Project Summary

Collecting representative Internet measurement data has remained a challenging and often elusive goal for the networking community. Obstacles include the Internet's scale and scope, technical challenges in capturing, filtering and sampling high data rates, difficulty obtaining measurements across a decentralized network with radically distributed ownership, cost of building and operating instrumentation, and political hurdles. Even (or especially) with all these obstacles, the demand for and importance of representative Internet data sets is increasing – which is good news for rigorous scientific Internet research. The primary driver of this demand is the now pervasive acknowledgement that we are unable to keep up with cybersecurity threats to various critical and increasingly interdependent infrastructures, and that a primary limiting factor in the escalating arms race is our surprisingly still primitive approach to sharing cyberinfrastructure data.

CAIDA has developed an Internet Measurement Data Catalog – IMDC – an index of information (metadata) about data sets and their availability under various usage policies. This catalog confronted a significant challenge in network science: reducing the cost of searching for data by organizing metadata about accessible Internet data sets into a single repository. We developed the underlying *DatCat architecture* and prototype software implementation to support the IMDC.

We propose to integrate the lessons we have learned during our research, development and operational experience with the IMDC to expand the underlying software capabilities to support the cybersecurity research and cyberinfrastructure development communities. Our three primary deployment goals are to: (1) reduce the burden on those contributing data via a streamlined interface and tools for easier indexing, annotation and navigation of relevant data; (2) convert from use of a proprietary database backend (Oracle) to a completely open source solution; and (3) to expand DatCat's relevance to the cybersecurity and other research communities. This last goal includes outreach activities such as workshops and demonstrations at security-related PI meetings, creating and indexing new data sets – ccTLD DNS zone files – which have been declared critically lacking by the cybersecurity community, and creation of public web forums for discussion of specific and broader data-sharing issues.

Although our focus for this SDCI project will be enhancing DatCat's utility for the cybersecurity and cyberinfrastructure research community, our proposed design objectives and outreach plans explicitly target a range of science and engineering communities. In particular, we believe the proposed software development can support and promote NSF's newly announced Data Sharing Policy, which as of January 2011 requires all proposals to include a plan for how researchers intend to share their data with other researchers.

Intellectual Merit. The proposed software development activities will support a range of measurable benefits to cyberinfrastructure research: maximizing the re-use of existing Internet data; decreasing the time spent collecting redundant data; reducing the effort needed to start a new study; promoting validation and reproducibility of analyses and results; enabling longitudinal and cross-disciplinary studies of the Internet; and **opening up new cross-domain areas of transformative networking research.**

Broader Impact. The broader impacts of this project are diverse. The success of the catalog and related workshops will facilitate wide dissemination of Internet measurement data to researchers and security experts across academic, commercial, and government sectors. By including education-oriented data collections in the catalog, this project creates an immediate link between research and education, and improves access to Internet research for underrepresented groups in computer science and engineering. Most importantly, the software created through this project will help other disciplines and sectors to develop their own catalog instances to support the type of data management plans now articulated as essential to NSF.

SDCI Sec: Metadata Management Software Tools to Support Cybersecurity Research and Development of Sustainable Cyberinfrastructure

Project Description

1 Introduction

Collecting and sharing representative Internet measurement data has remained a challenging and often elusive goal for the networking community. Obstacles include the Internet's scale and scope, technical challenges in capturing, filtering and sampling high data rates, difficulty in obtaining measurements across a network with radically distributed ownership, cost of building and operating instrumentation, and political hurdles [1]. Yet, as with other complex system sciences (climate, biology, sociology), data is, while not sufficient, absolutely necessary to advance our understanding of cyberinfrastructure and to facilitate its future development and growth.

The Internet research community has developed many novel measurement techniques in pursuit of understanding as well as empirical grounding for models of Internet structure and behavior. Operational providers also undertake measurement to protect or improve their infrastructure, and some providers are willing to establish formal privacy-sensitive sharing agreements with researchers. Yet because there is no legal framework that governs the sharing of network measurement data, data-sharing relationships that occur are market-driven or organically-developed. There is no standard procedures for network measurement data exchange, and effectiveness and benefits of existing inconsistent, ad hoc and/or opaque exchange arrangements are uncertain. A formidable consequence is the difficulty of justifying resources for research and other collaboration costs that would enable and support a sharing regime.

