) using one Italian keyword, produced 2737 results on average per year (10947 during the years 2012–2015) with 17.14% irrelevant documents (1876 during the years 2012–2015). In comparison, Taylor et al. (2015) identified 27 search terms for the English language, of them 8 must not occur in the document. In the case of the year 2006, these 27 search terms produced 711 documents of which 167 were irrelevant. Thus, the strategy to reduce irrelevant results by logical combination of search terms produced 23.50% irrelevant results (Table 5).

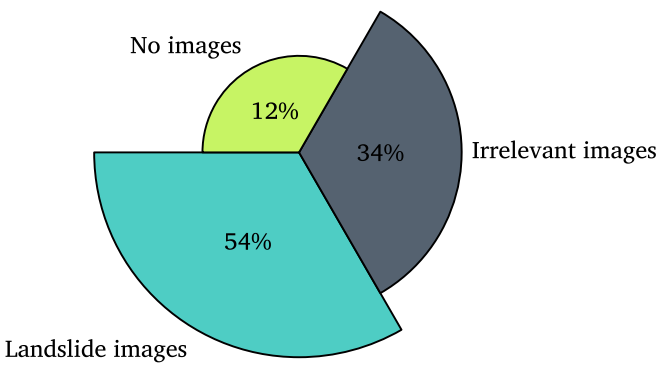


Fig. 4 Digitally acquired landslide documents ($n = 480$) with relevant, irrelevant, or no images

Compared with the present work, Innocenzi et al. (2017) report a lower number of irrelevant results in one year, even though they apply the same monitoring service and report a similar overall result count. In this case, only one Italian keyword was used to specifically produce such a low number of results. Corresponding to this strategy, if the present work would only use the most productive keyword with the least irrelevant results, i.e., "Hangrutsch" (scientific term for landslide in German), the relative number for irrelevant results would decrease by ≈50% (cp. Fig. 2). The relatively small difference in absolute numbers, i.e., one Italian keyword produces almost as much results as the four most productive German keywords, might be due to the larger absolute number of landslide events with impact in Italy. For example, the estimated annual losses caused by landslides are 3.9 billion Euros for Italy but only 0.3 billion Euros for Germany (Klose et al. 2016). The number of results for Italy would therefore increase if more than one keyword was used, especially keywords for different landslide processes. In this context, the respective keywords of the present work provide a diverse distribution of results. Thus, there are keywords that predominantly produce either relevant or irrelevant results (cp. Fig. 2). For example, as mentioned above, "Hangrutsch" (lit. trans. slope slide) is a scientific term for landslide in German which is rarely used colloquially; it produces ≈57.38% relevant results (22% from all relevant results). Its colloquial pendant is "Erdrutsch" (lit. trans. earth slide), which

Table 2 Confusion matrix for text classification with the naive Bayes classifier. The overall accuracy (= trace of the matrix) in italics in the lower right corner

n = 2768		Predicted (%)		
		Valid	Invalid	
Actual (%)	Valid	*13.40*	3.94	17.34
	Invalid	8.24	*74.42*	82.66
		21.64	78.36	***87.82***

produces eight times more irrelevant than relevant results; e.g., it is also used for the analogue to the English political term "landslide victory." In contrast, the term "Steinschlag" (lit. trans. rock fall) is a scientific as well as colloquial term; however, most of the time it refers to the damage in colloquial language but it always refers to the process in scientific language. As a result, many "Steinschlag" results come from documents which refer to object damages (specifically windshields) from gravel or stones flung by machines (specifically cars) or persons—it produces ≈97% irrelevant results (35% of all irrelevant results).

Other reasons for irrelevant results are mentions of slope security measures without a preceding landslide event, confusion of erosion or floods with landslide processes, warnings on possible landslides, advertisement for insurances of natural hazards, entertainment, and landslide events outside the region of interest. For example, in 2018, the above average number of irrelevant results for the weeks 21 to 23 (cp. Fig. 3) come from reports about deaths during landslide events triggered by monsoon rain events in (sub-)tropical regions.

