Concept for the development of an integrated landslide inventory system

Technical implementation and automation methods

Thomas M. Kreuzer

Dissertation zur Erlangung des Grades eines Doktors rer. nat.

Erstgutachter: Professor Dr. Bodo Damm
Zweitgutachter: Professor Dr. Roland Baumhauer

Vorgelegt November 2021

Acknowledgements

I would like to express my deep gratitude to Professor Dr. Bodo Damm for his patient guidance, encouragement and useful critiques of this research work. His advice did not just positively impact this work but also my personal development as well as the understanding of how to strive to be a better scientist. My grateful thanks are also extended to Professor Dr. Birgit Terhorst for her constructive and enduring support of my work and me as a person.

I also would like to thank my colleague Martina Wilde who was a great partner in discussing ideas and who helped me with the little details of this work that sometimes do have large impacts.

Special thanks should be given to the “Deutsche Forschungsgemeinschaft” (DFG) and the “Niedersächsisches Ministerium für Wissenschaft und Kultur” (MWK) for giving me the opportunity with their funding to do the required research for this work.

Finally, I wish to thank my wife Christine and my daughter Edda for burdening themselves to make sure I have the time to follow through with my work.

Contents

Kurzfassung

In der Rutschungsforschung sind Datenbanken (hier synonym zu Inventaren) von besonderer Bedeutung, da in ihnen Informationen erfasst und dokumentiert werden, die für statistische und prozessorientierte Analysen notwendig sind. Die dabei eingesetzten Datenbanken reichen in ihrer Art von analogen Dokumentenablagen bis hin zu komplexen Software-Anwendungen. Letztere werden in diesem Zusammenhang aufgrund ihrer technisch bedingten Eignung zur effizienten Datenverarbeitung bevorzugt eingesetzt. Unter den Software-Anwendungen haben sich in den letzten Jahren sogenannte „Relationale Datenbanksysteme" (RDBS) etabliert. Jedoch werden derzeit Rutschungsdaten weitgehend unabhängig vom Datenbanksystem erhoben und analysiert, obwohl durch ein RDBS ein zentraler Ort der Datenverarbeitung vorhanden wäre. Der Betrieb einer Datenbank findet somit auf zwei getrennten Ebenen statt: einerseits auf der Ebene der Datengewinnung und -analyse durch Betreiber und Nutzer der jeweiligen Datenbank und andererseits auf der Ebene der zentralen Datenspeicherung und -verteilung durch ein entsprechendes Datenbanksystem. Betreiber und Nutzer sind dadurch mit den Problematiken ihrer Betriebsebene konfrontiert, ohne eine Unterstützung durch das RDBS erhalten zu können. Eine besondere Herausforderung besteht an dieser Stelle darin, dass Rutschungen zeitlich sowie räumlich weit verteilt und durch komplexe Prozesse entstanden sind. Umfassende Datenerhebungen beinhalten infolgedessen einen hohen Arbeitsaufwand, der sich überdies auf Analysen auswirkt, die auf das Vorhandensein aktueller und zahlreicher Daten angewiesen sind. Für Betreiber und Nutzer einer Rutschungsdatenbank führt dies daher regelmäßig zu selbstauferlegten Einschränkungen in ihren Fragestellungen, um auf diese Weise den benötigten Datenumfang und, infolgedessen, den damit einhergehenden Aufwand der Datenerhebung zu begrenzen.

Das übergeordnete Ziel der vorliegenden Arbeit ist, der dargestellten Problematik entgegenwirkend, die Minimierung des Aufwands für den Betrieb und die Nutzung einer Rutschungsdatenbank, sodass selbstauferlegte Einschränkungen an Bedeutung verlieren. Zu diesem Zweck wird ein „Integriertes Rutschungsinventarsystem" (IRIS) entwickelt, welches die Ebene der Datenerhebung und -analyse mittels einer weitgehenden Automatisierung in ein RDBS integriert. Die Anwender dieses Systems werden nun in dem Maße entlastet, als sie lediglich die Funktionen überwachen müssen, die automatisiert umgesetzt werden.

Im Kontext dieser Zielsetzung wurde, im Rahmen der Publikation „A Landslide Inventory System as a Base for Automated Process and Risk Analyses", die technische Grundlage für das IRIS geschaffen. Hierfür sind zunächst die Anforderungen an ein derartiges Verfahren herausgearbeitet worden. Es galt eine Software zu finden, welche die gängigen Datenverarbeitungsmethoden eines RDBS umsetzt, zusätzlich räumliche Daten verarbeiten kann und bei der es möglich ist, Veränderungen an der Programmlogik vorzunehmen, um automatisierte Erhebungs-

und Analysemethoden zu integrieren. Daneben war darauf zu achten, dass es Betreibern der Datenbank weiterhin möglich ist, Daten — digital wie analog — aus dezentralen Erhebungen (bspw. Feldarbeit, manuelle Internetrecherche) in das System einzupflegen. Zu diesem Zweck musste die eingesetzte Software die Digitalisierung analoger Daten unterstützen, die daraufhin auch der automatisierten Datenverarbeitung zur Verfügung stehen können. Mit der Software „PostgreSQL" ist dementsprechend eine Umgebung gewählt worden, die diese Anforderungen erfüllt und damit ein RDBS darstellt, welches um gängige GIS-Funktionalität ergänzt wurde mittels der Erweiterung „PostGIS". PostgreSQL/PostGIS ist deshalb in der Lage, neben den primären Daten der Rutschungen, auch unterstützende Daten, wie beispielsweise digitale Karten und Geländemodelle, zu speichern und zu verarbeiten. Eine weitere Besonderheit im Kontext der gestellten Anforderungen besteht darin, dass die Software quelloffen zur Verfügung gestellt wird und beliebig modifiziert werden darf. Unter diesen Voraussetzungen konnte die Software um Möglichkeiten der Eingabe selbsterhobener Daten sowie um eine Analyse zur Risikobewertung erweitert werden. Dem folgend wurde anhand eines Fallbeispiels in der Fränkischen Alb, eine Karte automatisiert erzeugt, in der das Risiko für in dem im Untersuchungsraum vorhandene Infrastrukturobjekte dargestellt wird, durch in der Nähe befindliche, aktive Rutschungen bedroht zu sein. Dazu war es zunächst notwendig, analoge Rutschungsdaten aus vorangegangenen Arbeiten über die Eingabeschnittstelle zu digitalisieren sowie digitale Infrastrukturkarten und digitale Geländemodelle in das System einzuspeisen, wobei sämtliche Daten im System hinterlegt bleiben. Sobald dem System ergänzende/aktuellere Daten aus den unterschiedlichsten Erhebungsarten zur Verfügung stehen, kann daher auch die Analyse ohne weitere Bemühungen des Betreibers „per Knopfdruck" angepasst werden.

Mithilfe dieser technischen Grundlage, inklusive der automatischen Analysemöglichkeiten, gilt es weiter, den Betreiber einer Datenbank bei der Datenaufnahme zu unterstützen. Dies geschah zunächst im Zuge der Publikation „Automated Digital Data Acquisition for Landslide Inventories" durch die Entwicklung einer Prozesskette zur automatisierten Datenakquise digitaler Texte und den ihnen beiliegenden Bildern — die Texte und Bilder entstammen beispielsweise wissenschaftlichen Arbeiten, Polizeiberichten, Gutachten oder auch Zeitungsartikeln. Anhand einer weiteren Modifikation von PostgreSQL/PostGIS wurde eine Prozesskette in das IRIS integriert, um das System auf diese Art zentral und kontinuierlich mit möglichst aktuellen Daten zu versorgen. Diese Prozesskette selbst besteht dabei aus vier Gliedern, die schließlich wiederkehrend in bestimmen Zeitabständen, rutschungsrelevante Texte aus dem Internet sammeln und diese dem Betreiber einer Datenbank zur Verfügung stellen. Im Hinblick darauf ist es die Hauptaufgabe dieser Prozesskette, große Mengen anfallender und irrelevanter Texte auszusortieren und Textduplikate zu identifizieren, um die Datenmenge auf relevante Informationen zu begrenzen. Der Ablauf der Prozesskette gliedert sich dabei wie folgt: Zunächst wird jeder Text, der zum ersten Mal im Internet durch den Suchmaschinenbetreiber „Google" registriert wird, auf vorher festgelegte Schlagwörter (z. B. Erdrutsch, Hangrutsch, Steinschlag) und deren Flexionen überprüft. Das Vorhandensein eines der Schlagwörter in einem Text ist eine notwendige Bedingung für rutschungsbezogene Inhalte, deshalb werden nur solche Texte an das nächste Glied weitergereicht. Dieses überprüft dann, ob sich die gefundenen Schlagwörter in grammatikalisch vollständigen Sätzen befinden. Auf diese Weise wird sichergestellt, dass es

sich bei den rutschungsbezogenen Inhalten um abgeschlossene Informationseinheiten handelt. Zusätzlich werden in diesem Schritt auch vorhandene Bilder als weitere Informationseinheit extrahiert. Sämtliche gefundene Informationseinheiten werden daraufhin im nächsten Glied der Prozesskette mittels Methoden des maschinellen Lernens als relevant oder irrelevant bezüglich Rutschungen klassifiziert — irrelevant wäre beispielsweise ein Text über einen politischen „Erdrutschsieg" oder ein Bild einer zerstörten Windschutzscheibe aufgrund eines Steinschlags. Das letzte Glied entscheidet dann darüber, ob es sich bei einem vorher als relevant eingeordneten Text um ein Duplikat eines bereits erfassten Textes aus einer anderen Quelle handelt. Als Duplikat gilt, wenn es mittels einer Metrik zur Inhaltsähnlichkeit einen gewissen Schwellenwert überschreitet, wobei aufgrund von möglicherweise beinhaltenden Zusatzinformationen das identifizierte Duplikat nicht gänzlich verworfen, sondern zunächst vor dem Betreiber verborgen wird. Infolgedessen wird die Datenmenge weiter reduziert, es bleibt jedoch jederzeit die Möglichkeit, sich bei Bedarf Duplikate anzeigen zu lassen. Insgesamt wurden über die Testlaufzeit von 87 Wochen 4381 Dokumente mittels der implementierten Prozesskette analysiert und davon $90\,\%$ irrelevante Dokumente aussortiert. Infolgedessen konnten somit 385 Textquellen (exkl. Duplikate) zu Rutschereignissen direkt zur Verfügung gestellt werden.