Unfortunately, this barren data-sharing landscape forces many academic network researchers to choose scientific problems based on what data they can manage to scrape together rather than picking the most important problems to study and getting the data needed to thoroughly study them [2]. If word-of-mouth (or Google) has insufficiently propagated information about data ownership and access procedures, researchers may waste effort creating a similar new dataset, use a dataset inappropriate for a given problem, make up data, or abandon the research. New researchers often do not even know what data can be collected, much less how to navigate the policy issues in collecting and processing measurements.

Even when researchers find relevant data, it is rarely in a directly usable form. Researchers who invest time in processing and analyzing data for a specific investigation have no systematic method for publishing and sharing caveats about the data or noting idiosyncrasies they found in a given data set, much less finding out if similar data or knowledge already exists somewhere in the research community. Those who already have access to data might want to collaborate with researchers interested in that specific data, if there existed a mechanism to discover each other.

Notwithstanding all these obstacles, the demand for and importance of representative Internet data sets is increasing – which is good news for rigorous scientific Internet research. The primary driver of this demand is the now pervasive acknowledgement that we are unable to keep up with cybersecurity threats to various critical and increasingly interdependent infrastructures, and that a primary limiting factor in the escalating arms race is our surprisingly still primitive approach to sharing cyberinfrastructure data. Security researchers have affirmed the need for a more formal and responsible approach to data sharing [3], articulated most emphatically in the recent Conficker Working Group Lessons Learned Report [4, 5].

CAIDA has developed an Internet Measurement Data Catalog – IMDC – an index of metadata for existing datasets and their availability under various usage policies. This catalog tackles a significant obstacle to progress in network science: reducing the cost of searching for data by organizing information (metadata) about accessible Internet data sets into a single repository. A catalog of relevant high quality data sets supported by a user-friendly interface promises a range of measurable benefits to cyberinfrastructure research: maximizing the re-use of existing Internet data; decreasing the time spent collecting redundant data; reducing the effort needed to start a new study; promoting validation and reproducibility of analyses and results; enabling longitudinal and cross-disciplinary studies of the Internet; and opening up new cross-domain areas of transformative networking research.

We propose to integrate the lessons we have learned during our research, development and operational experience with the IMDC to expand its capabilities to support the cybersecurity research and cyberinfrastructure development communities. Our three primary deployment goals are: (1) reducing the burden on those contributing data via a streamlined interface and tools for automated indexing and easier annotation of relevant data; (2) converting from use of a proprietary database backend (Oracle) to a completely open source solution; and (3) engaging an expanded community of researchers. This project will focus on the cybersecurity research community, but our goal is to demonstrate how a modest investment in software development can measurably advance a field of research by enabling a self-perpetuating life cycle of data discovery, collection, and sharing. Our software design objectives and outreach activities aim to ensure that our architecture and implementation will easily transfer to a broader set of communities both in and out of academia. In particular, we are inspired by and in turn hope to promote NSF's newly announced Data Sharing Policy [6], under which "investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants."

Section 2 of this proposal describes the catalog system architecture, current status, and lessons learned in experimental deployment that motivate the proposed work. Section 3 details the proposed tasks listed above. Team qualifications and the project timeline are in Section **??**. Section 4 presents the broader impacts of this project including integrating research and education via intended classroom use of the catalog, and supporting diversity in cybersecurity research.

2 Internet Measurement Data Catalog

2.1 Supporting Network Science and SDCI programmatic goals

Building on its decade of experience with data collection, curation, and provision, CAIDA developed an Internet Measurement Data Catalog (IMDC) to support access to network research data. Our original vision for IMDC was reflected in a recommendation from a 2008 NSF-sponsored workshop on Network Science and design, reiterated in NSF's Network Science and Engineering Council's research agenda [7], namely to improve the "community's focus on the creation and association of metadata with any newly collected data set:"

> *We believe it is paramount to revisit the metadata concept, and develop it to the point where its utility becomes obvious and its implementation becomes as straightforward as possible. A first impediment is that there is today no "best practice" for collecting and maintaining metadata with a dataset. Such guidelines should be developed, with the aim that metadata description should include as much information as possible that is pertinent to the collection of the data*

and its future use by third parties. Ideally, this would include details about the measurement technique used, its shortcomings and limitations, and alternatives considered but not adopted. It should spell out in detail any issues concerning bias, completeness, accuracy, or ambiguity of the data that are known as a result of the data producer's in-depth understanding of the measurement and data collection effort [7].

