APPLIED RESEARCH LAB FOR INTELLIGENCE AND SECURITY

START
NATIONAL CONSORTIUM FOR THE STUDY OF TERRORISM AND RESPONSES TO TERRORISM

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

Datasets for Modeling and Mitigating Insider Risk Final Report and Codebook

September 30, 2021

Steve S. Sin[1]*, Kathryn A. Lindquist[1]
[1]National Consortium for the Study of Terrorism and Responses to Terrorism
*corresponding author, sinss@umd.edu

UWT
UNCONVENTIONAL WEAPONS AND TECHNOLOGY

The Applied Research Lab for Intelligence and Security (ARLIS) is The University Affiliated Research Center (UARC) for Sociotechnical Systems and Advantage in the Human Domain at the University of Maryland

ARLIS
7005 52nd Avenue
College Park, Maryland

EXECUTIVE SUMMARY

The breadth and scope of different variables that surround or are integral to understanding the risks that insiders may pose creates a complexity that requires additional applied and theoretical research across a wide range of topics and disciplines. Though data availability and privacy issues remain central issues in the field, data valuable for the study of insider threats, insider risk, detection, and mitigation do exist. The seedling project "Datasets for Modeling and Mitigating Insider Risk" (D-MInR) was designed specifically to identify extant datasets that may be useful for future Insider Risk efforts for the US government (USG) and its allies and partners, and to characterize those datasets in terms of their contents, location, relevance, accessibility, and other key parameters. D-MInR is the first known effort to track, catalogue, or characterize these potential resources across a wide variety of attributes and within multiple disciplines.

The result of the project was the development and completion of D-MInR Catalog Matrix and its accompanying final report and codebook. The catalog matrix lists the available data resources for insider threat and insider risk studies. The searchable and filterable catalog contains 107 total entries, with 45 in the main catalog and 62 in additional tabs to map more fully the data landscape. Information about each data source, including links and information about access as well as contents, data format, and other attributes, are coded for each resource in the catalog. The accompanying final report and codebook provides the approach and methodology used to develop the D-MInR Catalog Matrix, a quick start guide for using the catalog, and the codebook for the catalog.

One of the major findings from the catalog development was that there is a limited number of high quality, public-use datasets for the study of insider threat/ insider risk. While limited in numbers, these datasets do possess some important strengths and have contributed meaningfully to the advancement of the field; however, many are over a decade old and few include psychosocial measures alongside technical data, making them only partially suited to addressing the challenges of insider threat and insider risk circa 2021.

Another finding was that both the USG and a wide variety of other organizations have access to data on their personnel but limited capacity for turning that data into useful information about insider threat/insider risk. This is a problem shared by organizations writ large: Two in three organizations (66%) report a struggle with turning volumes of security activity and event data being collected into "intelligent, actionable insights." Many commercial products purportedly provide services to address this challenge, but the problem of integrating psychosocial information into tools that emphasize cyber security is largely unresolved.

As we anticipate that D-MInR Catalog Matrix will help researchers to improve their research designs and strategies, by leveraging these datasets, to conduct exploratory research (e.g., develop new hypotheses, discover new patterns, build new models or tools) and/or confirmatory research (e.g., test and validate existing hypotheses, tools, and models) that are crucial to enhancing the USG's understanding of and ability to mitigate insider risk, one should entertain the continuation of this seedling project so that the Catalog Matrix can be maintained and kept up to date. Additionally, to increase the accessibility of the D-MInR Catalog Matrix and its associated products to the countering insider threat enterprise, one should consider transitioning the catalog into a format and onto a platform that is conducive to wider research access.

Contents

INTRODUCTION

Insider threats are a real and persistent risk factor for organizations, especially organizations engaged in national security relevant activities. Carnegie Mellon (CERT) defines insider threat as "the potential for an individual who has or had authorized access to an organization's assets to use their access, either maliciously or unintentionally, to act in a way that could negatively affect the organization."[1] The breadth and scope of different variables that surround or are integral to understanding the risks that insiders may pose creates a complexity that requires additional applied and theoretical research across a wide range of topics and disciplines. Though data availability and privacy issues remain central in the field, valuable data for the study of insider threats, insider risk, detection, and mitigation do exist. Some of these are well known, readily available, and frequently used in research, such as the rich and valuable resources available through CERT. Other aspects of the data landscape for insider risk and insider threat research are more poorly mapped. At best, this reduces the efficiency of efforts to develop empirically based work in this field, and at worst, opportunities for impactful research and development are missed entirely.

