2019

Towards developing a goal-driven data integration framework for counter-terrorism analytics

Dapeng Liu
Virginia Commonwealth University

TOWARDS DEVELOPING A GOAL-DRIVEN DATA INTEGRATION FRAMEWORK FOR COUNTER-TERRORISM ANALYTICS

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

By Dapeng Liu

Supervisor: Dr. Victoria Yoon

Professor, Information Systems

Virginia Commonwealth University

Richmond, Virginia
July, 2019

Acknowledgement

First and foremost, I express my sincere gratitude and special thanks to my supervisor Dr. Yoon who have always been my strong and unchanging rock during my PhD study. I cannot imagine what I would be without her endless patience, warm inspirations, and solid support. She is the best supervisor and mentor.

I present my thanks to Dr. Allen Lee and Dr. Manoj Thomas for inspiring and motivating me, for teaching and mentoring me, and for giving me invaluable advice. Thank you, Dr. Christopher Whyte, for giving me advice which motivates me a lot.

Thank you, my mom and dad. You are the best in the world. Thank you, Meijun, my love, for everything we have together in the four years and you did for me.

Table of Contents

Abstract 6

Chapter 1. Introduction 8

1.1 Data Integration for Counter-terrorism Analytics 8

1.2 Research Motivation—Data Integration for Counter-terrorism Analytics 9

1.3 Challenges and Research Issues 12

1.4 Research Questions 14

1.5 Research Significance and Contributions 18

1.6 Organization of Dissertation 20

Chapter 2. Literature Review 22

2.1 Terrorism Data and Counter-terrorism Analytics 22

2.1.1 Terrorism 22

2.1.2 Data for Counter-terrorism Analytics 23

2.1.3 Terrorism Data and Counter-terrorism Analytics 30

2.2 Data Integration in IS 32

2.2.1 Data Integration 32

2.2.2 Existing Solutions for Data Integration 33

Chapter 3. Research Method 44

3.1 Research Objectives 44

3.2 Design Science Research Methodology 48

3.3 Theoretical Background 53

3.4 Goal-driven Data Integration Framework 57

3.5 Implementation 78

Chapter 4. Evaluation 82
4.1 Experimental Setting 83
4.2 Experiment 1—An Evaluation of Goal-relevant Data Identification 85
4.3 Experiment 2—An Evaluation of the Schema Mapper 88
4.4 Experiment 3—An Evaluation of the Instance Mapper 95
4.5 Experiment 4—A Sensitivity Analysis 100
Chapter 5. Discussion and Conclusion 104
5.1 Practical Implications 106
5.2 Theoretical Implications 108
5.3 Limitations and Future Directions 110
References 113
Appendices: 124
APPENDIX A: Highlights 124
APPENDIX B: Metadata Example 126
APPENDIX C: Design Potentials 128
APPENDIX D: Algorithm Performance Metrics (F-measure) 129

Abstract

TOWARDS DEVELOPING A GOAL-DRIVEN DATA INTEGRATION FRAMEWORK FOR COUNTER-TERRORISM ANALYTICS

By
Dapeng Liu

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

Virginia Commonwealth University, 2019

Director: Dr. Victoria Yoon
Professor, Information Systems

Terrorist attacks can cause massive casualties and severe property damage, resulting in terrorism crises surging across the world; accordingly, counter-terrorism analytics that take advantage of big data have been attracting increasing attention. The knowledge and clues essential for analyzing terrorist activities are often spread across heterogeneous data sources, which calls for an effective data integration solution. In this study, employing the goal definition template in the Goal-Question-Metric approach, we design and implement an automated goal-driven data integration framework for counter-terrorism analytics. The proposed design elicits and ontologizes an input user goal of counter-terrorism analytics; recognizes goal-relevant datasets; and addresses semantic heterogeneity in the recognized datasets. Our proposed design, following the design science methodology, presents a theoretical framing for on-demand data integration designs that can accommodate diverse

and dynamic user goals of counter-terrorism analytics and output integrated data tailored to these goals.

Keywords: data interoperability, data integration, data management, ontology

Chapter 1. Introduction

1.1 Data Integration for Counter-terrorism Analytics

Terrorist attacks can cause massive casualties and severe property damage, resulting in terrorism crises spreading across the world (Tutun et al. 2017); accordingly, counter-terrorism has been attracting increasing attention in the past two decades. Additionally, because terrorism recognizes no boundaries in terms of time, space, ideologies, perpetrators, targets, and motivations, counter-terrorism analytics is challenging—analyzing patterns in terrorist tactics and strategies; identifying terrorist communication networks; and predicting terrorists' behaviors, intentions, and future moves (Argomaniz et al. 2015).

In counter-terrorism analytics, data analysts often take advantage of various terrorism datasets and data analytics tools (Ding et al. 2017). Specifically, to facilitate counter-terrorism analytics, Computer Science (CS) and Information Systems (IS) professionals are dedicated to utilizing analytical services and techniques to discover, engineer, and manage terrorism knowledge, linking clues and information spreading across heterogeneous datasets (Ding et al. 2017). However, semantic heterogeneity existing in these terrorism datasets (e.g., both data schema and instances) precludes data analysts from fully integrating them and, accordingly, performing in-depth counter-terrorism analytics.

Therefore, to effectively and efficiently conduct counter-terrorism analytics and fully utilize these heterogeneous datasets to obtain a deeper understanding of terrorism,

scholars need to appropriately link and consolidate terrorism data (Behlendorf and LaFree 2014). For example, integrating the Global Terrorism Database (GTD) and Big Allied and Dangerous Database (BAAD) has enabled researchers to analyze terrorism patterns regarding perpetrator ideology (Forest 2012). Additionally, scholars and practitioners (e.g., the police and homeland security agents) are also calling for automated terrorism data integration solutions (Kurlander 2005), especially after the September 11 terrorist incident, arguing that if an intelligent data integration system had been in place, the isolated data and clues to suspicious activities might have been marked and reported, and the subsequent terrorist hijacking might have been forestalled (Popp et al. 2004).

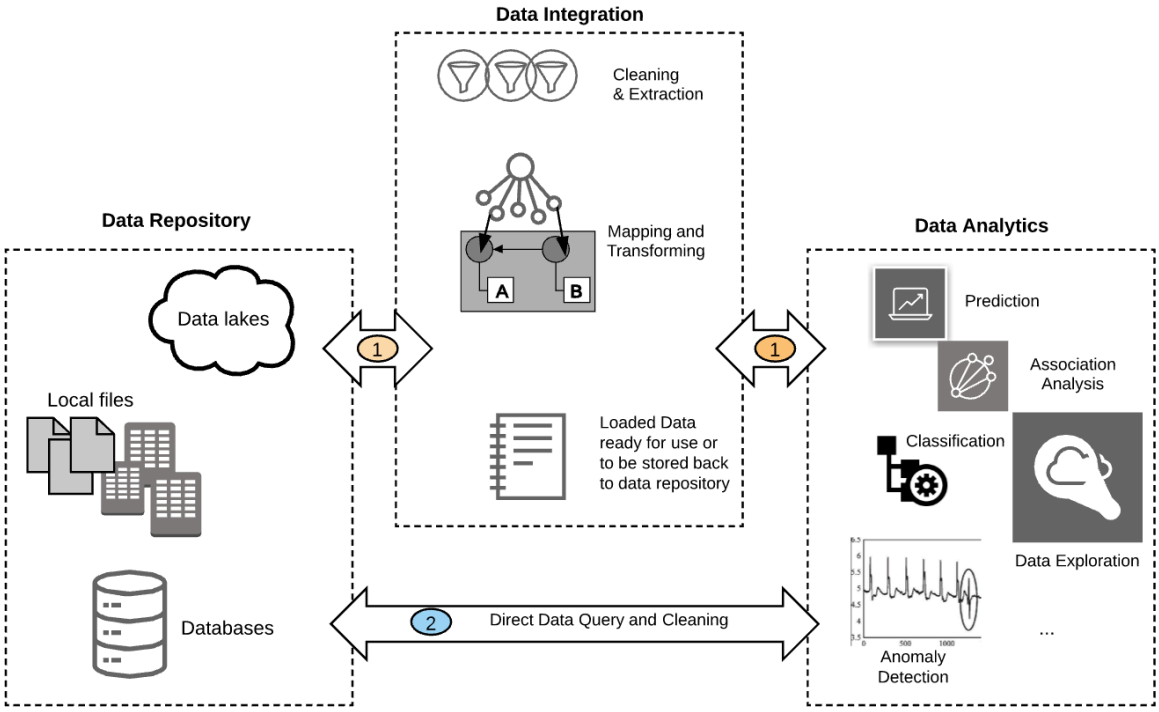
1.2 Research Motivation—Data Integration for Counter-terrorism Analytics

A massive range of terrorism data, stored in data repositories (e.g., databases, data lakes[1]) is now available publicly, enabling scholars, agents, and organizations to analyze terrorist attack patterns and to implement counter-terrorism intelligence; however, there is a lack of automated data integration solutions for counter-terrorism analytics, especially in relation to the tremendous heterogeneous data available in terrorism data lakes. A simple application of conventional data integration solutions to this case is problematic. First, counter-terrorism analytics tasks are often triggered by users' goals, which are distinct

[1] A data lake refers to a data repository that organizes and stores a vast amount of data (e.g., terrorism data) in its native format.

from one another; therefore, effective data integration for counter-terrorism analytics requires taking these various goals into consideration appropriately, thereby integrating and providing goal-relevant data based on users' interests. Second, rich metadata information in free text (e.g., title, keywords, and description) is available for terrorism datasets on the websites hosting these datasets and can assist an in-depth understanding of these datasets. Therefore, it is critical that we seek a data integration solution to take into consideration the two aspects listed above. To address this gap, in this study, we are motivated to follow the design science approach to design an intelligent goal-driven data integration solution for counter-terrorism analytics.

Data integration sits at the back end of a data analytics task (e.g., a counter-terrorism analytics task). It seeks to consolidate heterogeneous data from multiple sources for a particular use (Milo and Zohar 1998). In other words, data integration assembles those data particles that do not automatically fit together due to semantic heterogeneity, which initially precludes counter-terrorism analysts to find the linkage among data spread across multiple sources. Data integration can be a manual or computational process (or a hybrid) to provide end users with consolidated data from multiple sources based on users' needs. Specifically, it consists of analyzing the extant data from diverse sources; resolving the syntax and semantic heterogeneity; creating new datasets by reorganizing the aligned data based on a specific user's need; and sharing these newly generated datasets with the user.



(1) Data needed to conduct data analytics is of semantic heterogeneity; data integration is necessary.
(2) Data needed to conduct data analytics is from a single source and no heterogeneity exists.

Figure 1.1 Data Integration and Data Analytics

Automated data integration solutions have been extensively studied and developed by both practitioners and academics. Theoretically, data integration approaches can be categorized into one of two paradigms—"on-demand" and "in-advance" (Widom 1996). In the "on-demand" approach, the system accepts a user query, determines the appropriate set of information sources to answer the query, generates subqueries for each involved information source, retrieves results from each information source, and performs appropriate consolidation. In contrast, in the "in-advance" approach, data that may be of interest are extracted and consolidated appropriately in advance. When a query is received, the in-advance prepared data are returned as a response. The "in-advance" approach

assumes that "data of interests" are predictable. Therefore the "on-demand" approach is more applicable when users' needs are unpredictable (e.g., research goals, entities of interest, research scopes). For instance, a researcher might be interested in "kidnapping by terrorist groups, 1970–2010" (Forest 2012), whereas another one might be interested in "the risk of terrorist attacks at a global scale" (Ding et al. 2017). In such a situation, data integration will be carried out by both technology scientists and actual users, who may have various goals, interests, research emphases, and research scopes.

Accordingly, a goal-driven data integration solution is needed for counter-terrorism analytics, especially when tremendous heterogeneous data exist. As such, in this study, we propose an overarching research question: *how can we develop a goal-driven data integration framework for counter-terrorism analytics?* To answer our research question, in the next section, we will provide an anatomy of the existing challenges.

1.3 Challenges and Research Issues

The existing data integration solutions use ontological frameworks to tackle data heterogeneity in various disciplines, such as biomedical (Letunic et al. 2004; Pletscher-Frankild et al. 2015; Smith et al. 2007; Xiang et al. 2011), e-commerce (Fensel et al. 2001; Malucelli et al. 2006), network management (Wong et al. 2005), accounting (Chowdhuri et al. 2014), cloud services (Rodríguez-García et al. 2014), and public safety (Kaza and Chen 2008). The ontological method is preferred in a considerable number of IS studies due to

its ability to (a) annotate the knowledge entities, and (b) map and interoperate these entities. For counter-terrorism analytics, very limited efforts have been made to integrate terrorism data. One such effort is the Terrorism and Extremist Violence in the United States (TEVUS) portal from the Study of Terrorism and Responses to Terrorism (START), which integrates four terrorism-relevant open-source databases: the American Terrorism Study, GTD, U.S. Extremist Crime Database, and Profiles of Perpetrators of Terrorism in the United States. Through the TEVUS portal, users are allowed to submit queries to search terrorism events, perpetrators of incidents, groups, and/or court cases in the United States. It is important to note its contribution and capability to helping decision makers and data analysts to "make informed, data-driven decisions," (TEVUS 2017); accepting search queries; searching over a wealth of data; and returning search results to users who conduct counter-terrorism analytics (TEVUS 2017). However, the data integration solution in TEVUS is of serious limitations. It employs the "in-advance" approach to consolidate predefined data and provides no scalability (START 2017). Additionally, the TEVUS solution does not take into consideration the various users' data analytics goals, which are of critical importance for counter-terrorism analytics and cannot be predefined.

Current data integration methods are mostly driven by data management purposes rather than by the various user data analytics goals (Daraio et al. 2016; Lapatas et al. 2015). In other words, these solutions have not fully considered the fact that, in counter-terrorism analytics, the data integration process is initiated by analysts who have various goals, interests, research emphases, and research scopes; and they fail to model and incorporate these user goals in the data integration designs.

Moreover, existent data integration methods rely overly on data per se to be integrated, ignoring the available free-text information in codebooks or websites that interpret and host terrorism datasets. In particular, the metadata information (e.g., data topics, prior use, or data components) on the webpages or in the codebooks describes the basic principles guiding data collection, data organization, and subsequent data consumption. In other words, the descriptive and structural information in the free-text metadata can enrich the annotation and interpretation of terrorism datasets, thereby enhancing data integration performance.

To address the aforementioned gaps, an urgent need exists to design an intelligent goal-driven data integration solution that elicits users' goals, identifies goal-relevant data, retrieves the data over heterogeneous datasets, and integrates the data by resolving any semantic heterogeneity issues. Additionally, the solution should also have the capability to annotate and leverage free-text metadata of terrorism datasets (e.g., metadata information on webpages) to achieve better data integration performance.

1.4 Research Questions

The overarching and motivating research question of this study is *how can we develop a goal-driven data integration framework for counter-terrorism analytics?* Effectively developing a goal-driven data integration solution for counter-terrorism analytics requires the tackling of challenging tasks, such as eliciting analytical goals;

recognizing and extracting metadata from free text; identifying goal-relevant data; and addressing semantic heterogeneity through mapping the identified relevant data. Motivated to design an intelligent data integration framework for counter-terrorism analytics that is goal-driven and capable of utilizing metadata in free text, in this study we seek to satisfy four major design objectives that are of crucial importance for data integration: (a) metadata extraction and annotation—which extracts and annotates metadata describing the terrorism datasets stored in a data repository, enabling the data integration process to better understand the terrorism datasets and facilitating the search and retrieval of the terrorism data of interest; (b) goal elicitation and annotation—which elicits and annotates a clear vision of user goals for data analytics and ensures that these goals can be well understood and incorporated into the data integration process; (c) goal–dataset matching—which assesses whether a terrorism dataset in the data repository is relevant to a specific analytical goal through calculating the similarities between the goal and the metadata of the dataset, finally identifying all datasets that match the user goals; (d) data mapping (e.g., schema mapping and instance mapping), which consolidates the retrieved goal-relevant datasets, addressing semantic heterogeneity. In developing such a framework, as directed by the four design objectives, we specify seven research questions along with their key aspects and outcomes in Table 1.1. (The alignment of the four design objectives with the seven research questions is shown in Table 1.1.)

Table 1.1 Research Questions

Design Objective	Research Question	Key Design Aspect	Outcome
Metadata extraction and annotation	RQ1: How can we build a metadata catalogue that maintains metadata information of terrorism datasets and is useful for data integration?	Generation of a metadata catalogue for data integration, each item of which is of potential utility regarding data integration	A catalogue of metadata useful for data integration
	RQ2: How can we collect, extract, and manage the metadata of each terrorism dataset for the metadata items identified in the RQ1 catalogue?	Metadata information extraction and annotation, generating metadata ontology for terrorism datasets	A metadata ontology
Goal elicitation and annotation	RQ3: How can we elicit and represent user goals?	Elicitation and annotation of a user's analytical goal	A goal ontology
Goal–dataset matching	RQ4: How can we identify the datasets relevant to the given user goal?	Similarity matrix construction and goal-relevant datasets identification	Goal-relevant datasets recognized from the data repository
Schema mapping and instance mapping	RQ5: How can we extract relevant data from various data sources?	Data extraction and retrieval	Data slices
	RQ6: How can we conduct schema mapping on the datasets from RQ5 if semantic heterogeneity exists between schemas?	Schema mapping to reconcile the semantic heterogeneity between schemas of various datasets	Mapped schemas

	RQ7: How can we map the data instances?	Data instance mapping	Integrated data in response to a user's request

To ensure a comprehensive knowledge representation for a variety of counter-terrorism databases, we propose the first two RQs. To answer RQ1, we will build a metadata catalogue in which each item is of utility for data integration. For RQ2, we will build a metadata ontology, informed by extracted metadata information of terrorism datasets, along with the identified metadata catalogue from RQ1. The metadata ontology will serve as a metadata knowledge base in support of goal-driven data integration, keeping track of terrorism dataset information (e.g., authors, usage, topics, and structure). The information necessary to build up the metadata ontology will be collected and extracted from free text and then populated in our ontology for each terrorism dataset available in a data repository.

Data integration for counter-terrorism analytics is driven by various users' goals. In RQ3 and RQ4, we aim to automatically elicit an input goal and incorporate it into the data integration process. In response to RQ3, we will establish an ontological representation of the goal input by a user. In RQ4, we aim to identify datasets relevant to this goal.

In response to RQ5, RQ6, and RQ7, we will utilize a collection of data retrieval and integration techniques (e.g., data query, data mapping) to integrate the data returned in RQ4. Specifically, in RQ5, we recognize the appropriate datasets and retrieve the relevant data. In RQ6 and RQ7, intelligent agents will be built to conduct schema mapping and data instance mapping to reconcile data heterogeneity.

The answers to the listed questions RQ1–RQ7 will together allow us to establish an intelligent goal-driven data integration for counter-terrorism analytics. The proposed design can automatically generate and return a dataset that is comprehensive and appropriate for users with specific analytical goals.

1.5 Research Significance and Contributions

Because counter-terrorism analytics relies heavily on data-driven techniques, the success of data integration is of crucial importance for the advancement of counter-terrorism research and practice that use data analytics methods (Taipale 2003). For instance, when multiple relevant sources of data are available, integrated data can facilitate knowledge discovery and analysis through concatenating data in separately maintained datasets. Moreover, data integration and annotation are of key importance to improving the productivity of scientific research and data analytics (Lapatas et al. 2015). As posited by Yang and Wu (2006), researchers can build prediction models and discover patterns effectively using appropriately annotated and integrated data; however, the majority of the efforts are in the pre-processing stage, such as data integration. This calls for an intelligent terrorism data integration method through which terrorism researchers can save time when integrating datasets for counter-terrorism analytics.

Although data analysts often need data from various sources in order to extract and retrieve information of interests (e.g., incidents, profiles) for counter-terrorism analytics, it

is challenging to integrate the heterogeneous terrorism data, which are spread across various electronic datasets, each dealing with a specialized aspect of terrorism phenomena (Gordon 2004). It is imperative for IS scholars to make extensive efforts to design automated data integration solutions, facilitating data analytics and scientific research in counter-terrorism.

Existing data integration solutions are inapplicable to terrorism dataset due to their incapability of capturing various user goals of counter-terrorism analytics. Thus, standing on the continuously developing and maturing computational techniques such as semantic web and natural language processing (NLP), this study designs an intelligent data integration solution. Specifically, our proposed design, informed by the goal definition template (van Solingen and Berghout 1999), can computationally elicit and annotate user goals of counter-terrorism analytics. Moreover, our proposed design can address the interoperability issue in the heterogeneous terrorism datasets, thereby providing a customized and unified view of terrorism entities for counter-terrorism analytics. This study also provides a comprehensive evaluation of our proposed solution regarding its performance of goal-relevant data identification (in RQ4), schema mapping (RQ6) and instance mapping (RQ7) in terms of precision, recall, and F-measure.

In addition, this study has three major contributions. First, this study explicitly outlined the requirements needed for next-generation data integration (goal-driven data integration) flows and the research challenges involved in implementing them. In response, we design and instantiate a goal-driven data integration solution. Our design follows the

goal definition method in the Goal-Question-Metric (GQM) approach to model user goals in counter-terrorism analytics; progressively identifies goal-relevant datasets; implements data extraction, transformation, and mapping; and delivers integrated data to users.

Second, our study provides a baseline for developing an ontology-based goal-driven data integration model. We demonstrate that the use of goal ontology and metadata ontology enables users to discover and identify appropriate and goal-relevant datasets for data integration and, accordingly, data analytics. Our proposed approach reduces the semantic gap between the data consumers who perform data analytics and the terrorism datasets collected and managed for data analytics.

Third, our design, following the design science methodology, presents a theoretical framing that can guide scholars and practitioners to build automated goal-driven data integration solutions.

There are also some minor contributions. For instance, we instantiate an automated goal-driven data integration solution for counter-terrorism analytics, illustrating how to elicit and annotate user goals that are of crucial importance for data integration. Additionally, we leverage NLP techniques, along with the semantic web technique, to extract metadata information and establish metadata ontology, demonstrating how to handle and utilize free-text information in metadata to facilitate data integration.

1.6 Organization of Dissertation

The remainder of this paper is organized as follows: Section 2 provides a review of related works and a discussion of the challenges associated with data integration for counter-terrorism analytics and the shortcomings of existing systems. Section 3 elaborates the theoretical background of our design and research methodology, as well as the architecture and major modules of the proposed design artifact, called GoDa, which is a goal-driven data integration solution for counter-terrorism analytics. Section 4 provides the details of the artifact implementation and offers descriptions of a series of experiments to demonstrate the performance of GoDa. In section 5, we conclude our research with a discussion of its contributions, implications and limitations.