Our mission in developing the Internet Measurement Data Catalog was to promote data-sharing by publishing a wide variety of cyberinfrastructure-related metadata, in order to enhance scientific productivity and facilitate R&E collaborations, two explicit goals of the SDCI program. Our experience with IMDC yielded several valuable lessons (described in Section 2.4), which suggest necessary improvements to this software that are ideally aligned with this SDCI solicitation. Particularly strong feedback from early users of the catalog was that it was essential to lower the time investment required to index and share data in order to attract more contributors and users. Second, our use of proprietary software for the database backend was appropriate six years ago given the circumstances, but the relative price-performance characteristics have shifted now in favor of the use of open source software for the backend, which will also enable a broader set of communities to capitalize on our development investment. These two goals – open source and broadening the community of users – are explicitly targeted by the SDCI solicitation.

Finally, the proposed software development will make the catalog more useful to the cybersecurity community, with emphasis on several topics of the SDCI Cybersecurity focus area: enabling improved situational understanding, promoting data sharing across organizational boundaries, and supporting novel approaches to analyzing and presenting large scale net data, which could lead to transformative research on cybersecurity and cyberinfrastructure development. Common to SDCI and the NetSE Network Science program goals, the proposed software development contributes to the emergence of a discipline that can formalize and explain our observations and understanding of large-scale, complex networked systems [7]. Our proposed upgrade to the DatCat architecture underlying the IMDC catalog will extend this contribution to include the expanding field of cybersecurity research.

CAIDA is also an active participant in the PREDICT project, a DHS-sponsored framework for sharing network data useful for research on security products, models and strategies. PREDICT is oriented toward sharing data from private enterprises, and operates under a strict legal framework which raises the bar for participation. The IMDC and underlying DatCat architecture stems from the needs of academic community, and use of the catalog itself is AUP-free. CAIDA's participation in both projects will bring additional synergy will leverage the investments made by their respective sponsoring agencies.

2.2 DatCat Architecture

Using initial NSF funding and inspired by a previous design paper [8], we created the Internet Measurement Data Catalog (IMDC) [9], a flexible database capable of cataloguing and searching metadata from diverse network measurements, including details on tools, collection platform, location, experimental parameters, and relationships among datasets. To avoid the difficulties associated with ownership, privacy, and legal complications involved in storing data owned by others, the catalog stored only descriptive metadata, not the raw data itself. Current data provision models span a wide range of access controls, from public domain to requiring in-person analysis to send-code-to-data models. Many interesting cybersecurity and other network research papers are published using "unavailable" data. A catalog entry documenting a dataset's existence provides a starting point to broker a relationship or collaboration that may result in a significant scientific

contribution.

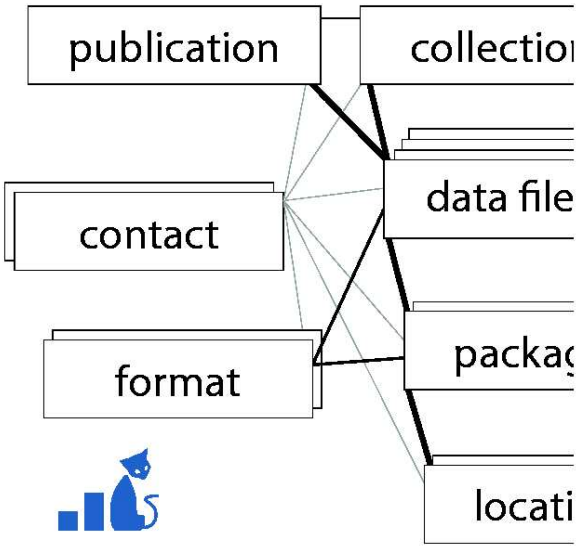

Figure 1: *Object relationships in current DatCat architecture. (Compare to proposed changes, Figure 3)*