The seedling project "Datasets for Modeling and Mitigating Insider Risk" (D-MInR) was designed specifically to identify extant datasets that may be useful for future Insider Risk efforts for the US government (USG) and its partners, and to characterize those datasets in terms of their contents, location, relevance, accessibility, and other key parameters. D-MInR is the first known effort to track, catalogue, and/or characterize these potential resources across a wide variety of attributes and within multiple disciplines. In so doing, D-MInR will help researchers to improve their research designs and strategies, by leveraging these datasets, to conduct exploratory research (e.g., develop new hypotheses, discover new patterns, build new models or tools) and/or confirmatory research (e.g., test and validate existing hypotheses, tools, and models) that are crucial to enhancing the USG's understanding of and ability to mitigate insider risk.

The project adopts a broad understanding of insider risk, and makes the key assumption that behavioral and social factors, not just technical or internal systems-based ones, impact the degree and nature of insider threats. It provides an easy-to-navigate snapshot of the landscape of potentially available sources for those seeking to test and develop theories, tools, and models to better assess insider risk, and detect and mitigate threats. Most importantly, the D-MInR project enables applied research efforts that do more than provide narrow answers to specific questions ("Does algorithm X increase the detection of anomaly Y in specific context Z?") and can instead undergird a growing body of empirically-based, data-driven knowledge about insider risk in interconnected human and technical networks that accounts for a broad range of sociotechnical factors.

[1] Briguglio, Frank. 2019. "Insider Threat: How to Properly Govern Identities & Identify Nefarious Actors". Security. https://www.securitymagazine.com/articles/91051-insider-threat-how-to-properly-govern-identities-identify-nefarious-actors.

APPROACH AND METHODOLOGY

Catalog Methodology

The search procedures for surveying possible resources and identifying datasets used a phased, domain-based approach. Initially, we established a set of broad domains, to include academia, think tank, USG, commercial, and online resources and data repositories. Preliminary reviews of commercial resources suggested that data identified in these searches were unlikely to be publicly available to future researchers; subsequent interviews with a number of researchers in the field as well as USG personnel confirmed this. Within each of the other broad domain, a few specific avenues likely to yield the best and more relevant results during the projects brief period of performance were identified and systematically reviewed.

Academic and Think Tank Domain

Academic review articles: There are a number of review articles on insider threat which provide surveys and overviews of either the field as a whole or certain areas of the existing body of applied research, such as intrusion detection.[2,3] One of these, the review of Homoliak et al., (2019)[4], served as an initial review of the academic works in this field. Many of the entries in the main catalog are from this domain, including the authors' appendix. Additional review articles were not systematically pursued as preliminary reviews indicated that cited publications had either already been cataloged or did not actually contain replication or publicly available datasets. Numerous resources associated with think tanks and research institutes like CERT, MITRE, and the Lincoln Lab were identified using this approach.

START's Data and Past Research: START has a long history of research, including the development, maintenance, and housing of datasets and resources with relevance to national security. A number of START's own datasets were included as a result of reviewing in this domain. Recent START work to catalog datasets for a different topic area also resulted in a number of the entries on the "Rich Data Sources" tab (See Rich Data Sources section below for explanation of this tab). Additionally, past data cataloging efforts associated with the Counter-Terrorism Net Assessment Data Structure (CT-NEADS)[5] project were reviewed and relevant entries in that catalog were included.

---

[2] See Ivan Homoliak, Flavio Toffalini, Juan Guarnizo, Yuval Elovici, and Martín Ochoa. 2019. "Insight Into Insiders and IT: A Survey of Insider Threat Taxonomies, Analysis, Modeling, and Countermeasures". ACM Comput. Surv. 52(2), Article 30.

[3] Al-Mhiqani, Mohammed N., Rabiah Ahmad, Z. Zainal Abidin, Warusia Yassin, Aslinda Hassan, Karrar H. Abdulkareem, Nabeel S. Ali, and Zahri Yunos. 2020. "A Review of Insider Threat Detection: Classification, Machine Learning Techniques, Datasets, Open Challenges, and Recommendations" Applied Sciences 10(15): 5208.