Chapter 2. Literature Review

In this section, we present a thematic literature review regarding two interrelated themes—(a) terrorism data and counter-terrorism analytics and (b) data integration solutions in IS, which constitute our research objective in this study. For the first theme, terrorism data and counter-terrorism analytics, we review (1) the academic concept of terrorism, based on which our research context is defined; (2) the concepts of data and metadata and their relationship; and (3) terrorism data for counter-terrorism analytics. Additionally, we provide a list of terrorism datasets that are popularly used for counter-terrorism analytics, as well as the features of these datasets, and illustrate how these datasets are used in counter-terrorism analytics research, implying the importance of data integration in counter-terrorism analytics. For the second theme, we review the relevant academic works that design and implement data integration solutions. This second portion focuses on the (1) central concepts and constructs for data integration; (2) IS designs for data integration; and (3) features of these designs. As a summary, we present a synthesis of research gaps in the literature that explicitly motivate our research into data integration for counter-terrorism analytics.

2.1 Terrorism Data and Counter-terrorism Analytics

2.1.1 Terrorism

Although a variety of definitions of terrorism exist, that of Schmid and Jongman (1988), based on terrorism researchers' responses to a comprehensive questionnaire, gives an inclusive and detailed portrait of terrorism—the features of terrorism actors; the reasons for which actors commit terrorist attacks; and the relationships between actors, targets, and victims. Additionally, this definition highlights the social effects of the terrorist act—the communicative intention of terror—and recognizes "semi-clandestine individual, group, or state actors" as potential perpetrators. According to this definition, "idiosyncratic, criminal, or political reasons" could be the motivating factor for initiating a terrorist action (p.28).

When analyzing terrorist activities, researchers are often interested in the various features of terrorism, such as the weapon, attack target, attack type, facility, terrorism perpetrator, perpetrator ideology, region, and attack motivation. In recent decades, these features have been continually changing, and a growing amount of data has been accumulated for terrorism research, resulting in increased difficulties in observing and analyzing terrorism. Although scholars' understanding and research are evolving, Schmid and Jongman's (1988) definition is still well accepted and used to recognize terrorism, forming the basis upon which scholars collect terrorism data. This research is mainly focused on terrorism data integration—that is to say, the integration of data that describe and record the terrorism or are relevant to counter-terrorism analytics and use Schmid and Jongman's (1988) quoted definition as an operative meaning of "terrorism."

2.1.2 Data for Counter-terrorism Analytics

Since the late 1960s, terrorism research has been gradually gaining attention from academicians and terrorism data are increasingly collected and prepared (Bahgat and Medina 2013). Especially after 9/11, terrorism research and counter-terrorism analytics saw a dramatic rise, requiring not only access to multiple data sources but also the capability to compare and integrate the data in these sources (Ding et al. 2017). It is often observed that the data from various data sources are heterogeneous in both terminology and data structure. This heterogeneity significantly impedes the understanding and integration of the data from various sources in an analytical task.

The datasets that are essential for analyzing terrorism are often separately created and maintained by different research groups for various purposes based on different standards and fashions. For example, the National Consortium for START at the University of Maryland has been continuously collecting data on terrorism events worldwide since 1970 with annual. The research group behind the Project on Violent Conflict (PVC) curates the BAAD database of terrorist organization information, such as organization ideology, organization location, funding, and sponsors. With the assistance of the aforementioned datasets, extant studies have contributed to terrorism risk assessment (Piegorsch et al. 2007), ideology analysis (Asal and Rethemeyer 2008), and pattern recognition for terrorist behavior prediction (Tutun et al. 2017). Specifically, to assess terrorist incidents, previous researchers have extensively used the GTD dataset, which identifies and describes the terrorist incidents in time series. In other instances, BAAD data

have been used to study the ideology or network characteristics of terrorist groups. These datasets typically serve as the basis of counter-terrorism studies.

Data and Metadata

To better understand data integration for counter-terrorism analytics, it is essential to clarify the concept of data and its constructs. What are data? Numerous scholars and practitioners have offered answers encompassing a broad range of definitions in an array of contexts. From the perspective of data generation, data are referred to as "materials generated or collected during the course of conducting research" (NEH 2018). The National Institutes of Health (NIH) defines data as "recorded factual material commonly accepted in the scientific community as necessary to validate research findings" (Arzberger et al. 2004). Data are essential not only to individual researchers but also to research teams and the research community when conducting scientific inquiry and validating research results. Various types of data exist (i.e., documentation, citations, geospatial coordinates, reports, and articles, according to the note from the NEH) that go far beyond quantitative datasets.

In brief, data are organized in various forms, from numeric to textual, from laboratory collections to on-field acquisitions. Researchers and practitioners from different disciplines or backgrounds may perceive and use data in different ways. This research will situate within the key concepts established in the aforementioned definitions from the NEH and NIH. The basic data constructs include not only the content (e.g., data instances) but also the structure (e.g., schemas).

In academic research, datasets can be identified and described via their features. These features (i.e., context, structure, research utility, and heterogeneity) are defined as metadata. In other words, metadata are data that provide information about a specific dataset (Sen 2004). Various types of metadata exist (Baca 2008): descriptive metadata, structural metadata, and administrative metadata (Figure 2.1). Metadata, the data about data, are an essential concept in data management (DM) (Gilliland 2008, p. 1). More specifically, it allows scholars and practitioners to more easily retrieve, manage, and/or use data. In other words, metadata are of critical value for DM and integration because enterprise IS will gain better efficiency to query, identify, annotate, and interpret data with the support of metadata.

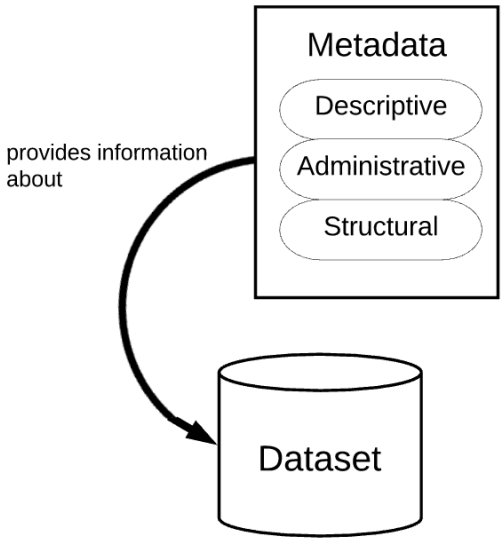


Figure 2.1 Data and Metadata (from the description in NISO (2004))

To use metadata, which are sometimes accessible in free text, the prerequisite step is to collect, structure, and deposit metadata of interest into formal representations (e.g., an

ontological representation) on the basis of industry standards, thereby facilitating the data identification, management, and integration operations in IS implementations. Although a variety of metadata standards exist (Dublin Core, DDI, EML, MINSEQE, FITS, etc.), there are three commonly accepted types of metadata: descriptive metadata, administrative metadata, and structural metadata (NISO 2004). Drawing on the literature and available metadata for terrorism datasets, we provide a list of metadata (Table 2.1).

Table 2.1 Metadata Categories (NISO 2004)

Category	**Definition**	**Metadata Examples**
Descriptive	Metadata items describing data topics, time coverage, usage, and so on	title, description, keyword, topic, related publication, time period covered, geographic coverage, geographic unit, unit of analysis
Administrative	Metadata items essential in administering data collections and resources	dataset persistent id, producer, author, publication date, production date, grant information, distributor, distribution date, date of collection, sampling procedure
Structural	Metadata items indicating how data components (i.e., columns) are organized together	schema, structure format

In the catalogue, descriptive metadata provide descriptive information about terrorism datasets and can be used for data interpretation. Descriptive information consists of the title, author, subjects, keywords, publisher, geographic description, historical usage description, and so on (NISO 2004). Specifically, descriptive metadata on usage provides

information indicating prior usage for data analytics. Published research reports, conference papers, journal articles, and books are a reflective embodiment of data usage. As the amount of usage metadata grows, we would build up knowledge of how data are exploited for data analytics—the topics users engage with and the techniques they use. Administrative metadata refer to metadata information that can be used in administering data collections and resources, including file type, time created, and methods used to create and preserve it (NISO 2004); administrative metadata can be used to facilitate version and access control (NISO 2004). In this paper, we will not put a great deal of effort into such legal and ethical data integration issues as data access control or copyrights. Structural metadata indicate the structure-relevant information (e.g., the components of a dataset and how they are organized). They also describe the relationships between the entities mentioned in the data (Arms et al. 1997). The metadata in the catalogue can be collected, extracted, and stored in ontological format.

This work focuses on how to conduct data integration to address heterogeneity through analyzing and understanding both the datasets of interest and the metadata of these datasets, based on the assumption that both will be heuristic in discovering data usage for analytical tasks.

Terrorism Data

To drive their scientific study of terrorism and counter-terrorism analytics, academicians and practitioners build valuable datasets as the starting point for decision-making (Ding et al. 2017). The knowledge and clues that are essential for identifying and analyzing terrorist activities may be spread across several data sources (Ding et al. 2017).

These datasets are often independently built and maintained by different research groups for distinct research purposes, using various standards and methods. Some terrorism datasets in use are listed in Table 2.2.

Table 2.2 Examples of terrorism datasets for data analytics

Data Source	Contributor	Description	Focus	Period Covered
Global Terrorism Database (GTD)	START	A collection of terrorist incidents around the world	Incidents	1970–2010
Big Allied and Dangerous (BAAD) Database	PVC	A comprehensive database of terrorist organizations	Terrorist group ideology	1998–2005
Minorities at Risk Organizational Behavior (MAROB)	CIDCM	A set of records of "the characteristics of the ethnopolitical organizations most likely to employ violence and terrorism in the pursuit of their perceived grievances with local, national, or international authority structures"	Organizations of 22 ethnopolitical groups	1980–2004
Profiles of Perpetrators of Terrorism in the United States (PPT-US)	START	A collection of detailed profile information on organizations that have conducted terrorist activities on the U.S. homeland	Ideology, perpetrators	1970–2016
Radiological and Nuclear Non-State Adversaries Database (RANNSAD)	START	A collection of "profiles of all former non-state users and attempted users of radiological and nuclear weapons" toward explaining "Who are the most likely radiological or nuclear non-state threat actors?'	Incident, perpetrator, ideology	1974–2008

Note: To keep the descriptions accurate, the information is extracted from the database hosting webpages

2.1.3 Terrorism Data and Counter-terrorism Analytics

Terrorism is a difficult research subject that involves a large variety of actors and activities over time and across the globe (Silke 2004). The characteristics of terrorism have also been changing continuously—e.g., terrorist groups depend less on formal leadership that maintains traditional organizational hierarchy (Ellis 2008). As such, the difficulties of counter-terrorism analytics for researchers and practitioners are constantly multiplying. Meanwhile, accumulated efforts are made for counter-terrorism analytics and knowledge discovery, using a variety of analytical tools for risk assessment (Piegorsch et al. 2007), ideology analysis (Asal and Rethemeyer 2008), and pattern recognition for behavior prediction (Tutun et al. 2017).

In general, counter-terrorism analytics is a process that employs appropriate tools and datasets to derive meaningful insights that improve counter-terrorism performance and decision-making. This process can help assess risks, identify terrorism dynamics, gauge the effectiveness of terrorist motivations, predict perpetrator actions, and so on. Data analytics utilizing historical data may help governments or other agencies outline the way forward, developing new policies or counter-measures (Asal and Rethemeyer 2008).

The essence of counter-terrorism analytics is data analytics using terrorism-relevant data. Data analytics is becomingly increasingly important in academic and other communities (e.g., homeland security and business) over the past two decades, especially in relation to big data. It uses technological means to manage, analyze, visualize, and

extract useful information from diverse and heterogeneous datasets to "increase understanding of human and social processes and interactions" (Chen et al. 2012). In this sense, data analytics is a practice that requires the identification of which data to analyze, clean, and prepare before organizing those data into the right format (Provost and Fawcett 2013). Once the data are transformed into a useful form, a certain algorithmic process can be applied to it for machine learning or other statistical processes to solve given research problems.

IS scholars have designed various information systems to support counter-terrorism analytics. For example, Xu and Zhang (2008) propose a terrorist analysis system that employs data distortion methods and metrics. Elovici et al. (2008) instantiate a content-based detection system of terrorists browsing the web. Oh et al. (2011), leveraging content analysis of Twitter postings, suggest a conceptual framework for Twitter information analysis and control in the context of terrorism. Choi et al. (2014) present a method whereby, through calculating word similarity using WordNet n-gram data frequency, web documents can be classified to reveal their relevance to terrorism. Conlon et al. (2015) propose a terrorism information extraction method for online reports.

A considerable amount of existent data analytics on terrorism access a single data source (Silke 2004). For example, to assess terrorist incidents, prior studies have extensively used GTD that identifies and describes the terrorism incidents in time series. In other instances, BAAD data have been used to study the ideology or network characteristics of terrorist groups. However, in the big data era, countering terrorism

imperatively requires taking stock of the extant knowledge; properly integrating data; and discovering and sharing knowledge, so that we can develop credible recommendations to help manage the terrorism threat (Ellis 2008). Although we acknowledge the contributions of prior studies and the critical importance of data integration for counter-terrorism, we notice that there is a lack of a terrorism data integration system that is intelligent enough to integrate and manage terrorism datasets, assisting both practitioners and researchers to discover and link the pertinent "dots" to generate useful hints or warnings—in other words, an intelligent terrorism data integration system for terrorism analysis.

2.2 Data Integration in IS

2.2.1 Data Integration

Data integration consolidates heterogeneous data from multiple sources for a particular use (Milo and Zohar 1998). In other words, data integration assembles data particles that do not originally fit together due to semantic heterogeneity. Regarding counter-terrorism analytics, we define data integration as the solution that provides the end users with consolidated data from multiple heterogeneous sources based on their analytical goal. Data integration is the process of analyzing the existent data in diverse sources; resolving the syntax and semantic heterogeneity issues; creating new datasets by reorganizing the aligned data toward the user's goal; and sharing these newly generated datasets with the scientific community or counter-terrorism analysts.

In the extant terrorism data, semantic heterogeneity precludes counter-terrorism analysts to join the knowledge that is essential to identify and analyze terrorism entities and activities that may spread across multiple sources. The success of counter-terrorism analytics relies heavily on data-driven techniques. Data integration is crucial for the advancement of counter-terrorism research and practice, especially in terms of knowledge discovery (Taipale 2003). When multiple relevant data sources are available, integrated data can improve knowledge discovery and analysis. Furthermore, data integration and annotation are essential for the reproducibility and interpretation of scientific research and data analytics (Lapatas et al. 2015).

As stated in Yang and Wu (2006), while researchers build prediction models and discover patterns in data analytics, a large amount of efforts is put in the pre-processing stage, such as data integration. In other words, it is pressing to develop an intelligent terrorism data integration system, through which terrorism researchers can save time in manually integrating datasets for counter-terrorism analytics.

2.2.2 Existing Solutions for Data Integration

"On-demand" vs. "In-advance"

Data integration has two modes—namely "eager" and "lazy" (Widom 1996). The difference between the two approaches is the way the data get integrated. In the eager approach, the data are managed in a federate schema and stored in a central data repository, whereas in the lazy mode the data are managed in a manner whereby a federated schema is not required and allows the distribution of data storage (Widom 1996).

Accordingly, technical solutions for data integration, which are being extensively studied and developed by both practitioners and academicians, address data integration based on two approaches—"on-demand" and "in-advance" (Widom 1996). In the "on-demand" approach, the system works in the "lazy" mode. It accepts a user query, determines the appropriate set of data sources that can answer the query, generates subqueries for each involved data source, retrieves relevant data, and performs appropriate data consolidation.

In contrast, in the "in-advance" approach, data integration systems work in an "eager" mode, whereby data of interest are extracted and consolidated in advance and managed in a federate repository (Widom 1996). When a query is received, the in-advance prepared data are returned as a response to the user query. The "in-advance" approach assumes that "data of interest" are predictable.

Therefore, the "on-demand" approach is more applicable when the user's needs are unpredictable (e.g., research goals, entities of interest, research scopes). In counter-terrorism analytics, researchers may have various goals. For example, in a project aimed at analyzing the patterns in Al Qaeda's actions toward the United States, analysts are only interested in "incidents [that] happened in the U.S. which is relevant with Al Qaeda." In comparison, when analyzing the ideology of right-wing organizations, researchers will be interested in all the incidents that are motivated by right-wing ideologies and beliefs, including those arising from Islamophobia, anti-communism, and so on. In such a situation, the success of data integration will not only rely on the technical implementation

but also be driven by the actual users, who may have various goals, interests, research emphases, and research scopes.

Ontology-based Data Integration—Single vs. Multiple

Ontologies, as formal representations of explicitly defined concepts and the relationships between these concepts, can serve to tackle semantic heterogeneity in heterogeneous data sources (Uschold and Gruninger 1996).

Specifically, ontology representation enables the accurate interpretation of data through an explicit definition. It facilitates queries about entities—i.e. through ontology-mediated query formulation (Wang et al. 2009). The data-to-ontology mapping is an essential step for integrating data from multiple sources.

Ontology-based data integration approaches can be categorized into three types (Table 2.3): single ontology approach, multiple-ontology approach, and hybrid approach (Wache et al. 2001). The single ontology approach employs a global ontology as a shared reference describing concepts and/or instances in all data sources (Wache et al. 2001). In some contexts, it may be difficult to develop a consistent and comprehensive global ontology wherein data spreads from disparate sources, especially in cases where data sources suffer from frequent change. A multiple-ontology approach may use one ontology per dataset (Wache et al. 2001). Such an approach allows for the use of flexible and changing data sources. The hybrid approach may incorporate both the global and dataset ontologies (Wache et al. 2001). This approach can also be considered a special case of the

multiple-ontology approach. In other words, the multiontology approach and the hybrid approach can work only when the individual dataset ontologies are available.

Table 2.3 Ontology-based data integration methods

Ontology-based Data Integration Variants	**Layers**			**Weakness**
	Local Ontology Layer	**Global Ontology Layer**	**Application Layer**	
Single-ontology		Relies on a single global ontology to integrate all data sources	Communicates with global ontology layer	Vulnerability to data change
Multiple-ontology	A local ontology per dataset and mapping between the local ontologies		Communicates with local ontology layer	Mapping complexity
Hybrid	A local ontology per dataset	A global vocabulary that maps to the local ontologies	Communicates with global ontology layer (global ontology or a global vocabulary)	Predefined vocabulary

As a synthesis, there are four essential layers in ontology-driven data integration systems: local data layer, local ontology layer, global ontology layer, and application layer. The heterogenous local data repositories need to be integrated. The local ontology layer contains so-called “local ontologies,” which are ontological storages of the individual data

repositories. The global ontology layer contains a "global ontology" or a shared vocabulary, which provides semantic sufficiency for integrated data. Finally, the application layer represents the applications that build upon the data integration. In a data integration solution, not all the four layers necessarily appear together. For example, in a "single ontology" solution, the local ontology repository layer may be absent.

In the case of counter-terrorism analytics, a considerable volume of terrorism datasets exists. Meantime, it is hard to design the ontological representation for each dataset. Therefore, the multi-ontology approach does not fit the data integration for counter-terrorism analytics; designing a single-indexing ontology will be helpful for annotating various terrorism datasets and their metadata.

IS Designs of Data Integration

IS scholars have made great efforts to develop various ontological data integration methods. Table 2.4 synthesizes some representative efforts regarding data integration application areas and approaches.

Table 2.4 Ontology-based data integration IS exemplars

Source	**System Description**	**DM Oriented**	**User-Centered**	**Application Area**	**Data Mapping Approach**	**Global Ontology**	**Metadata and NLP support**
Kaza and Chen (2008)	Data integration solution that uses WordNet and mutual information	X		public safety data sources	data instances and secondary sources	Not needed	Mutual information calculation

	between data instances to map ontologies.				(WordNet)		for data content
Lee et al. (2011)	A knowledge exchange framework that integrates semantically heterogeneous learning objects	X		learning objects	WordNet	X	NLP-supported data content analysis (lower casing, stemming, stop-word removing, etc.)
Arch-int and Arch-int (2013)	A learning resource integration system leveraging conflict detection and resolution techniques addressing both semantic and structural conflicts; the semantic bridge ontology has been employed as a core component for generating mapping rules	X		learning resource	WordNet	X	Rule-based mapping for data content
Nedersti gt et al. (2014)	FLOPPIES, a framework capable of semiautomatic ontology population of tabular product information from Web stores	X		product, e-commerce	product numeric values and textual values	X	Use manually annotated metadata

Chowdhuri et al. (2014)	XBRL filings, Integration solution leveraging, XBRL-concept mapping; XBRLOnt serves to address semantic heterogeneities between different XBRL filings		X	accounting for XBRL filing	financial concepts extracted from XBRL reports	X	NLP-supported content analysis
Etudo et al. (2017)	XBRL filings integration solution leveraging, XBRL-concept mapping; especially, the XBRL Calculation Linkbases are used to improve mapping ability		X	accounting for XBRL filing	XBRL Calculation Linkbases	X	Use metadata in calculation links, NLP-supported data content comparison
Hawalah and Fasli (2015)	User profile and interest integration solution for Web personalization; a reference ontology is used		X (user interest)	web personalization	User interest as a reference ontology; web personalization mapping through user interest	X	NLP-supported information retrieval at word level
Kouskouridas et al. (2015)	A design that can integrate the industry objects and demanding tasks of	X		industry object	mapping between visual stimuli and	X	not needed for the industry object

	autonomous object manipulations				motor commands		
Ebrahimipour et al. (2010)	An ontology solution that can overcome the limitations of text-based descriptions, leveraging FMEA ontology for data integration for failure modes and effects analysis			project development	a global ontology based on ISO-15926	X	NLP-supported, no metadata used
Santipantakis et al. (2017)	A general data integration solution for heterogeneous sources; an ontology-based distributed framework for accessing, integrating, and reasoning with disparate data sources	X		general purpose	Content and schema	X	Not mentioned or emphasized
This paper	Ontology-driven data integration solution for counter-terrorism analytics that can catch and model user goals and leverage metadata to enhance data integration capability		X (user goal)	Terrorism		X	A combination of NLP techniques for metadata extraction, retrieval, and manipulation

In the existing ontology-based data integration solutions, scholars have frequently adopted a data-management-oriented approach and followed the "in-advance" mode (Kaza and Chen 2008; Lee et al. 2011; Nederstigt et al. 2014; Santipantakis et al. 2017). However, there is a growing interest in developing user-focused designs (Etudo et al. 2017; Hawalah and Fasli 2015) that put more efforts on the human-machine interfaces. This may stem from a well-documented belief that is influencing IS design science researchers: "research in the information systems field examines more than just the technological system, or just the social system, or even the two side by side; in addition, it investigates the phenomena that emerge when the two interact" (Lee 2001, p. iii). IS designs are increasingly giving respect to the interactions and interfaces between humans and information technologies and between users and designed artifacts.