Im Hinblick auf die zweigleisige Verwendungsmöglichkeit des IRIS (dezentral/manuell und zentral/automatisiert, s. o.), wurde im Zusammenhang mit der Publikation „Quantitative Assessment of Information Quality in Textual Sources for Landslide Inventories" eine quantitative Bewertung der Nützlichkeit verschiedenster textlicher Quellenarten (bspw. Zeitungsartikel, Polizeibericht, wissenschaftliche Publikation, technisches Gutachten) durchgeführt, um insbesondere die manuelle Datenaufnahme zu optimieren. Dem liegt zugrunde, dass eine manuelle Sichtung möglicher Quellen einen hohen Aufwand bedeutet und dieser durch eine, auf Nützlichkeit basierende Vorauswahl der Quellenart gesenkt werden kann. Es stellt sich insbesondere die Frage, ob eine bestimmte Quellenart nützlich für Rutschungsinventare ist, wenn deren enthaltene Informationen nicht von Rutschungsexperten stammen, sondern beispielsweise von Journalisten, Polizisten oder Förstern. Zur Beantwortung dieser Frage wurde eine „Nützlichkeit" definiert, die der quantitativen Wahrscheinlichkeit entspricht, festgelegte Rutschungsinformationen, gewichtet nach deren jeweilig auftretenden Detailgraden, zu finden. Häufiges Vorkommen eines hohen Detailgrades schlägt sich dementsprechend in einer höheren Nützlichkeit nieder verglichen mit Quellenarten, die zwar die gleiche Art Information beinhalten, aber häufiger mit einem niedrigeren Detailgrad. Da die Nützlichkeit hier einer mathematischen Wahrscheinlichkeit entspricht, gelten gleichfalls die bekannten Regeln der Kombinatorik. Auf diese Weise kann die Nützlichkeit nicht nur für eine Quellenart angegeben werden, sondern auch für deren beliebige Kombinationen. Beispielhaft wurde ein Datensatz eines deutschen Rutschungsinventars untersucht, der neben ausgewählten Rutschungsinformationen zu einzelnen Rutschprozessen, auch deren originäre Quellenart beinhaltet. Konkret wurden die vermerkten Quellenarten nach Inhalten zu Lokation, Datum und Prozesstyp einer Rutschung in verschiedensten Detailgraden analysiert. Es zeigte sich, dass die drei nützlichsten Quellenarten in Kombination eine über $86\,\%$ige Wahrscheinlichkeit zur Findung verwertbarer Informationen aufweist. Bei den drei Quellenarten handelt es sich in absteigender Reihenfolge der einzelnen Nützlichkeit um: Zeitungsartikel, Gutachten und administrative Dokumente. Weiter zeigte sich,

dass die Einbindung weiterer Quellenarten diese Wahrscheinlichkeit lediglich logarithmisch erhöhen würde, sodass, bezüglich eines effizienten Einsatzes vorhandener Ressourcen, zunächst darauf verzichtet werden kann.

Die drei angeführten Arbeiten bilden zusammen das technologische und konzeptionelle Fundament des IRIS. Dieses Fundament ermöglicht es, die vormals getrennte Ebene des Betreibens und Nutzens einer Datenbank mit der Ebene der Datenverarbeitung zu verbinden, wobei die Automatisierung der Datenakquise sowie die Risikoanalyse erhobener Daten in das relationale Datenbanksystem integriert wurden. Das Wissen um die Nützlichkeit verschiedener Quellenarten ermöglicht die effiziente Steuerung und Fokussierung der manuellen, aber auch digitalen Datenaufnahme. Es handelt sich bei IRIS folglich um ein quasi-abgeschlossenes, erweiterbares und autarkes System, welches durch den Betreiber kontrolliert wird und die Verwaltung großer und kontinuierlich anfallender Rutschungsdaten erlaubt. Zukünftige Arbeiten bezügliche einer erweiterten Datenerhebung könnten die automatisierte Erfassung der, in den gefundenen Textquellen enthaltenden Informationen sein und/oder die Integration automatisierter Landformenerkennung mittels fernerkundlicher Methoden. Bezüglich der Datenanalyse würde eine zukünftige Integration weiterer etablierter Analysemethoden die Möglichkeiten zur Gefahrenerkennung und -bewertung steigern. Zusammen erscheint in dieser Form ein vollautomatisches, „lebendes" Rutschungsinventar möglich, das auf der regionalen bis globalen Ebene kontinuierlich aktuelle und umfassende Informationen und Prognosen zu Rutschereignissen liefern kann.

Abstract

In landslide research, databases (here synonymous with inventories) are of particular importance, as they are used to record and document information necessary for statistical and process-oriented analyses. The databases used in these circumstances range in type from analogue document repositories to complex software applications. The latter are preferred in this context due to their technical suitability for efficient data processing. Among software applications, so-called “relational database systems” (RDBS) have distinguished themselves in recent years. However, landslide data are currently collected and analysed largely independently of such a database system, although an RDBS would provide a central location for data processing. The operation of a database thus takes place on two separate levels: on the one hand, on the level of data acquisition and analysis by operators and users of the respective database and, on the other hand, on the level of central data storage and distribution by a corresponding database system. Operators and users are thus confronted with problems of their operational level without being able to receive support from a RDBS. A particular challenge at this point is that landslides are widely distributed in time as well as in space and are the result of complex processes. Comprehensive data collection consequently involves a large amount of work, which moreover affects analyses that depend on the availability of up-to-date and numerous data. For operators and users of a landslide database, this therefore regularly leads to self-imposed restrictions in their problem definitions in order to limit the required scope of data, and hence the associated effort of data collection.

The overall objective of the present work is to counteract the presented problems by minimizing the effort for operation and use of a landslide database, so that self-imposed restrictions become less important. For this purpose, an “Integrated Landslide Inventory System” (IRIS) is developed, which integrates the level of data collection and analysis into an RDBS by means of automation. The users of this system are thus relieved to the extent that they only have to monitor automated processes.

In the context of this objective, the technical basis for IRIS was created within the framework of the publication ‘A Landslide Inventory System as a Base for Automated Process and Risk Analyses’. For this purpose, the requirements for such a technical basis were first worked out. It was necessary to find a software that implements the common data processing methods of an RDBS, can additionally process spatial data and for which it is possible to make changes to the program logic in order to integrate automated collection and analysis methods. In addition, it had to be ensured that it is still possible for database operators to enter data — digital as well as analogue — from decentralised surveys (e.g. fieldwork, manual internet research) into the system. Therefore, the applied software solution had to support the digitisation of analogue

data, which could then be made available for automated data processing. The software “PostgreSQL” was chosen accordingly, which fulfills these requirements and thus represents a RDBS, which was further enhanced for common GIS functionality by means of the extension “PostGIS”. PostgreSQL/PostGIS is therefore able to store and process not only the primary landslide data but also supports data such as digital maps and digital terrain models. Another special feature in the context of the requirement set is that the software is made available as Open Source and may be modified as desired. Under these conditions, the software was extended to include the possibility of entering self-collected data as well as an automated analysis for risk assessment. Following this, a case study in the Franconian Alb was used to automatically generate a map showing the risk of infrastructure objects being threatened by active landslides in the vicinity. In this respect, it was first necessary to digitise analogue landslide data from previous work via the input interface and to feed digital infrastructure maps and digital terrain models into the system, with all data remaining stored in the system. As soon as supplementary or more up-to-date data from the various types of surveys are available to the system, the analysis can therefore also be updated “at the touch of a button” without any further effort on the part of the operator.

In accordance with the overarching goal and after the establishment of a technical basis, including the automatic analysis possibilities, it is further necessary to support the operator of a database in data acquisition. This was initially done in the course of the publication ‘Automated Digital Data Acquisition for Landslide Inventories’ by developing a process chain for the automated data acquisition of digital texts and their accompanying images — the texts and images originate, for example, from scientific papers, police reports, expert opinions, or even newspaper articles. Using a further modification of PostgreSQL/PostGIS, a process chain was integrated into IRIS in order to supply the system centrally and continuously with the most up-to-date data possible. This process chain itself consists of four links, which finally, recurrently in certain time intervals, collect landslide-relevant texts from the internet and make them available to the operator of a database. In view of this, the main task of this process chain is to sort out large quantities of accumulating and irrelevant texts and to identify text duplicates in order to limit the data to relevant information. The process chain is structured as follows: First, each text that is registered for the first time on the Internet by the search engine operator “Google” is checked for predefined keywords (e. g., landslide, mudflow, rockfall) and their inflections. The presence of one of the keywords in a text is a necessary condition for landslide-related content, so only such texts are passed on to the next link, which then checks whether the keywords found are in grammatically complete sentences. In this way, it is ensured that the landslide related content is a self-contained information unit, in addition, existing images are extracted as further information units. Using machine learning methods, all information units found are then classified in the next link of the process chain as relevant or irrelevant with respect to landslides — irrelevant would be, for example, a text about a political “landslide victory”, or a picture of a destroyed windscreen due to a rockfall. The final link then decides whether a text previously classified as relevant is a duplicate of an already recorded text from another source. A duplicate is considered to be a duplicate if it exceeds a certain threshold using a content similarity metric, however, due to additional information that may be included, the

identified duplicate is not discarded entirely but is just hidden from the operator. As a result, the amount of data is further reduced, but the ability to view duplicates remains possible at all times. In total, over the test period of 87 weeks, 4381 documents were analyzed using the implemented process chain and $90\,\%$ of these irrelevant documents were sorted out, with the result that 385 text sources (excl. duplicates) on slide events could be made directly available to the operator of IRIS.

With regard to the two-pronged use of IRIS (decentralized/manual and centralized/automated, see above), a quantitative evaluation of the usefulness of various textual source types (e.g. newspaper article, police report, scientific publication, technical report) was carried out in connection with the publication 'Quantitative Assessment of Information Quality in Textual Sources for Landslide Inventories', particularly to optimize manual data acquisition. This is because a manual sifting of possible sources means a high effort and this effort can be reduced by a preselection of the source type based on usefulness. In particular, the question arises whether a certain type of source is useful for landslide inventories if the information it contains does not come from landslide experts but, for example, from journalists, police officers, or foresters. To answer this question, a "usefulness" was defined, which corresponds to the quantitative probability of finding specified landslide information, weighted according to their respective degrees of detail. Frequent occurrence of a high level of detail accordingly translates into higher usefulness compared to source types that contain the same type of information but more frequently with a lower level of detail. Since usefulness here corresponds to a mathematical probability, the well-known rules of combinatorics also apply. In this way, usefulness can be specified not only for one type of source, but also for any combination. As an example, a data set of a German landslide inventory was investigated, which contains not only selected landslide information on individual landslide processes, but also their original source type. Specifically, the noted source types were analyzed according to the content of location, date and process type of a landslide in various degrees of detail. It was found that the three most useful source types had a greater than $86\,\%$ probability of finding the required information when combined. The three source types, in descending order of individual usefulness, are: newspaper articles, expert opinions, and administrative documents. It was further shown that the inclusion of additional source types would only increase this probability logarithmically, so that, with regard to an efficient use of available resources, it can be dispensed with for the time being.

Together, the three works listed above form the technological and conceptual foundation of IRIS. This foundation makes it possible to link the previously separate level of operating and using a database with the level of data processing, whereby the automation of data acquisition and the risk analysis of collected data have been integrated into a relational database system. Thus, knowledge of usefulness of different types of sources enables the efficient control and focus in manual, as well as digital, data acquisition. Consequently, the IRIS is a quasi-closed, extensible, and self-sufficient system controlled by the operator that allows for the management of large and continuously accumulating landslide data. Future work to extend the system with respect to data acquisition could be the automated extraction of information contained in the retrieved text sources and/or the integration of automated landforms recognition using remote

sensing methods. With respect to data analysis, future integration of other established analysis methods would increase hazard detection and assessment capabilities. Together, a fully automated, “living” landslide inventory appears possible in this form, which can continuously provide up-to-date and comprehensive information and forecasts on landslide events on the regional to global scale.