Table 3 Confusion matrix for image classification with the logistic regression classifier. The overall accuracy (= trace of the matrix) in italics in the lower right corner

n = 432		Predicted (%)		
		Valid	Invalid	
Actual (%)	Valid	*58.99*	2.23	61.22
	Invalid	4.09	*34.69*	38.78
		63.08	36.92	***93.68***

Taylor et al. (2015) do not report in what quantity irrelevant results could be reduced by their logical keyword combination. However, their relative number for the reported irrelevant results of 23.50% for the year 2006 and 19.24% for the "landslide year" 2012 exceeds the 8.24% of the naive Bayes classifier from the proposed automated process chain (cp. Table 2). The same can be stated for Innocenzi et al. (2017) with 17.14% of irrelevant results. Furthermore, the naive Bayes classifier discards 3.94% of results that were falsely classified as irrelevant (Table 2). In contrast, Innocenzi et al. (2017) report 8% unidentified landslide events, and Taylor et al. (2015) do not report a possible loss of landslide documents due to the logical combination of search terms.

Since no naive Bayes classifier has been implemented specifically for landslide documents, the overall accuracy can only be compared with the results from other applications. For example, a common application of a naive Bayes classifier is email spam detection. In this context, the result of 87.82% overall accuracy for the classifier in the present work (Table 2) is worse than many reported spam filter classification accuracies that are mostly in the mid-90% range (Rusland et al. 2017). Although spam documents are designed to actively avoid identification as such, documents about landslides are not. Here, the main difference between both applications of the classifier is based on the fact that the landslide classifier considers only landslides from a specific geographic region as valid, while spam filters generally do not operate with such a geographic restriction. Specifically, the applied process chain ("Automated digital data acquisition and processing" section) produced 286 documents that contained information on landslides that occurred in other countries than Germany and where thus manually classified as invalid. Then, during the training phase of the classifier, the classifier decreases the probability of words that actually indicate a landslide event because the event is not within the region of interest. This affects all landslide document classifications and is probably one reason for the relative underperformance compared with email spam filters. Another source of error could be because of the text selection for the landslide classifier. Here, the process chain selected grammatically complete sentences that contain any of the predefined keywords (cp. Fig. 1). Thus, documents that contain only grammatically incomplete headlines, e.g., news aggregators, are principally avoided. Furthermore, the keyword requirement ensures the topic of the sentence is on landslides; however, this restriction could miss other landslide-relevant sentences, which otherwise would increase the performance of the classifier.

The image detection of the automated process chain performed generally better than the textual part. Compared with the naive Bayes classifier with an accuracy of 87.82%, the logistic regression detected relevant images with an accuracy of 93.68%. This meets the expected range of results according to image logistic regression classifications as used in other applications (León et al. 2007). Since only images from relevant documents are classified, the image classifier is not restricted by its geographic location; it solely decides whether an image depicts a landslide event or not. This fact further underlines the assumption that the naive Bayes classifier would perform better without a geographic restriction for the relevancy of the documents. A manual review of the irrelevant images showed that they are usually logos or archive footage.