This study notices that scholars in the terrorism research domain have also made efforts to integrate terrorism data in relation to open data that may be stored with heterogeneity. START has built up the TEVUS Portal, which integrates four related open-source databases. Through the portal, users submit search queries in four data types—namely specific events, perpetrators of an incident, groups, and/or court cases in the United States. This portal enables practitioners to "make informed, data-driven decisions" (TEVUS 2017). Being a data integration solution, TEVUS employs the "in-advance" approach and combines data "related to terrorist incidents, pre-incident activities, and extremist crimes in the United States" (START 2017). Consequently, it has very limited

scalability and does not take into consideration various user goals for counter-terrorism analytics.

Summary

The "in-advance" data integration approach has roused a large volume of studies which have developed design artifacts to integrate data for a variety of domains. In these solutions, NLP techniques have served as the fundamental construct for ontology-based data integration methods, assisting data content analysis. However, we have seen limited efforts to incorporate metadata to enhance data integration solutions, whereas the NLP-supported metadata retrieval and extraction can largely support data integration, facilitating data understanding and annotation.

Further, we discern a lack of research on data integration in relation to counter-terrorism analytics. The application of existing data-management-oriented data integration solutions to counter-terrorism analytics, as characterized by diverse users' data analysis goals, is problematic. One reason, as identified in prior sections, is that the "on-demand" approach is more appropriate because user goals of data analytics are unpredictable and cannot be predefined. Effective data integration for counter-terrorism analytics requires appropriately modeling these goals. Another reason is that rich metadata information exists in free-text format on the websites that host terrorism datasets, which is valuable in interpreting these terrorism datasets and can be incorporated into the data integration process. Therefore, we argue that existing data integration solutions are inapplicable to the case of counter-terrorism analytics, because these solutions, by using the "in-advance"

approach, assume that “data of interest” are predictable and lack the capability of identifying and modeling user goals.

To address the aforementioned limitations of existing solutions, we will develop an innovative design artifact that will integrate terrorism datasets based on users’ analytical goals. Our solution is capable of eliciting, modeling, and annotating user goals and extracting and analyzing the metadata (in free text) of terrorism datasets, thereby implementing an automated goal-driven data integration process.

Chapter 3. Research Method

This chapter elaborates on a set of design objectives for an IS artifact addressing goal-driven data integration for counter-terrorism analytics. Subject to these objectives, this chapter proposes an innovative design informed by design science methodology. The detailed structure of this chapter is as follows: First, this chapter lists the design objectives from which the research questions derive. Second, this chapter introduces the design science methodological paradigm into which this work fits and expounds on the information systems design theory (ISDT) components that frame our design process and artifact. Third, this chapter presents the kernel theory that informs the design and its alignment with this research's objectives. Finally, drawing on the informing theory and design objectives, this chapter presents and explains the resultant design.

3.1 Research Objectives

The overarching objective of this research is to develop a goal-driven data integration framework for counter-terrorism analytics. This objective is motivated by the well-documented belief that an appropriate consolidation of terrorism datasets can facilitate counter-terrorism analytics. In relation to data analytics, data integration is driven by a variety of user goals that should be elicited and modeled in the process of data integration. Stemming from this objective and addressing the limitations of existent data

integration solutions, four design objectives have emerged for our research, and we accordingly propose seven specific research questions (also shown in chapter 1):

Objective 1: As existing solutions provide inefficient metadata exploitation, especially the ineffective use of collective information in natural languages (e.g., dataset descriptions on hosting-websites), our design seeks to crawl and extract the useful free-text metadata and annotate it in ontological representation which serve as a knowledge base of terrorism datasets. To implement this design objective, this study proposes to address two research questions as follows:

RQ1: How can we build a metadata catalogue that maintains metadata information of terrorism datasets and is useful for data integration?

RQ2: How can we collect, extract, and manage the metadata of each terrorism dataset for the metadata items identified in the RQ1 catalogue?

RQ1 and RQ2 aim to build up a comprehensive knowledge representation for counter-terrorism datasets. To answer RQ1, this study builds a metadata catalogue in an ontological format, in which each metadata item is of utility for data integration. For RQ2, this study, for each terrorism dataset, collects and extracts the metadata information corresponding to the listed items in the RQ1 catalogue. Afterwards, the metadata information instances are populated in the metadata ontology, which serves as a metadata knowledge base supporting goal-driven data integration. Specifically, the metadata information, which constitutes the proposed metadata ontology, is automatically collected

and extracted from free text through NLP techniques and populated in the ontology to keep track of descriptive, structural and administrative information (e.g., data usage, author, keywords, and subjects) of critical importance for data integration. Using the output of this phase, we construct a complete metadata ontology—*MeDaOnto*—with both concepts and instances, a knowledge representation of the metadata of terrorism datasets, which will operationally support goal-relevant data recognition later on.

Objective 2: Because users have diverse, dynamic, and unpredictable analytical goals, this design seeks to extract and ontologize user goals and incorporate them into the data integration process to identify goal-relevant datasets. To implement this design objective, we propose to address the following research question:

RQ3: How can we elicit and represent user goals?

As the data integration for counter-terrorism analytics is driven by various user goals, RQ3 aims to automatically elicit user goals and represent these goals in a structured format. These elicited and annotated user goals drive the data integration process, assisting our proposed design in recognizing terrorism data of relevance and users' specific interests. This output at this phase will construct a goal ontology—*GoOnto*.

Objective 3: Because returning an integration of all datasets may overload system users, this design seeks to assess whether a terrorism dataset in the data repository is relevant to a specific analytical goal input by a user, through calculating the similarities between a user's goal and the metadata of a dataset, identifying datasets that matches a

specific user's goal. To implement this design objective, this study proposes to address the following research question:

RQ4: How can we identify the datasets relevant to the given user goal?

In RQ4, this study seeks to identify datasets that are relevant to the user's analytical goal, utilizing *GoOnto* and *MetaOnto*. Given the user goal annotated in *GoOnto*, we will use the metadata information in *MetaOnto* to calculate the relevance of terrorism datasets, generating relevance indices to help recognize goal-relevant datasets. These goal-relevant datasets to be integrated are output at this phase.

Object 4: Semantic heterogeneity exists in both the data schemas and instances of terrorism datasets. Our design seeks to retrieve the previously identified goal-relevant datasets and address data heterogeneity through data schema mapping and instance mapping.

RQ5: How can we extract relevant data from various data sources?

RQ6: How can we conduct schema mapping on the datasets from RQ5 if semantic heterogeneity exists between schemas?

RQ7: How can we map the data instances?

In response to RQ5, RQ6 and RQ7, we use a collection of data retrieval and data integration techniques to integrate the datasets output from the answers of RQ4. Specifically, in RQ5, we recognize the relevant datasets and retrieve data segments of user's interests from the terrorism data repository. To address RQ6 and RQ7, intelligent

agents will be built to conduct schema mapping and data instance mapping to reconcile data heterogeneity.

The answers for RQ1-RQ7 together establish a solution for an intelligent goal-driven data integration framework for counter-terrorism analytics. The proposed design artifact provides a seamless continuum that will generate a dataset sufficiently comprehensive and appropriate for users with specific goal of data analytics.

3.2 Design Science Research Methodology

Simon (1996) equates design science to the science of the artificial which is "a body of knowledge about artificial (man-made) objects and phenomena designed to meet certain desired goals." Hevner et al. (2004) articulate the guidelines for design science in information systems research. *Design* can be deemed as both a *product* and a *process* (Abbasi et al. 2010; Hevner et al. 2004). Specifically, a design process consists of the steps and procedures taken to develop an artifact. The design process starts with an awareness of a problem to which a suggested solution is to be found and applied (Takeda et al. 1990). A suggested solution is then proposed, developed, and evaluated. The notion of *product* is essential within the design science research paradigm, where the artifact is the resultant deliverable of the design process and an embodiment of justificatory knowledge relevant to research problems (Gregor and Hevner 2013). In our study, the design of a goal-driven

data integration framework can be viewed through both the design artifact and design process.

The major objective of this dissertation is to develop a goal-driven data integration framework for counter-terrorism analytics. Our work follows the design science methodology (Gregor and Hevner 2013; Gregor and Jones 2007; Hevner et al. 2004). Specifically, we develop our design artifact and evaluate its utility highlighting the following eight imperative components in regards to an Information Systems Design Theory (ISDT): (1) purpose and scope, (2) constructs, (3) principles of form and function, (4) artifact mutability, (5) testable propositions, (6) justificatory knowledge, (7) principles of implementation, and (8) expository instantiation (Gregor and Jones 2007). These eight components govern the design process of capturing, articulating, justifying, and communicating our design knowledge. Table 3.1 presents a summary of the eight components of the ISDT in relation to our study.

Table 3.1 Eight Information Systems Design Theory Components

Core Components	**Description**
Purpose and scope	In the context of counter-terrorism analytics, users need to appropriately consolidate terrorism datasets for various tasks of data analytics (e.g., to predict terrorist behavior, to analyze terrorist ideology). Although user goals in these data analytics tasks are of critical importance in determining which data should be integrated for use, existent data integration designs are incapable of capturing users' various goals. This study aims to design an innovative solution—a goal-driven data (GoDa) integration framework for counter-terrorism analytics.

Constructs	The underlying constructs in GoDa are the user's goals of data analytics that drives the data integration process and the metadata information that assists in describing, administering and structuring the terrorism data. Moreover, terrorism data (e.g., schema and instance) per se is the basic construct and bearer of the data integration process.
Principle of form and function	The principles of the goal definition template by the GQM approach guides goal identification and conceptualization in our design. We have incorporated an authoritative metadata catalog in our design to help annotate terrorism data.
Artifact mutability	The proposed artifact design is scalable. Although we build our artifact with the 10 extensively used terrorism datasets, this artifact can expand its utility by incorporating additional datasets and metadata information. Further, the artifact is capable of adjusting to the dynamic environment within which various users' analytical goals exist.
Testable propositions	Drawing on the literature and the goal definition template, a set of testable propositions are proposed.
Justificatory knowledge	It is shown how the goal-driven data integration works by referencing the existing literature and the underlying goal definition approach.
Principles of implementation	Guidelines are given on how to produce a goal-driven data integration solution regarding goal extraction and conceptualization, metadata annotation, concept inference, as well as schema and instance mapping.
Expository instantiation	The illustration of a working instantiation is provided.

The first component articulates the purpose of the designed artifact (Gregor and Jones 2007). We focus on designing an artifact that helps to consolidate datasets for various goals of data analytics (e.g., predicting terrorist behavior and analyzing terrorism patterns) in the context of counter-terrorism analytics. Our review of the literature suggests that existent data integration solutions cannot effectively capture users' various goals,

which inform the data integration process. Therefore, this study aims to design an innovative solution—a goal-driven data integration framework for counter-terrorism analytics—that addresses the semantic heterogeneity in terrorism data for counter-terrorism analytics.

The second component focuses on constructs that represent the entities of interest in the design artifact (Gregor and Jones 2007). This study turns to both the research problem and the literature to better understand the constructs relevant to our designed artifact—a goal-driven data integration framework for counter-terrorism analytics. The review of the existent solutions in the literature helps to identify the critical concepts that both researchers and practitioners wanted to better understand in the context of counter-terrorism analytics. In particular, the underlying constructs in GoDa are users' goals of data analytics, which drive the process of data integration, and the metadata information of terrorism datasets that assists in describing, administering and structuring the terrorism data. Additionally, terrorism data per se is a basic construct and bearer of the data integration process.

The third component concentrates on the architecture that explains the artifact under development (Venkatesh et al. 2017). This component outlines the underlying principles that describe the IS construction (Venkatesh et al. 2017). Our design follows the goal definition principles from the GQM approach to guide goal identification and elicitation (Basili et al. 1994; van Solingen et al. 2002; van Solingen and Berghout 1999).

We have incorporated an authoritative metadata catalog into our design to help annotate terrorism data efficiently.

The fourth component focuses on the dynamic of the phenomenon under study (Gregor and Jones 2007). Users' goals of counter-terrorism analytics are inherently complex; terrorism databases and associated metadata are dynamically increasing as terrorism data collection continues. Although our artifact is instantiated with the 10 extensively used terrorism datasets, it can expand its utility by incorporating additional datasets and metadata information. Thus, we believe our proposed artifact is potentially mutable and has the capability of adjusting to the dynamic environment within which various analytical goals and terrorism datasets exist. Therefore, it can be integrated into practical operations and adapted to a variety of data analytics contexts.

The fifth component focuses on propositions/hypotheses development, which will assist design science researchers in identifying the key results relevant to the proposed artifacts (Gregor and Jones 2007). Drawn upon the existent literature and the GQM approach, we propose testable propositions. The evaluative results testing these propositions will be discussed in the evaluation section.

The sixth component "justificatory knowledge" relates to kernel theory (Gregor and Jones 2007). As Kuechler and Vaishnavi (2008) posit, kernel theories often advise design solutions, bringing in "possibility of refinement or development." Design science methodology requires that researchers demonstrate how and why the designed artifact is derived from existent justificatory knowledge and lead to desired outcomes (Gregor and

Hevner 2013). Informed by the real need of counter-terrorism analysts, this study explains how the goal-driven data integration works, referencing the literature and underlying GQM approach.

The seventh component focuses on the guidelines given to direct the artifact implementation (Gregor and Jones 2007). Our study illustrates how to produce a goal-driven data integration solution that incorporates goal extraction and conceptualization, metadata annotation, goal-dataset matching and schema and instance mapping.

The eighth component focuses on the instantiation of the proposed design artifact. The major purpose of design instantiation is to build an instance of our proposed design, based on which the artifact can be evaluated (e.g., in terms of efficacy and efficiency). The specifications of design instantiation will be provided in section 3.5.

Major design science contributions include design principles and a demonstration of the feasibility of those principles (Hevner et al. 2004). In this research, we will propose a framework and a set of design guidelines advocating the development of our goal-driven data integration system (GoDa) for counter-terrorism analytics. Both these guidelines and the resultant IT artifact constitute research contributions.

3.3 Theoretical Background

An appropriate integration of terrorism datasets would enhance counter-terrorism analytics capabilities, such as terrorist behavior prediction and pattern analysis (Behlendorf and LaFree 2014). However, an oversupply of data, integrating every piece of terrorism data available, would simply overload users and complicate the decision-making process. Users would appreciate customized datasets that include data relevant to their goals of counter-terrorism analytics. To effectively assist users in analyzing terrorism data, data integration in relation to data analytics should take users' analytical goals into account. In other words, counter-terrorism analytics tasks are diverse, thereby precluding a versatile dataset that data integration scientists can integrate and predefine; accordingly, a goal-driven data integration approach is necessary for data analytics.

Goals can be defined for "any object, for a variety of reasons, with respect to various models of quality, from various points of view and relative to a particular environment"(Basili et al. 2014, p. 39). One extensively used goal definition method comes from the GQM approach developed by Basili et al. (Basili et al. 1994; Mashiko and Basili 1997). GQM is based upon the belief that engineering projects must be constructed based upon an explicitly stated goal. This approach has been frequently used for characterizing, categorizing, decomposing, and structuring goals and related measures (Li et al. 2017). Using the GQM approach, a goal can be defined in two phases—defining questions that need to be answered to interpret the goal and then defining measures that answer the questions. We employ the goal definition template from van Solingen and

Berghout (1999) into our paper (see Table 3.2), clearly articulating the five components essential for defining a goal.

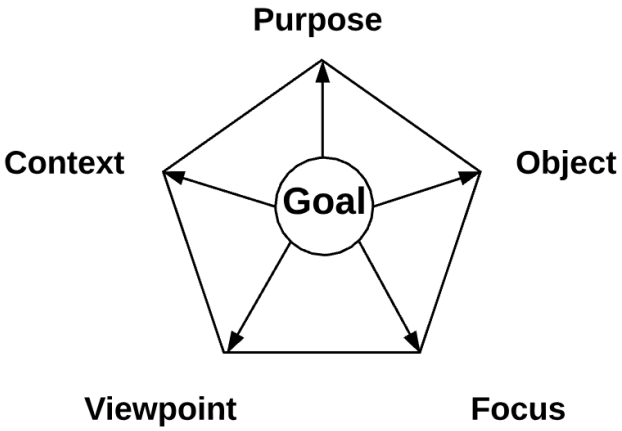


Figure 3.1 Five Components of a Goal

To effectively capture and model user's goals of data analytics, this study draws on the goal definition template from the GQM approach (Basili et al. 1994; van Solingen et al. 2002; van Solingen and Berghout 1999) for user goal elicitation and modeling. Specifically, in our design, the goal definition template is adapted in the context of counter-terrorism analytics. This research adopts the goal definition template on the following propositions:

Proposition 1: The identified components of a user goal of counter-terrorism analytics enable our design to identify the goal-relevant datasets needed for data integration.

Proposition 2: Using a complete set of goal components will more effectively assist in identifying goal-relevant datasets than only using a part of it.

Table 3.2 Goal Components in Relation to Counter-terrorism Analytics, Adapted from van Solingen and Berghout (1999)

Goal Component	Description
Purpose	Motivation of counter-terrorism analytics
Object	Entity of interests in counter-terrorism analytics
Focus	A particular angle or set of angles of the entities for data analytics
Viewpoint	From whose perspective the counter-terrorism analytics is conducted
Context	Scope of counter-terrorism analytics

Informed by the goal definition template, the goal formulation for counter-terrorism analytics consists of information in the following five dimensions (see Figure 3.1): (1) purpose (motivation behind counter-terrorism analytics), (2) object (entity under data analytics), (3) focus (a particular angle or a set of angles of the entities for data analytics), (4) viewpoint (from whose perspective the counter-terrorism analytics is conducted), and (5) context (scope of counter-terrorism analytics). These five important goal dimensions reflect a series of user interests in counter-terrorism analytics, and assist in customizing the data integration process to retrieve and consolidate data relevant to a specific user goal.

This study will elicit user goals in terms of each of the five components listed in Table 3.2 and represent them in the ontology *GoOnto*.

In summary, the goal definition template will help us to model user goals of data analytics and to retrieve goal-relevant data. Our design—a goal-driven data integration—underpinned by the template enables users to input their goal descriptions and statements and returns goal-relevant data as data integration output.

3.4 Goal-driven Data Integration Framework

Developing a goal-driven data integration solution for counter-terrorism analytics requires tackling challenging tasks, such as eliciting users' goals, identifying and extracting metadata from free text, recognizing goal-relevant data, and addressing semantic heterogeneity in the recognized goal-relevant datasets through schema mapping and instance mapping. Motivated to design such an intelligent goal-driven data integration framework for counter-terrorism analytics and implement it for instantiation, this study highlights four major design objectives that constitute the necessary and essential tasks in our data integration design. These four design objectives are (1) metadata extraction and annotation—which extracts and annotates metadata describing terrorism datasets, allowing the data integration process to incorporate the metadata information of terrorism datasets, which facilitates data understanding, search and retrieval; (2) goal elicitation and annotation—which elicits and annotates a clear vision of user goals of data analytics and

enables these goals to be taken into account, well-structured and incorporated into data integration process; (3) goal-dataset matching—which assesses whether a terrorism dataset in the data repository is relevant to a specific analytical goal, calculating the distance between the user goal and metadata to identify the terrorism datasets that match the given goal; and (4) schema and instance mappings—which retrieve the relevant datasets and consolidates them, addressing semantic heterogeneity in the goal-relevant data (both schema heterogeneity and instance heterogeneity). The alignment of the four design objectives with the seven research questions are shown in Table 1.1. Figure 3.2 lists the detailed functions that implement each design objective, concreting our goal-driven data integration framework (GoDa) for counter-terrorism analytics.

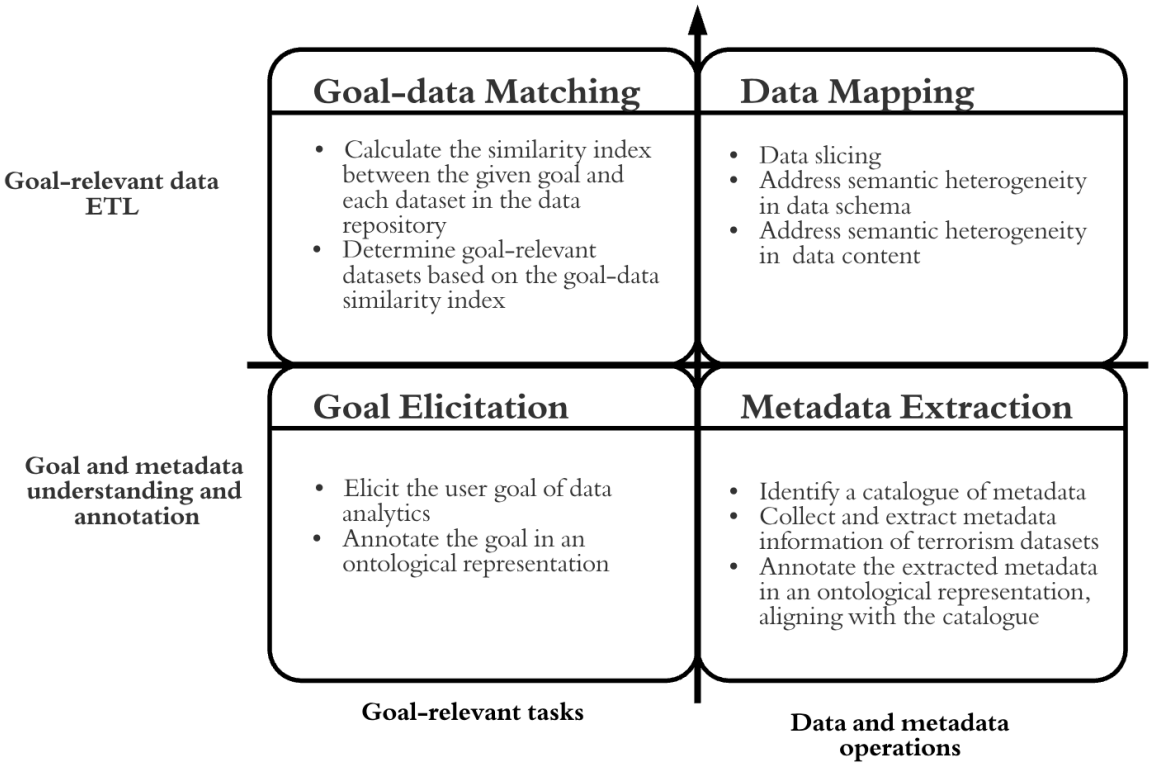


Figure 3.2 Design Objectives of Goal-driven Data Integration Framework

3.4.1 Architecture

Based on our proposed framework, we design and construct a system architecture that consists of four major functional modules, as is shown in Figure 3.3—namely, the Metadata Collection and Management Module (MetaMM), the Goal Extractor (GoalEx), the Goal-relevant Dataset Identification Module (DIM), and the Schema-mapping and Instance-mapping Module (SIMappingM). The MetaMM collects and extracts metadata information from free text and annotates the metadata information of terrorism datasets into *MeDaOnto*—an ontology of terrorism datasets. The GoalEx extracts and represents user goals of data analytics in an ontology called *GoOnto*. The DIM calculates the match index between a goal and terrorism datasets in the data repository and recognizes the goal-relevant datasets to be integrated. The SIMappingM retrieves the goal-relevant datasets identified in the prior step which matches the elicited goal and consolidates them through schema and instance mappings. These four components work together seamlessly toward implementing our four design objectives.