Acronyms

ARIM Automated Risk Identification Module

DBMS Database Management System

DEM Digital Elevation Model

ELFA Extended Landslide Flow Area

GIS Geographic Information System

ILIS Integrated Landslide Inventory System

IR Information Retrieval

NLP Natural Language Processing

OSM Open Street Map

PDF Portable Document Format

POS Part Of Speech

RDBMS Relational Database Management System

SQL Structured Query Language

SRTM Shuttle Radar Topography Mission

Chapter 1
Introduction

Landslide research chiefly relies on inventories for a multitude of spatial, temporal, and/or process analyses (Damm et al., 2010; Neuhäuser et al., 2012; Van Den Eeckhaut & Hervás, 2012). Such landslide inventories can be anything from as trivial as a shelf for paper notes to a complex database software system. However, regardless of the backend technology, landslide inventories generally are two layered systems: the first layer represents the user's responsibility to acquire and analyse data, and the second layer is the inventory technology that stores and distributes data for the user. Thus, the two-layered inventory system consists of three components: data acquisition, data analysis, and inventory technology (Figure 1.1).

All the three components have their unique challenges and are important for the functioning of the inventory system. Strategies for *data acquisition* can be put into two categories that are applied either alone or in combination (Guzzetti et al., 2012): one, morphological examination, including field surveys, analysis of remote sensing products, topographic or geological maps (Ayenew & Barbieri, 2005; Carrara & Merenda, 1976; Meinhardt et al., 2015), and two, archival work, i.e., sighting of private and public archives for textual information in source types like scientific publications, reports from varying entities (e.g., police, fire department, road construction office, or private contractors), as well as newspaper articles (Guzzetti et al., 1994; Liu et al., 2013; Valenzuela et al., 2017). Both strategies generally involve high effort, time and cost intensive tasks if applied manually, specifically on analog data (Guzzetti et al., 1994; Klose, 2015; Wohlers et al., 2017).

Data analysis generally depends on auxiliary data, i.e., not directly landslide related data. For example, a susceptibility analyses of landslides requires information for the respective preparatory factors, e.g., Digital Elevation Models (DEMs) for topological and geology maps for lithological information (Manzo et al., 2013; Reichenbach et al., 2018; Wilde et al., 2018). Users of two-layered inventory systems either have to acquire and store such auxiliary data

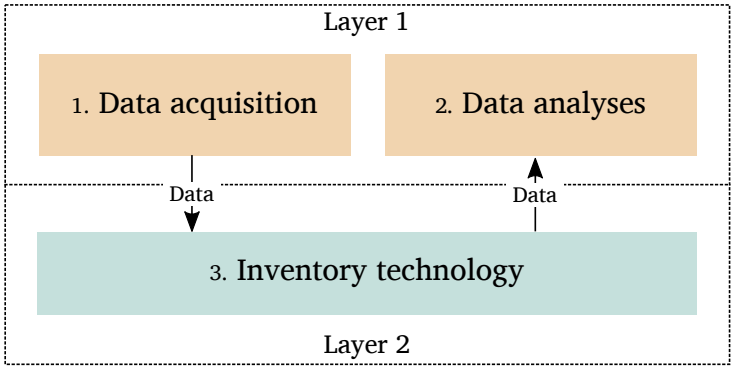


Figure 1.1 Two-layered/three-component inventory system. Layer 1 represents the user's responsibility, layer 2 the inventory technology that stores and distributes data.

themselves or receive the required data from the inventory, which commonly involves copyright issues and large data transfers in case of digital data (Kreuzer et al., 2017).

In this context, the *inventory technology* must be able to store and distribute the data for the user. Generally, those two requirements are detrimental to each other. For example, a shelf can readily store paper notes but it is difficult to distribute those notes to far-away users, in contrast, a software for a Database Management System (DBMS) can seamlessly distribute data all over the world through a network but storage requires high maintenance for complex data structures and user interfaces (Petkovic, 2016).

The question arises whether a two-layered landslide inventory system can be integrated into a one-layered system with the help of automation, where digital inventory technology aids the user and thus reduces his responsibilities, therefore putting him primarily into a controlling position and avoiding, to a large degree, the aforementioned problems of *data acquisition* and *analyses*.

To date, there exists no integrated system for landslide inventories in the literature that merges both layers of a traditional system through means of automation. However, there are studies that concern themselves with automation for *data acquisition*, in general those studies focus on *morphological examination*, commonly involving remote sensing products (Behling et al., 2014; Gaidzik et al., 2017; Hölbling et al., 2015). For *archival work*, there only exist studies proposing a semi-automated form of data acquisition where the user remains strongly involved and self-imposed restrictions are put into place to reduce the accruing data volume (Battistini et al., 2013; Innocenzi et al., 2017; Taylor et al., 2015).

Automated *data analyses* in the context of landslide inventories are bound to the inventory technology because, in this case, the automation of analyses is primarily about data provision and the possibility to integrate analysis algorithms with the used technology. No study has concerned itself with this problem regarding landslide inventories. In regard to inventory technology, Relational Database Management Systems (RDBMSs) are generally the technology of choice for digital landslide inventories (Foster et al., 2012; Jäger et al., 2018; Sabatakakis et al., 2013; Van Den Eeckhaut & Hervás, 2012). This technology is either open source or proprietary, where the latter is, by design, closed to extensive customization (Raymond, 2001).

The overarching goal of the present work is to present a methodological approach and a technological basis for an (one-layered) Integrated Landslide Inventory System (ILIS). This includes a method for automated data acquisition from *archival work* as well as an integration of *data analysis* into the selected inventory technology. This way, ILISs can be a comprehensive, effective, up-to-date base for spatial, temporal, and/or process analyses that automatically adapt to the changing data availability from the acquisition process. Another objective is to preserve the possibility to use the system in a two-layered manner to be able to manually supplement landslide information, specifically from analog sources that are not covered by the automation process.

The following presents a technological foundation of an ILIS with an automated analysis module that assess landslide risks — in this thesis, risk refers to objects of interest which are exposed to hazards, thus following the definition according to Ropeik (2002). Furthermore, an automated process chain for the acquisition of landslide information from textual documents

and their images is presented. In addition, source types from the automated and manual acquisition processes are assessed on their suitability for landslide inventories to enable a focus on the most relevant source types for a given problem, may it be analog or digital. Together, these steps form the basis on which the user's responsibilities can be transferred to the underlying inventory technology, thus creating a one-layered inventory system.

Chapter 2
Technical implementation

The technical implementation is the foundation for any ILIS, specifically for automation of digital *data acquisition* and *analyses*. However, this does not mean that analog data is excluded. Therefore, the technology needs to fulfil the following requirements: assist the digitalization of analog data for the inventory; have network access for effective data communication with one or more users; ensure consistent data structure of landslide information for data storage and analyses; have large enough storage capacity to include auxiliary data; and it is required to be able to integrate analysis algorithms with direct data access as well as to provide an interface for the user to execute those analyses. All those requirements are met with hardware and software implementation based on Kreuzer et al. (2017).

On the hardware side, a system of multiple redundant and extendable hard drives ensures capacity for dynamic storage requirements while reducing the risk of failure. For comparison, stored landslide properties are generally not the limiting factor, millions of landslide properties can be stored in a few hundred MB, yet a single DEM of Germany with a $1\,\mathrm{m}$ resolution may take up to $4.4\,\mathrm{TB}$. Additionally, the system has access to the Internet, today a standard, to establish a wide reaching network of potential users. More technical specifics are not described any further because they do not pose an implementation challenge and advance rapidly over time. The technical minimum requirements can be derived from the minimum requirements of the used database software, other than that, given no financial restrictions, the notions of "newer is always better" and "more is better than less" apply (Lynch, 2008; Mattmann et al., 2016).

The foundation on the software side is PostgreSQL. As a Relational Database Management System (RDBMS), PostgreSQL is customized to process large amounts of complex data with numerous internal relations (Sumathi & Esakkirajan, 2007). It provides high standards in terms of stability, capability, memory, and most important, the option of geodata processing. PostgreSQL stores single datasets in different tables, with operators establishing internal linkage. It is in synchronization with Structured Query Language (SQL) standards, providing all functions of SQL as a database language for generating, managing, and manipulating stored data as well as data queries. Although PostgreSQL supports handling geometric data, it is not capable to process and to store large volumes of spatial data. Therefore, PostGIS, an extension for PostgreSQL, is used for an ILIS. PostGIS does not only improve the storage of spatial data in the DBMS, but offers spatial operators, functions, data types, and indices for rapid spatial applications. Beside the creation of new geometric data, users may generate morphometric measurements and establish spatially related links between different objects (Mitchell et al., 2008; Obe & Hsu, 2011). Furthermore, PostGIS enables editing, analysing, and saving of spatial data (e.g. point, polylines or polygons) in a database without the usage of an external

Geographic Information System (GIS). According to Obe and Hsu (2011), the extension of PostGIS transforms an object-relational DBMS into a spatial database.

The PostgreSQL software package contains a server-client model, which distributes data over a network as text. PostgreSQL/PostGIS is open source and free of charge. It can easily be accessed from external software tools. The nature of open-source guarantees independence from any single company over the development of the software, thus, modifications and extensions are always possible (Raymond, 2001). Virtually, all common open source GIS and analysis applications, as well as some selected proprietary ones, are able to handle and to visualize spatial information of a PostgreSQL/PostGIS database. Therefore, experts can access data by their own working tools or through standardize SQL commands. As such, PostgreSQL/PostGIS is readily extensible with custom modules, which can be programmed in a variety of supported languages, i.e., C, Python, pgSQL, Perl, and Tcl. In this present work, such modules are used for automated data analyses because they have direct access to the inventory's data and do not require data transfers. Moreover, PostGIS provides GIS functionality to custom modules, in combination with third-party libraries from the respective programming languages, a comprehensive GIS is available for the creation of such modules, which is comparable to commercial solutions.

The proposed technological system is also intended for optional two-layered inventory usage, where the user feeds the inventory system with analog data. For the reason of convenient digitalization, a standardized, digital data sheet in the Portable Document Format, is proposed after Jäger et al. (2018). The user fills the data sheet with landslide information and uploads it to the database via a programmed interface (Kreuzer et al., 2017). However, with the help of spatial auxiliary data from the inventory, the input process is semi-automatable to aid the user, i.e., spatial information can be automatically deduced from auxiliary data and is inserted into the data sheet. In Figure 2.1 a semi-automatable, schematic data input with a digital data sheet is seen, the actual implementation is realized as an extension module to PostgreSQL.

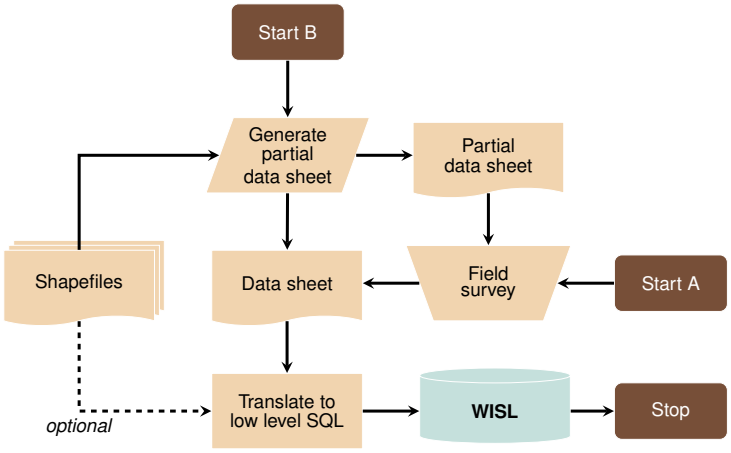


Figure 2.1 Flow chart on how to enter data to the inventory with two separated entry points **A** and **B**. **A** represents the work flow of an empty data sheet before field observations, spatial information is optional. **B** represents the work flow with pre-calculated (semi-automated) values in the data sheet, those pre-calculated values require spatial, auxiliary data (Kreuzer et al., 2017).