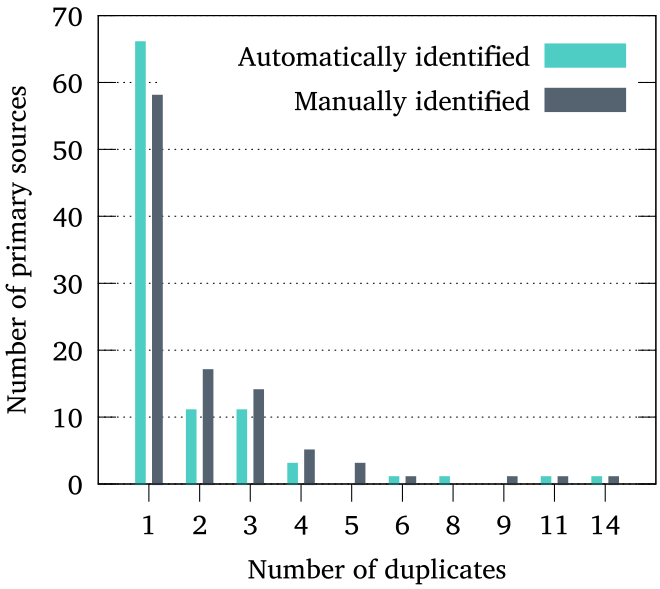


Fig. 5 Duplicate distribution in absolute numbers for the manually classified landslide documents ($n = 480$) divided into manual (101 primary sources with 209 duplicates in total) and automatic (95 primary sources with 172 duplicates in total) duplicate identification

The duplicate detection could reduce the data volume about 35.73% with 86.12% accuracy (cp. "Automated process chain" section). However, the duplicate detection algorithm is not designed to identify different documents that are about the same landslide event; it assesses the textual similarity of documents with each other. The algorithm works under the assumption that documents are often secondary sources and thus have a strong textual resemblance if they utilized the same original source. This is in contrast to Taylor et al. (2015) and Innocenzi et al. (2017) who report on manually identified "same event documents" that can be worded very differently. Taylor et al. (2015) found 43.30% "same event documents" and Innocenzi et al. (2017) found 86.86%. Given the different principals of the presented algorithm and manual identification, it follows that manual duplicate identification performs always better. Yet, the here-presented duplicate identification succeeds in its purpose of data reduction. Additionally, duplicates are optional data reductions, since they are not discarded like irrelevant results but are attached to the primary source. The data analyst can decide for himself whether the duplicates are accessed for cross-checking or not. Furthermore, in the amount of the 172 automatically identified duplicates in our study, there are 113 (65.70%) whose tokens were identical with the primary source. This usually indicates the reuse of content provided by news agencies. After all, the predominant source type (≈80%) of the monitoring service are news articles, which are also known to primarily report on "landslides with consequences" (Guzzetti et al. 2003, p. 472). On the one hand, this introduces a data bias; on the other hand, the amount of identified duplicates correlates with the involvement of the consequences. Specifically, the highest duplicated documents with 9, 11, and 14 duplicates (cp. Fig. 5) report on the reopening of a large, touristy attractive bathing lake after it was closed because of a landslide event, an important railway obstructed by a landslide, and a landslide event with fatalities. In comparison, a randomly selected document with only one duplicate is about a closed hiking trail. Images have not been tested for duplication, since it is assumed that duplicated images result from duplicated documents.

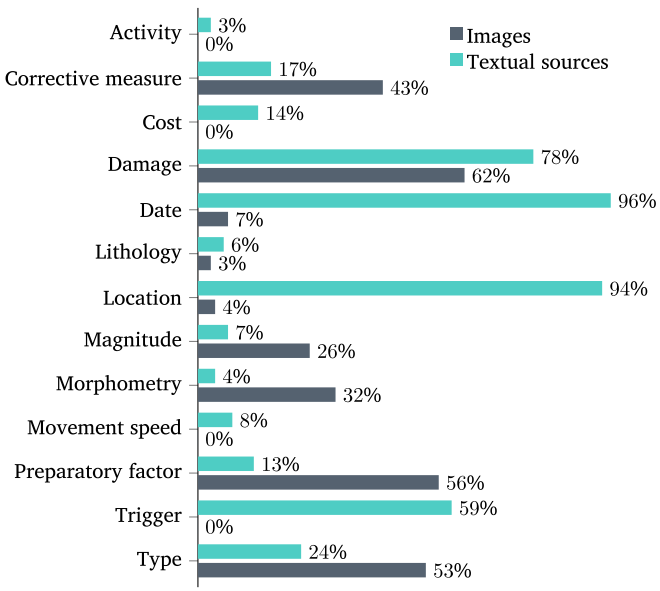


Fig. 6 Valid textual sources ($n = 480$) and images ($n = 264$) from automated digital data acquisition that contain information about the respective attribute class after Table 1