[4] Ivan Homoliak, Flavio Toffalini, Juan Guarnizo, Yuval Elovici, and Martín Ochoa. 2019. "Insight Into Insiders and IT: A Survey of Insider Threat Taxonomies, Analysis, Modeling, and Countermeasures". ACM Comput. Surv. 52(2), Article 30.

[5] Koven, Barnett S., Katy Lindquist, and Max Erdemandi. 2020. "Counterterrorism Net Assessment Data Structure." College Park, MD: START (April). https://www.start.umd.edu/pubs/START_CTNeads_Overview_2020.pdf.

Online Data Repositories Domain

ICPSR Social Science Data Repository. The Inter-University Consortium for Political and Social Research (ICPSR) is an international consortium with several hundred member institutions. The ICPSR Social Science Data Repository at the University of Michigan also houses one of the largest repositories of available data from empirical research on social issues. As part of the cataloging process, the research team performed keyword searches on its archive of more than 250,000 files, which includes special collections on a variety of fields including criminal justice and terrorism.[6] Several promising resources, including the STARRS Army collection of studies, were identified in this repository.

Open source computer science/machine learning data repositories. Large datasets for training and testing AI and machine learning algorithms have been made available to the general public through these tools. The review of this domain included queries of Kaggle, Github, Datamart, the Institute for Electrical and Electronics Engineers (IEEE) Dataport, Zenodo, FigShare, Dryad, and the University of California-Irvine Machine Learning Repository. Google's own Database search function remains in beta mode, pulling from a range of other resources, including those indicated here. These databases were searched using a list of key terms developed by the research team. **Table 1**, below, provides the search terms used for the search:

Table 1: Search Terms Utilized

Search Terms			
insider threat	vulnerability	keystroke	threat analysis
malicious insider threat	masquerade	insider threat masquerader	anomaly detection
insider threat anomaly detection	intrusion detection	insider threat intrusion detection	insider threat psychological
insider threat psychological	unix commands	malicious cyber activity	hacking
user search behavior	search behavior	workplace violence	workplace shooting

Relatively few data resources were identified from this portion of the online repository domain. IEEE's data repository returned just one hit despite the dozens upon dozens of journals, conferences, and workshops related to insider threat that are affiliated with or hosted by the organization and its members. Datamart was queried but searches yielded no results, and the research team's direct communication with Datamart ultimately confirmed that they do not have any datasets relevant to insider threat analysis. Similarly, Zenodo returned 15 publications and 0 datasets for the term "insider threat." FigShare initially yielded hits for 55 datasets, of which just two were related to insider threat; one was a table for a review article identified in the first domain, and the other was a recent upload/mirror of CERT's test dataset. Dryad returned no hits on "insider threat"

[6] "About ICPSR." ICPSR: Sharing Data to Advance Science. Accessed September 17, 2021. https://www.icpsr.umich.edu/web/pages/about/.

U.S. Government Resources Domain

USG Administrative Data, USG Insider Threat Resources, and **Other Resources Identified via Interviews:** Internal subject matter expertise at ARLIS and START, and information and insights from interviews with data users, researchers, and other stakeholders both in and outside the USG were used to inform the review of this domain. Most of the datasets identified for this domain were included in the dataset catalog under the "Vetting and Investigative Sources" tab (See Vetting and Investigative Sources section below for explanation of this tab). Others were included under the "Main Catalog" tab (See Main Catalog section below for explanation of this tab).

Coding Methodology

Inclusion/Exclusion Criteria

Accessibility and relevance were the primary dimensions in determining inclusion and exclusion of a dataset in the catalog. Sources that were both highly relevant (as evidenced by frequent use or citation in the insider risk field) and that had data generally available to researchers were assigned to the "Main Catalog" tab of the catalog. Published studies that were relevant (focused on insider threat) but did not have accessible or publicly available data were assigned to the "Academic Publications without Available Data" tab of the catalog. Conversely, rich data sources, especially in the social sciences, which could be employed in future insider risk and insider threat studies, were included in the "Rich Data Sources" tab of the catalog. **Table 2**, below, provides a visualization of the inclusion criteria for the dataset catalog:

Table 2: Dataset Catalog Inclusion Criteria

		Relevance	
		High	Low
Accessibility	Generally Accessible to Researchers	**Include:** Main Catalog	**Limited Inclusion:** Rich Sources
	Generally Unavailable to Researchers	**Limited Inclusion:** Academic Publications without Available Data	**Exclude**