Chapter 3
Material and Methods

3.1 Automated data analysis module

The following describes the custom module that is integrated into the technical foundation for the purpose of automated risk analysis (cf. Kreuzer et al., 2017). The material for the illustration of Automated Risk Identification Module (ARIM) consists of: location, extent, and activity state of landslides gathered by preliminary field work in the Franconian Alb (Jäger et al., 2018; Sandmeier et al., 2013); a DEM derived from the freely available Shuttle Radar Topography Mission (SRTM); as well as georeferenced infrastructure objects from the Open Street Map (OSM) project. All information from the material is stored in a landslide inventory, based on the technical implementation from chapter 2, to fulfil ARIM's data requirements for the region of the Franconian Alb.

ARIM follows the objective to generate completely automated landslide susceptibility/risk maps. In summary, the module categorizes the hazard exposition in the proximity of landslides that were recorded as active slope movements in the inventory. Therefore, the module calculates the area of possible flow direction of (potentially) active landslides and further identifies exposed objects (e.g., buildings and infrastructure) located in the flow path. Possible flow directions are statistically determined from all available data, hence a raster map of landslide accumulations is evaluated together with a slope map for the same region, i.e., the Franconian Alb. During this evaluation, downward facing slope cells adjacent to landslide cells are counted to create a frequency distribution. On this basis, the inclination of the third quartile is used as the module's minimal slope inclination. Then, in a given region of interest, the module identifies adjacent raster cells to (potentially) active landslide masses that have an inclination equal or above the statistically determined level and models the outflow path in form of an arbitrarily shaped area, termed Extended Landslide Flow Area (ELFA). All intersections of ELFA with georeferenced objects, like settlements and infrastructure, are marked. These modelling results are indicated as risk areas, where the identified objects are exposed to possible landslide damages from the (potentially) active mass.

In order to present first results of the ARIM, an area defined as the rectangle between the lower left coordinates of $11.4519°$E and $49.5273°$N and the upper right coordinates of $11.5050°$E and $49.5568°$N is examined with a derived minimal inclination from the whole Franconian Alb dataset. The area is approximately $40\,\mathrm{km}$ east of Nuremberg in Germany and comprises three (potentially) active landslide records in the material.

3.2 Automated data acquisition

The proposed method for automated data acquisition after Kreuzer and Damm (2020) focuses on textual documents and their images. This is because required landslide information like date of occurrence, movement speed, costs and damages are not readily derived from morphological examinations (Klose et al., 2015). To automatically process digital texts, established methods from the field of Information Retrieval (IR) are used for the creation of a software module for the ILIS. 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).

Generally, as implicated in the above quote, IR is meant for textual analysis, and thus relies on methods from Natural Language Processing (NLP). Particularly: *tokenization* to decompose text into tokens (words, sentences; Aggarwal and Zhai, 2012); *Part Of Speech (POS) tagging* to determine word classes (e.g., nouns, verbs) based on the word position in a sentence (Schmid, 1999); as well as *word vectoring* to quantitatively describe the semantic distance (similarity) of words (Kusner et al., 2015). 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 analysed 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*, in this case the publicly available "Leipzig Corpora Collection" for the German language (Goldhahn et al., 2012).

The automated process chain is shown in Figure 3.1. It consists of five steps (**A**-**E**) that together provide the methodical approach on automated data acquisition for landslide inventories. Step **A** is the only manual step in this process chain, it requires the definition of keywords that are expected to occur in document texts on landslides. In this case, keywords come from science (e.g., Cruden & Varnes, 1996) and findings from manually conducted searches in media (Damm & Klose, 2015). Moreover, keywords are defined as singular nouns, the process chain will use the respective word stems in document searches and thus also include plural occurrences. These document searches are performed in step **B**. For this reason, step **B** relies on the monitoring service from Google LLC named *Google Alert*. It checks any document that newly registers with *Google Search* for keyword presence and reports document sources that contain any of the predefined keywords back to the process chain with additional meta information, i.e., title, short description, and publication date. These preselected documents are further processed for additional information in step **C**. Thereby, based on NLP methods, the process chain extracts images and complete sentences (with subject, predicate, object) that also contain any of the predefined keywords. The extracted sentences and images are classified as landslide relevant or irrelevant in step **D**. For the classification of the respective sentences the multinomial naive Bayes algorithm is applied (Zhang, 2005), and the logistic regression

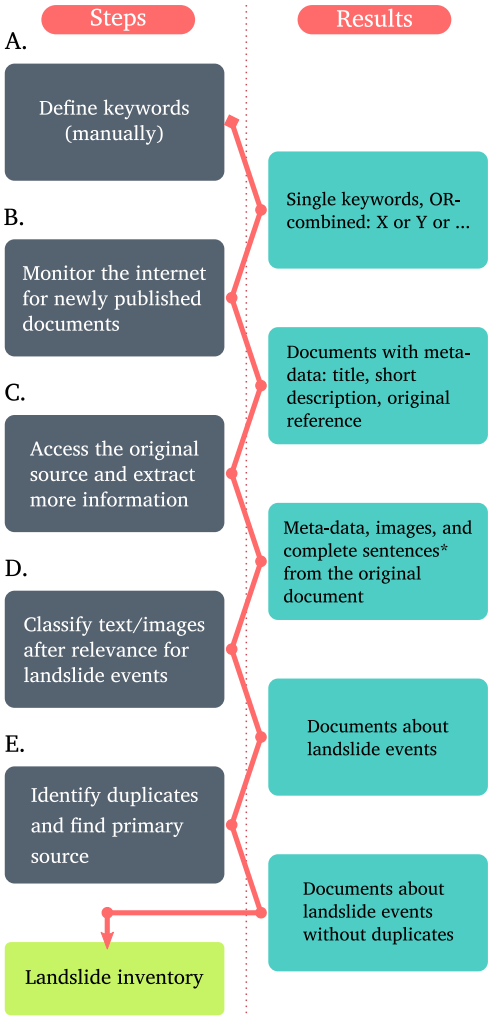


Figure 3.1 Outline of the automated process chain for digital data acquisition. A. is the only manual process step and it is not continuously repeated. Steps B.-E. are programmatically executed every hour (Kreuzer & Damm, 2020).

algorithm for the respective images (León et al., 2007). The last step **E** then compares all acquired textual information (sentences, meta information) from relevant landslide documents for similarity. In case two or more texts are as similar as to be considered duplicates, the document with the most textual content on landslides is declared the primary source and all other duplicates are hidden from the final list of results. However, they are not discarded and can still be retrieved for further review. Furthermore, duplicate detection is designed to not produce any false positives (Kreuzer & Damm, 2020).

Except for step **A**, the process chain is repeated hourly in an ILIS to not miss any alerts on potentially new landslide documents. The document results are automatically added to the inventory and have to be manually reviewed to insert their landslide information into the respective ILISs

For evaluation of the process chain, digital textual sources have been collected between July 2017 and February 2019. The fully automated steps **B**-**E** are assessed through comparison with results from manual review of the collected documents. Furthermore, these documents are checked for the type of landslide attributes they provide to assess their suitability for landslide inventories. For this reason, thirteen attribute classes commonly used in landslide inventories

Table 3.1 Common landslide attribute classes whose presence is examined in texts and images from the automated process chain to assess the respective document's suitability for landslide inventories (Kreuzer & Damm, 2020)

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

have been selected whose frequency of occurrence is determined in the landslide documents (cf. Table 3.1).

3.3 Quality assessment of source types

Data acquisition through archival work, be it analog or digital, gathers information from a variety of source types (cf. *archival work*, chapter 1). The following proposes a method after Kreuzer et al. (2022) to assess source types on the basis of the landslide information they provide. This way, data acquisition can focus on source types with the most promising results for specific requirements of an ILIS or its data analysis modules. A significantly large dataset that recorded the source types for each landslide event has to be present in the respective ILISs to perform the proposed assessment.

In this case, the material to be examined is a dataset on landslides provided by the German landslide database (cf. Damm & Klose, 2015). The database mainly collects information on landslides from various textual sources, including results from scientific publications of field surveys (cf. Bibus & Terhorst, 2001; Hardenbicker & Grunert, 2001; Jäger et al., 2018; Klose et al., 2014; Schmidt & Beyer, 2003; Von der Heyden, 2004) and automated digital data acquisition (cf. Kreuzer & Damm, 2020). For the analysis of the present work, the dataset's information of interest are the following landslide parameters: (a) location, (b) date, (c) process type, and (d) type of source where the aforementioned information was found.

Items (a), (b), and (c) are thereby of varying degree of detail. The detail of the location is specified by three spatial confidence descriptors modified after Calvello and Pecoraro (2018)

Table 3.2 Source types from the German landslide database and examples of their respective content providers (Kreuzer et al., 2022).

Source type	Content providers
Admin. document	Forestry, road construction
Archive record	Court proceeding, deed, map commentary
Compilation	Chronicle, database, yearbook
Expert opinion	Construction opinion, technical report
Mission report	Civil protection, fire or police department
News article	Supraregional-, regional, or local newspaper
Scientific pub.	Research on landslides or subsidiary subjects

that describe the accuracy of the position: less than $100\,\mathrm{m}$ (Sd1); less than $1\,\mathrm{km}$ (Sd2); equal or grater than $1\,\mathrm{km}$ (Sd3). The date's detail is its specified exactness as either of day, month, year, or historic/prehistoric. The landslide's process type is generally differentiated between "falls" and "slides" (Cruden & Varnes, 1996; Dikau et al., 1996), and also contains additional details (e.g., size of falling material or depth of a slide's slip surface) in some cases. Thus, parameter (a) has 3 classes of detail degree (Sd1, Sd2, Sd3), (b) has 4 (day, month, year, relict), and (c) has 2 (general, detailed).

Item (d) describes a total of seven different source types from which the dataset is compiled: administrative document, archive record, compilation, expert opinion, mission report, news article, and scientific publication. The source types with examples of their content providers are listed in Table 3.2.

The method to analyse the material is based on probability and game theory. The "usefulness" for source types is defined on the basis of probability, thereby usefulness describes the probability to find adequately detailed information on landslide parameters in a given source type. The required detail for parameter information can be controlled with the conception of detail classes according to preference. In this case, economic game theory provides the framework to quantify the preference for a parameter's detail class. Therefore, it provides a mathematical *utility function* that maps detail classes to numerical values according to preference for greater detail. Thereby, high detail classes are assigned high values, and low detail classes low values, thus providing a ranking mechanism for the respective classes. After Choi and Ruszczyński (2011) and Pratt (1964) an exponential utility function best describes risk-aversion in preferences. Originally, risk-aversion meant the aversion of financial risks but in this case, it means the aversion of the risk to not find any useful information on landslides in textual documents. For this reason, the proposed method uses an exponential relation as utility function (cf. Kreuzer et al., 2022). This means the utility (or preference) of a parameter decreases exponentially with every step towards a lesser detail class.