The underlying architecture for IMDC, which we now call *DatCat*, is organized around seven conceptual objects (Figure 1): Data Files, Formats, Packages, Locations, Collections, Publications and Contacts. A *Data object* is a real-world file of a certain format that a researcher has identified as "data". Data objects are grouped in *Packages* (e.g. a tar or zip file) and *Locations* identify instances of a Package object, possibly at multiple mirror sites. Data objects can be grouped into *Collections*, e.g., as part of the same measurement event, project, theme, or related to a specific question. *Publications* are specialized *Collections* that group data files and associated metadata specifically used in a paper. Finally, *Contacts* represent people – data creators, contributors, and users. We allow all required contact information to be hidden from public view (though we do publish names and/or affiliations of data users in our reports to funding agencies). All core objects have associated keywords that can be used in searches.

The catalog's search features include the ability to browse featured data collections and/or search across all indexed fields to find data sets matching characteristics of interest. The catalog also provides access to a centralized list of each contribution's status, feedback from moderators, and auto-generated errors on submissions.

One of the most powerful features of the catalog architecture is its system of Annotations, which allows contributors to supply rich structured versatile metadata, and also allows any user of the data to contribute documentation as they learn problems, features, or other useful information about a given dataset. Sufficiently rich metadata and annotations could conceivably support meaningful scientific queries against such metadata itself, without access to the actual dataset. Unfortunately, our commitment to that vision increased the complexity of the system, as it led us to treat individual files as atomic units of storage, encryption, and compression. It also created a large burden on the data contributor, who had to submit detailed meta data for each file they indexed, described further in Section 2.4.1.

2.3 History of IMDC Data Contributions

IMDC opened for public browsing on June 12, 2006 with 4.8 terabytes of data from two organizations. By August 1, 2009, it indexed 121 datasets (more than 26 TB of data) from over 30 organizations (Figure 2). By that time nearly 500 researchers had created accounts, a significant vote of confidence in IMDC's future, as accounts were not required to view

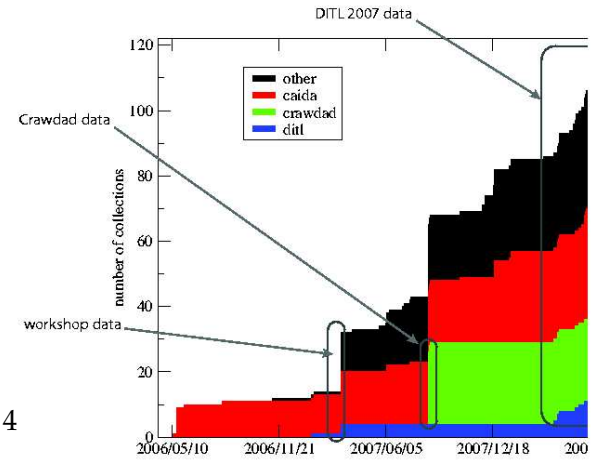

Figure 2: *Number of collections contributed to datcat.*

data offerings. To help contributors of large data sets automate the submission process, in 2007-2008 we developed the *data-to-yaml* conversion utility (described in Section 2.4.1) and related tools, which we introduced to researchers at our first Data Catalog Workshop in March 2007. We began supporting public contribution via the web in late 2007, followed by several workshops to introduce its features to the community and solicit feedback on how to improve the catalog [10, 11, 12, 13]. A total of 37 researchers have added their metadata to IMDC.