The automated process chain is implemented with a monitoring service under the assumption of a "living" landslide inventory, which continuously gets new landslide events added as they happen. Thus, the time series of detected landslide documents presented in Fig. 3 approximately corresponds to the time series of landslide occurrences. This enables immediate investigations of the reported landslide in a fresh state, e.g., by field survey, in order to acquire as much information as possible (Dikau et al. 1996; Lu et al. 2011). The timely manner of the monitoring service also very likely increases the performance of duplicate detection with the assumption that the likelihood of two documents being duplicates decreases with an increasing time gap between them. Nevertheless, the monitoring service works only in a timely manner for digital publishing, yet the initial publishing date of a document and its digital publishing date do not always coincide. Other temporal detachments are (news) updates on past landslide events, e.g., repairs of a landslide damage that have been finished years after the event. This temporal detachment exemplifies that a timely monitoring service is not a prerequisite for the application of steps C–E (cp. Fig. 1). Therefore, step B can be replaced by any keyword search and thus increase the possibilities for the process chain's application. For instance, Pennington et al. (2015) applied the keyword search proposed in Taylor et al. (2015) to social media content.

The results of the quality assessment described in the "Document quality assessment" section give an overview of the information type that can be expected from digital data acquisition for landslides. Specifically, the results shown in Fig. 6 underline the importance of image analysis as a complementary information source to text analysis. For example, corrective measures can be seen 2.5 times more often on images than in texts. Given that 34% of all images are irrelevant (Fig. 4), images need to be filtered as

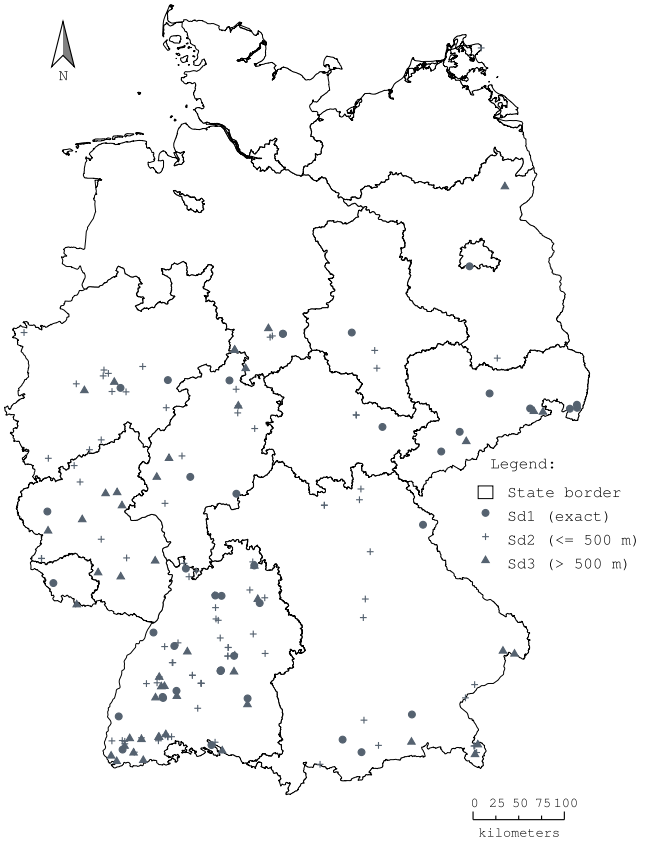
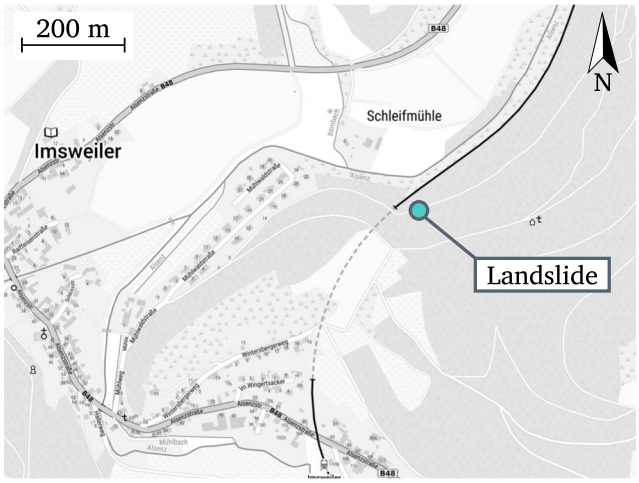