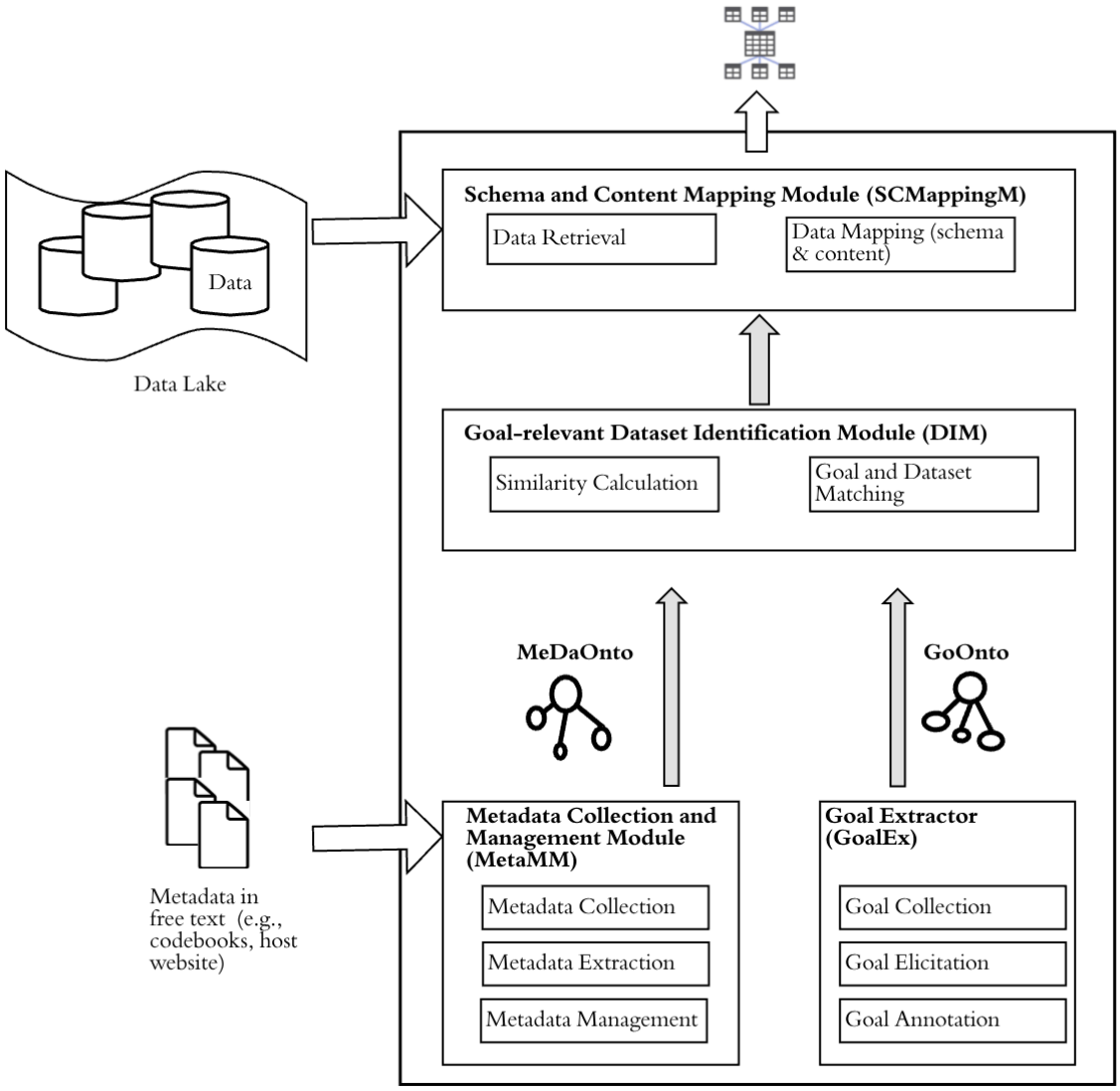


Figure 3.3 GoDa System Architecture

3.4.2 Metadata Collection and Management

MetaMM aims to locate, extract and manage the metadata information of terrorism datasets. MetaMM crawls the metadata available in free text on dataset hosting website (Figure 3.4). The metadata crawled provides descriptive, administrative, and structural

information of terrorism datasets, enabling dataset users to understand these digital datasets and to query and retrieve data of interest. Therefore, after crawling and extracting the metadata, this study annotates and manages the metadata in the *MeDaOnto*, an ontological representation of metadata information.

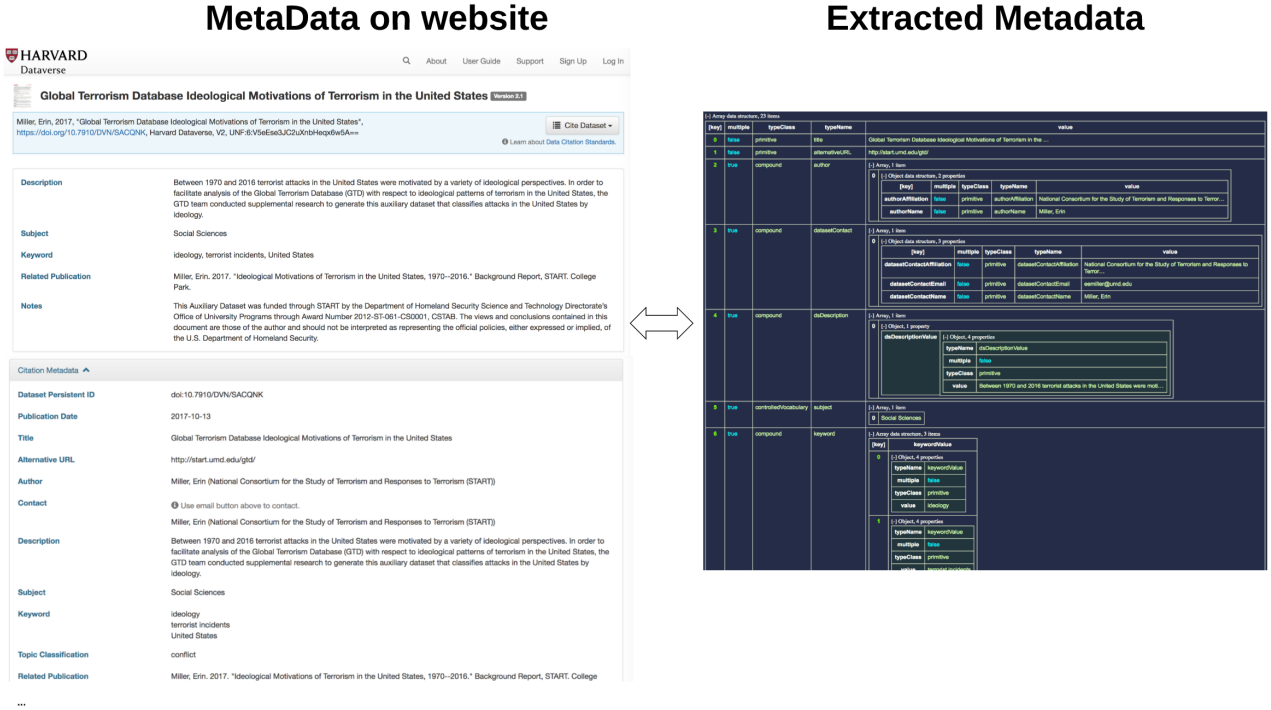


Figure 3.4 Metadata Crawler

Although there exists a variety of metadata organization standards (e.g., Dublin Core, DDI, EML, etc.), there are three types of commonly accepted metadata: descriptive metadata, administrative metadata and structural metadata (NISO 2004). Accommodating the reachable metadata of terrorism datasets from their hosting websites, we structure their metadata in the ontological representation. Ontology is a formal representation of a body of domain knowledge, representing the entities, the properties of entities, and the

relationships linking these entities. Ontology is proven to be efficient for knowledge annotation, sharing and reuse.

The metadata ontology *MeDaOnto* annotates and shares a body of metadata knowledge of terrorism datasets, structuring the descriptive, administrative and structural metadata information in the form of ontological representation. In the ontological catalogue, there are three metadata entities: *DescriptiveMetaData, AdministrativeMetaData, and StructuralMetaData*. *DescriptiveMetaData* provides the descriptive information about terrorism datasets and can be used for terrorism data understanding, e.g., the topics, prior usage and keywords (NISO 2004). Specifically, our design includes, in *DescriptiveMetaData*, a comprehensive list of descriptive information retrieved from the dataset hosting webpages, such as the title, alternative title, keywords, publications in which the dataset is used, period covered, and unit of analysis. *AdministrativeMetaData* annotates and manages metadata information that is mainly used to manage and administrate data collections and resources, including the authors, date of collection, date of deposit, and other metadata about methods used to create and preserve it. *StructuralMetaData* includes the metadata indicating structure-relevant information (e.g., the components of a dataset and how these components are organized).

Our design automates the ontology generation process, utilizing NLP techniques to collect, analyze, extract and transform the metadata information into ontological representation—*MeDaOnto* (Figure 3.6). The extracted metadata information of terrorism

datasets is shown in Figure 3.5.

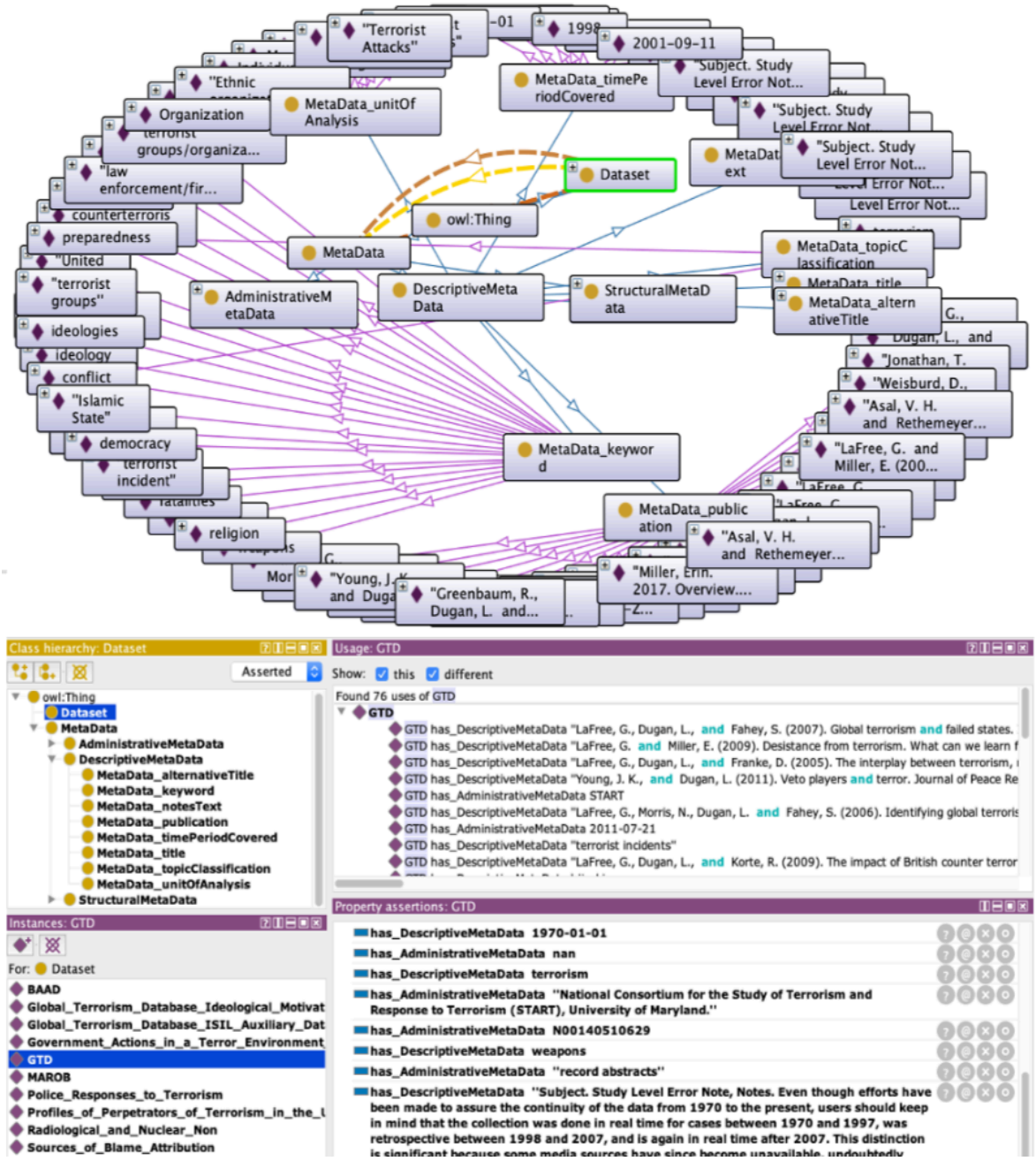


Figure 3.5 Extracted Information in MeDaOnto

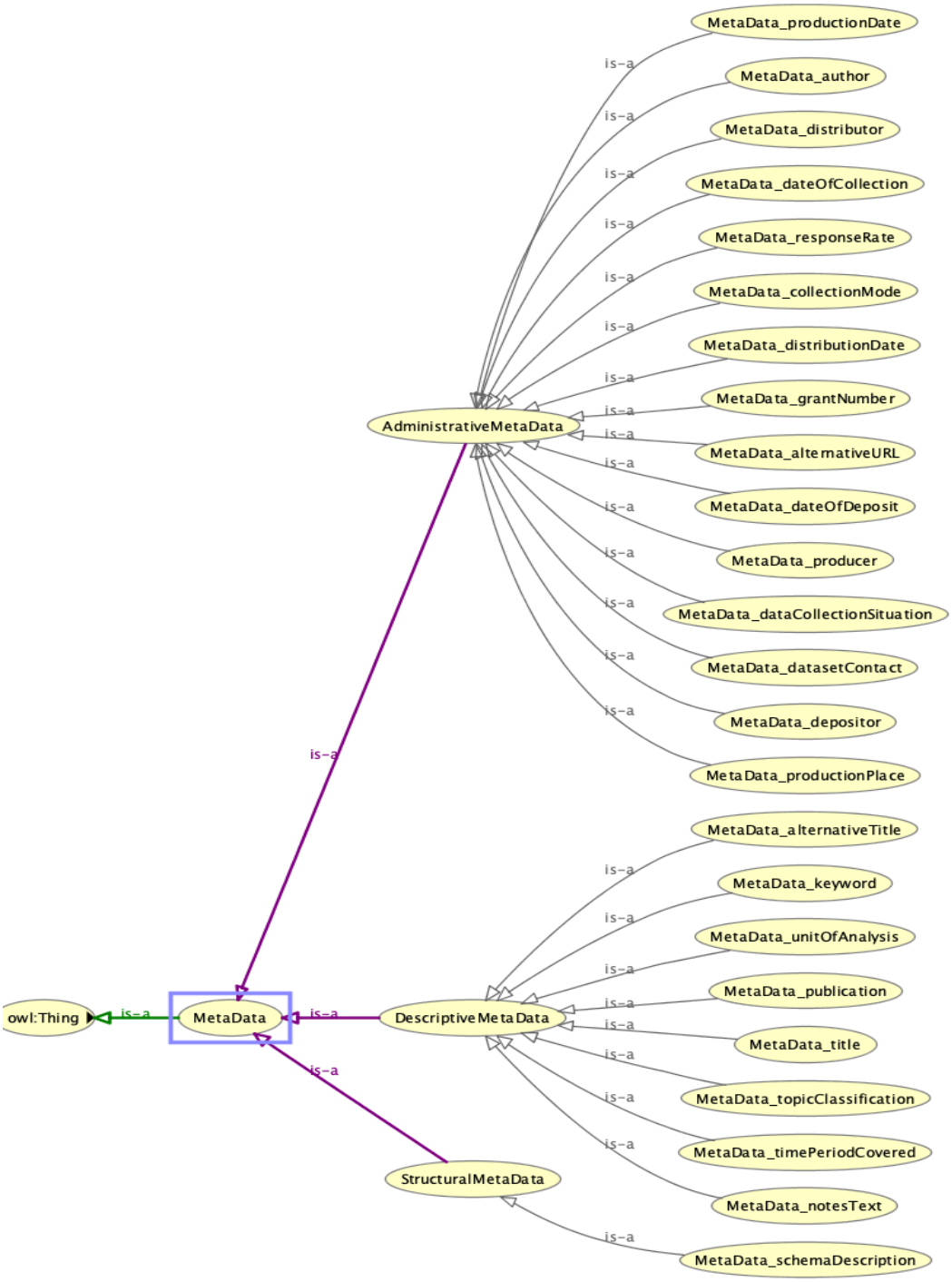


Figure 3.6 MeDaOnto Entities

3.4.3 Goal Extraction

GoalEx seeks to elicit and frame user goals of counter-terrorism analytics. The intended users are data analysts or terrorism researchers who are involved in counter-terrorism analytics or those who participate in terrorism research projects with specific analytical goals. These goals drive the subsequent process of data analytics, which determines terrorism datasets to use if some applicable terrorism datasets exist. Informed by the domain-specific ontology design guidelines (Uschold and Gruninger 1996), this study identifies two purposes that the goal ontology (*GoOnto*) should serve. First, it aims to build an ontological representation of users' analytical goals based on users' statements. Second, it is designed to calculate the relevance of terrorism datasets in relation to counter-terrorism analytical goals.

Table 3.3 Question Template for Goal Elicitation

Goal Component	**Dimension of Interest**	**Question Item**
Purpose	Purpose of analytical task	Which of the following most closely matches your current data analytics task? (select all that apply)
	Self-described purpose	Please describe the purpose of your current data analytics task.
Object	Objects in metadata	Which of the following entities most closely describe your research interest or task? (select all that apply)
	Supplementary objects from user self-description	Please describe other entities that are of interest to you or are relevant to your research task.

Focus	Focus specification	There will be a set of questions (a question for each entity selected by a user in the Object Section) asking about the user's focus.
View of point	Historical research	Please list studies that you think are relevant to your counter-terrorism analytical task (you can submit the list as an attachment).
Context	Time	Time span: Your counter-terrorism analytics will use the data that spreads from ____ to ____. (If not applicable, please leave it blank)
	Geographical coverage	Geographical coverage: Your counter-terrorism analytics will cover the below territories (e.g., country, cities, states/provinces, and others) ____. (If not applicable, please leave it blank)

Informed by the goal definition method from the GQM approach, we extract and annotate the purpose, object, focus, viewpoint, and context (see Table 3.2 and Table 3.3). Counter-terrorism analysts who use our system will describe, through the user interface, the five aforementioned goal components. These extracted components will help to capture and conceptualize the goals of counter-terrorism analytics input into our system. In *GoOnto*, the concept of the *Purpose* is represented by the CTAPurpose class, *Object* by the CTAObject class, *Focus* by the CTAFocus class, *Viewpoint* by the CTAViewpoint, and *Context* by the CTAContext (Figure 3.7). Furthermore, the DMGoal class is inferred from the CTAPurpose, CTAObject, CTAFocus, CTAViewpoint, and CTAContext.

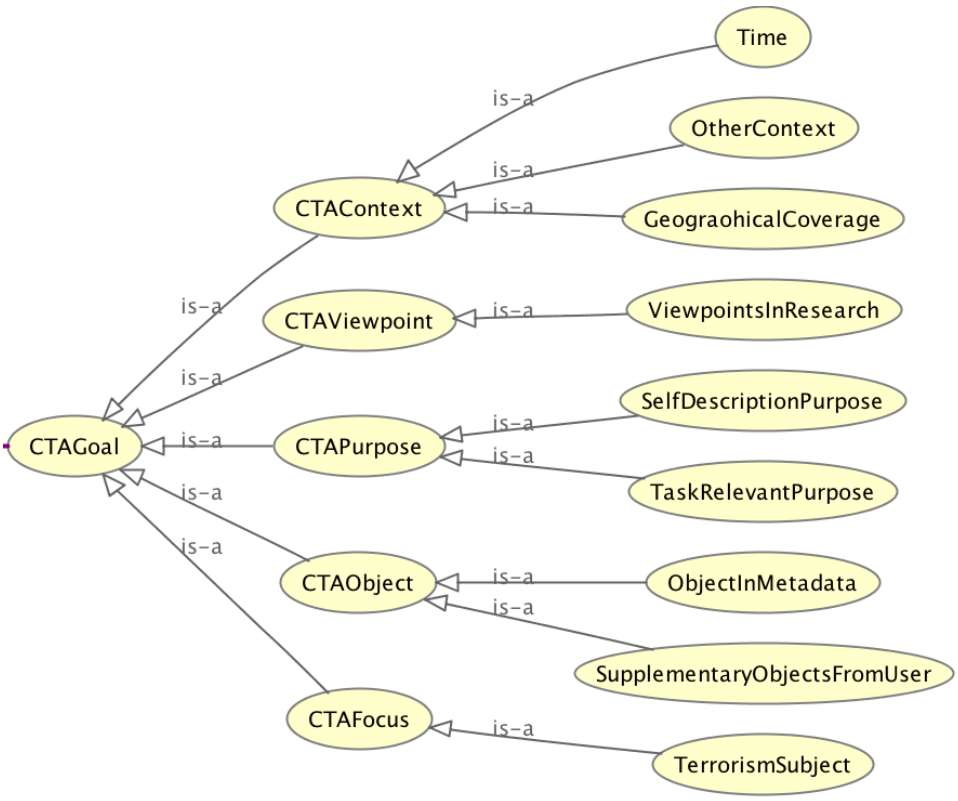


Figure 3.7 GoOnto

Counter-terrorism analytics purpose (class:** CTAPurpose**)

Counter-terrorism analytics purpose is related to the data analytics tasks. The task-relevant data analytics purpose may fall into one or more of six widely accepted categories —data exploration, data classification, association analysis, cluster analysis, prediction analysis, and anomaly detection (Hand 2007; Larose 2005). Noticeably, while a dataset can be used for a variety of data analytics tasks (e.g., GTD is used for prediction analysis and data classification), the annotation of these task-relevant purposes is of value for two-fold reasons: First, our design employs a comprehensive set of metadata information including such task-relevant metadata items as "relevant studies," in which the usage history of

terrorism datasets has been stored. Second, the collaboration of the identified task-relevant purpose and the increasing collection of task-relevant metadata will enable our system to identify the most relevant datasets regarding the user's task-relevant purpose.