The usefulness of source types is defined as probability and thus can be logically combined according to the same stochastic rules (Baldi, 2017), e.g., a combined probability that two events occur simultaneously, or that either event occurs exclusively. On this basis, the method proposes three degrees of usefulness to assess the quality of source types according to landslide information: the first degree describes the usefulness of a single source type regarding a single

parameter (*single-parameter usefulness*), the second degree regarding a single source type with arbitrary number of parameters (*overall usefulness*), and the third degree regarding an arbitrary number of source types and parameters (*combined usefulness*).

The first degree is the foundation for the following degrees, it is defined as the "anticipated utility function" after Quiggin (1985), principally this is a probability weighting function (Koida, 2018). In this case, the weights are represented by the exponential relation mentioned above, and the probabilities refer to the frequency of occurrence of the landslide paramater's detail classes in the respective source types. It is then a matter of combinatorics to determine the usefulness for degree two and three based on the respective usefulness's from degree one (Kreuzer et al., 2022).

The examination of the given dataset's source types is performed by yet another custom written module of the ILIS, organized into four different parameter cases. All cases demand the location parameter because it is a requirement for any spatial analysis. The cases are:

Case 1 → Location
Case 2 → Location *and* Date
Case 3 → Location *and* Process type
Case 4 → Location *and* Date *and* Process type

For all four cases, the *overall usefulness* for specific source types is examined, as well as all additional combinations of an increasing number of source types. This means, first one source type is examined, then all combinations of two source types, then of three, and so forth — the corresponding mathematical relation is determined with a least-square fit of the combinations' mean values (Guest & Guest, 2012). With this analysis, the number of required source types to reach high possibilities to find useful information is determined. In this study, possibilities of greater than $75\,\%$, i.e., within the fourth quartile, are considered as high.

Chapter 4
Results

4.1 The Automated Risk Identification Module (ARIM)

The automated statistical analysis of the material on the Franconian Alb examined 444 (semi-) active landslides with a corresponding number of 8602 adjacent downward facing slope cells. It was found that 75 % (third-quartile) of the examined landslides are situated in slope positions with adjacent inclinations equal to or less than 12°. On the basis of this result, the ARIM created ELFAs for six landslides in the defined region of interest, which are up to twice as large as the original accumulations. Furthermore, the ARIM identified an endangered rail track in two of three ELFAs. In Figure 4.1 the region of interest with its landslides, ELFAs, and endangered rail track is shown.

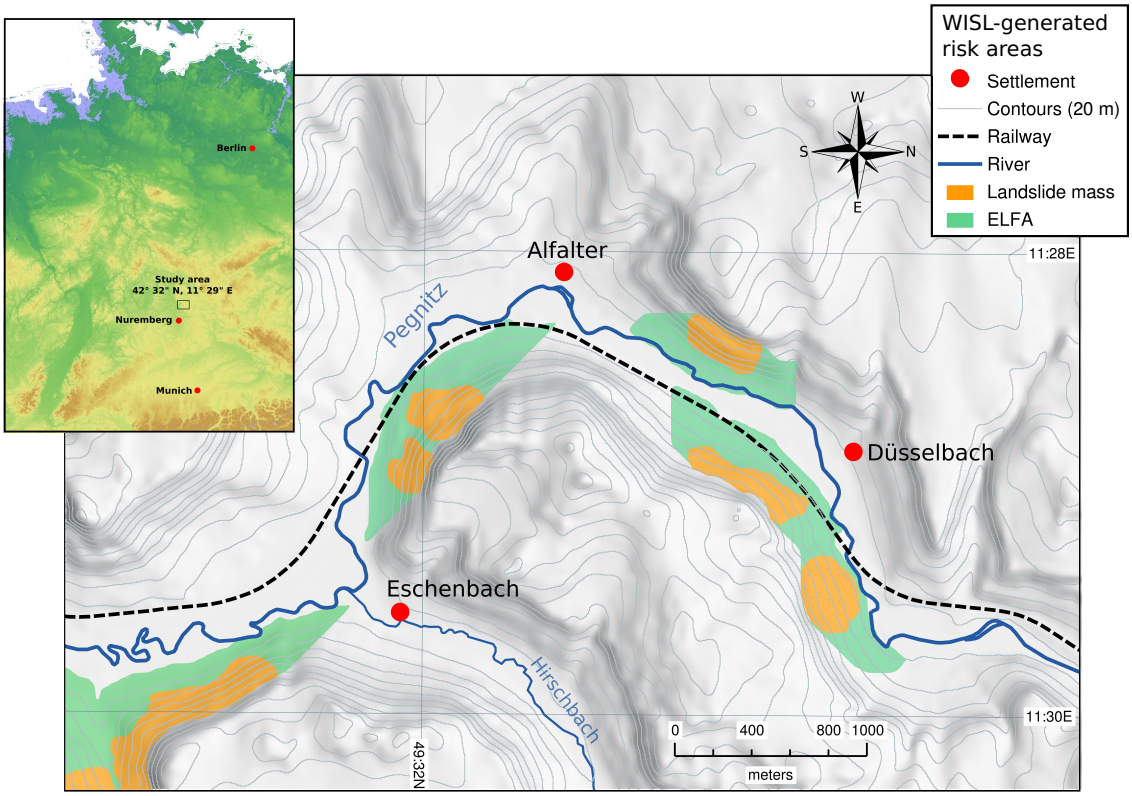


Figure 4.1 Railway between Nuremberg and Neuhaus (Pegnitz) in Northern Bavaria and the potential landslide run-out (ELFA), with a minimum of 12° inclination for the flow path, pictured on a shaded relief. The railway intersects the ELFA east of Düsselbach and west of Eschenbach and is thus marked as object at risk (Kreuzer et al., 2017).

4.2 Process chain for digital, textual sources

During the course of 87 weeks, from July 2017 to February 2019, ten keywords (German: Bergrutsch, Bergsturz, Erdrutsch, Felsrutsch, Felssturz, Gerölllawine, Hangrutsch, Schlammlawine, Schuttlawine, Steinschlag; step **A**) produced 4381 documents that were automatically collected from *Google Alert* (step **B**). Out of these documents, a share of 37 % were filtered out due to the absence of a complete sentence (step **C**), another 53 % were classified as irrelevant and thus discarded (step **D**). Duplicate detection hid 4 % from the user (step **E**), in sum this leaves 6 % of *valid* landslide documents (cf. Figure 4.2).

In case of step **C**, manually and randomly reviewed samples of the discarded and acknowledged documents were always in agreement with the automated decision process.

The classification results from the naive Bayes classifier for texts and the logistic regression for images (step **D**) 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 are 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 4.1 shows the confusion matrix for the naive Bayes classifier and Table 4.2 for results of the logistic regression on images.

Finally, primary documents and their respective duplicates were manually and automatically identified (step **E**). In the total sum of 480 detected landslide documents, a percentage of 43.51% are actual duplicates. For 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, because the algorithm does not produce any falsely identified duplicates. It is emphasized that 93% of the primary sources

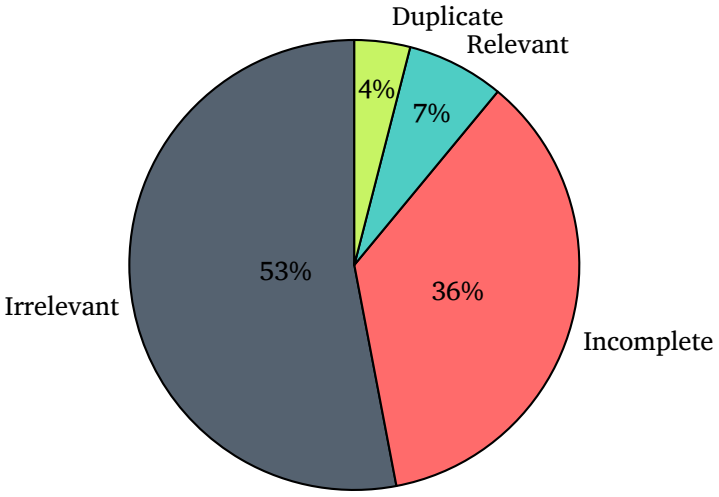


Figure 4.2 Document type distribution as determined by the automated process chain during a data collection period of 87 weeks, from July 2017 to February 2019 (n = 4381).

Table 4.1 Confusion matrix for text classification with the naive Bayes classifier. The overall accuracy (= trace of the matrix) in bold in the lower right corner.

$n = 2768$	Predicted valid	Predicted invalid	
Actual valid	**13.40%**	3.94%	17.34%
Actual invalid	8.24%	**74.42%**	82.66%
	21.64%	78.36%	**87.82%**

Table 4.2 Confusion matrix for image classification with the logistic regression classifier. The overall accuracy (= trace of the matrix) in bold in the lower right corner.

$n = 432$	Predicted valid	Predicted invalid	
Actual valid	**58.99%**	2.23%	61.22%
Actual invalid	4.09%	**34.69%**	38.78%
	63.08%	36.92%	**93.68%**

have 3 duplicates or less.

For the 6 % relevant documents, the information distribution determined by manual review is presented in Figure 4.3. Almost all texts contained information on location and date of the landslide event (94% and 96% respectively). Date and location information for images (Figure 4.3) 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, the other 10 cases provided geographic names.

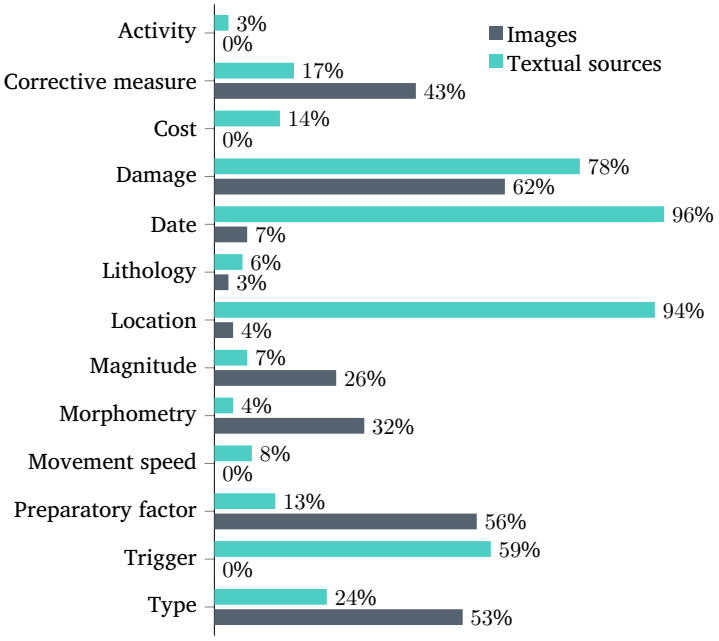


Figure 4.3 Valid textual sources ($n = 480$) and images ($n = 264$) from automated digital data acquisition that contain information about the respective attribute class (Kreuzer & Damm, 2020).

4.3 Expectation of useful landslide information in source types

The foundation to determine usefulness is the observed frequency of occurrence, which is synonymous to probability, of landslide information in source types. All in all, the examined dataset contains 4856 records that describe 4658 events reported by 4556 sources (including multiple mentions), together with location, date, and process type. The three parameters' observed probabilities for their respective detail classes are shown in Figure 4.4, differentiated after source type. Notably, scientific publications are the only source type with historic/prehistoric landslides and are the penultimate source type for detailed process type descriptions. In contrast, news articles provide the most sources with detailed process type descriptions. For clarification, it is coincidental that all three parameters have a 100 % probability to occur in any source type.