Main Catalog

This tab contains datasets that are broadly and publically available. As such, they are coded in the greatest detail, as researchers were able to access data and/or codebooks to provide a thorough overview of the resource. This tab contains commonly used datasets in insider risk and insider threat research, such as the CERT suite of datasets, as well as valuable social science data that can be used in isolation for theory development, or to augment studies that aims to test existing theories, typologies, and profiles. There are low barriers to accessibility of all of the datasets listed in the Main Catalog tab, though terms of use must be followed and researchers may have to obtain permissions. In nearly all cases, however, these barriers were deemed sufficiently low (i.e. terms of use that prevented use for profit, filling out a "contact us" form, etc) that they should generally be considered free and available resources for insider threat/insider risk researchers.[7]

[7] Some items in the main catalog were temporarily unavailable, but had been made available in the past and are likely to be available again in the future. For instance, [MN_09 and MN_11].

An exception of note is the Army STARRS project data. While these data are available for research purposes, there is a more extensive set of procedures for obtaining, using, and analyzing the data, which require an investigator's institution and project team to have in place a relatively high number of data security measures. Although it may require additional investment and effort from both investigators and sponsors, this is potentially a phenomenally fruitful avenue of research, as the Army STARRS suite of datasets is available to researchers and includes a large volume of "real world" data, including psychometric data, survey data, and, for many records associated administrative data.

Academic Publications without Available Data

Academic publications provide a useful insight into both theoretical and practical issues associated with insider threat and insider risk; however, replication data and other ancillary files are seldom made available alongside these publications. This portion of the catalog is based largely on recent data and literature review work and provides a description and summary of datasets that authors have used in their publications. For the datasets contained in this portion of the catalog, some descriptions are sparser than others based on the content of the original publications and the response (or mostly lack thereof) from the authors to the research team's inquiries.

While many additional entries could be included in this tab, the research team only included 1) those from the initial Academic domain and 2) items in other domains that were initially presented as having data but did not. The choice not to thoroughly log entries from other domains was made due to the relative inutility of datasets that are not available to future researchers. Additionally, the choice to retain the records from the first academic review article domain was made to underscore one of the key findings about replicability of publications and replication datasets (see above). Many of the datasets listed here could be useful for future research efforts, and their findings have been included in major review articles and field publications, but cannot be independently verified or examined, nor used to build a foundation of growing empirical knowledge in the field, as the datasets are not (easily) made available to others.

Rich Data Sources

As indicated above, these are data sources with great potential for researchers, but without clear previous application to insider risk and insider threat studies. Some of these links connect to a library of resources, such as the Social Security Administration data library or the University of California-Irvine Machine Learning Repository. Of particular note is the Global Barometer Surveys, which provides high-quality individual-level survey data across multiple world regions, and is a hugely underutilized resource in the social sciences. Individual-level data on attitudes and preferences has a strong potential to benefit insider risk studies, where better analysis and understanding of the individual's behaviors and attitudes are crucial.

Vetting and Investigative Sources

This portion of the catalog contains a list of the data sources available to those tasked with protecting against insider risk in the DOD. These are included to encourage future researchers to

consider the data that USG users have to work with, and to develop tools, algorithms, and other research products that maximizes the information that can be derived from these data sources.

Coding and Describing Data Resources

Datasets included in the main catalog were given an overall description. Many of these categories were determined using previously completed cataloging efforts for CT-NEADS, with some additional categories suggested by academic review articles and internal SME expertise at START and ARLIS. Descriptions of the data, ownership, related publications, and links to the data were included as basic information about the resource overall. Further details about the focus or purpose of the set, its reputation in the literature, and other notes were included in longer-form text box entries. Dataset contents, like the key variables included, the number of variables and number of observations, and the data file types were also coded.