IMDC currently holds metadata for seventeen datasets:

- Day in the Life of the Internet collection [2006] (7,810 files)
- Day in the Life of the Internet collection [2007] (9,544 files)
- UCSD SIGCOMM Wireless Traces (25 files)
- AOL 500k User Session Collection (10 files)
- Router Adjacency Data (2 files)
- Autonomous System (AS) Adjacency Data (2,451 files)
- Autonomous System (AS) Relationships Data (79 files)
- Autonomous System (AS) Taxonomy Data (113 files)
- CAIDA Dataset on the Code-Red Worms (14 files)
- CAIDA Dataset on the Witty Worm [public version] (7 files)
- Skitter Macroscopic Topology Data (60,478 files)
- OC48 peering point traces (119 files)
- CAIDA Dataset on the Witty Worm [restricted, raw traffic traces] (132 files)
- Denial-of-Service Backscatter-TOCS [2001-2004] (231 files)
- Denial-of-Service Backscatter 2004-2005 (63 files)
- Denial-of-Service Backscatter 2006 (168 files)
- DNS RTT Dataset (5,047 files)

Essential metadata recorded for all datasets includes: dataset name, file format, start and end times, creation process, creators, primary contact, description of known anomalies during collection, measurement platform, time zone, geographic location, topological location, organizational location, and clock accuracy parameters [14]. Additional meta-data fields depend on the type of data. Traffic data sets may be indexed with metadata for workload statistics such as packet and byte counts, unique IP addresses, port number distributions, both for the whole data set and over smaller time intervals, e.g., per second. Metadata for macroscopic IP topology data sets could include number of paths probed per data set, distribution of path lengths, reachability and path stability data, or more sophisticated computed statistics [15], as well as results of DNS lookups of probed addresses. Annotations may be added subsequent to the original data indexing, expanding the potential range of metadata indefinitely.

Lack of funding and increased Oracle database licensing cost at SDSC required that we disable the IMDC temporarily while we integrate lessons learned into our transition from this research prototype to the proposed increased operational capabilities. The *http://datcat.org/* web site still retains much of the information describing the original project goals, and the common object types currently supported.

2.4 Lessons Learned in Experimental Deployment

The lessons we have learned from the metadata catalog development and usage thus far revealed the need for additional functionality and framework to increase the utility of this resource to broader science and engineering communities. Our most significant lesson was that entering fine-grained metadata is too time-consuming for all but the most passionate users. Our second lesson was that the cost of supporting the proprietary Oracle database backend inhibited sustainability of the project even for us, much less other communities interested in leveraging our technology. Third, we gained less traction than expected due in part to casting a broad net across the network research community, rather than focusing on the sub-community of network research with the most acute need to share as current data as possible, and the most likely to be able to generate concrete and relevant queries of metadata in the catalog. We review some details of these lessons which motivate our proposed new functionality in Section 3.

2.4.1 Lesson 1: Fine-grained Metadata is Unwieldy

The IMDC catalog hosts several mature data submission tools. The `subcat` tool provides an easy-to-use command-line interface to submission. To submit a dataset, a user writes text files describing the dataset, runs `subcat` on the text files to generate an XML submission file, and then uses a web browser to upload the XML submission file to the catalog server through a submission web page. We tried to make the input files as easy and intuitive to write as possible by distilling the format to a simple syntax of key-value pairs, using the open-ended annotation system mentioned above. No programming knowledge or other technical skills are required to write the input files. To further reduce manual labor, we provide the `data-to-yaml` command-line tool that the user can run on the actual data files to automatically generate well-formed `subcat` input files readable by the catalog. For known file types, `data-to-yaml` can extract detailed metadata such as counts of packets, bytes, and flows, and start/end times, minimizing the effort needed by the user to enter baseline metadata.

At our March 2007 Data Catalog Contribution Workshop we launched an early version of `subcat` and `data-to-yaml`. Hands-on instruction at the workshop enabled a sharp increase in submissions, shown in Figure 2. But despite our best efforts to simplify the submission process with `subcat`, our success was modest: the time required to install and use the catalog submission tools, especially to index fine-grained metadata for large data sets or collections, still exceeded what busy researchers are likely to spend. Even navigating such fine-grained metadata when browsing the catalog to find a certain data set turned out to be tedious. Section 3.1 proposes our first Task: developing several new software features to lower the bar for indexing data.