On the basis of the observed frequency of occurrence the *single parameter usefulness* for each parameter and source type is calculated after Kreuzer et al. (2022). Here, scientific publications have the highest probability to find a useful location, and news articles have the highest

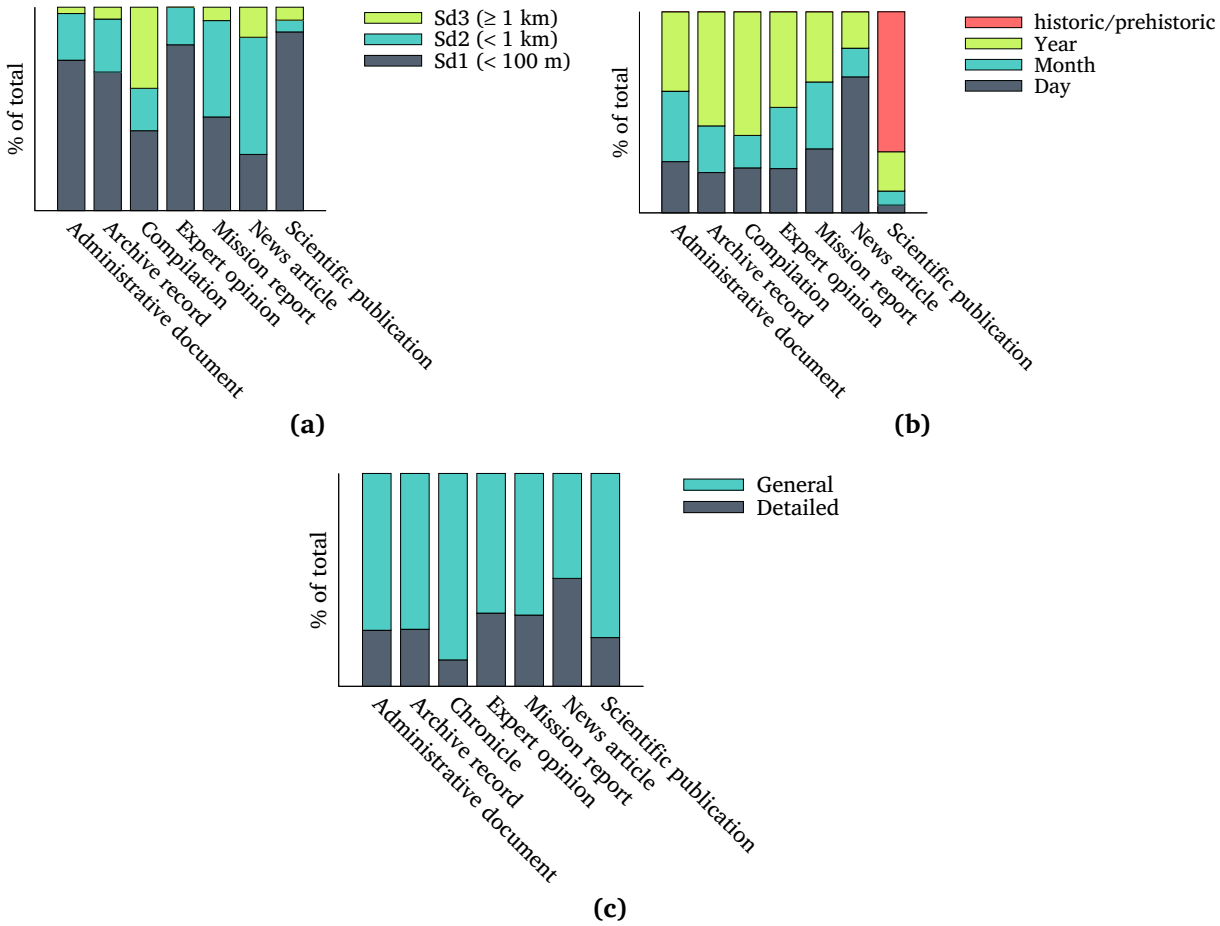


Figure 4.4 Observed probabilities, i.e., frequency of occurrence, for different degrees of detail (colouring) differentiated after examined source types for parameter (a) Location, (b) Date, and (c) Process type (Kreuzer et al., 2022).

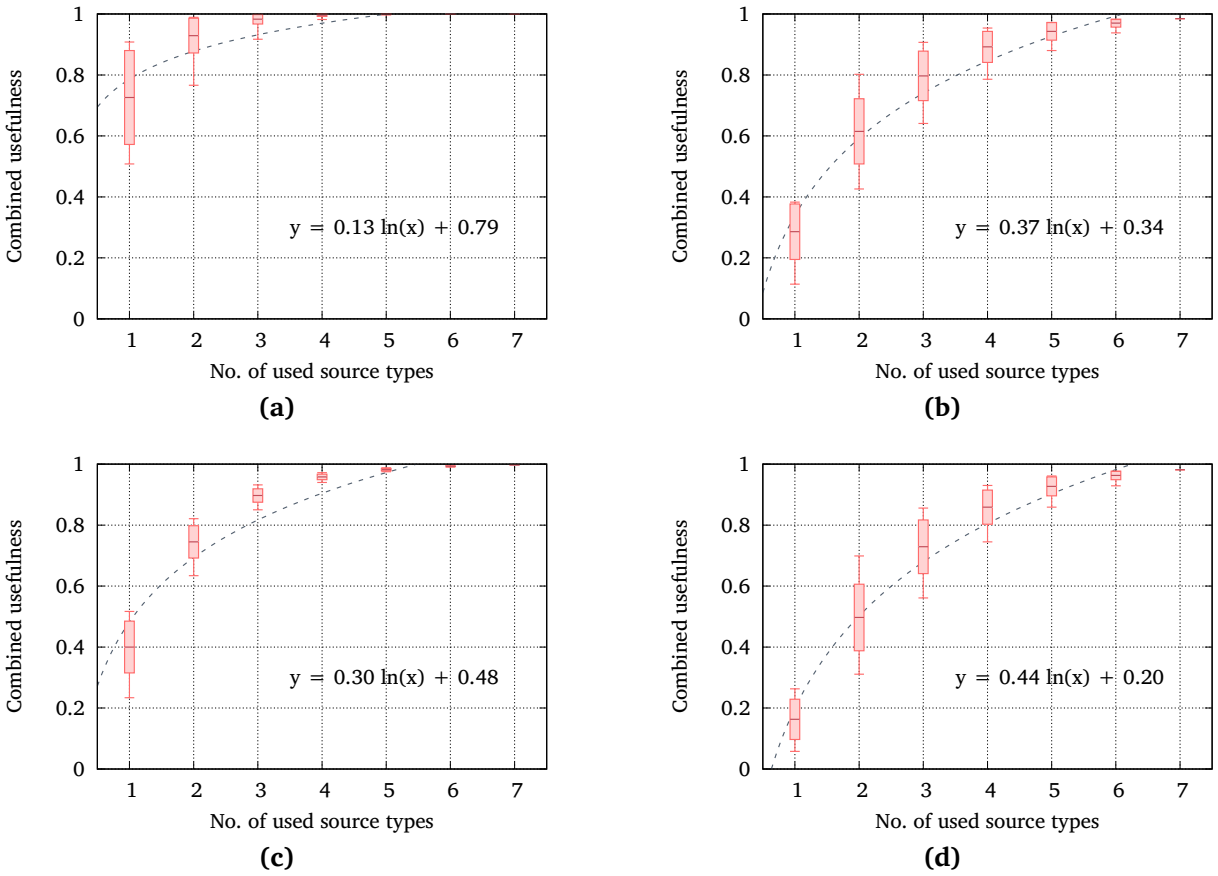


Figure 4.5 *Combined usefulness's* for (a) Case 1 (Location), (b) Case 2 (Location *and* Date), (c) Case 3 (Location *and* Process type), and (d) Case 4 (Location *and* Date *and* Process type). The x-axis shows the number of sources that are combined from the pool of all seven sources. The "candlestick" corresponding to a specific number of sources shows the minimum and maximum (lower and upper end of line), mean (horizontal bar), and standard deviation band (box) of all possible usefulness-combinations for the given number of sources; the dashed line shows the logarithmic relation (Kreuzer et al., 2022).

probability to find a useful date or process type. The average usefulness across all source types is 0.74 for location, 0.38 for date, and 0.55 for process type. "Compilation" is the only source type that is below average for all three parameters and "expert opinion" is the only source type that is above average for all three parameters. For location, the complete statistical description (minimum, maximum, mean, standard deviation band) is shown in Fig. 4.5a at x-axis location $x = 1$. Moreover, on the same x-axis location, Figure 4.5b-d show statistical descriptions of the *overall usefulness* for the specific cases, i.e., the intersection of probabilities for location *and* date (Fig. 4.5b), location *and* process type (Fig. 4.5c), and location *and* date *and* process type (Fig. 4.5d). The highest *combined usefulness* for each case and number of combinations seen in Figure 4.5 is always the combination of the respective number of source types with the highest *overall usefulness* for the desired case. It can be seen that the more parameters are required the less likely it is to find useful information on all of them in one source type. For this reason, a higher usefulness is achieved through combination of source types as is shown in Figure 4.5. Furthermore, the least-square fit proved a logarithmic relation for the increase of the *combined usefulness* with the number of combined source types N, i.e., $\tilde{U} = a\ln(N)+b$. Here, the intercept

b corresponds to the initial value of usefulness for 1 source type, and the gradient a represents the logarithmic growth rate for every additional source type (Figure 4.5).

Chapter 5
Discussion

The proposed technical foundation for an ILIS is a flexible, integrative, spatial database, which allows for convenient registration procedures, data storage of general spatial information, statistical analyses, and custom-made modules. Furthermore, it enables data transfers to external users on the base of results, without raw data movement, in contrast many scientific data systems still exchange large amounts of data (Mattmann et al., 2016). The implementation with PostgreSQL/PostGIS corresponds to a low-budget-solution which provides independence of corporate support. The disadvantage is that there is no party with legal obligations for maintenance and warranty of operation. However, PostgreSQL/PostGIS is a tried-and-true RDBMS deployed in many projects, which rely on the functioning of the software. Maintenance is thus a community effort potentially distributed among all users of the software. Projects in other scientific fields have already good experience with open source databases (Craig et al., 2004). The competence of the system lies, in general, on its high flexibility, openness and integration capability for further modules.

The data sheet for landslide registration is conceived for measurable and interpretative data based on descriptions of field results, moreover, it serves as the digital user interface for ILISs (Jäger et al., 2018; Kreuzer et al., 2017). However, data interpretation is not obligatory for data input, and thus is suitable for laymen and experts (Bechtold & Patterson, 2005).

The integration and application of custom analysis modules (i.e., ARIM) illustrated the capabilities of an ILIS. The elucidated case study shows that rapid spatial analyses of registered data is possible (cf. Figure 4.1). The ARIM as a topographic analysis module can be seen as one step towards automatically generated risk analyses. In this context, ARIM is applied in order to expose challenges of the selected approach related to automated analyses (van Westen et al., 2008; Wieczorek, 1984). The decision to implement the ARIM as a statistical model was made due to the unavailability of process data for a wide area of interest. However, the module is designed for data scalability: as additional information is obtained in the inventory, e.g. reholitic parameters, lithographic maps etc., respective physical models can complement ARIM. Thus, the module is a common base for statistically and physically based disposition and procedural modelling.