Furthermore, users may prefer to disclose more details about the purposes of counter-terrorism analytics than those that are task-relevant. Therefore, in the *GoOnto* ontology, CTAPurpose is an entity that consists of two dimensions - TaskRelevantPurpose which includes six distinct ontology individuals and SelfDescribedPurpose which includes ontology individual generated from users' input (see Figure 3.8). Each individual depicts one specific counter-terrorism analytics purpose.

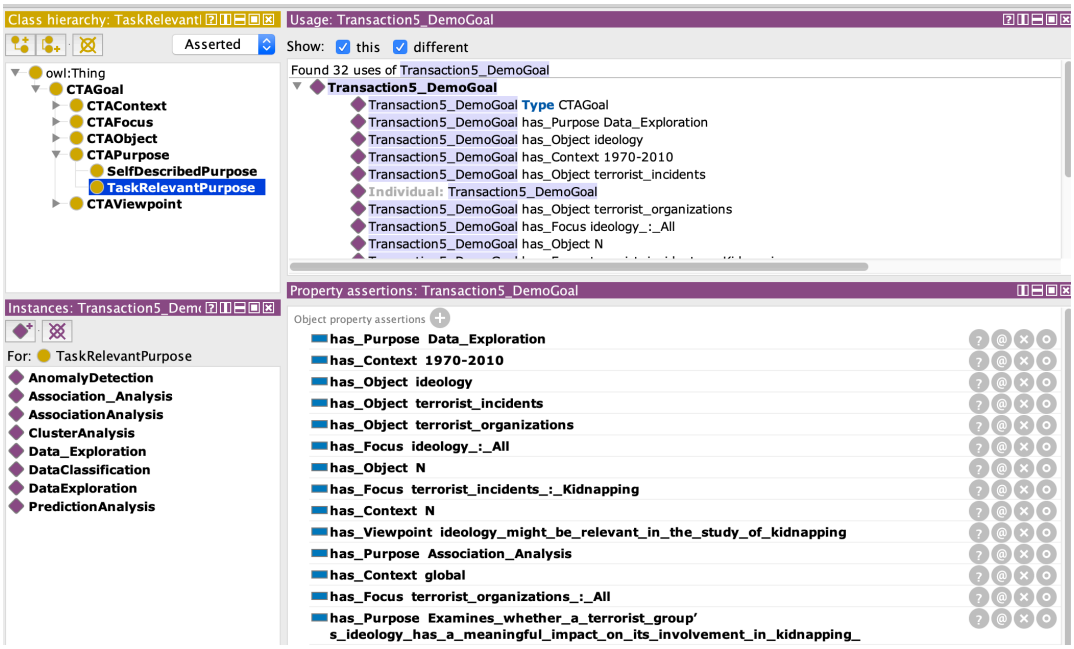


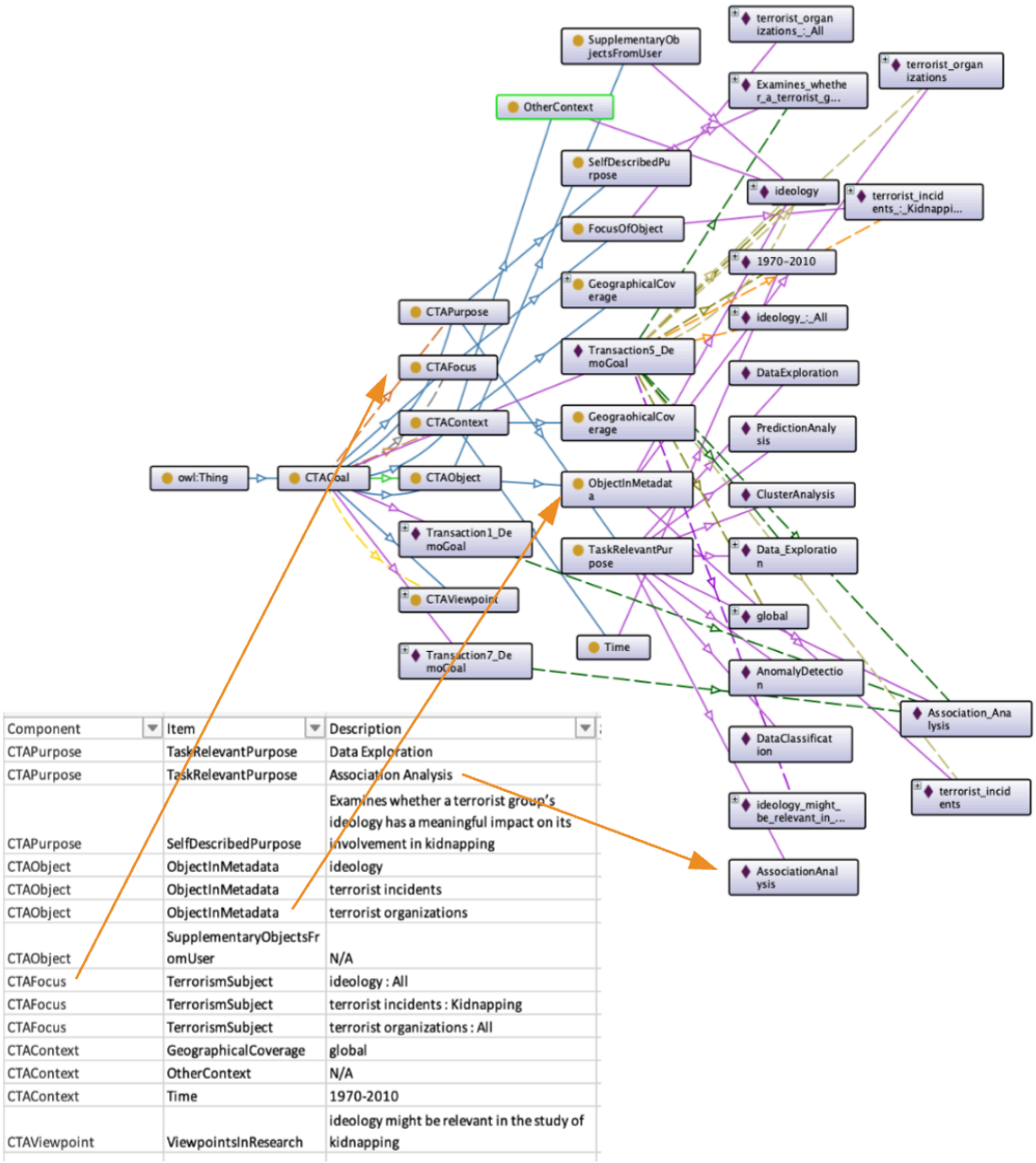
Figure 3.8 Counter-Terrorism Analytics Purpose of a Goal

***Counter-terrorism analytics object (class:** CTAObject)*

Counter-terrorism analytics object is the terrorism entity under investigation. A terrorism-relevant entity is an entity in a terrorism dataset that corresponds *to terrorism event*, *ideology*, *terrorist group/organization*, *religion*, *conflict*, etc. Our system synthesizes the topic entities from the metadata of the terrorism databases in our data repository and builds up a list of terrorism objects, enabling the users to click and select numerous objects of his/her interests for counter-terrorism analytics. As a user may be interested in objects that are not reflected by topic entities, supplementary objects are allowed from user-provided statements.

Counter-terrorism analytics focus (class:** CTAFocus**)

Counter-terrorism analytics focus depicts a particular angle or a set of angles of the entities under study. As has been mentioned previously, the terrorism-relevant entities include *terrorism event*, *ideology*, *terrorist group/organization*, *religion*, *conflict*, etc. The focuses may be different for different studies. For example, in a study analyzing terrorist events, the analysts may focus only on "airline hijackings" (Dugan et al. 2005). Our system enables users to input numerous focuses (e.g., *ideology*, *events type*, and *terrorist type*) corresponding to their input in the *counter-terrorism analytics object*.



Component	Item	Description
CTAPurpose	TaskRelevantPurpose	Data Exploration
CTAPurpose	TaskRelevantPurpose	Association Analysis
CTAPurpose	SelfDescribedPurpose	Examines whether a terrorist group's ideology has a meaningful impact on its involvement in kidnapping
CTAObject	ObjectInMetadata	ideology
CTAObject	ObjectInMetadata	terrorist incidents
CTAObject	ObjectInMetadata	terrorist organizations
CTAObject	SupplementaryObjectsFromUser	N/A
CTAFocus	TerrorismSubject	ideology : All
CTAFocus	TerrorismSubject	terrorist incidents : Kidnapping
CTAFocus	TerrorismSubject	terrorist organizations : All
CTAContext	GeographicalCoverage	global
CTAContext	OtherContext	N/A
CTAContext	Time	1970-2010
CTAViewpoint	ViewpointsInResearch	ideology might be relevant in the study of kidnapping

Figure 3.9 CTA Goal Component Extraction and Annotation

***Viewpoint of counter-terrorism analyst (class:** CTAViewpoint)*

Counter-terrorism analyst viewpoint depicts from whose perspective the counter-terrorism analytics is conducted. The individual viewpoint may influence individual

preferences and individual data use intentions and can be affected by individual identity and research experience (Kotonya 1999). As disclosure of identity or individual experience may subject system users to privacy concerns, this study modestly encourages users voluntarily to disclose their viewpoint information, asking about their research experience regarding counter-terrorism analytics.

***Counter-terrorism analytics context (class:** CTAContext)*

Counter-terrorism analytics context describes the scope within which the counter-terrorism analytics is conducted. Because terrorists' interests and the cumulative effect of terrorism can change significantly over time regarding space, time and geographical coverage are two types of contextual information that are essential and helpful for analyzing terrorism phenomena (Cronin 2003; Cutter 2014; Lake 2002). Additionally, other contextual information may also exist and be of value. For example, Stanton (2013) analyzes terrorism in the context of civil wars, recognizing a pattern of terrorism usage in rebel groups. To capture the information of *counter-terrorism analytics context*, our system enables the users to provide their specifications of time, geographical coverage, and other relevant contextual information.

3.4.4 Goal-relevant Dataset Identification

DIM aims to match a goal and its relevant datasets through similarity calculation and ontology inference. This task can be implemented in a two-step process: (1) similarity calculation and (2) inference informed by similarity scores and inference rules. The similarity calculation sub-module measures the degree of relevance between a goal *g* and a

dataset *d.* For a given dataset *d*, this study determines its degree of relevance by calculating the semantic distance between the input goal *g* and the metadata *m* of *d,* which is a collection of the metadata items for the dataset *d*. As in our *GoOnto*, g is depicted in a 5-tuple vector $g = (g_p, g_o, g_f, g_v, g_c)$, which consists of the purpose, the object, the focus, the viewpoint, and the context, respectively. Using the aggregate function, this study calculates the similarity between a goal component and the dataset metadata, similarity $(g_i, m) = \sum_{m_j \in m} \text{sim}(g_i, m_j)$, where *m* is an n-tuple $(m_1, m_2, \ldots m_n)$, in which n can be determined by the count of metadata concepts caught by our metadata extractor. These similarity coefficients are annotated as data properties into the *GoOnto* for a given goal instance. The goal-dataset matching phase draws on the similarity coefficients identified in the aforementioned phase. In the aftermath, a learner is built incorporating rules derived from the listed rule principles in Table 3.4, which assist in identifying goal-relevant dataset. The learner matches relevant datasets to a given goal.

Table 3.4 Inference Rule Generation Principles

Principles	**Formal Representation**
If a topic entity e in a dataset d matches an object o that exists in the goal g, then d is a relevant dataset.	$\exists o \in g$ matches $\exists e \in d \rightarrow$ rel(d,g) = 1
If a column c in a dataset d matches an object o that exists in the goal g, then d is a relevant dataset.	$\exists o \in g$ matches $\exists c \in d \rightarrow$ rel(d,g) = 1
If the similarity sim between the metadata m of d and the goal g is high	sim (m, d) > threshold $\rightarrow$ rel(d,g) = 1

(e.g., sim > threshold), then d is a relevant dataset.	
If no instance in a dataset d matches a focus f that exists in the goal g, then d is filtered out (i is an instance).	$\exists f \in g$ does not match $\forall i \in d \rightarrow$ rel(d,g) := 0
If a dataset d does not satisfy a context item c in the goal g, then d is filtered out.	$\exists c \in g$ is not covered by $d \rightarrow$ rel(d,g) := 0

3.4.5 Schema and Instance Mapping

SCMappingM is designed for retrieving the data that matches the goal and consolidating the data. This module will first retrieve the goal-relevant datasets, where there may be semantic heterogeneity in the data schemas and data instances. For semantically heterogeneous data, SCMappingM implements two mapping tasks—schema mapping and instance mapping.

Table 3.5 Schema Name Similarity Matrix

Schema Name Similarity		**S**			
		Organization name	**Eventid**	**Place**	**Description**
T	Terrorist Organization	0.333	0	0	0
	Ideology	0	0	0	0

Schema mapping is important, as schemas are often created by different people whose styles (e.g., coding style and naming style) are different. For two terrorism datasets (e.g., S and T), the schema mapping aims to identify data columns that refer to an identical entity (e.g., S. *perpetrator_name* mapping to T. *perpetrator*), assigning to each pair (e.g., S.column_i and T.column_j) a number in [0,1] (1 for matching, 0 for not). In our schema matching, the mapper identifies string columns of heterogeneity and recognizes the mapping relations based on similarity score: (1) schema-name similarity score, which only uses the schema name to calculate the semantic similarity between two schema elements (Table 3.5), and (2) individual-based similarity score (Table 3.6). These two similarity scores are combined into one similarity matrix through the maximum combiner (Figure 3.10). A high similarity score between two elements indicates a high confidence that these two elements match together.

Table 3.6 Individual-based Similarity Matrix

Instance Similarity		**S**			
		Organization Name	**Eventid**	**Place**	**Description**
T	Terrorist Organization	0.978	0	0	0.001
	Ideology	0	0	0	0

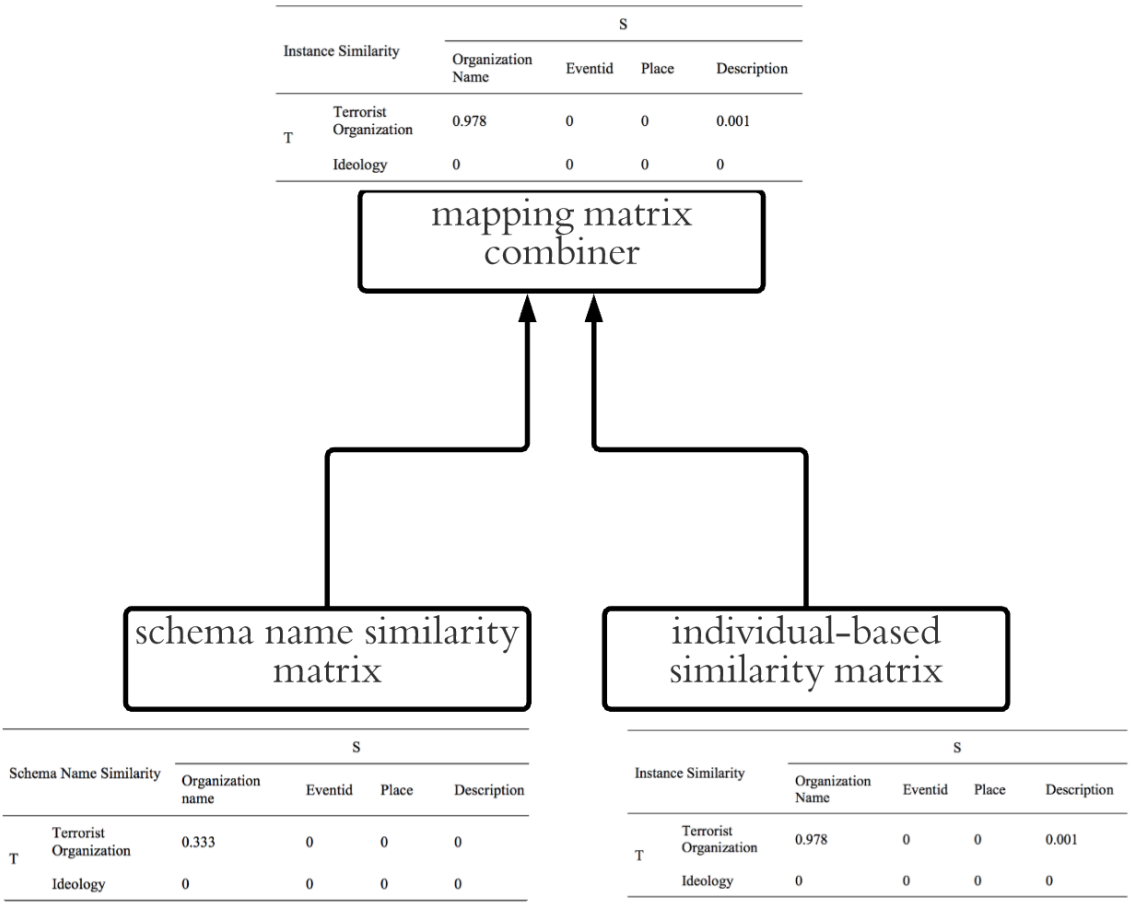


Figure 3.10 Schema Mapping Process

Instance mapping is the second phase of the mapping task. To integrate terrorism data, our framework cares not only about the schemas but also the dataset instances. Instance mapping finds data instances that refer to the same entity (e.g., ISIL and Islamic State of Iraq and the Levant).

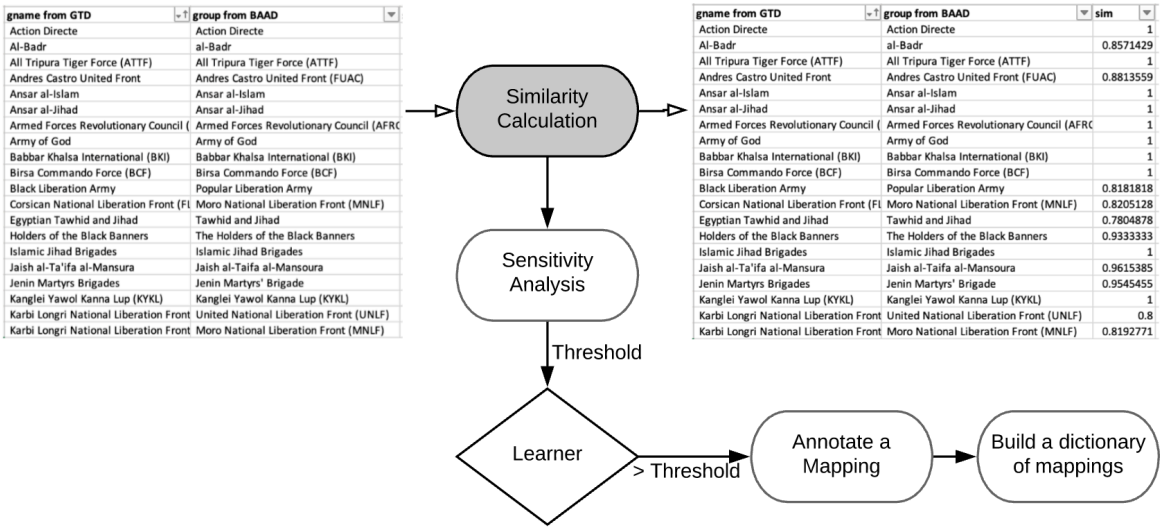


Figure 3.11 Similarity Calculation and Instance Mapping Learner

There exist a variety of string similarity calculation methods (e.g., set-based algorithms—Jaccard, Cosine, and Masi; sequence-based algorithms—Longest Common Substring Matcher; edit distance-based algorithms—Levenshtein and Hamming). These algorithms are developed upon different theories, thereby following their unique operations and generating different results. For example, given two strings “Kuki Liberation Army (KLA)” and “Kosovo Liberation Army (KLA)”, both of which are data instances of perpetrators, the Jaccard Similarity algorithm gives a similarity score of 0.8000; the Cosine Similarity algorithm rates the similarity as 0.8524; the Levenshtein Similarity algorithm calculates a score of 0.8518.

Table 3.7 provides a list of widely used string similarity calculation algorithms. String similarity calculation algorithms are categorized based on their operations properties.

Table 3.7 Instance Mapping Algorithms

Category	Similarity Calculation Algorithms
Edit distance based	Levenshtein Similarity, Hamming Similarity, Damerau-Levenshtein, Jaro-Winkler, Needleman-Wunsch
Sequence based	Longest Common Substring Matcher, Ratcliff-Obershelp similarity
Token based	Jaccard Similarity, Cosine Similarity, Masi Similarity, Tanimoto distance

Token-based algorithm: Algorithms falling in this category, accepting strings as input, transform a string into a set of tokens and compute the similarity using the tokens. The more common the tokens, the larger the similarity coefficient. For example, the Jaccard Similarity locates the similar tokens in the two sets of tokens transformed from two strings and then divides the count of common tokens by the count of the union of two sets (Wang et al. 2014).

Edit-distance-based algorithm: Algorithms falling into this category calculate the number of edit operations (e.g., minimum edit operations) needed to transform one string into the other. The lower the number of needed edit operations, the higher the similarity between two given strings (Wang et al. 2014).

Sequence-based algorithm: Sequence defines a common substring between two strings. The sequence-based algorithms locate the sequence(s) presented in two given

strings. The longer these sequences, the more similar are the two strings, according to sequence-based algorithms (pypi 2019).

Hybrid algorithms: Other algorithms exist that extend the basic algorithms in the listed three types. For example, scholars may apply fuzzy logic to string matching (Wang et al. 2014). Fuzzy string matching aims to find strings that approximately match a pattern (or approximately match one another). Fuzzy string matching methods can be built on the Levenshtein Distance algorithm (e.g., fuzzywuzzy lib in Python).

In our scenario, we will test a set of six string similarity methods and select one for final artifact implementation based on two criteria: (1) the outstanding performance in terms of F-measure and (2) the lower response time.