2.4.2 Lesson 2: Proprietary Database Backend Limits Sustainability of Catalog

From the beginning, we understood the design of our Internet Measurement Data Catalog required a full-featured, industrial strength relational database for the backend storage of the metadata. At the start of the project in 2003, CAIDA's home department, SDSC, offered hosted services for both IBM's DB2 and Oracle Advanced Database. IBM provided an academic no-cost license and Oracle required user licensing fees in the form of a recharge for services. Early attempts to make use of DB2 fell short of expectations and so the group made the decision to switch to Oracle services. The effort to implement the backend in Oracle resulted in success. For several years, the metadata had been served from these SDSC hosted Oracle services. At the conclusion of the previous supporting grant, we realized that we could no longer pay for the SDSC hosting fees.

Also, since the time of the original design, open source database software such as Postgresql and MySQL have matured to the point where these open solutions now provide the full-featured reliability of commercial database solutions. Migrating the metadata from the catalog to an open solution (Task 2, Section 3.2) will lower the costs of support and enable replication, expansion, and access to a broader set of communities/countries.

2.4.3 Lesson 3: Penetration would benefit from momentum in a narrow community

Although we captured the passion and interest of a few dedicated meta-data enthusiasts, our early ambitions with the catalog were too broad, and the network research community incentives to share data still too limited for our Internet Measurement Data Catalog to gain wide traction in the community. Recent developments make now a perfect time to launch a strategy grounded by these insights. First, the public and private sectors have finally begun to acknowledge, independently [16, 4] and together [17], that our inability to defend against cybersecurity threats to various critical infrastructures is perpetuated by our persistent lack of progress in effective sharing of cyberinfrastructure data. Security researchers years ago affirmed the need for a more scalable, responsible approach to data sharing [3], articulated most emphatically in the recent Conficker Working Group Lessons Learned Report [4, 5]. Our Task 3 (Section 3.2) will refocus our cataloging effort on the cybersecurity research community, with new cybersecurity-relevant data sets (DNS zone files), outreach to the NSF-funded cybersecurity (Trustworthy Computing) research community, and hosting a public forum for community-directed discussion of data sharing issues.

3 Proposed Work: Applying Lessons Learned to Improve Data Sharing Landscape

Based on the lessons we have learned during the development and operation of IMDC, we propose to upgrade the underlying DatCat architecture to the next level with three substantial tasks: upgrading data capabilities to streamline the user experience; moving from use of a proprietary database backend (Oracle) to a completely open source solution; and expanding the community of the catalog users to a broader range of cybersecurity and other researchers.

3.1 Task 1: Expanding DataCat Capabilities to Streamline the User Experience

Our primary goal behind this task is to reduce the burden on those sharing data.

While rich per-file metadata has advantages, including answering some research questions without having to download and analyze raw data, it is prohibitively costly to support in all cases. We will still pursue rich metadata for selected large-scale, ongoing datasets (including all of our own), but we will add functionality to the catalog to create standalone collections that point directly to an entire data set rather than requiring a pointer to a package of individually indexed data files. We will also augment our command-line submission tools with a web-form based submission tool for easier indexing of single files and collections of related files. Finally, we will im-

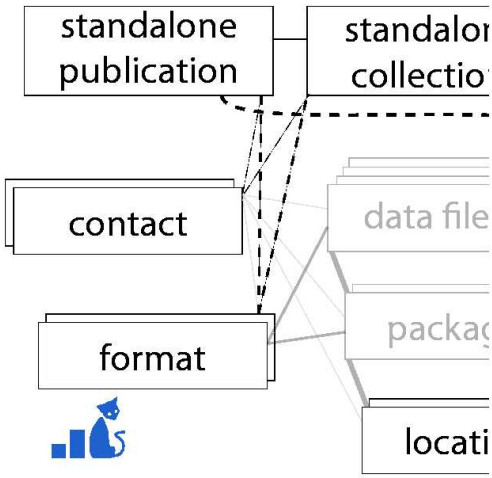

Figure 3: *Relationships between existing DatCat objects and proposed new Standalone objects (compare to Figure 1)*

prove our comment and annotation system to take advantage of these simplifications.