To facilitate searching and filtering, datasets were also coded from a finite set of options along a set of highly relevant attributes. These were determined both by the research team's own experience in cataloging (e.g. including 'discipline' as an easy-to-filter dimension), and based on feedback from interviews (e.g. type of data generating process as a key attribute.). **Table 3**, below, provides an illustration of the four variables included in the catalog to facilitate searching and filtering:

Table 3: Catalog Variables included to Facilitate Searching and Filtering

Catalog Variable	Variable Values
Dimension of Insider Risk	Cyber, Psychosocial, Physical world, Admin data
Type of Analysis	Theory Development, Theory Testing, Training Set, Test Set, Case Studies
Type of Data Generating Process	Synthetic; part synthetic[8]; real-world
Discipline[9]	Computer Science (CS), Psychology, Biology, Cyber Security (Cyber), Social Sciences

After an initial review of the domain to collect a list of potentially includable items, individual members of the research team were assigned to particular sources. Because inclusion was based on relevance and accessibility of data, items that were not ultimately included in the dataset were sometimes partially coded. During the initial coding stage, individuals became the 'expert' on that particular data resource, though difficult cases were discussed with the rest of the team, especially if one person had a more relevant disciplinary background. This round of coding was an iterative process, wherein some dimensions were added based on their relevance to a high number of entries (for instance, file type) and some were eliminated (for instance, temporal unit of analysis).

[8] Data from simulations in which real people participated, including real-time simulations or wargames that included "red team" actors, were considered "Part-synthetic". Datasets into which malicious activity was inserted were considered "Synthetic".
[9] Researchers with a background in specific discipline would be most familiar and comfortable employing datasets in their disciplines and related fields.

When the initial round of coding was completed, a basic quality control procedure was implemented. For the quality control, a random set of 7-8 entries (roughly 10% of the total entries in the Main and Academic Publication portions of the catalog) that had originally been coded by another team member was assigned to each researcher. They were asked to use the citation and dataset name and independently code the remaining attributes and descriptions. After this review resulted in no more than a handful of minor edits to improve clarity, the lead researcher provided a final review of all entries.

FINDINGS ON LIMITATIONS ON INSIDER RISK-RELEVANT DATA, AND POSSIBLE REMEDIES

Finding 1

There are a limited number of high-quality, public-use data sets for the study of insider risk. These datasets have some important strengths and have contributed meaningfully to the advancement of the field; however, many are over a decade old and few include psycho-social measures alongside technical data, making them only partially suited to addressing the challenges of insider risk and insider threat circa 2021. This issue has several key components:

1) Companies and organizations actually attempting to counter insider threats must consider the rights and privacy of their employees and members, making access to data for outside researchers challenging.
2) Academic and think tank researchers do not routinely publish or provide replication data. This is related to the larger issue of transparency and replicability of scholarly works in general. The research team found that relatively few journals, even high-impact and well-respected scholarly journals, require that authors make datasets, algorithms, or other ancillary files available.[10] Furthermore, many oft-cited publications in the insider threat space are actually non-peer reviewed conference papers, proceedings, and white papers.

Possible Remedies

- Encourage researchers to partner with companies, or to get access to data from their own organizations (academic institutions, think tanks), and prioritize funding for those researchers that are able to provide letters of support and data collection plans that include details about what data will be used for the research.
- Require some types of sponsored research to publicly release the data collected for research. The National Institute for Justice at the Department of Justice generally requires that de-identified datasets collected with NIJ funds be made available via the ICPSR data repository. Requirements could be limited, so that only a subset of the data, or a single test set/validation set pairing, must be shared. Replication data could be housed in existing libraries such as IEEE's DataPort, or with individual journals, if sponsoring entities are not positioned to serve as a repository themselves.

Finding 2

Both the USG and a wide variety of other organizations have access to data on their personnel but limited capacity for turning that data into useful information about insider risk. This is a problem shared by organizations writ large: Two in three organizations (66%) report a struggle with turning volumes of security activity and event data being collected into "intelligent,

[10] The *Journal of Network and Computer Applications* for instance, has this note on research data: "**Research data:** This journal encourages and enables you to share data that supports your research publication where appropriate, and enables you to interlink the data with your published articles. Research data refers to the results of observations or experimentation that validate research findings. To facilitate reproducibility and data reuse, this journal also encourages you to share your software, code, models, algorithms, protocols, methods and other useful materials related to the project (see https://www.elsevier.com/journals/journal-of-network-and-computer-applications/1084-8045/guide-for-authors)

actionable insights."[11] Many commercial products provide services to address this challenge, but the problem of integrating psychosocial information into tools that emphasize cyber security is largely unresolved. This research identified three aspects of this issue for the USG and DOD in particular:

1) Lack of timely access to relevant data. Cases may be referred to insider threat teams only after cases or investigations have been closed in other departments and offices.
2) Lack of access to information streams outside of an active investigation. Privacy concerns and individual rights limit the ability of DOD insider threat teams from using a variety of data sources unless they have cause – namely, unless it is part of an active investigation into a potential threat.
3) Lack of automated systems or integration processes for disparate data sources. Relatively few processes in analyzing and assessing insider risk are automated. The issue is both with the technology itself – as data integration is a hard problem and reliable, proven algorithms are lacking – and the technical quantitative or data analytics expertise of analysts and staff, which is not uniformly suited to routinizing these tasks.

Possible Remedy

- Encourage researchers to develop evidence-backed tools that help integrate the available data sources that USG already has at its disposal, and to automate analysis. Specifically, processes that can be run on de-identified versions of data such as financial and travel data – so that individual privacy rights are not violated – could increase the ability of insider threat programs to identify high-risk activities and individuals in a timely fashion. It could also help focus investigator time and energy on the 'most likely' cases, and focus on data like personnel files or other data streams that are more challenging to integrate.
- (Foster better communication between vetting and insider risk components)

CATALOG QUICK START GUIDE AND CODEBOOK

Getting Started

For the first time users of the dataset catalog, we recommend they begin by familiarizing themselves with the various datasets included in the catalog by reading each dataset's description. This will provide the users an idea of what types of datasets are included in the catalog. All datasets included in the catalog includes a description.

If a user is interested in a specific dimension of the datasets, the user has the option to filter the catalog based on the specific dimension(s) of interest. For instance, if the user is interested in the datasets that contain cyber dimension of the insider risk, then the user can filter the catalog by the "Dimension of Insider Risk" variable and select "Cyber" to see only those datasets included in the catalog that contains Cyber insider cases. Same can be done for other variables, such as the discipline where the datasets originate from, the unit of analysis, and the data generating process used for each dataset, just to name a few.

[11] "Research Confirms Organizations Continue to Struggle with Insider Threat Detection" *Security Magazine (online),* https://www.securitymagazine.com/articles/90654-research-confirms-organizations-continue-to-struggle-with-insider-threat-detection, dated August 2, 2019, accessed July `3, 2021

Finally, a variable worth noting is the "Type of Analysis" variable that is contained in the Main Catalog. This variable provides the user information on how each dataset included in the Main Catalog could be used in relation to modeling and mitigating insider risk. For example, a dataset where the "Type of Analysis" is listed as "Training set" should be used to train detection tools such as a machine-learning algorithm but should not be used for psychological analyses of the insider behavior.

Main Catalog

All datasets contained in the Main Catalog were coded to capture 26 independent characteristics of each dataset. These characteristics are grouped into five (5) categories. **Table 4**, below, provides the list of these categories:

Table 4: Main Catalog Dataset Characteristic Categories

Dataset Characteristic Categories
Dataset Basic Information
Dataset Sustentative Overview
Data Structure and Stats
Accessibility
Citation

Dataset Basic Information

1. **ID Code:** An identification number assigned to a dataset. The ID code provides the source of the dataset followed by the sequential number that corresponds to the order the dataset was added to the catalog.

2. **Datasource Name:** This field provides the name of the dataset.

3. **Author/Owner:** This field provides the name(s) of the author(s) or identified owner(s) of the dataset.

4. **Description:** This field provides a short description of the dataset.

5. **Link:** This field provides last known good Internet link to the dataset.

6. **Source Domain:** This field provides the source domain that the dataset belongs to as explained earlier in this report.

Dataset Sustentative Overview

7. **Additional Coverage or Use Information:** This field provides additional information on the dataset as well as how the dataset could be used in addition to the dataset description.

8. **Dimension of Insider Risk:** This field provides the dimension of insider risk the dataset covers (e.g., cyber, physical, psychosocial, etc.).

9. **Variable (count):** This field provides the number of variables contained in the dataset.

10. **Key Variables:** This field provides a list of key variables contained in the dataset (e.g., user, pc, activity, Survey results, etc.).

11. **Type of Analysis:** This field provides the type of analysis the dataset is best suited for (e.g., training set, theory development, etc.).

12. **Type of Data Generating Process:** This field provides how the data was generated for the dataset (e.g., synthetic, real world, etc.).