3.5 Implementation

This study employs a combination of techniques (e.g., regular expression, XPath, semantic web, database, and NLP) to instantiate our design (Table 3.8). To automatically collect the free-text metadata available on the dataset-hosting webpages, our instantiation uses Selenium WebDriver in Python. Selenium is a web automation tool that is originally and widely utilized by website developers to automate website testing (Bruns et al. 2009). We use Selenium WebDriver in the implementation to automatically start and control a web browser (specifically, our implementation uses the Firefox driver). The combination of Selenium WebDriver and Python enables our metadata collection and management

module to intelligently mimic the human actions of visiting a website, clicking the relevant tabs, and collecting the metadata of terrorism datasets from web pages.

XPath and regular expression are employed to locate and extract metadata information elements of interests, as well as the values corresponding to these elements. XPath, a major component in the eXtensible Stylesheet Language Transformations (XSLT) standard, uses path expressions to navigate through our collected XML/HTML document to recognize desirable attributes and elements (Clark 1999). Regular expression, a technique developed on the basis of formal language theory, defines a sequence of characters (or say, a search pattern) to help string search algorithms identify the strings that match a defined pattern.

To generate the ontologies—*GoOnto* and *MeDaOnto*, our implementation uses Owlready2, a Python library for ontology-oriented programming, including the utilities of importing ontologies, exporting ontologies, manipulating ontology entities, individuals, and properties (data properties and object properties). In addition, it is compatible with the RDFlib library to perform SPARQL queries in our generated ontologies. In addition, we use Protégé 5.17 to review the generated ontologies.

To analyze text information in the metadata and user goals, this study utilizes NLP techniques (e.g., entity extraction) provided by spaCy. In particular, to determine whether a dataset is relevant to a given goal of counter-terrorism analytics, we build an unsupervised learner to calculate semantic similarity between the goal and the metadata of terrorism datasets, using Python Pandas. Pandas provides a toolkit for data manipulation and

analysis. It provides a rich list of built-in functions which facilitates data extraction, transformation, and loading (ETL) tasks and can handle data quickly and efficiently. Additionally, it supports data input and output with SQL databases.

In the data consolidation phase, we implement the schema mapper and data instance mapper with Python, building up agents and learners conducting similarity calculation and comparison. A set of Python libraries (e.g., difflib, nltk) are utilized for similarity calculation and data mapping. PostgreSQL is used to store and manage the original terrorism datasets to protect data integrity.

The implementation includes a data storage of 10 publicly available terrorism datasets commonly used in counter-terrorism analytics and retrieved from the Harvard Dataverse data archive (Dataverse 2018).

Table 3.8 Implementation Specifications

Functional Utility	**Implementation Specifications**
Collect metadata	Selenium WebDriver in Python
Extract metadata	XPath and Regulation Expression in Python, Pandas (for data manipulation)
Metadata annotation and goal annotation	Owlready2 for Python (for ontology generation), Protégé (for manual review)
Goal-relevant data identification	spaCy for Python (for named entity extraction), other Python libraries to build a non-supervised learner
Data schema mapping	Python libraries to build a schema learner

Data instance mapping	Python libraries to build an instance learner

Chapter 4. Evaluation

In the previous section, we discussed our proposed design artifact: a goal-driven data integration framework for counter-terrorism analytics, as well as the specifications (e.g., techniques and settings) of the artifact implementation. The main task of the implementation phase is to build an instance of the proposed design, based on which the design can be evaluated in terms of efficacy and efficiency. Moreover, the disclosure and specifications of design implementation will provide the necessary details, following which future studies can replicate or extend our research.

This section includes a series of experiments. Experiment 1 evaluates the efficacy of the elicited goal components in retrieving goal-relevant datasets to confirm the importance of using the holistic set of goal components rather than a single goal component. Experiment 2 evaluates the effectiveness of our schema mapper. Experiment 3 accesses the efficacy of the data instance mapper (e.g., string mapping); specifically, as there exists a variety of string similarity calculation algorithms, our research presents a comparative evaluation of algorithms widely used in IS and CS research. Experiment 4 provides a sensitivity analysis against our proposed mappers, illustrating the algorithm tuning process and the stability of the experimental results in experiment 3. Through the four experiments, we demonstrate the worth of our design, addressing criteria such as effectiveness, quality and stability. The evaluation results assessing the implemented artifact will inform future scientific inquiries (Hevner et al. 2004).

4.1 Experimental Setting

Our evaluation phase assesses the utilities of the proposed artifact and informs the future research for improvement (Hevner et al. 2004). Specifically, this research proposes an artifact that is tailored to a real-world problem—data integration for counter-terrorism analytics. Notably, our proposed data integration solution is goal-driven and takes the metadata of terrorism datasets into account. This study conducts a series of experiments to examine the utilities of our proposed artifact. To evaluate our artifact, we would investigate its effectiveness—how effective the proposed artifact enables users to obtain integrated data that can be used in successive counter-terrorism analytics, implementing a goal-driven data integration process.

A considerable amount of counter-terrorism analytics cases is well documented in the literature. In one such case, Forest (2012) examines the impact of a terrorist group's ideology on its involvement in kidnapping incidents. Specifically, this analytical study incorporates the terrorism incident data on a global level from 1970 to 2010 and the data of ideological attributes of terrorist organizations. The analytical results indicate that the terrorist organization's ideology "does not play an important role in determining the likelihood of a group's involvement in kidnapping." This analytical finding has implications for counter-terrorism policies to fight "kidnapping" by terrorist organizations. We propose an evaluation of our design artifact, using the case of Forest (2012) as a reference framework for two reasons: (1) this study clearly articulated its goal in counter-terrorism analytics which can be extracted and used as user input for data integration; and

(2) this study clearly described the usage of terrorism datasets (e.g., dataset names, analytical tasks).

In the case, our evaluation in a series of experiments assesses how our implemented artifact could effectively and efficiently help Forest (2012) to obtain the desirable dataset for his research "Kidnapping by Terrorist Groups, 1970–2010: Is Ideological Orientation Relevant?" We extensively test each of the essential features in our design effort. Specifically, we fully test the mappers (which are unsupervised learners) in terms of precision, recall, and F-measure.

Table 4.1 A Comparison Between Data Integration Using GoDa and That in Forest (2012)

Feature	**GoDa**	**Forest (2012)**
Goal for counter-terrorism analytics	GoDa requires users to be clear about their goals for data analytics.	Forest may explore the data without a goal at the beginning.
Metadata of terrorism datasets	Users can access the metadata of terrorism datasets stored and annotated in our system.	Forest may need to search for and retrieve dataset-relevant information (e.g., metadata) manually.
Dataset schema mapping process	Automatic mapping	Forest may manually recognize the mapped schema and determined which can be used for the data consolidation.
Data instance mapping process	Automatic mapping	Forest may manually map data instances.

4.2 Experiment 1—An Evaluation of Goal-relevant Data Identification

Experiment 1 aims to evaluate the efficacy of the elicited goal components in relation to retrieving the goal-relevant terrorism datasets. In our evaluation, we use Forest's (2012) goal, the five components of which are listed in Table 4.2 as the input to the proposed artifact. All responses corresponding to the goal component items use the original words or statements from the author, thereby reflecting the author's research goal as closely as possible. Moreover, when eliciting the goal, we follow the principles proposed by van Solingen and Berghout (1999) to ensure that the goal are "defined in an understandable way and should be clearly structured" (p. 51).

These recognized and structured goal components are utilized to match the annotated metadata information of terrorism datasets, calculating the semantic similarity (in literature, researchers often use semantic distance and semantic distance interchangeably) between the given goal and the metadata of each datasets.

Table 4.2 Five Goal Components Recognized from Forest (2012)

Component	Item	Response	Source
CTAPurpose	TaskRelevantPurpose	Data Exploration	p. 777
	TaskRelevantPurpose	Association Analysis	p. 769
	SelfDescribedPurpose	"Examines whether a terrorist group's ideology has a meaningful impact on its involvement in kidnapping"	p. 769

CTAObject	ObjectInMetadata	Ideology	p. 769
	ObjectInMetadata	Terrorist Incidents	p. 769
	ObjectInMetadata	Terrorist Organizations	p. 769
	SupplementaryObjectsFromUser	N	
CTAFocus	TerrorismSubject	Ideology: All	p. 769
	TerrorismSubject	Terrorist Incidents: Kidnapping	p. 769
	TerrorismSubject	Terrorist Organizations: All	p. 769
CTAContext	GeographicalCoverage	Global	p. 769
	OtherContext	N	
	Time	1970-2010	p. 769
CTAViewpoint	ViewpointsInResearch	"ideology might be relevant in the study of kidnapping"	p. 773

While the goal definition template has been well adopted and used in research and practice, the literature yields little empirical evidence demonstrating the importance of formulating a goal using the complete set of five components. To confirm the critical importance of formulating and utilizing a holistic set of goal components in a goal-driven data integration process rather than using a single selected component, we built a benchmark solution which uses only one goal component, "Object," for goal-dataset matching. Table 4.3 presents the comparative results.

Table 4.3 Goal-Dataset Matching Results

Solution	Dataset Matched
Benchmark	GTD
	Profiles of Perpetrators of Terrorism in the United States
	Sources of Blame Attribution
	Government Actions in a Terror Environment
	Global Terrorism Database Ideological Motivations of Terrorism in the United States
	Global Terrorism Database ISIL Auxiliary Dataset
	BAAD
	RANNSAD
GoDa	GTD
	BAAD
Manual Solution by Forest (2012)	GTD BAAD

The evaluation shows that the goal-relevant dataset identification module in our proposed artifact retrieves two goal-relevant datasets, GTD and BAAD, which are the datasets recognized and used in the Forest (2012) case for data analytics. In other words, the observation complies with our proposition that the identified components of a user goal of counter-terrorism analytics enable our design to identify the goal-relevant datasets needed for data integration.

In contrast, the benchmark system, which uses only one goal component, retrieves eight datasets that satisfy a matching. In other words, a complete definition of a goal with five components in the goal definition template from the GQM approach will be more effective than a fragmented goal definition. This complies with our proposition 2.

In summary, our goal-dataset matching solution performs very well, recognizing the desirable datasets based on the calculated degree of relevance between a given goal and the terrorism datasets in our data repository. In other words, the goal-dataset matching task, at the pivot part of our design, effectively links the utility of goal modeling with successive goal-driven data extraction, transformation and loading.

4.3 Experiment 2—An Evaluation of the Schema Mapper

Experiment 2 evaluates the effectiveness of our schema mapper. Schema mapping is an antecedent phase for instance mapping. The accuracy of schema mapping sets the ceiling for the reliability and validity of the aftermath instance mapping. This study evaluates the schema mapper in terms of precision, recall and F-measure.

Table 4.4 illustrates how to construct a confusion matrix to evaluate the effectiveness of a mapping solution. In a typical mapping task, each mapping relation (e.g., schema or instance) can be measured utilizing a given score or scale, on the assumption that the score or measurement is an indicator that classifies a pair (e.g., schema concepts, data instances) into one of two states (e.g., Positive/Negative, where Positive indicates the pair is matching, and Negative indicates that the pair is not matching).

Table 4.4 Confusion Matrix for Mappers

		Predicted mapping pairs	
		Mapping	**Not Mapping**
Actual mapping pairs	**Mapping**	TP	FP
	Not mapping	FN	TN

Researchers may use sensitivity analyses (see Experiment 4) to determine a threshold that serves as a cutoff score to classify the condition as Positive / Negative (or say Mapping/Not-Mapping). For instance, if the similarity score is greater than the cutoff, the predicted result is Positive (in other words, the elements in the pair are mapping to each other). Otherwise, the predicted result is Negative. Although a learning (e.g., supervised or unsupervised) algorithm seeks to determine the optimal threshold value that serves as a decision cutoff, the reality is that the predicted results do not always reflect the true condition (e.g., the mapping relationship of the elements in a pair). In this sense, there exist four possible outcomes: true positive, true negative, false positive, and false negative. The true positive shows that the predicted result indicating a mapping relationship is a true prediction. The false positive means that the predicted result indicating a mapping relationship is a false prediction; in other words, the algorithm misclassifies a non-mapping relationship as a mapping relationship. The true negative shows that the result predicting a

non-mapping relationship is a true prediction. The false negative means that the result predicting a non-mapping relationship is a false prediction.

Based on the confusion matrix, the performance measures—precision, recall, and F-measure—can be defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \times 100\% \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \times 100\% \quad (2)$$

$$F - measure = 2 * \left(\frac{Precision * Recall}{Precision + Recall} \right) \times 100\% \quad (3)$$

Precision is the ratio of true positive predictions (e.g., the correct prediction of mapping relations) to the total predicted positive observations (e.g., the total prediction of mapping relations). This metric value indicates how precisely a mapper makes a positive prediction.

Recall is the ratio of true positive predictions (e.g., the correct prediction of mapping relations) to the total actual positive observations (e.g., the total actual mapping relations). This metric value indicates the capability of the mapper to identify mapping relations.

The F-measure, a weighted harmonic mean of precision and recall, takes into calculation both false positives and false negatives. F1 is particularly informative when false positives and false negatives are significantly different. The complete set of precision,

recall, and F-measures will enable our evaluation to comprehensively review a mapper's performance.

The schema mapper in this study identifies, from data schemas to be matched (e.g., in Forest's case, schemas of GTD and BAAD, which are returned in the goal-relevant data identification phase), the columns that consist of string instances (see Table 4.5). These columns, consisting of semantic heterogeneity, are mapped based on their similarity.

Table 4.5 String Columns in GTD and BAAD

Dataset	String Columns
GTD	addnotes, alternative_txt, approxdate, attacktype1_txt, attacktype2_txt, attacktype3_txt, city, claimmode_txt, claimmode2_txt, claimmode3_txt, corp1, corp2, corp3, country_txt, dbsource, divert, gname, gname2, gname3, gsubname, gsubname2, gsubname3, hostkidoutcome_txt, kidhijcountry, location, motive, natlty1_txt, natlty2_txt, natlty3_txt, propcomment, propextent_txt, provstate, ransomnote, region_txt, related, resolution, scite1, scite2, scite3, summary, target1, target2, target3, targsubtype1_txt, targsubtype2_txt, targsubtype3_txt, targtype1_txt, targtype2_txt, targtype3_txt, weapdetail, weapsubtype1_txt, weapsubtype2_txt, weapsubtype3_txt, weapsubtype4_txt, weaptype1_txt, weaptype2_txt, weaptype3_txt, weaptype4_txt
BAAD	cowmastercountry, group

We calculate similarity scores for the columns identified in the above table. For example, a pair of columns—*addnotes* and *cowmastercountry* (the first column is from

GTD, whereas the other is from BAAD)—has a similarity score 0.009. The similarity matrix is constructed as in Table 4.6.

Table 4.6 Schema Similarity Matrix

GTD \ BAAD	**cowmastercountry**	**group**
addnotes	0.009	0.047
alternative_txt	0.012	0.003
approxdate	0.000	0.010
attacktype1_txt	0.000	0.000
attacktype2_txt	0.000	0.003
attacktype3_txt	0.000	0.003
city	0.001	0.007
claimmode_txt	0.000	0.003
claimmode2_txt	0.000	0.003
claimmode3_txt	0.000	0.000
corp1	0.014	0.049
corp2	0.040	0.067
corp3	0.020	0.042
country_txt	0.427	0.038
dbsource	0.051	0.005
divert	0.316	0.020
gname	0.033	0.189

gname2	0.038	0.107
gname3	0.009	0.030
gsubname	0.027	0.084
gsubname2	0.011	0.006
gsubname3	0.000	0.000
hostkidoutcome_txt	0.000	0.000
kidhijcountry	0.403	0.040
location	0.010	0.022
motive	0.010	0.055
natlty1_txt	0.375	0.039
natlty2_txt	0.466	0.037
natlty3_txt	0.343	0.029
propcomment	0.001	0.034
propextent_txt	0.000	0.000
provstate	0.009	0.027
ransomnote	0.026	0.058
region_txt	0.034	0.006
related	0.000	0.000
resolution	0.000	0.000
scite1	0.011	0.040
scite2	0.012	0.043
scite3	0.015	0.048

summary	0.004	0.020
target1	0.005	0.021
target2	0.011	0.046
target3	0.028	0.035
targsubtype1_txt	0.005	0.020
targsubtype2_txt	0.000	0.019
targsubtype3_txt	0.000	0.012
targtype1_txt	0.000	0.008
targtype2_txt	0.000	0.006
targtype3_txt	0.000	0.006
weapdetail	0.006	0.033
weapsubtype1_txt	0.000	0.002
weapsubtype2_txt	0.000	0.002
weapsubtype3_txt	0.000	0.002
weapsubtype4_txt	0.000	0.000
weaptype1_txt	0.000	0.000
weaptype2_txt	0.000	0.002
weaptype3_txt	0.000	0.002
weaptype4_txt	0.000	0.000

Informed by the similarity scores in the similarity matrix, the proposed schema mapper identifies the equivalent columns. We construct the confusion matrix in Table 4.7

and present the performance of the schema mapper as follows: we obtained a precision of 1, a high precision indicating a low false positive rate. The recall is 0.899; and the F score is 0.941. In the case of Forest (2012), the proposed mapper identifies the key columns for joining GTD and BAAD: the *gname* column from GTD and the *group* column from BAAD.

Table 4.7 Confusion Matrix for the Schema Mapper

		Predicted mapping pairs	
		Mapping	**Not Mapping**
Actual mapping pairs	**Mapping**	8	1
	Not mapping	0	107

4.4 Experiment 3—An Evaluation of the Instance Mapper

Experiment 3 evaluates the efficacy of the data instance mapper (e.g., string mapping). The instance mapper matches structured data instances that refer to the identical entity. The performance of the instance mapper is subject to the accuracy of the string similarity calculation algorithm.

Table 4.8 Confusion Matrix for the Instance Mapper (Using Levenshtein Similarity)

		Predicted mapping pairs	
		Mapping	**Not Mapping**
Actual mapping pairs	**Mapping**	23	1
	Not mapping	5	157977

As there exists a variety of string similarity calculation algorithms, our research presents a comprehensive comparative evaluation of six algorithms which are widely adopted in IS and CS research. Our evaluation constructs confusion matrix (see Table 4.8) and calculates precision, recall and F-measure, examining and comparing the performance of these string similarity calculation methods in our proposed mapper. The algorithm with best overall performance (e.g., F-measure in our case) is utilized to implement the instance mapper.

Table 4.9 Evaluation of Instance Mapper

Type	**Similarity Calculation Algorithm**	**Precision**	**Recall**	**F-measure**
Token based	Jaccard Similarity	0.800	**0.857**	0.828
Token based	Masi Similarity	0.800	**0.857**	0.828
Token based	Cosine Similarity	0.913	0.750	0.824

Sequence based	Longest Common Substring Matcher	0.958	0.821	0.**885**
Edit distance based	Hamming Similarity	**1.000**	0.714	0.833
Edit distance based	Levenshtein Similarity	0.958	0.821	0.**885**

In experiment 3, we utilized instances in the *gname* column from GTD and *group* column from BAAD for testing and generates the evaluation metrics in Table 4.9. By comparing these evaluation metrics, we assessed the mapping capability of the six algorithms. These metrics indicate the extent to which the instance mapper can resolve the heterogeneity in terrorism data instances to address interoperability.

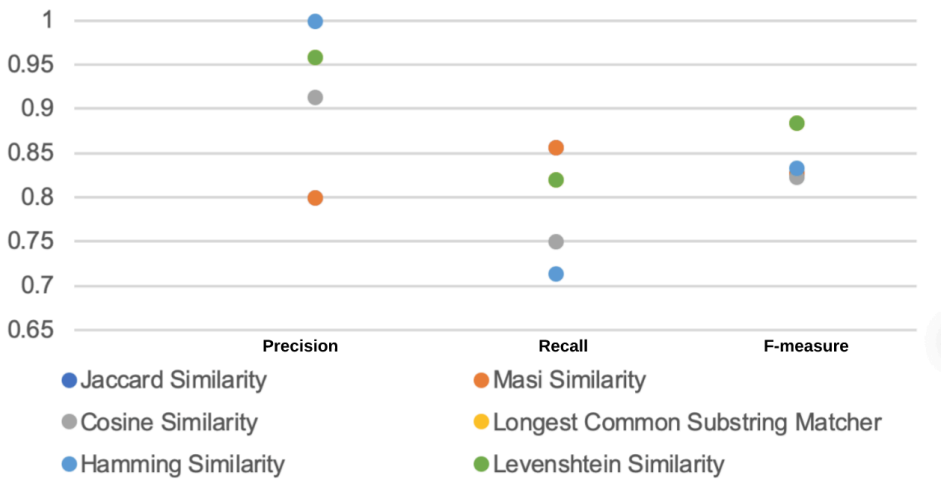


Figure 4.1 Plotted Performances of the Six Algorithms

The comparison indicates that the Levenshtein Similarity algorithm and the Longest Common Substring Matcher algorithm have outstanding overall performance, with F scores of 0.885, over the other algorithms. We then step further to compare the time efficiency between the two and record the experimental results in Table 4.10. The Levenshtein Similarity algorithm is more time-efficient than the Longest Common Substring Matcher algorithm in our case of mapping terrorism-relevant strings, thereby shortening the response time; therefore, our final implementation uses the Levenshtein Similarity.

Table 4.10 Time Need of the Instance Mapper

Case	**Strings**	**Response Time**	
		Levenshtein Similarity	**Longest Common Substring Matcher**
1	st1 = 'National Democratic Front-Bicol (NDF-Bicol)' st2 = 'National Democratic Front of Bodoland (NDFB)'	.00103ms	.25568ms
2	st1 = 'National Socialist Council of Nagaland-Unification (NSCN-U)' st2 = 'Ansar al-Jihad'	.00228ms	.25647ms
3	st1 = 'National Socialist Council of Nagaland-Unification (NSCN-U)' st2 = 'Ansar al-Jihad'	.00224ms	.17814ms
4	st1 = 'Action Directe' st2 = 'Action Directe'	.00048ms	.05303ms

Using the Levenshtein Similarity Algorithm (threshold = 0.84), the instance mapper can identify the mappings in Table 4.11.