Fig. 7 Landslide inventory map for Germany, compiled from 201 locations discovered in landslide documents of the process chain during the testing period of 19 months. Differentiated after spatial confidence descriptors (Sd) with the respective accuracy in brackets ("Document quality assessment" section).



(a)

(b)

Fig. 8 **a** Exact landslide location (Sd1), inferred from information that the north end of the tunnel (dashed line) in "Imsweiler" was buried by debris of a landslide process. **b** Landslide location approximated (accuracy/radius ≈ 200, i.e., Sd2) after information that a landslide occurred at road "K6924" between "Börstingen" and "Eckenweiler" (not on map); in this case, the circle encloses the only terrain with a topographic situation suitable for landslide processes near the specified road segment. Source: modified after "Federal Agency for Cartography and Geodesy"

well as texts for an effective data reduction method. Moreover, image metadata inspection promised a simple and automatable way of information extraction; however, the results are underwhelming since only 6.56% of images contained relevant metadata (see "Quality assessment" section). Although it is reasonable to assume that the majority of digital cameras can provide metadata for their images, it is thus assumed that the absence of metadata in the extracted images is intentional but for unknown reasons.

In general, the documents have the same principal information problems as any other textual or visual source (Wills and McCrink 2002; Guzzetti et al. 2003). For example, attributes that were found in less than 10% of the documents are activity, lithology, and movement speed. These classes need expert assessment that is often not available to the authors of the documents.

Notwithstanding the aforementioned problems, an inventory map was compiled with 22.40% exact landslide locations (Sd1) that may be appropriated for detailed process analyses (Ermini et al. 2005; Terhorst and Kreja 2009). Moreover, of all detected locations, 73.77% (Sd1 + Sd2) provide an accuracy commonly used for dispositional analyses (Aleotti and Chowdhury 1999; Neuhäuser et al. 2012; Manzo et al. 2013; Klose et al. 2014). Since other works successfully use inventories with an accuracy at the scale of municipalities, even Sd3 locations can be considered for susceptibility assessments (Battistini et al. 2013; Calvello and Pecoraro 2018).

Conclusions

The present work introduces a method to automatically filter large amounts of irrelevant data during digital data acquisition for landslide inventories. Compared with manual methods, the

Table 4 Processing steps, without image classification, of the process chain (Fig. 1) differentiated in manual and automated application for this study, Innocenzi et al. (2017) and Taylor et al. (2015). ◦= manual step, ×= automated step

Work	Step A	B	C	D	E
This study	◦	×	×	×	×
Innocenzi et al.	◦	×	◦	◦	◦
Taylor et al.	◦	◦	◦	◦ (×)	◦

Table 5 Number of results for keyword searches during one year: this study for the year 2018, Innocenzi et al. (2017) annual average of the years 2012–2015, and Taylor et al. (2015) for the year 2006

Work	Keywords	All results	Irrelevant results
This study*	10	3172	2855 (90.01%)
Innocenzi et al.*	1	2737	470 (17.14%)
Taylor et al.†	27	711	167 (23.50%)

*Results from a pool of all documents registered with Google Search via Google Alert

† Results from Nexis UK (newspaper) archive

automated process chain eliminates personal idiosyncrasies and human error and replaces it with a quantifiable machine error. The applied individual algorithms for natural language processing, information retrieval, and classification have been tried and tested in their respective fields. Thus, they are widely available as program libraries and can be easily utilized for custom software of the process chain. Furthermore, all languages on which these algorithms are applicable can be used with the proposed method. Additionally, inventories who do not primarily rely on textual data can still profit from the timely design of the process chain, since it enables prompt morphological investigations (e.g., field survey or remote sensing data) to expand on the document results.