With the technical base of the ILIS in place, it is made possible to implement multiple methods for automated data acquisition that continuously provide input for the ILIS. In this case, the automated process chain provides a method to acquire digital, textual data on landslides automatically. During its testing period, $94\,\%$ of the gathered data was discarded or hidden (Figure 4.2), this share of unused documents requires context. Particularly relevant for the

Table 5.1 Number of results for keyword searches during one year: Kreuzer and Damm (2020) 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
Kreuzer and Damm*	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

evaluation of the proposed method in Kreuzer and Damm (2020) are the works of Taylor et al. (2015) and Innocenzi et al. (2017). Both publications provide a detailed evaluation of a similar acquisition process.

In general, the keyword search is language specific, however, common principles apply and motivate a deeper comparison with studies of different languages. The results of this comparison are listed in Table 5.1. Compared to the present work, Innocenzi et al. (2017) reports 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 irrelevant results, while accepting to miss out on landslide documents that could have been found with other landslide specific keywords. The relatively small difference in absolute numbers, i.e. one Italian keyword produces almost as many results as ten 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. Taylor et al. (2015) relied on document preselection of a specific news archive during their acquisition process, this limited the number of overall results as well as irrelevant results. Thus, neither strategy of the compared works provides a comprehensive approach.

Generally, double meaning of keywords, specifically of scientific and colloquial use, produce irrelevant results. For example, 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. 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.

Furthermore, the process chain filtered $37\,\%$ of all documents that did not contain a complete sentence with any of the predefined keywords (cf. Figure 4.2). Thus, documents that con-

tain only grammatically incomplete headlines, e.g. news aggregators, are principally avoided. Moreover, 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 following classification step.

The $53\,\%$ irrelevant documents are determined through naive Bayes classification for texts (cf. Figure 4.2). Since no naive Bayes classifier has been implemented specifically for landslide documents the overall accuracy can only be compared to 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 4.1) 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, while 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, landslide documents are manually classified as invalid if they are outside the region of interest. 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 to email spam filters.

The image detection of the automated process chain performed generally better than the textual part. 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.

Regarding duplicate detection, this lead to a $4\,\%$ data volume reduction with $86.12\,\%$ accuracy (cf. Figure 4.2). 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 principles 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.

The results of the quality assessment give an overview of the information type that can be expected from digital data acquisition for landslides. Specifically, the results shown in Fig-

ure 4.3 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.

To increase the effectiveness of data acquisition, both digital and analog, for an ILIS, a method was proposed to assess the usefulness of textual sources used in the acquisition process. The method assesses textual sources regarding information on landslide parameters in three degrees:

1. *single parameter usefulness* (one source type, one parameter)
2. *overall usefulness* (one source type, arbitrary number of parameters)
3. *combined usefulness* (arbitrary number of source types and parameters)

Moreover, for every landslide parameter there have to be established detail classes. In this case, the detail classes of the examined material resulted from the given data structure and were not altered. Here, the lowest detail classes from all examined parameters define a wide quality range that includes virtually any parameter information. For example, Sd3 from "location" describes an accuracy of equal or greater than $1\,\mathrm{km}$, this means any location information can at least be Sd3. This kind of inclusive detail class is no prerequisite for the method, and in this case, adds to the occurrence of $100\,\%$ chances in Figure 4.4.

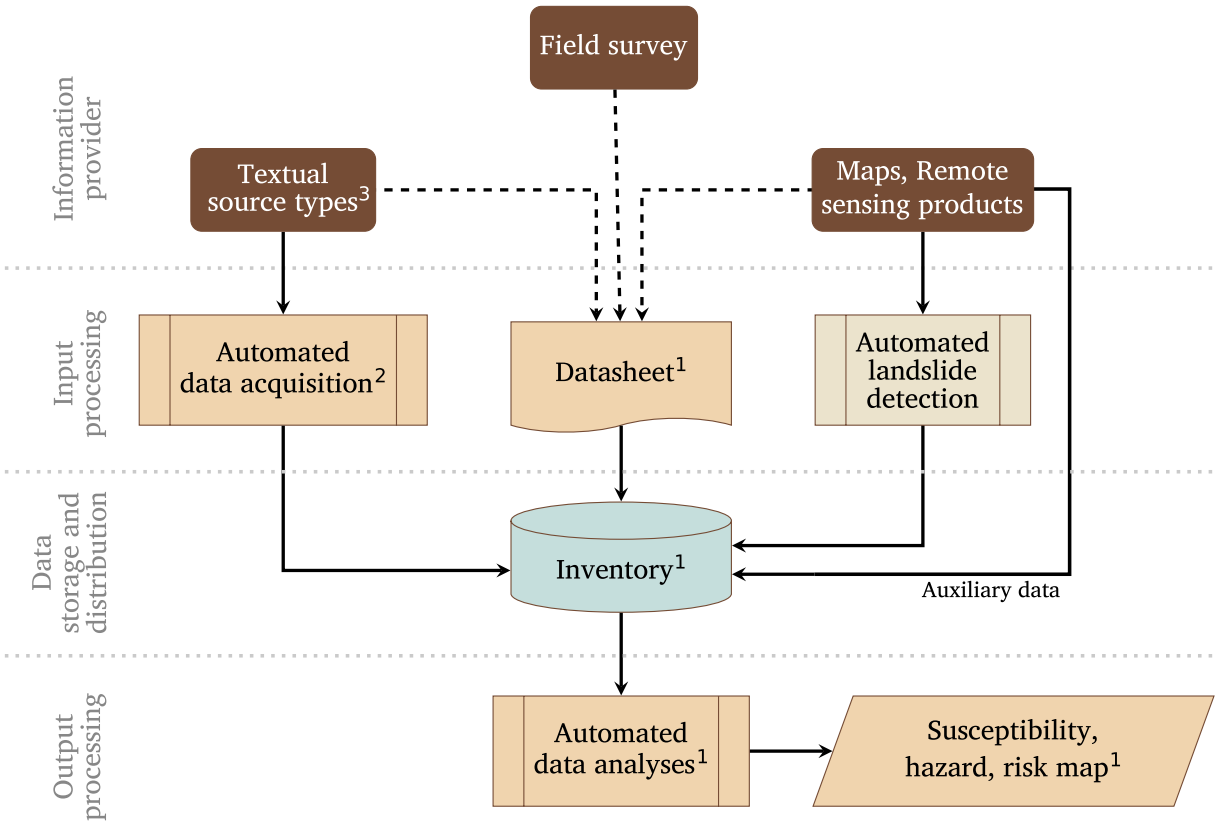
On this basis, the frequency of occurrence for information quality in scientific publications was found to be strongly heterogeneous in this study. On one hand, they provide a location information with the best quality "Sd1" for more than $87\,\%$ of all events (Figure 4.4a). On the other hand, it is the only source type that contains the weakest detail class "historic/prehistoric" for date, and this in $70\,\%$ of all cases (Figure 4.4b). Presumably, this disparity is due to lack of bias in scientific research towards contemporary landslide events, which are in most regions less numerous than relict landslides that were amassed over time. In contrast, it is likely that all other source types have a bias towards contemporary landslide events, as exemplified by the high frequency ($68\,\%$, Figure 4.4b) of day-exact dates in news articles. Additionally, news articles contain the highest frequency of detailed process type descriptions (Figure 4.4c), which reinforces the conclusion that they are mainly concerned with the impact of landslide events.

The aforementioned particularities of scientific publications and news articles are correctly reflected in the *single parameter usefulness*. Here, scientific publications are the most "useful" source type for location information, and news articles for date and process type. In this context, numbers for the "overall usefulness" give the consequent qualitative result: scientific publication, expert opinion, and news article are most "useful" for the four devised cases. Quantitatively, however, the numbers are roughly halved with every additional parameter requirement, i.e., one parameter has an average usefulness of 0.74 (Case 1), two parameters 0.34 (mean of Case 2 and 3), and three parameters 0.16 (Case 4). This means, in case source types are to be examined regarding even more parameters, a single source type is unlikely to produce useful information on all of them. In contrast to the "overall usefulness's" decline with every additional parameter, the opposite is true for the "combined usefulness": the more parameters are used the higher the logarithmic growth rate (Figure 4.5). Consequently, every additional source type that is combined increases the usefulness more strongly the more parameters are

used, this is a weak counterbalance: linear decline vs. logarithmic growth.

Chapter 6
Conclusions and outlook

The present thesis proposes a set of methods and a technical implementation that operate on four sequentially interdependent levels to create an ILIS: (1) information provider, (2) input processing, (3) data storage and distribution, and (4) output processing (Figure 6.1). Specifically, textual data acquisition in level (1) is made more effective with a focus on the most promising source types after the assessment of Kreuzer et al. (2022); in level (2), the proposed data sheet is the focal point for all analogously collected data, while digital data is continuously processed by the automated data acquisition method from Kreuzer and Damm (2020). Principally, digital and analog data alike converges in a central data processing location provided by the technical foundation laid in Kreuzer et al. (2017) (cf. chapter 2). It is also this foundation



1: T.M. Kreuzer et al. (2017)
2: T.M. Kreuzer & B. Damm (2020)
3: T.M. Kreuzer et al. (2022)

Figure 6.1 Schematic of the proposed Integrated Landslide Inventory System (ILIS). Dashed arrows represent analog, and solid arrows digital data flow.

that allows for automated analyses of the collected data, as exemplified in Kreuzer et al. (2017) (cf. section 4.1).

Together, a semi-automated ILIS is established where the user's primary role can be reduced to control the selection of information, as well as the output products. This way, a continuous and automated flow of landslide information is possible, i.e., a "living" landslide database that updates its analyses results based on the latest acquired landslide information.

The next step is a full automation on digital data acquisition that not only delivers sources for further manual evaluation but readily identifies and automatically extracts information for specified landslide parameters in those sources. In this context, the data already collected in the course of this work can be reused. For example, the available texts can train artificial neural networks to identify specific information on landslides in newly collected texts. Furthermore, integration of established methods for automated landslide detection from maps and remote sensing products would complement the ILIS. Here, too, integration is readily possible, particularly because digital maps and remote sensing products are stored in the inventory as auxiliary data, and therefore remain available for further analyses (Figure 6.1).

On the output level of the ILIS research on analyses with increased detail and accuracy is another important step. That is because many risk/hazard/susceptibility analysis methods are primarily data driven, i.e., the available data defines the possible methods. In an environment with continuously changing data this presents another challenge that should be addressed in future works.

All in all, if data acquisition and analyses are fully automated as well as aligned adequately, such a system could make landslide inventories available to anyone independent of required geographic scale and expert knowledge in data processing. It could keep the data continuously updated and consistent in a timely manner, which in itself opens up new possibilities for integration in early warning systems.

Bibliography

Aggarwal, C. C., & Zhai, C. (Eds.). (2012). *Mining Text Data*. Springer US.

Ayenew, T., & Barbieri, G. (2005). Inventory of Landslides and Susceptibility Mapping in the Dessie Area, Northern Ethiopia. *Engineering Geology*, *77*(1), 1–15.

Baldi, P. (2017). *Stochastic Calculus: An Introduction Through Theory and Exercises*. Springer International Publishing.