13. **Discipline:** This field provides the discipline in which the dataset is situated (e.g., cyber/CS, social sciences, etc.).

Data Structure and Stats

14. **Observations (number):** This field provides the number of observations contained in the dataset.

15. **Unit of Analysis:** This field provides information on the level of analysis the dataset is designed for (e.g., Unix command, attack events, individual person, etc.).

16. **Date Published/Created/Collected:** This field provides the last known year the dataset was published, created, and/or collected.

17. **Last Update:** This field provides the last known year the dataset was updated.

18. **Reputation/Concerns:** This field provides general reputation of the dataset in the field as well as any strengths and/or weaknesses with the dataset.

Accessibility

19. **File Type/Software Needed:** This field provides the file type the dataset is instantiated in as well as any specific software required to open the dataset.

20. **Cost:** This field provides the cost associated with accessing the dataset.

21. **Who contact to get it:** This field provides the point of contact to gain access to the dataset.

22. **Overall barriers to access (low/med/hi):** This field provides an assessment of the ease/difficulty associated with gaining access to the dataset.

23. **Specific Permissions/Restrictions:** This field provides any specific permissions, requirements, and/or restrictions associated with gaining access to the dataset.

24. **Contains data on identifiable U.S. persons:** This field provides information whether the dataset contains any personally identifiable information.

Citation

25. **What is Cited:** This field provides information on what is cited for the dataset (e.g., a journal article that utilizes the dataset; the codebook for the dataset; etc.).

26. **Citation:** This field provides the citation associated with the dataset.

Academic Publications without Available Data

All datasets contained in the Academic Publications without Available Data were coded to capture 12 independent characteristics of each dataset. These characteristics are grouped into four (4) categories. **Table 5**, below, provides the list of these categories:

Table 5: Main Catalog Dataset Characteristic Categories

Dataset Characteristic Categories
Dataset Basic Information
Dataset Sustentative Overview
Data Structure and Stats
Citation

Dataset Basic Information

1. **ID Code:** An identification number assigned to a dataset. The ID code provides the source of the dataset followed by the sequential number that corresponds to the order the dataset was added to the catalog.

2. **Author/Owner:** This field provides the name(s) of the author(s) or identified owner(s) of the dataset.

3. **Description:** This field provides a short description of the dataset.

4. **Link:** This field provides last known good Internet link to the dataset.

5. **Source Domain:** This field provides the source domain that the dataset belongs to as explained earlier in this report.

Dataset Sustentative Overview

6. **Dimension of Insider Risk:** This field provides the dimension of insider risk the dataset covers (e.g., cyber, physical, psychosocial, etc.).

7. **Key Variables:** This field provides a list of key variables contained in the dataset (e.g., user, pc, activity, Survey results, etc.) as gleaned from the academic source that mentions/utilizes the dataset.

8. **Type of Data Generating Process:** This field provides how the data was generated for the dataset (e.g., synthetic, real world, etc.).

9. **Discipline:** This field provides the discipline in which the dataset is situated (e.g., cyber/CS, social sciences, etc.).

Data Structure and Stats

10. **Date Published/Created/Collected:** This field provides the last known year the dataset was published, created, and/or collected.

Citation

11. **What is Cited:** This field provides information on where the dataset was cited (e.g., a journal article that utilizes the dataset; conference paper; white paper; etc.).

12. **Citation:** This field provides the citation associated with the dataset.

Rich Data Sources

For datasets contained in the "Rich Data Sources" tab, each dataset was coded for three (3) independent characteristics that a user would need to locate the data source in the future. Since this tab only has three (3) characteristics, they are not subdivided into categories like the two previous tabs.

1. **Datasource Name:** This field provides the name of the dataset.

2. **Website/Link:** This field provides last known good website address or an Internet link where the dataset is hosted.

3. **Description:** This field provides a short description of the dataset.

Vetting and Investigative Sources

For datasets contained in the "Vetting and Investigative Sources" tab, each dataset was coded for five (5) independent characteristics to provide the user enough information about each data source. Since this tab only has five (5) characteristics, like the immediately previous tab, this tab does not contain any categorization of data source characteristics.

1. **Datasource Name:** This field provides the name of the dataset.

2. **Description:** This field provides a short description of the dataset.