Even though the results of the automated process chain leave room for improvement, the process chain is already suitable for practical application with minimal information loss and strong data reduction. Thus, self-imposed restrictions to avoid large data volumes during digital data acquisition become less important.

The remaining problem pertains to a subpar classification performance for texts. This is supposedly due to geographic restrictions, but without a geographic restriction, there is still a language restriction on the documents. This means that the majority of the results are indirectly steered to be within language borders. Thus, users of the process chain who want a country-specific restriction, therefore, operate with a language that does not transcend national borders will most likely have better classification performances.

In general, future works should investigate the performance of other classification algorithms, e.g., from the family of neural networks, and analogous, other similarity metrics for duplicate identification should be tested. The final step for a completely automated digital data acquisition is the automatic extraction of the document's information. A prerequisite for this step is a clear definition of the information type in the document search; here established attribute classes provide an applicable starting point. On behalf of information extraction itself, a variety of methodologies, for example, rule learning-based, classification-based, or sequential labeling–based methods exist. The classification-based methods can be implemented with the here presented classification algorithms. In the case of a successful application, landslide inventories are automatically populated with data and promise to be a comprehensive, highly effective, up-to-date base for landslide research.

Funding information

Open Access funding provided by Projekt DEAL. This research project was supported by the German Research Foundation (DFG, project DA 452/5-1/5-2) and the Lower Saxonian Ministry for Science (MWK, project MWK 76ZN1504 2016-2020 - "Niedersächsisches Vorab"). We gratefully acknowledge the support.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Aggarwal CC, Zhai C (eds) (2012) Mining Text Data. Springer US, Boston, MA

Aleotti P, Chowdhury R (1999) Landslide hazard assessment: summary review and new perspectives. Bull Eng Geol Environ 58(1):21–44

Battistini A, Segoni S, Manzo G, Catani F, Casagli N (2013) Web data mining for automatic inventory of geohazards at national scale. Appl Geogr 43:147–158

Calvello M, Pecoraro G (2018) Franeltalia: a catalog of recent Italian landslides. Geoenvironmental Disasters 5(1):13

Cruden DM, Varnes DJ (1996) Landslide types and processes: Chapter 3. In: Landslides-Investigation and Mitigation. National Academy Press, Washington DC, pp 36–75

Damm B, Klose M (2015) The landslide database for Germany: closing the gap at national level. Geomorphology 249:82–93

Devoli G, Morales A, Høeg K (2007) Historical landslides in Nicaragua – collection and analysis of data. Landslides 4(1):5–18

Dikau R, Brunsden D, Schrott L, Ibsen ML (1996) Landslide recognition: identification, Movement and Causes. In: International Association of Geomorphologists. Wiley, Chichester [u.a.]

Ermini L, Catani F, Casagli N (2005) Artificial neural networks applied to landslide susceptibility assessment. Geomorphology 66(1-4):327–343

Foster C, Pennington CVL, Culshaw MG, Lawrie K (2012) The national landslide database of Great Britain: development, evolution and applications. Environ Earth Sci 66(3):941–953

Goldhahn D, Eckart T, Quasthoff U (2012) Building large monolingual dictionaries at the Leipzig Corpora Collection: from 100 to 200 languages. Proceedings of the Eighth International Conference on Language Resources and Evaluation, Istanbul, Turkey, vol 29:31–43