Table 4.11 Mapped Perpetrator Instances (for the Kidnapping Data Fragments)

gname from GTD	**group from BAAD**	**Levenshtein**
Birsa Commando Force (BCF)	Birsa Commando Force (BCF)	1.000
Action Directe	Action Directe	1.000
Army of God	Army of God	1.000
Popular Resistance Committees	Popular Resistance Committees	1.000
Babbar Khalsa International (BKI)	Babbar Khalsa International (BKI)	1.000
Popular Revolutionary Army (Mexico)	Popular Revolutionary Army (Mexico)	1.000
All Tripura Tiger Force (ATTF)	All Tripura Tiger Force (ATTF)	1.000
Ansar al-Islam	Ansar al-Islam	1.000
Muttahida Qami Movement (MQM)	Muttahida Qami Movement (MQM)	1.000
Andres Castro United Front	Andres Castro United Front (FUAC)	0.881
Holders of the Black Banners	The Holders of the Black Banners	0.933
Islamic Jihad Brigades	Islamic Jihad Brigades	1.000
Armed Forces Revolutionary Council (AFRC)	Armed Forces Revolutionary Council (AFRC)	1.000

Maoist Communist Center (MCC)	Maoist Communist Center (MCC)	1.000
Kuki Liberation Army (KLA)	Kosovo Liberation Army (KLA)	0.852
Kuki Liberation Army (KLA)	Kuki Liberation Army (KLA)	1.000
Protectors of Islam Brigade	Protectors of Islam Brigade	1.000
United People's Democratic Solidarity (UPDS)	United People's Democratic Solidarity (UPDS)	1.000
Purbo Banglar Communist Party	Purbo Banglar Communist Party (PBCP)	0.892
Ansar al-Jihad	Ansar al-Jihad	1.000
Al-Badr	al-Badr	0.857
Kanglei Yawol Kanna Lup (KYKL)	Kanglei Yawol Kanna Lup (KYKL)	1.000
Jaish al-Ta'ifa al-Mansura	Jaish al-Taifa al-Mansoura	0.962
Jenin Martyrs Brigades	Jenin Martyrs' Brigade	0.955

4.5 Experiment 4—A Sensitivity Analysis

Experiment 4 provides a sensitivity analysis of the instance mappers evaluated in the Experiment 3, assessing the stability of our algorithm in the tuning process and demonstrating how the optimal experimental results are achieved in experiment 3.

The instances mapping will not be capable of retrieving a mapping relation for a pair of strings from the terrorism data instances, when the two strings in the pair do not

have a similarity larger than the threshold *t*. To conclude that a data instance (e.g., string) in a dataset (e.g., GTD) is mapping to a data instance in another dataset (e.g., BAAD), the similarity index between the two data instances must exceed the specified threshold *t*. In other words, a proper threshold assists mappers to appropriately justify a mapping relation between two strings, thereby filtering out the non-mapping relations and affecting the precision, recall and F ratios of mappers.

A larger threshold indicates that a mapper needs a larger similarity score to conclude a mapping relationship between two data instances. For example, the Levenshtein Similarity score of *Purbo Banglar Communist Party* and *Purbo Banglar Communist Party (PBCP)* is 0.892; if the threshold is set to 0.9, the mapper will not catch the mapping relationship between these two strings. In this case, the mapper would have a low recall ratio. In contrast, if a threshold is small, the learner will map two strings even when there is no mapping relation between them. For example, the Levenshtein Similarity score of *Kuki Liberation Army (KLA)* and *Kosovo Liberation Army (KLA)* is 0.851; if the mapper sets a threshold of 0.850, a mapping relation will be concluded between the two given strings. In other words, a higher threshold set by the learner may result in a lower precision ratio.

In summary, an algorithm may fail to appropriately identify mappings when the threshold is set either too high or too low. Therefore, this study conducts a sensitivity analysis, in order to find the ideal threshold that trades off precision and recall and accordingly achieves the best overall performance.

This experiment carefully assesses the effect of threshold on the experimental metrics—precision, recall, and F-measure (Figure 4.2)—starting with a threshold of 0.1 (the figure only shows the analysis results with a threshold > 0.65). In the figure, we calculate and plot the precision, recall, and F-measure with a corresponding threshold. It is observed that different thresholds result in different performance scores. This study takes account of the F-measure, an index that measures the instance mapper's overall performance, and finds that the threshold of 0.84 results in the highest performance score, 0.885, with the Levenshtein algorithm. Thus, we chose 0.84 as the threshold for our instance mapper implemented with the Levenshtein Similarity algorithm.

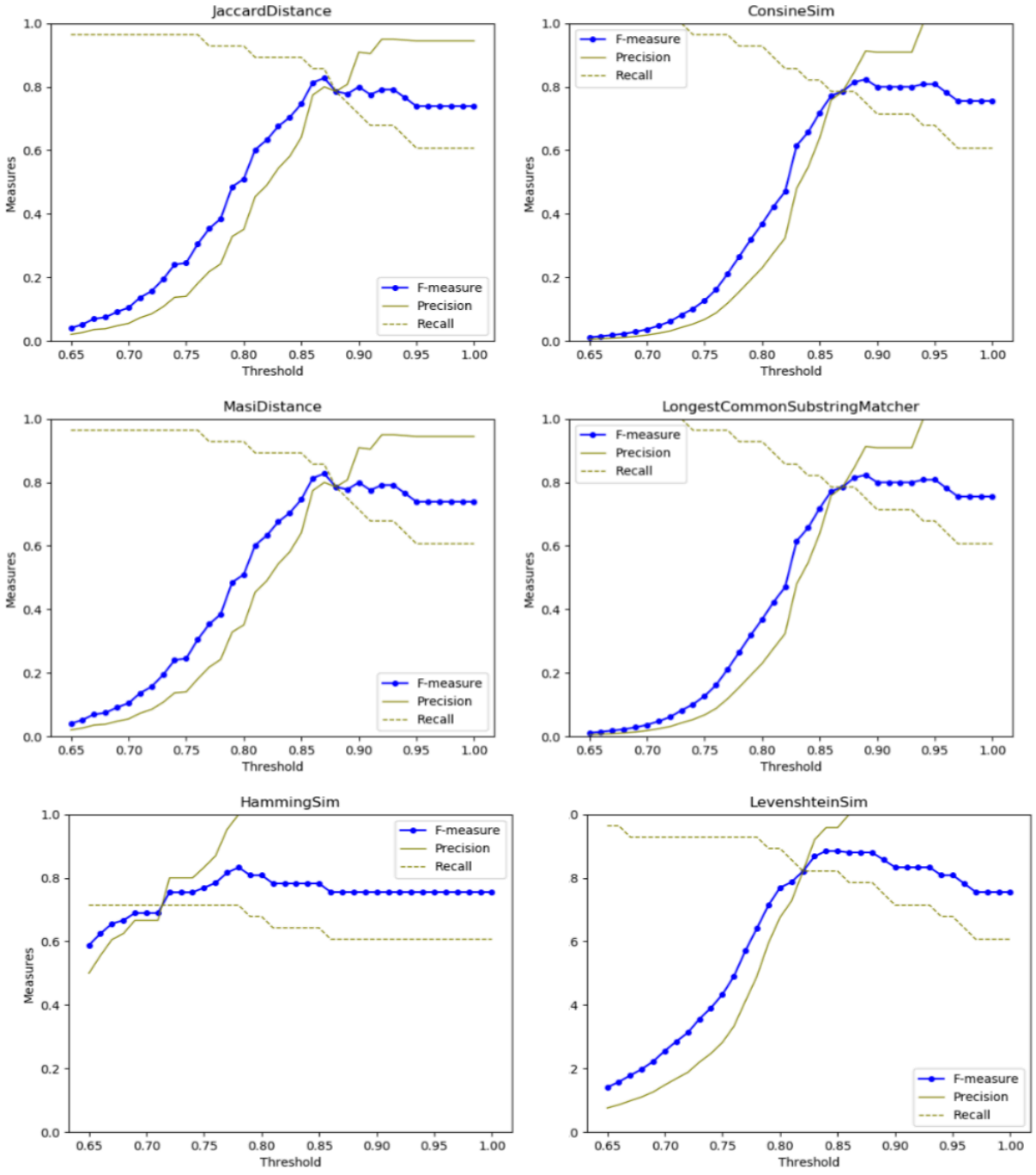


Figure 4.2 Sensitivity Analysis

Chapter 5. Discussion and Conclusion

Data integration resides in the back end of a data analytics task, being typically labor-intensive and effort-consuming, thereby attracting organizational resources and requiring high development costs. The success and quality of data integration directly determine the success of data analytics. In other words, successful data integration is of critical importance for modern organizations in relation to business intelligence, government intelligence, and the like.

Consequently, there is an abundance of design research aiming to automate the data integration process. However, the major body of existent technical data integration solutions are data-management oriented, providing little support for systematically capturing users' needs (e.g., users' goals behind data integration) and translating these needs into optimized designs. In other words, in relation to data analytics, existent data integration solutions are far from satisfactory.

In the context of counter-terrorism analytics, although a tremendous amount of data has been collected and structured for analytical use, a scalable data integration solution is lacking to integrate the heterogeneous terrorism data for counter-terrorism analytics. To the best of our knowledge, few designs have been successful, except for the portal TEVUS (TEVUS 2017). However, the data integration in TEVUS employs the "in-advance" approach, preparing data only for studies interested in "terrorist incidents, pre-incident activities, and extremist crimes in the United States" (START 2017). Additionally, users' goals of counter-terrorism analytics are diverse and dynamic and cannot be predefine by

data integration system developers. In this sense, the "in-advance" data integration approach used by TEVUS has serious limitations and inadequate scalability. Meanwhile, conventional solutions in the data integration literature provide no pertinent solutions. On the one hand, data integration for counter-terrorism analytics requires a framework that can accommodate various analytical goals of data users. On the other hand, existent data-management oriented methods for data integration mainly focus on leveraging data mapping to achieve data interoperability, ignoring the urgent need to take data consumers' goals into account. Therefore, enhancing current data integration solutions to consider user goals in counter-terrorism analytics is a pressing concern, especially for users who lack the technical skills that are necessary to perform data integration, because the user-centered approach can offer data integration process customized output for various users' goals of data analytics to facilitate data analysts demanding data at the front end.

Our study takes an important step towards introducing goal modeling in data integration process. We have designed the *GoOnto* ontology to annotate users' goals of data analytics. Meanwhile, our design collects, extracts and structures the metadata of terrorism datasets, presenting an ontological metadata knowledge base—*MeDaOnto*. The incorporation of the ontological metadata knowledge, with the support of NLP techniques, can facilitate data integration process. For instance, the administrative metadata information consists of principles for data collection and consequent data consumption; and the descriptive metadata information assists in understanding the topics and usages of terrorism datasets, thereby enabling goal-relevant dataset identification.

To implement our proposed data integration solution, we address a series of issues regarding heterogeneous data interoperation and management, because a large portion of counter-terrorism analytics requires access to multiple data sources. Data from various data sources are heterogeneous with respect to terminology and data structure (Brazhnik and Jones 2007). This heterogeneity complicates the process of meaningfully integrating data from multiple sources for an analytical task (Brazhnik and Jones 2007). Therefore, our implementation consists of a list of data ETL utilities, through which data heterogeneity can be addressed.

5.1 Practical Implications

This research makes practical contributions in areas of data integration and ontology design. First of all, in this research, we explicitly outline the requirements we see for next-generation data integration flows (goal-driven data integration, which is user centered) and the challenges in implementing them. In response, our design uses the goal definition method in the GQM approach to collect and model a user goal for counter-terrorism analytics, progressively identifies the goal-relevant datasets, and then implements ETL. This paper proposes conceptual and architectural designs and implements them with a physical artifact. In the design, the suite of goal components annotated in ontological representations drives the holistic process of data integration and enables our design to deliver customized data. Additionally, our research evaluates the design, testing the

feasibility, effectiveness, and efficiency of the proposed artifact, and discussed a combination of techniques for producing optimized designs.

Second, our design technically links the data integration process with users' goal of data analytics, which is a driver and determinant in initiating a data integration request. In other words, comparing with the conventional designs that often highlight and encompass the utility of data management, our proposed design, which takes users' goals of data analytics into account, is a better reflection of the real practice of data integration, thereby adding practical value for instantiating the user-centered data integration approach. While the proposed artifact is designed and implemented in the area of counter-terrorism analytics, the design principles and constructs can be applied across many industries (e.g., financial, retail, healthcare, manufacturing, and banking), especially when there is a need for self-served data analytics that requests an integration over heterogeneous data.

Third, our study provides a baseline for developing an ontology-based goal-driven data integration solution. We demonstrate that the use of goal ontology and metadata ontology enables users to discover and identify goal-relevant datasets for data integration and, accordingly, data analytics. Our proposed approach helps to reduce the semantic gap between the data consumers who perform data analytics and the terrorism datasets collected and managed for data analytics. In the area of the semantic web, the metadata ontology is contributive, serving as a knowledge base through which machines can understand terrorism datasets better; the goal ontology serves as an extensible semantic model for dataset selection. Practitioners and researchers, taking advantage of our metadata

ontology and goal ontology, can cumulatively build knowledge toward metadata-supported goal-driven data integration and management.

5.2 Theoretical Implications

This research makes valuable theoretical contributions. First, our design, following the design science research methodology, presents a theoretical framing, which can guide scholars and practitioners to build automated goal-driven data integration solutions. In addition, we identify a list of tasks to support the goal-driven data integration process; thus, our research can also serve as prescriptive knowledge toward building a technical goal-driven data integration artifact, potentially benefiting other researchers who are interested in human-centered data integration designs, especially those falling into the "on-demand" paradigm. Moreover, our theoretical framing, informed by the socio-technical thinking[2], which provides the idea at its core that IS issues can be better described and prescribed when 'social' and 'technical' perspectives are taken into account and treated as interacting components of a complex system (Lee 2004), highlights not only the techniques

[2] Socio-technical is a term devised to circumvent simplistic technological determinism or social determinism (Baxter and Sommerville 2011). Technology determinism assumes that technological development and constructs determine the social structure and cultural values, as well as human action and thought, whereas social determinism assumes that social interactions and constructs are the determinant factors.

supporting and implementing the data integration, but also the user goals that interact with the technological part to drive the data integration process.

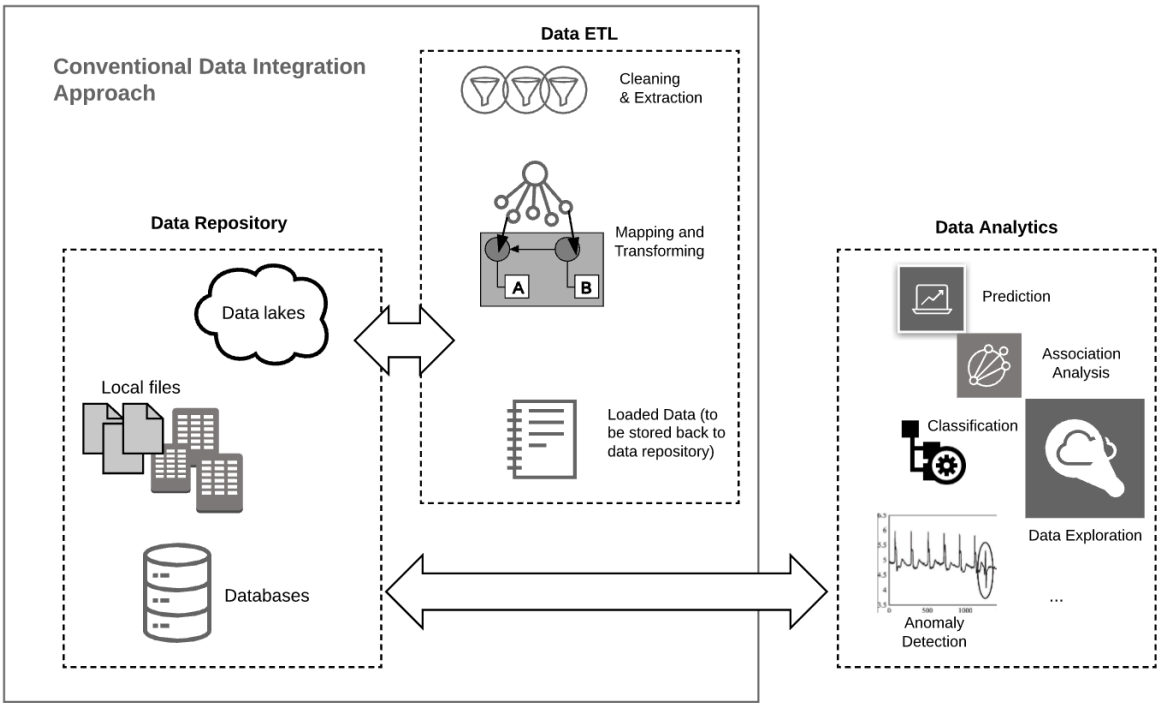


Figure 5.1 Conventional Approach of Data Integration

Second, from a process perspective, this research innovatively incorporates goal modeling into both the design theory framing and design implementation (Figure 5.2), thereby better reflecting the real practical data integration process (e.g., data consumers, driven by their goal of data analytics, request data integration and demand customized integrated data). Compared with the conventional technical designs of data integration which are data-management oriented and assume "in advance" mode (Kaza and Chen 2008; Lee et al. 2011; Nederstigt et al. 2014; Santipantakis et al. 2017), our design theory framing helps IS designers to better accommodate users' goals of data analytics into the

data integration artifact. More specifically, informed by the goal definition template from the GQM approach, our design follows "on demand" mode to propose an automated data integration solution, which consolidates heterogeneous data from multiple sources based on users' particular needs. We portray the next generation of data integration artifacts for data analytics, which are capable of eliciting users' goals and building a formal representation of these goals, identifying goal-relevant data of semantic heterogeneity in diverse sources, resolving the semantic heterogeneity issues, creating and delivering unified views of the aligned data to satisfy users' need (Figure 5.2).

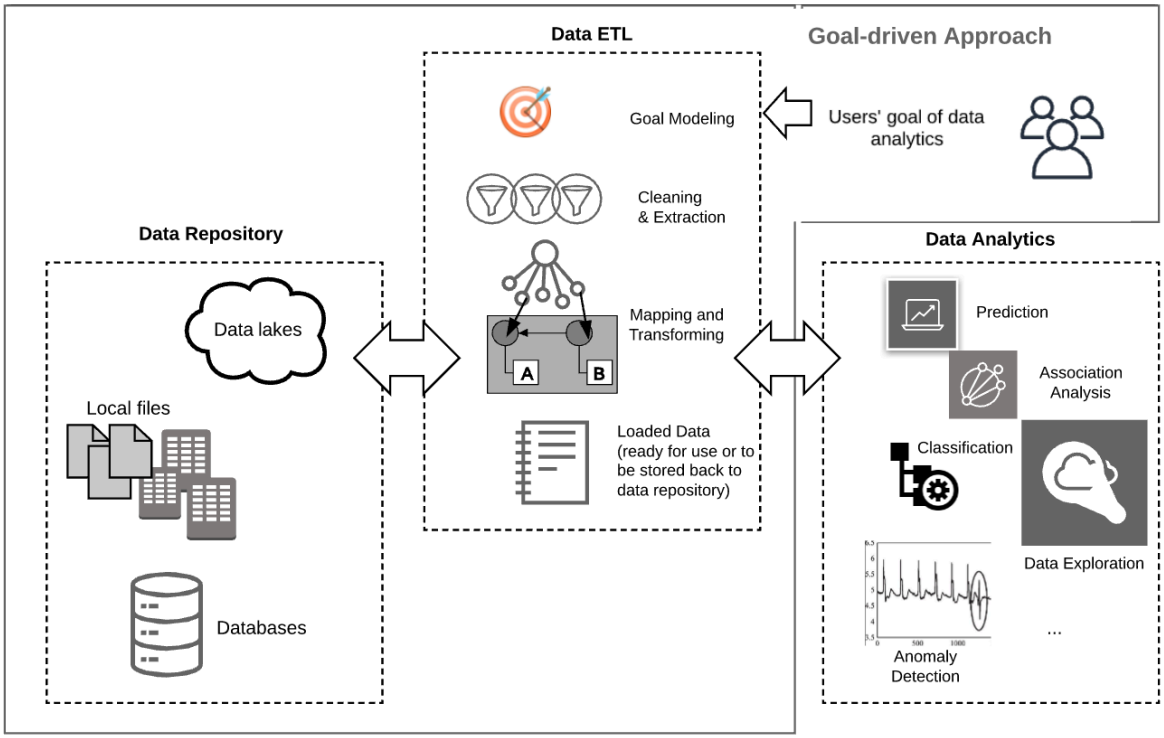


Figure 5.2 Goal-driven Approach for Data Integration

5.3 Limitations and Future Directions

This research has limitations and several improvements can be pursued to further enhance our design artifact as well as its instantiation and deployment. First, similarity calculation algorithms for mappers (e.g., data instance mappers) manifest trade-offs produced by thresholds in measures such as precision, recall, and F-measure. The sensitivity test in our study assess the effect of the threshold on system performance measures of our focus. However, other criteria can be incorporated to further enhance and extend this sensitivity analysis. Our evaluation does not consider these trade-offs systematically, but future studies may use the multiple criteria decision analysis (e.g., weighted sum, analytic hierarchy process) to evaluate these trade-offs to finalize a satisfactory solution (Li et al. 2017).

Second, data access control and authorization are important activities in managing integrated data. Access control and authorization define what access and resources we should and should not give users. Ideally, our design needs to prevent users from accessing data that they should not be able to access. In our current design and implementation, such factors as data control are not considerations. Future research can build more complicated designs to comprehensively address a list of data access control issues for better user experience and data security.

Third, our proposed design provides utilities to elicit and annotate user goals of counter-terrorism analytics, which are of critical importance in retrieving goal-relevant data to facilitate subsequent data analytics. However, the theoretical framing and

implementation in our design focus only on data integration utilities. We call for future research to extend our design and build design artifacts seamlessly combining utilities for both data integration and data analytics to fully automate the pipeline of counter-terrorism analytics. We also notify that the proposed design is implemented and evaluated in the context of counter-terrorism analytics. Future research is needed to assess its applicability in other domains.

Last, data integration for business intelligence imposes more challenging requirements than addressed in this research (e.g., data reuse, integration of unstructured data, and query within big data repository). Our study sets a direction to address these issues in Appendix C.

References

Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., and Nunamaker Jr, J. F. 2010. "Detecting Fake Websites: The Contribution of Statistical Learning Theory," *MIS Quarterly* (34:3), pp. 435-461.

Arch-int, N., and Arch-int, S. 2013. "Semantic Ontology Mapping for Interoperability of Learning Resource Systems Using a Rule-Based Reasoning Approach," *Expert Systems with Applications* (40:18), pp. 7428-7443.

Argomaniz, J., Bures, O., and Kaunert, C. 2015. "A Decade of Eu Counter-Terrorism and Intelligence: A Critical Assessment," *Journal Intelligence and National Security* (30:2-3), pp. 191-206.

Arms, W. Y., Blanchi, C., and Overly, E. A. 1997. "An Architecture for Information in Digital Libraries," *D-Lib Magazine* (3:2).

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P., and Wouters, P. 2004. "Promoting Access to Public Research Data for Scientific, Economic, and Social Development," *Data Science Journal* (3), pp. 135-152.