3. **Data Type:** This field provides a description of the type of data contained in each dataset (e.g., financial data, law enforcement data, etc.)

4. **Access:** This field provides information on how one can gain access to the dataset.

5. **Location:** This field provides information on what organization a user could contact to inquire about gaining access to the dataset.

ACKNOWLEDGEMENTS

This report was prepared for [Office of the Undersecretary of Defense for Intelligence and Security (OUSD(I&S)), United States Department of Defense (DoD)] under the following agreement:

HQ003420F0655, *University of Maryland,* "Insider Threat and Personnel Vetting."

DISCLAIMERS

Any views, opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of an official United States government position, policy, or decision. Additionally, neither the United States government nor any of its employees make any warranty, expressed or implied, nor assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, product, or process included in this publication.

Certain commercial entities, equipment, or materials may be identified in this document to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the Applied Research Laboratory for Intelligence and Security (ARLIS), the University of Maryland, or the United States government, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

ABOUT ARLIS

Applied Research Laboratory for Intelligence and Security (ARLIS) is a UARC based at the University of Maryland College Park and established in 2018 under the auspices of the OUSD(I&S). ARLIS is intended as a long-term strategic asset for research and development in artificial intelligence, information engineering, acquisition security, and social systems. One of only 14 designated United States Department of Defense (DoD) UARCs in the nation, ARLIS conducts both classified and unclassified research spanning from basic to applied system development and works to serve the U.S. Government as an independent and objective trusted agent.

ABOUT START

The National Consortium for the Study of Terrorism and Responses to Terrorism (START) is a university-based research, education and training center comprised of an international network of scholars committed to the scientific study of terrorism, responses to terrorism and related phenomena. Led by the University of Maryland, START is a Department of Homeland Security Emeritus Center of Excellence that is supported by multiple federal agencies and departments. START uses state-of-the-art theories, methods, and data from the social and behavioral sciences to improve understanding of the origins, dynamics, and effects of terrorism; the effectiveness and impacts of counterterrorism and CVE; and other matters of global and national security. For more information, visit www.start.umd.edu or contact START at infostart@umd.edu.

Technical Points of Contact:

PI: Adam Russell, D.Phil.
Chief Scientist, ARLIS
301.226.8834; arussell@arlis.umd.edu

Co-PI: Kelly Jones, Ph.D.
Assistant Research Scientist, ARLIS
301.226.8850; kjones@arlis.umd.edu

Task Lead: Steve Sin, Ph.D.
Director, Unconventional Weapons and Technology Division, START
301.405.6656; sinss@umd.edu

Administrative Points of Contact:

Ms. Monique Anderson
Contract Officer, Office of Research Administration
Assistant Director, ARLIS
301.405.6272; manders1@umd.edu

REFERENCES

"About ICPSR." ICPSR: Sharing Data to Advance Science. Accessed September 17, 2021. https://www.icpsr.umich.edu/web/pages/about/.

Al-Mhiqani, Mohammed N., Rabiah Ahmad, Z. Zainal Abidin, Warusia Yassin, Aslinda Hassan, Karrar H. Abdulkareem, Nabeel S. Ali, and Zahri Yunos. 2020. "A Review of Insider Threat Detection: Classification, Machine Learning Techniques, Datasets, Open Challenges, and Recommendations" Applied Sciences 10(15): 5208.

Briguglio, Frank. 2019. "Insider Threat: How to Properly Govern Identities & Identify Nefarious Actors". Security. https://www.securitymagazine.com/articles/91051-insider-threat-how-to-properly-govern-identities-identify-nefarious-actors.

Ivan Homoliak, Flavio Toffalini, Juan Guarnizo, Yuval Elovici, and Martín Ochoa. 2019. "Insight Into Insiders and IT: A Survey of Insider Threat Taxonomies, Analysis, Modeling, and Countermeasures". ACM Comput. Surv. 52(2), Article 30.

Koven, Barnett S., Katy Lindquist, and Max Erdemandi. 2020. "Counterterrorism Net Assessment Data Structure." College Park, MD: START (April).

"Research Confirms Organizations Continue to Struggle with Insider Threat Detection" *Security Magazine (online),* https://www.securitymagazine.com/articles/90654-research-confirms-organizations-continue-to-struggle-with-insider-threat-detection, dated August 2, 2019, accessed July `3, 2021