Guzzetti F, Cardinali M, Reichenbach P (1994) The AVI project: a bibliographical and archive inventory of landslides and floods in Italy. Environ Manag 18(4):623–633

Guzzetti F, Reichenbach P, Cardinali M, Ardizzone F, Galli M (2003) The impact of landslides in the Umbria region, central Italy. Nat Hazards Earth Syst Sci 3(5):469–486

Guzzetti F, Mondini AC, Cardinali M, Fiorucci F, Santangelo M, Chang KT (2012) Landslide inventory maps: New tools for an old problem. Earth Sci Rev 112(1):42–66
Hess J, Rickli C, McArdell B, StÃ¤hli M (2014) Investigating and managing shallow landslides in Switzerland. In: Landslide Science for a Safer Geoenvironment. Springer, Cham, pp 805–808
Innocenzi E, Greggio L, Frattini P, de Amicis M (2017) A web-based inventory of landslides occurred in Italy in the period 2012–2015. In: Mikos M, Tiwari B, Yin Y, Sassa K (eds) Advancing Culture of Living with Landslides. Springer International Publishing, Cham, pp 1127–1133
Japan Electronics and Information Technology Industries Association (2019) Exchangeable Image File Format for Digital Still Cameras: Exif Version 2.32. Camera & Imaging Products Association
Klimeš J, Stemberk J, Blahut J, Krejčí V, Krejčí O, Hartvich F, Kycl P (2017) Challenges for landslide hazard and risk management in 'low-risk' regions, Czech Republic-landslide occurrences and related costs (IPL project no. 197). Landslides 14(2):771–780
Klose M, Gruber D, Damm B, Gerold G (2014) Spatial databases and GIS as tools for regional landslide susceptibility modeling. Zeitschrift für Geomorphologie 58(1):1–36, library Catalog: www.ingentaconnect.com
Klose M, Damm B, Highland L (eds) (2015) Geohazard databases: concepts, development, Applications locations [Special Issue], Geomorphology 249
Klose M, Maurischat P, Damm B (2016) Landslide impacts in Germany: a historical and socioeconomic perspective. Landslides 13(1):183–199
Kreuzer TM, Wilde M, Terhorst B, Damm B (2017) A landslide inventory system as a base for automated process and risk analyses. Earth Sci Inf 10(4):507–515
Kusner MJ, Sun Y, Kolkin NI, Weinberger KQ (2015) From word embeddings To document distances. In: Proceedings of Machine Learning Research, Lille, France, vol 37, pp 957–966
León T, Zuccarello P, Ayala G, de Ves E, Domingo J (2007) Applying logistic regression to relevance feedback in image retrieval systems. Pattern Recogn 40(10):2621–2632
Liu C, Li W, Wu H, Lu P, Sang K, Sun W, Chen W, Hong Y, Li R (2013) Susceptibility evaluation and mapping of China's landslides based on multi-source data. Nat Hazards 69(3):1477–1495
Lu P, Stumpf A, Kerle N, Casagli N (2011) Object-oriented change detection for landslide rapid mapping. IEEE Geoscience and Remote Sensing Letters 8(4):701–705, conference Name: IEEE Geoscience and Remote Sensing Letters
Manning C, Raghavan P, Schuetze H (2009) Introduction to information retrieval. Cambridge University Press, Cambridge [England] ; New York
Manzo G, Tofani V, Segoni S, Battistini A, Catani F (2013) GIS techniques for regional-scale landslide susceptibility assessment: the Sicily (Italy) case study. Int J Geogr Inf Sci 27(7):1433–1452
Neuhäuser B, Damm B, Terhorst B (2012) GIS-based assessment of landslide susceptibility on the base of the Weights-of-Evidence model. Landslides 9(4):511–528
Pennington C, Freeborough K, Dashwood C, Dijkstra T, Lawrie K (2015) The National Landslide Database of Great Britain: acquisition, communication and the role of social media. Geomorphology 249:44–51