Asal, V., and Rethemeyer, R. K. 2008. "Dilettantes, Ideologues, and the Weak: Terrorists Who Don't Kill," *Conflict Management and Peace Science* (25:3), pp. 244-263.

Baca, M. 2008. *Introduction to Metadata*. Getty Publications.

Bahgat, K., and Medina, R. M. 2013. "An Overview of Geographical Perspectives and Approaches in Terrorism Research," *Perspectives on Terrorism* (7:1).

Basili, V., Trendowicz, A., Kowalczyk, M., Heidrich, J., Seaman, C., Münch, J., and Rombach, D. 2014. *Aligning Organizations through Measurement: The Gqm+ Strategies Approach*. Springer.

Basili, V. R., Caldiera, G., and Rombach, H. 1994. "The Goal Question Metric Approach," in *Encyclopedia of Software Engineering*. Wiley.

Baxter, G., and Sommerville, I. 2011. "Socio-Technical Systems: From Design Methods to Systems Engineering," *Interacting with Computers* (23:1), pp. 4-17.

Behlendorf, B., and LaFree, G. 2014. "Building a Unified Infrastructure for Data Integration on Political Violence and Conflict." Retrieved Feb. 2nd, 2018, from http://www.start.umd.edu/research-projects/building-unified-infrastructure-data-integration-political-violence-and-conflict

Brazhnik, O., and Jones, J. F. 2007. "Anatomy of Data Integration," *Journal of Biomedical Informatics* (40:3), pp. 252-269.

Bruns, A., Kornstadt, A., and Wichmann, D. 2009. "Web Application Tests with Selenium," *IEEE Software* (26:5), pp. 88-91.

Chen, H., Chiang, R. H., and Storey, V. C. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly* (36:4), pp. 1165-1188.

Choi, D., Ko, B., Kim, H., and Kim, P. 2014. "Text Analysis for Detecting Terrorism-Related Articles on the Web," *Journal of Network and Computer Applications* (38), pp. 16-21.

Chowdhuri, R., Yoon, V. Y., Redmond, R. T., and Etudo, U. O. 2014. "Ontology Based Integration of Xbrl Filings for Financial Decision Making," *Decision Support Systems* (68), pp. 64-76.

Clark, J. 1999. "Xsl Transformations (Xslt)."

Conlon, S. J., Abrahams, A. S., and Simmons, L. L. 2015. "Terrorism Information Extraction from Online Reports," *Journal of Computer Information Systems* (55:3), pp. 20-28.

Cronin, A. K. 2003. "Behind the Curve: Globalization and International Terrorism," *International Security* (27:3), pp. 30-58.

Cutter, S. L. 2014. *The Geographical Dimensions of Terrorism*. Routledge.

Daraio, C., Lenzerini, M., Leporelli, C., Moed, H. F., Naggar, P., Bonaccorsi, A., and Bartolucci, A. 2016. "Data Integration for Research and Innovation Policy: An Ontology-Based Data Management Approach," *Scientometrics* (106:2), pp. 857-871.

Dataverse. 2018. "Harvard Dataverse." from https://dataverse.harvard.edu/dataverse/start

Ding, F., Ge, Q., Jiang, D., Fu, J., and Hao, M. 2017. "Understanding the Dynamics of Terrorism Events with Multiple-Discipline Datasets and Machine Learning Approach," *PloS one* (12:6), p. e0179057.

Dugan, L., LaFree, G., and Piquero, A. R. 2005. "Testing a Rational Choice Model of Airline Hijackings," *International Conference on Intelligence and Security Informatics*: Springer, pp. 340-361.

Ebrahimipour, V., Rezaie, K., and Shokravi, S. 2010. "An Ontology Approach to Support Fmea Studies," *Expert Systems with Applications* (37:1), pp. 671-677.

Ellis, J. O. 2008. "Countering Terrorism with Knowledge," in *Terrorism Informatics: Integrated Series in Information Systems*. pp. 141-155.

Elovici, Y., Shapira, B., Last, M., Zaafrany, O., Friedman, M., Schneider, M., and Kandel, A. 2008. "Content-Based Detection of Terrorists Browsing the Web Using an Advanced Terror Detection System (Atds)," *Terrorism Informatics*), pp. 365-384.

Etudo, U., Yoon, V., and Liu, D. 2017. "Financial Concept Element Mapper (Fincem) for Xbrl Interoperability: Utilizing the M 3 Plus Method," *Decision Support Systems* (98), pp. 36-48.

Fensel, D., Ding, Y., Omelayenko, B., Schulten, E., Botquin, G., Brown, M., and Flett, A. 2001. "Product Data Integration in B2b E-Commerce," *IEEE Intelligent Systems* (16:4), pp. 54-59.

Forest, J. J. 2012. "Kidnapping by Terrorist Groups, 1970-2010: Is Ideological Orientation Relevant?," *Crime & Delinquency* (58:5), pp. 769-797.

Gilliland, A. J. 2008. "Setting the Stage," in *Introduction to Metadata*. Los Angeles: Getty Publications.

Gordon, A. 2004. "The Effect of Database and Website Inconstancy on the Terrorism Field's Delineation," *Studies in Conflict & Terrorism* (27:2), pp. 79-88.

Gregor, S., and Hevner, A. R. 2013. "Positioning and Presenting Design Science Research for Maximum Impact," *MIS Quarterly* (37:2).

Gregor, S., and Jones, D. 2007. "The Anatomy of a Design Theory," *Journal of the Association for Information Systems* (8:5), pp. 312-323,325-335.

Hand, D. J. 2007. "Principles of Data Mining," *Drug safety* (30:7), pp. 621-622.

HARVARD. 2018. "Big Allied and Dangerous (Baad) Database 1 - Lethality Data, 1998-2005version 3.0." Retrieved Feb, 1st, 2018, from https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl%3A1902.1/16062

Hawalah, A., and Fasli, M. 2015. "Dynamic User Profiles for Web Personalisation," *Expert Systems with Applications* (42:5), pp. 2547-2569.

Hevner, A., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75-105.

Kaza, S., and Chen, H. 2008. "Evaluating Ontology Mapping Techniques: An Experiment in Public Safety Information Sharing," *Decision Support Systems* (45:4), pp. 714-728.

Kotonya, G. 1999. "Practical Experience with Viewpoint-Oriented Requirements Specification," *Requirements Engineering* (4:3), pp. 115-133.

Kouskouridas, R., Amanatiadis, A., Chatzichristofis, S. A., and Gasteratos, A. 2015. "What, Where and How? Introducing Pose Manifolds for Industrial Object Manipulation," *Expert Systems with Applications* (42:21), pp. 8123-8133.

Kuechler, B., and Vaishnavi, V. 2008. "On Theory Development in Design Science Research: Anatomy of a Research Project," *European Journal of Information Systems* (17:5), pp. 489-504.

Kurlander, N. 2005. "Fighting Crime and Terrorism through Data Integration," *The Police Chief* (72:2), pp. 29-30.

Lake, D. A. 2002. "Rational Extremism: Understanding Terrorism in the Twenty-First Century," *Dialogue IO* (1:1), pp. 15-28.

Lapatas, V., Stefanidakis, M., Jimenez, R. C., Via, A., and Schneider, M. V. 2015. "Data Integration in Biological Research: An Overview," *Journal of Biological Research-Thessaloniki* (22:1), p. 9.

Larose, D. T. 2005. *Introduction to Data Mining*. Wiley Online Library.

Lee, A. 2001. "Editor's Comment," *MIS Quarterly* (25:1), pp. iii-vii.

Lee, A. 2004. "Thinking About Social Theory and Philosophy for Information Systems," in *Social Theory and Philosophy for Information Systems*. Chichester, UK: John Wiley & Sons, pp. 1-26.

Lee, M. C., Tsai, K. H., and Hsieh, T. C. 2011. "A Multi-Strategy Knowledge Interoperability Framework for Heterogeneous Learning Objects," *Expert Systems with Applications* (38:5), pp. 4945-4956.

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., Ponting, C. P., and Bork, P. 2004. "Smart 4.0: Towards Genomic Data Integration," *Nucleic acids research* (32:suppl_1), pp. D142-D144.

Li, Y., Thomas, M. A., and Osei-Bryson, K.-M. 2017. "Ontology-Based Data Mining Model Management for Self-Service Knowledge Discovery," *Information Systems Frontiers* (19:4), pp. 925-943.

Malucelli, A., Palzer, D., and Oliveira, E. 2006. "Ontology-Based Services to Help Solving the Heterogeneity Problem in E-Commerce Negotiations," *Electronic Commerce Research and Applications* (5:1), pp. 29-43.

Mashiko, Y., and Basili, V. R. 1997. "Using the Gqm Paradigm to Investigate Influential Factors for Software Process Improvement," *Journal of Systems and Software* (36:1), pp. 17-32.

Milo, T., and Zohar, S. 1998. "Using Schema Matching to Simplify Heterogeneous Data Translation," *VLDB*: Citeseer, pp. 24-27.

Nederstigt, L. J., Aanen, S. S., Vandic, D., and Frasincar, F. 2014. "Floppies: A Framework for Large-Scale Ontology Population of Product Information from Tabular Data in E-Commerce Stores," *Decision Support Systems* (59), pp. 296-311.

NEH. 2018. from https://www.neh.gov/files/grants/data_management_plans_2018.pdf (retrieved on Feb 8th, 2018)

NISO. 2004. *Understanding Metadata*. MD, USA: NISO (National Information Standards Organization) Press.

Oh, O., Agrawal, M., and Rao, H. R. 2011. "Information Control and Terrorism: Tracking the Mumbai Terrorist Attack through Twitter," *Information Systems Frontiers* (13:1), pp. 33-43.

Piegorsch, W. W., Cutter, S. L., and Hardisty, F. 2007. "Benchmark Analysis for Quantifying Urban Vulnerability to Terrorist Incidents," *Risk Analysis* (27:6), pp. 1411-1425.

Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X., and Jensen, L. J. 2015. "Diseases: Text Mining and Data Integration of Disease–Gene Associations," *Methods* (74), pp. 83-89.

Popp, R., Armour, T., and Numrych, K. 2004. "Countering Terrorism through Information Technology," *Communications of the ACM* (47:3), pp. 36-43.

Provost, F., and Fawcett, T. 2013. "Data Science and Its Relationship to Big Data and Data-Driven Decision Making," *Big Data* (1:1), pp. 51-59.

pypi. 2019. "Textdistance." Retrieved 11th, Jan, 2019, from https://pypi.org/project/textdistance/

Rodríguez-García, M. Á., Valencia-García, R., García-Sánchez, F., and Samper-Zapater, J. J. 2014. "Ontology-Based Annotation and Retrieval of Services in the Cloud," *Knowledge-Based Systems* (56), pp. 15-25.

Santipantakis, G., Kotis, K., and Vouros, G. A. 2017. "Obdair: Ontology-Based Distributed Framework for Accessing, Integrating and Reasoning with Data in Disparate Data Sources," *Expert Systems with Applications* (90), pp. 464-483.

Schmid, A. P., and Jongman, A. J. 1988. *Political Terrorism: A New Guide to Actors, Authors, Concepts, Data Bases, Theories, and Literature*. Transaction Publishers.

Sen, A. 2004. "Metadata Management: Past, Present and Future," *Decision Support Systems* (37:1), pp. 151-173.

Silke, A. 2004. *Research on Terrorism: Trends, Achievements and Failures*. Routledge.

Simon, H. A. 1996. *The Sciences of the Artificial*. Cambridge, MA: MIT press.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., and Mungall, C. J. 2007. "The Obo Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration," *Nature Biotechnology* (25:11), p. 1251.

Stanton, J. A. 2013. "Terrorism in the Context of Civil War," *The Journal of Politics* (75:4), pp. 1009-1022.

START. 2017. "Tevus Portal." from http://www.start.umd.edu/tevus-portal

Taipale, K. A. 2003. "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data," *Columbia Science and Technology Law Review* (5:2), pp. 1-83.

Takeda, H., Veerkamp, P., and Yoshikawa, H. 1990. "Modeling Design Process," *AI Magazine* (11:4), p. 37.

TEVUS. 2017. "Tevus Analyst Portal." from https://tap.cast.uark.edu/about

Tutun, S., Khasawneh, M. T., and Zhuang, J. 2017. "New Framework That Uses Patterns and Relations to Understand Terrorist Behaviors," *Expert Systems with Applications* (78), pp. 358-375.

Uschold, M., and Gruninger, M. 1996. "Ontologies: Principles, Methods and Applications," *The Knowledge Engineering Review* (11:2), pp. 93-136.

van Solingen, R., Basili, V., Caldiera, G., and Rombach, H. D. 2002. "Goal Question Metric (Gqm) Approach," in *Encyclopedia of Software Engineering*.

van Solingen, R., and Berghout, E. 1999. *The Goal/Question/Metric Method: A Practical Guide for Quality Improvement of Software Development*. McGraw-Hill.

Venkatesh, V., Aloysius, J. A., Hoehle, H., and Burton, S. 2017. "Design and Evaluation of Auto-Id Enabled Shopping Assistance Artifacts in Customers'mobile Phones: Two Retail Store Laboratory Experiments," *MIS Quarterly* (41:1).

Wache, H., Voegele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., and Hübner, S. 2001. "Ontology-Based Integration of Information-a Survey of Existing Approaches," *IJCAI-01 workshop: ontologies and information sharing*: Citeseer, pp. 108-117.

Wang, J., Li, G., and Feng, J. 2014. "Extending String Similarity Join to Tolerant Fuzzy Token Matching," *ACM Transactions on Database Systems (TODS)* (39:1), p. 7.

Wang, J., Lu, J., Zhang, Y., Miao, Z., and Zhou, B. 2009. "Integrating Heterogeneous Data Source Using Ontology," *Journal of Software* (4:8), pp. 843-850.

Widom, J. 1996. "Integrating Heterogeneous Databases: Lazy or Eager?," *ACM Computing Surveys* (28A:4).

Wong, A. K. Y., Ray, P., Parameswaran, N., and Strassner, J. 2005. "Ontology Mapping for the Interoperability Problem in Network Management," *IEEE Journal on Selected Areas in Communications* (23:10), pp. 2058-2068.

Xiang, Z., Mungall, C., Ruttenberg, A., and He, Y. 2011. "Ontobee: A Linked Data Server and Browser for Ontology Terms," *ICBO*.

Xu, S., and Zhang, J. 2008. "Data Distortion Methods and Metrics in a Terrorist Analysis System," in *Terrorism Informatics: Knowledge Management and Data Mining for Homeland Security*. pp. 347-364.

Yang, Q., and Wu, X. 2006. "10 Challenging Problems in Data Mining Research," *International Journal of Information Technology & Decision Making* (5:04), pp. 597-604.

Appendices:

APPENDIX A: Highlights

1. Research Question

How can we develop a goal-driven data integration framework for counter-terrorism analytics?

This study follows design science methodology to design an artifact for terrorism data integration.

2. Motivation

- Counter-terrorism has become increasingly important in the past decade as the terrorism crisis surges across the world.
- To propel the scientific study of terrorism and counter-terrorism analytics, academics and practitioners take advantages of valuable datasets as the starting point for decision-making. However, the knowledge and clues that are essential for identifying and analyzing terrorist activities may spread across semantically heterogeneous data sources.
- Consolidating appropriate datasets to enhance the technological capability of counter-terrorism has become vitally important.
- There is a lack of research on data integration for counter-terrorism analytics.

3. Special Design Features

- Goal-driven—In counter-terrorism analytics, the analysts may have various goals, interests, research emphasis, and even research scopes. In other words, effective data integration in the terrorism domain requires appropriately modeling the user's goals, which can assist with effectively identifying and retrieving the data necessary for terrorism analytics.
- NLP supported—Hosting websites of terrorism datasets contain rich metadata in free-text format that are essential to understanding the terrorism data. Our artifact leverages the NLP techniques and semantic web to enhance data integration performance.

4. Research Questions

Chapter 1 – P7 Table 1.1

5. Our Design Objectives

The effective development of a goal-driven data integration solutions for counter-terrorism analytics requires tackling four challenging tasks:

- Metadata extraction and annotation, which extracts and annotates the metadata of terrorism datasets, allowing the data integration process to understand the terrorism datasets and facilitating search and retrieval of the terrorism data in a data repository;
- Goal elicitation and annotation, which formulates and annotates a clear vision of user goals of data analytics and ensures that these goals can be well understood and incorporated into the data integration process;
- Goal-dataset matching, which assesses whether a terrorism dataset in the data repository is relevant to a specific analytical goal by calculating the similarities

between a given analytical goal and the metadata of terrorism datasets, aiming to identify the terrorism data that matches a user's goal; and

- Schema mapping and instance mapping, which retrieves the relevant datasets and consolidates them, addressing semantic heterogeneity in the goal-relevant data (the alignment of the four design objectives with the seven research questions are shown in Table 1.1).

Chapter 3 – P10

6. System Architecture

Chapter 3 – P11

APPENDIX B: Metadata Example

Example of Metadata Information Crawled from HARVARD (2018)

Metadata Item	**Example (BAAD) From Harvard Dataverse Website**	**Metadata Type**
Title	Big Allied and Dangerous (BAAD) Database 1 - Lethality Data, 1998–2005	Descriptive
Author	Victor Asal (University at Albany, State University of New York. Rockefeller College of Public Affairs & Policy. Assistant Professor of Political Science.)	Administrative
	R. Karl Rethemeyer (University at Albany, State University of New York. Rockefeller College of Public Affairs & Policy. Assistant Professor of Public Administration & Policy.)	

	Ian Anderson (University at Albany, State University of New York. Rockefeller College of Public Affairs & Policy. Project on Violent Conflict. Research Director.)	
Keyword	Fatalities	Descriptive
	Ideologies	
	Religion	
	Terrorist groups	
Topic Classification	Conflict	Descriptive
Related Publication	Asal, V. H., & Rethemeyer, R. K. (2008). Dilettantes, ideologues and the weak: Terrorists who don't kill. *Conflict Management and Peace Science, 25*(3), 244–263.	Descriptive
	Asal, V. H., & Rethemeyer, R. K. (2008). The nature of the beast: Terrorist: The organizational and network characteristics of organizational lethality. *Journal of Politics, 70*(2), 437–449.	
Notes	The study uses primarily media sources, which could vary in their definition of a terrorist attack, and in their data collection methods. The study only accounted for slightly more than half of all fatalities attributed to acts of terrorism between 1998 and 2005. The other significant number of fatalities are acts that went unclaimed, either by terrorist organizations in the database, unknown terrorist entities, or terrorist entities not classified as organizations by the database and therefore would not be part of the study. Also, some organizations may have been excluded because they were deemed religious in nature. This would mean the findings on religion and lethality are understated. Finally, of the 499 organizations that committed an act, MIPT had no information or information that could be duplicative on 104 of	Descriptive

	the organizations, leaving the possibility of inaccurate data.;	
Time Period Covered	Start: 1998-01-01; End: 2005-12-31	Descriptive
Unit of Analysis	Terrorist organizations	Descriptive
Producer	Project on Violent Conflict	Administrative
Production Date	2009	Administrative
Grant Information	U.S. Department of Homeland Security Science and Technology Directorate, Office of University Programs, administered through the U.S. Office of Naval Research.: N000410510629	Administrative
Distributor	Project on Violent Conflict	Administrative
Distribution Date	2009	Administrative
Date of Deposit	6/6/11	Administrative
Date of Collection	Start: 2007; End: 2007	Administrative
Schema Description	Available in the coding book file	Structural

Notes: The information in Column 2 are from BAAD dataset hosting website, crawled and extracted by MetaMM

APPENDIX C: Design Potentials

Potentials	**Future Research or Deployment**
Improve performance	Develop a hybrid algorithm for instance mapping

Enhance security and reuse	Incorporate access control and authorization, using AWS data lake S3 and AWS IAM role
Improve ETL performance	Incorporate AWS Glue
Enhance scalability	Utilize micro services for deployment
Enhance data inclusiveness	Collect more datasets and metadata

APPENDIX D: Algorithm Performance Metrics (F-measure)

Threshold	F-Measure Scores					
	Jaccard	**Cosine**	**Masi**	**Longest Common Substring**	**Hamming**	**Levenshtein**
0.65	0.041	0.012	0.041	0.012	0.588	0.141
0.66	0.053	0.015	0.053	0.015	0.625	0.158
0.67	0.070	0.019	0.070	0.019	0.656	0.179
0.68	0.075	0.023	0.075	0.023	0.667	0.198
0.69	0.092	0.029	0.092	0.029	0.690	0.223
0.7	0.105	0.037	0.105	0.037	0.690	0.256
0.71	0.137	0.048	0.137	0.048	0.690	0.286
0.72	0.157	0.062	0.157	0.062	0.755	0.313
0.73	0.194	0.083	0.194	0.083	0.755	0.356
0.74	0.241	0.102	0.241	0.102	0.755	0.391
0.75	0.245	0.127	0.245	0.127	0.769	0.433

0.76	0.305	0.162	0.305	0.162	0.784	0.491
0.77	0.354	0.212	0.354	0.212	0.816	0.571
0.78	0.385	0.265	0.385	0.265	0.833	0.642
0.79	0.486	0.319	0.486	0.319	0.809	0.714
0.8	0.510	0.369	0.510	0.369	0.809	0.769
0.81	0.602	0.424	0.602	0.424	0.783	0.787
0.82	0.633	0.471	0.633	0.471	0.783	0.821
0.83	0.676	0.615	0.676	0.615	0.783	0.868
0.84	0.704	0.657	0.704	0.657	0.783	0.885
0.85	0.746	0.719	0.746	0.719	0.783	0.885
0.86	0.814	0.772	0.814	0.772	0.756	0.880
0.87	0.828	0.786	0.828	0.786	0.756	0.880
0.88	0.786	0.815	0.786	0.815	0.756	0.880
0.89	0.778	0.824	0.778	0.824	0.756	0.857
0.9	0.800	0.800	0.800	0.800	0.756	0.833
0.91	0.776	0.800	0.776	0.800	0.756	0.833
0.92	0.792	0.800	0.792	0.800	0.756	0.833
0.93	0.792	0.800	0.792	0.800	0.756	0.833
0.94	0.766	0.809	0.766	0.809	0.756	0.809
0.95	0.739	0.809	0.739	0.809	0.756	0.809
0.96	0.739	0.783	0.739	0.783	0.756	0.783
0.97	0.739	0.756	0.739	0.756	0.756	0.756

0.98	0.739	0.756	0.739	0.756	0.756	0.756
0.99	0.739	0.756	0.739	0.756	0.756	0.756
1	0.739	0.756	0.739	0.756	0.756	0.756