3.1.1 Standalone Publications and Collections

We will create standalone versions of our existing *Collections* and *Publications*, with sufficient metadata that they no longer need contain entire *Data objects*, but can use a limited number of *Data object* fields in addition to fields that describe aspects of the specific collection. In addition to streamlining the data indexing process for most users, this feature will also considerably ease navigation of the catalog. Whereas now users must navigate from the *Collection* to a given *Data object*, then onto a *Package*, until finally arriving at a *Location*, the new standalone objects will provide direct links from *Collections* to locations, which will no longer refer to individual files, but instead to the location of the whole dataset.

3.1.2 Web-based Submission and Search

We will implement an entirely web-based submission interface, consisting of simple forms that guide the user through what fields are needed for each type of object. The interactivity and greater display estate will allow the tools to provide context-dependent help and other support for the submission workflow. A point-and-click function will be able to perform some tasks, such as listing creators or file formats of a dataset. A web-based approach allows us to provide a flexible multi-level submission interface, with separate categories of fields: *absolutely required*; *please provide if available*; and *desirable*. This approach ensures that contributors with just 15 minutes can still make a useful contribution, while allowing dedicated submitters to provide richer metadata.

Two related features identified by users as desirable but lacking in DatCat are: (1) versioning of dataset metadata, so that contributors can incrementally enter metadata for subsequent iterations of a related or same data set; and (2) the ability to dump search results for use by subsequent automated (rather than interactive) processing. We will implement both features via the web-based interface.

3.1.3 Comments and Annotations

The most powerful yet underutilized feature of DatCat is its ability to annotate already catalogued datasets. With annotations, users can share findings about data, express concerns, and identify caveats. Data contributors can currently annotate their dataset with quantitative statistics as they index it. The underutilization of annotations derives directly from the primary problem addressed in this task: too many objects in the database to annotate, a result of the overengineering imposed by indexing individual data files. We will add support for more flexible granularity in annotations – annotating entire collections, a subset of files, specific sections of files. Annotations can be added based on characteristics within the file, e.g., time, sources, destinations, topology, geography. As an example, it is often challenging to detect and identify individual attacks in traffic data, e.g., which IP addresses are participating in a DDoS attack, over what time period, to what destinations. If a data user does ascertain such an attack, simplified manual annotations will make easy to share the information via the catalog. In aggregate, a set of annotations on several data sets could yield an general Internet even timeline, allowing correlation of anomalies across time and space.

3.2 Task 2: Moving to completely open source solution

The IMDC web interface and application code all came of open source solutions, however, the backend database relied on a proprietary solution (see Section 2.4.2 and we finally had to disable DatCat because of Oracle database licensing costs. SDSC currently enjoys an unlimited academic department license to run the Oracle Advanced Database server for internal use. Under this limited license, we will regain access to the metadata in the catalog. CAIDA system administrators will (i) install the Oracle Advanced Database server on existing hardware; (ii) operate the services long enough to restore the IMDC data from backups; and (iii) migrate the IMDC Oracle schema/database to an open source database platform. In its final form, the IMDC Oracle implementation made minimal use of proprietary functionality such as subroutines (PL/SQL) or other vendor-specific features. Therefore, we propose to accomplish migration and conversion to an open source database platform with the following subtasks:

1. replication of the IMDC schema in an alternative platform such as MySQL[18] or Postgresql[19] (CAIDA has experience supporting both.);
2. migration of the data to the newly created database/schema;
3. modifications and updates to the IMDC web application to convert any vendor-specific database connection and SQL to the selected open solution.

Tools we develop for this project will use the GNU General Public License (GPL)[20]; the open source database tools we choose will use various licenses, including GPL and The PostgreSQL License (TPL)[21].

3.3 Task 3: Expanding DatCat's Community to Cybersecurity and other fields

Our plans to expand DatCat's relevance to the cybersecurity and other research communities include four subtasks: (1) present and demonstrate new web-based submission interface at security-related PI meetings to promote indexing of datasets used by NSF-funded projects; (2) create and index new data sets – ccTLD zone files – which have been declared critically lacking by the cybersecurity community; (3) host annual workshops at UCSD (Year 1) and Indiana University (Year 2); (4) support public web forums for discussion of specific and broader data-sharing issues.