Battistini, A., Segoni, S., Manzo, G., Catani, F., & Casagli, N. (2013). Web Data Mining for Automatic Inventory of Geohazards at National Scale. *Applied Geography*, *43*, 147–158.

Bechtold, W. A., & Patterson, P. L. (2005). *The enhanced forest inventory and analysis program - national sampling design and estimation procedures* (Gen. Tech. Rep. SRS-80). U.S. Department of Agriculture, Forest Service, Southern Research Station. Asheville, NC.

Behling, R., Roessner, S., Kaufmann, H., & Kleinschmit, B. (2014). Automated Spatiotemporal Landslide Mapping over Large Areas Using RapidEye Time Series Data. *Remote Sensing*, *6*(9), 8026–8055.

Bibus, E., & Terhorst, B. (2001). Mass Movements in Southwest Germany. Analyses and Results from the Tübingen Work Group of the MABIS Project. *Zeitschrift für Geomorphologie*, *125*, 93–103.

Calvello, M., & Pecoraro, G. (2018). FraneItalia: A Catalog of Recent Italian Landslides. *Geoenvironmental Disasters*, *5*(13), 1–16.

Carrara, A., & Merenda, L. (1976). Landslide Inventory in Northern Calabria, Southern Italy. *GSA Bulletin*, *87*(8), 1153–1162.

Choi, S., & Ruszczyński, A. (2011). A Multi-Product Risk-Averse Newsvendor with Exponential Utility Function. *European Journal of Operational Research*, *214*(1), 78–84.

Craig, R., Cortens, J. P., & Beavis, R. C. (2004). Open Source System for Analyzing, Validating, and Storing Protein Identification Data. *Journal of Proteome Research*, *3*(6), 1234–1242.

Cruden, D. M., & Varnes, D. J. (1996). Landslide Types and Processes: Chapter 3. In *Landslides-Investigation and Mitigation* (pp. 36–75). National Academy Press.

Damm, B., Becht, M., Varga, K., & Heckmann, T. (2010). Relevance of Tectonic and Structural Parameters in Triassic Bedrock Formations to Landslide Susceptibility in Quaternary Hillslope Sediments. *Quaternary International*, *222*(1), 143–153.

Damm, B., & Klose, M. (2015). The Landslide Database for Germany: Closing the Gap at National Level. *Geomorphology*, *249*, 82–93.

Dikau, R., Cavallin, A., & Jäger, S. (1996). Databases and GIS for Landslide Research in Europe. *Geomorphology*, *15*(3), 227–239.

Foster, C., Pennington, C. V. L., Culshaw, M. G., & Lawrie, K. (2012). The National Landslide Database of Great Britain: Development, Evolution and Applications. *Environmental Earth Sciences*, *66*(3), 941–953.

Gaidzik, K., Ramírez-Herrera, M. T., Bunn, M., Leshchinsky, B. A., Olsen, M., & Regmi, N. R. (2017). Landslide Manual and Automated Inventories, and Susceptibility Mapping Using LIDAR in the Forested Mountains of Guerrero, Mexico. *Geomatics, Natural Hazards & Risk*, *8*(2), 1054–1079.

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*, *29*, 31–43.

Guest, P. G., & Guest, P. G. (2012). *Numerical Methods of Curve Fitting*. Cambridge University Press.

Guzzetti, F., Cardinali, M., & Reichenbach, P. (1994). The AVI Project: A Bibliographical and Archive Inventory of Landslides and Floods in Italy. *Environmental Management*, *18*(4), 623–633.

Guzzetti, F., Mondini, A. C., Cardinali, M., Fiorucci, F., Santangelo, M., & Chang, K.-T. (2012). Landslide Inventory Maps: New Tools for an Old Problem. *Earth-Science Reviews*, *112*(1), 42–66.

Hardenbicker, U., & Grunert, J. (2001). Temporal occurrence of mass movements in the Bonn area. *Zeitschrift für Geomorphologie*, *125*, 13–24.

Hölbling, D., Friedl, B., & Eisank, C. (2015). An object-based approach for semi-automated landslide change detection and attribution of changes to landslide classes in northern Taiwan. *Earth Science Informatics*, *8*(2), 327–335.

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 M. Mikos, B. Tiwari, Y. Yin & K. Sassa (Eds.), *Advancing Culture of Living with Landslides* (pp. 1127–1133). Springer International Publishing.

Jäger, D., Kreuzer, T., Wilde, M., Bemm, S., & Terhorst, B. (2018). A Spatial Database for Landslides in Northern Bavaria: A Methodological Approach. *Geomorphology*, *306*, 283–291.

Klose, M. (2015). *Landslide Databases as Tools for Integrated Assessment of Landslide Risk*. Springer International Publishing.

Klose, M., Damm, B., & Highland, L. (2015). Databases in Geohazard Science: An Introduction. *Geomorphology*, *249*, 1–3.

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.

Klose, M., Maurischat, P., & Damm, B. (2016). Landslide Impacts in Germany: A Historical and Socioeconomic Perspective. *Landslides*, *13*(1), 183–199.

Koida, N. (2018). Anticipated Stochastic Choice. *Economic Theory*, *65*(3), 545–574.

Kreuzer, T. M., & Damm, B. (2020). Automated Digital Data Acquisition for Landslide Inventories. *Landslides*, *17*(9), 2205–2215.

Kreuzer, T. M., Damm, B., & Terhorst, B. (2022). Quantitative Assessment of Information Quality in Textual Sources for Landslide Inventories. *Landslides*, *19*, 505–513.

Kreuzer, T. M., Wilde, M., Terhorst, B., & Damm, B. (2017). A Landslide Inventory System as a Base for Automated Process and Risk Analyses. *Earth Science Informatics*, *10*(4), 507–515.

Kusner, M. J., Sun, Y., Kolkin, N. I., & Weinberger, K. Q. (2015). From Word Embeddings To Document Distances. *Proceedings of Machine Learning Research*, *37*, 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 Recognition*, *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. *Natural Hazards*, *69*(3), 1477–1495.

Lynch, C. (2008). Big data: How do your data grow? *Nature*, *455*, 28–29.

Manning, C., Raghavan, P., & Schuetze, H. (2009). *Introduction to Information Retrieval*. Cambridge University Press.

Manzo, G., Tofani, V., Segoni, S., Battistini, A., & Catani, F. (2013). GIS Techniques for Regional-Scale Landslide Susceptibility Assessment: The Sicily (Italy) Case Study. *International Journal of Geographical Information Science*, *27*(7), 1433–1452.

Mattmann, C. A., Cinquini, L., Zimdars, P., Joyce, M., & Khudikyan, S. (2016). A topical evaluation and discussion of data movement technologies for data-intensive scientific applications. *Earth Science Informatics*, *9*(2), 247–262.

Meinhardt, M., Fink, M., & Tünschel, H. (2015). Landslide Susceptibility Analysis in Central Vietnam Based on an Incomplete Landslide Inventory: Comparison of a New Method to Calculate Weighting Factors by Means of Bivariate Statistics. *Geomorphology*, *234*, 80–97.

Mitchell, T., Christl, A., & Emde, A. (2008). *Web-Mapping mit Open Source-GIS-Tools*. O'Reilly.

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.

Obe, R. O., & Hsu, L. S. (2011). *PostGIS In Action*. Manning Publications.

Petkovic, D. (2016). Temporal Data in Relational Database Systems: A Comparison. In Á. Rocha, A. M. Correia, H. Adeli, L. P. Reis & M. Mendonça Teixeira (Eds.), *New Advances in Information Systems and Technologies* (pp. 13–23). Springer International Publishing.

Pratt, J. W. (1964). Risk Aversion in the Small and in the Large. *Econometrica*, *32*(1/2), 122–136.

Quiggin, J. (1985). Subjective Utility, Anticipated Utility, and the Allais Paradox. *Organizational Behavior and Human Decision Processes*, *35*(1), 94–101.

Raymond, E. S. (2001). *The Cathedral & the Bazaar. Musings on Linux and Open Source by an Accidental Revolutionary*. O'Reilly.

Reichenbach, P., Rossi, M., Malamud, B. D., Mihir, M., & Guzzetti, F. (2018). A Review of Statistically-Based Landslide Susceptibility Models. *Earth-Science Reviews*, *180*, 60–91.

Ropeik, D. (2002). *Risk: A practical guide for deciding what's really safe and what's really dangerous in the world around you*. Houghton Mifflin Harcourt.

Rusland, N. F., 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*, 012091.

Sabatakakis, N., Koukis, G., Vassiliades, E., & Lainas, S. (2013). Landslide Susceptibility Zonation in Greece. *Natural Hazards*, *65*(1), 523–543.

Sandmeier, C., Büdel, C., & Schwindt, D. Multi-methodological investigation of a mass movement in the cuesta landscape of the northeastern Franconian Alb, Germany [EGU2013-3298]. In: In *Egu general assembly conference abstracts*. *15*. EGU2013-3298. Vienna, 2013, 3298.

Schmid, H. (1999). Improvements in Part-of-Speech Tagging with an Application to German. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann & D. Yarowsky (Eds.), *Natural Language Processing Using Very Large Corpora* (pp. 13–25). Springer Netherlands.

Schmidt, K.-H., & Beyer, I. (2003). High-Magnitude Landslide Events on a Limestone-Scarp in Central Germany: Morphometric Characteristics and Climatic Controls. *Geomorphology*, *49*(3), 323–342.

Sumathi, S., & Esakkirajan, S. (Eds.). (2007). Objected-Oriented and Object Relational DBMS. In *Fundamentals of Relational Database Management Systems* (pp. 477–558). Springer.

Taylor, F. E., Malamud, B. D., Freeborough, K., & Demeritt, D. (2015). Enriching Great Britain's National Landslide Database by Searching Newspaper Archives. *Geomorphology*, *249*, 52–68.

Ting, K. M. (2017). Confusion Matrix. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 260–260). Springer US.

Valenzuela, P., Domínguez-Cuesta, M. J., Mora García, M. A., & 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.

van Westen, C. J., Castellanos, E., & Kuriakose, S. L. (2008). Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview. *Engineering Geology*, *102*(3–4), 112–131.

Von der Heyden, D. (2004). *Rutschungen an Den Malmschichtstufen Der Nordwestlichen Frankenalb: Untersuchungen Zu Formenschatz, Alter Und Ursachen* (Diss.). Univ. Bamberg.

Wieczorek, G. F. (1984). Preparing a Detailed Landslide-Inventory Map for Hazard Evaluation and Reduction. *Environmental and Engineering Geoscience*, *21*(3), 337–342.

Wilde, M., Günther, A., Reichenbach, P., Malet, J.-P., & Hervás, J. (2018). Pan-European Landslide Susceptibility Mapping: ELSUS Version 2. *Journal of Maps*, *14*(2), 97–104.

Wohlers, A., Kreuzer, T., & Damm, B. (2017). Case Histories for the Investigation of Landslide Repair and Mitigation Measures in NW Germany. In K. Sassa, M. Mikǒs & Y. Yin

(Eds.), *Advancing Culture of Living with Landslides* (pp. 519–525). Springer International Publishing.

Zhang, H. (2005). Exploring Conditions for the Optimality of Naïve Bayes. *International Journal of Pattern Recognition and Artificial Intelligence*, *19*(2), 183–198.