Pereira S, Zêzere JL, Quaresma ID, Bateira C (2014) Landslide incidence in the North of Portugal: Analysis of a historical landslide database based on press releases and technical reports. Geomorphology 214:514–525
Piacentini D, Troiani F, Daniele G, Pizziolo M (2018) Historical geospatial database for landslide analysis: the Catalogue of Landslide OCcurrences in the Emilia-Romagna Region (CLOCkER). Landslides 15(4):811–822
Raska P, Klimes J, Dubisar J (2015) Using local archive sources to reconstruct Historical landslide occurrence in selected urban regions of the Czech Republic. Land Degrad Dev 26(2):142–157
Rosser B, Dellow S, Haubrock S, Glassey P (2017) New Zealandâ€™s National Landslide Database. Landslides 14(6):1949–1959
Rupp S, Wohlers A, Damm B (2018) Long-term relationship between landslide occurrences and precipitation in southern Lower Saxony and northern Hesse. Zeitschrift für Geomorphologie 61(4):327–338
Rusland NF, Wahid N, Kasim S, Hafit H (2017) Analysis of Naïve Bayes algorithm for email spam filtering across multiple datasets. IOP Conference Series: Materials Science and Engineering 226:012,091
Schmid H (1999) Improvements in Part-of-speech tagging with an application to German. In: Armstrong S, Church K, Isabelle P, Manzi S, Tzoukermann E, Yarowsky D (eds) Natural Language Processing Using Very Large Corpora. Text, Speech and Language Technology, Springer Netherlands, Dordrecht, pp 13–25
Taylor FE, Malamud BD, Freeborough K, Demeritt D (2015) Enriching Great Britain's National Landslide Database by searching newspaper archives. Geomorphology 249:52–68
Terhorst B, Kreja R (2009) Slope stability modelling with SINMAP in a settlement area of the Swabian Alb. Landslides 6(4):309–319
Ting KM (2017) Confusion Matrix. In: Sammut C, Webb GI (eds) Encyclopedia of machine learning and data mining. Springer US, Boston, MA, pp 260–260
Valenzuela P, Domínguez-Cuesta MJ, Mora García MA, Jiménez-Sánchez M (2017) A spatio-temporal landslide inventory for the NW of Spain: BAPA database. Geomorphology 293:11–23
Van Den Eeckhaut M, Hervás J (2012) State of the art of national landslide databases in Europe and their potential for assessing landslide susceptibility, hazard and risk. Geomorphology 139-140:545–558
Voumard J, Derron MH, Jaboyedoff M (2018) Natural hazard events affecting transportation networks in Switzerland from 2012 to 2016. Nat Hazards Earth Syst Sci 18(8):2093–2109
Weichelt B, Salzwedel M, Heiberger S, Lee BC (2018) Establishing a publicly available national database of US news articles reporting agriculture-related injuries and fatalities. Am J Ind Med 61(8):667–674
Weissweiler L, Fraser A (2018) Developing a stemmer for German based on a comparative analysis of publicly available stemmers. In: Rehm G, Declerck T (eds) Language Technologies for the Challenges of the Digital Age, Springer International Publishing, Cham, vol 10713, pp 81–94
Wills CJ, McCrink TP (2002) Comparing landslide inventories: the map depends on the method. Environ Eng Geosci 8(4):279–293
Wohlers A, Kreuzer T, Damm B (2017) Case Histories for the Investigation of Landslide Repair and Mitigation Measures in NW Germany. In: Sassa K, MikC's M, Yin Y (eds) Advancing culture of living with landslides. Springer International Publishing, Cham, pp 519–525
Zhang H (2005) Exploring conditions for the optimality of naïve bayes. Int J Pattern Recognit Artif Intell 19(2):183–198

T. M. Kreuzer (✉) · B. Damm
University of Vechta,
Vechta, Germany
Email: thomas.kreuzer@mail.uni-vechta.de