3.3.1 Indexing Datasets Used by NSF-funded Cybersecurity Projects

We propose to create a survey of NSF-funded Internet measurement researchers, a response to which will automatically index their data into DatCat, including annotations describing data protection (e.g., anonymization) methods. (If a particular data item is shown as "anonymized" in the catalog, it will include a required annotation referencing the anonymization method.) We will present the survey at NSF Trustworthy Computing PI meetings. CAIDA's participation in the DHS Cybersecurity program will also allow us to do outreach to that community at PI meetings.

Widespread use of this tool will improve the accountability and legacy of federally sponsored network research. Information about collected data that might otherwise be lost as soon as the sponsored project is over, will be preserved and will continue to serve the community. We envision the software developed for this project as supporting NSF's newly announced Data Sharing Policy [6]: as of January 2011, all proposals must include a 2-page "Data Management Plan". Easily portable architectures for data indexing could potentially have a near-term benefit to communities across all NSF directorates, as well as a long-term improvement in sustainability of federally funded research.

3.3.2 Expanding Data Catalog to Include More Cybersecurity-Relevant Data Sets

While there are many cybersecurity-relevant data sets, we propose to focus our attention on a set of data of demonstrated long-term interest to the security research and operations community: Top Level Domain *zone files* containing domains on the Internet and their DNS servers. We propose to extend beyond the typical information contained in zone files to derive a more security-relevant data set, and index these files into the catalog to advertise their access to the wider community.

Domains on the Internet fall into two categories: generic top-level domains (gTLDs), such as .com and .net, and country-code TLDs (ccTLDs), such as .de (Germany) and .hk (Hong Kong). Researchers, including PI Gupta, often use TLD zone files – which enumerate the domains contained in a TLD – to investigate DNS aspects of cyberfraud infrastructures. Unfortunately, access to zone files today is problematic. While zone files for many of the prominent gTLDs, including .com and .net, can be obtained by researchers on a nightly basis through contracts with registries, the availability of ccTLD zone files to researchers is practically non-existent, which dramatically limits insight into roughly half of the 183 million domains in the Internet today [22]. Some ccTLDs, including .cn, .de, .uk, and .eu, are among the largest contributors to current growth in domain registrations. Lack of visibility into ccTLD DNS activity is a recurrent theme at security workshops, particularly in discussions of the availability of security data to the research community.

We will derive ccTLD zone files from several sources, and index relevant metadata about them into DatCat. The first data source is *passive DNS* data, which contains DNS requests and responses witnessed by local DNS servers (resolvers) of clients from around the world. We will receive this data from at least two passive DNS installations: the Security Information Exchange (SIE), which collects passive DNS data from 15 large ISPs and commercial DNS service providers around the world, including a U.S.-based Tier1 ISP, two US Cable/DSL access providers, and four U.S.-based universities; and similar instrumentation with fewer collectors at the University of Auckland [23]. We will further populate the TLD zone files using reverse DNS lookups on IP addresses collected from other passive as well as active sources in CAIDA's ongoing data collections [24]. In addition to the standard metadata fields listed in section 2.3, we will index metadata indicating the specific source of data for various domains in the derived zone file, as well as cybersecurity-relevant metadata such as hosts and sub-domains observed with unusual frequency or patterns, either in queries or responses. To validate the accuracy of our derived zone files, we will test them against full zone files available through Verisign, OARC, and other sources. We will add annotations for this validation process into the catalog entries for these data sets, such as indicating what percent of domains our methods captured for a given zone file. We will also leverage the power of annotations to gather feedback from researchers regarding the completeness, accuracy, and utility of the zone files we create, as well as annotations for additional applications of the data.

3.3.3 Workshops to Engage with Broader Communities of Users

To ensure strong channels for feedback, we propose to organize two workshops, one in each year of the project. The first workshop will introduce the newly developed functionality and solicit feedback from users. The second one will be in collaboration with REN-ISAC at Indiana University, where we will explore how DatCat can best serve the cybersecurity community's operations research needs. The Research and Education Networking - Information Sharing and Analysis Center (REN-ISAC) acts as the security information collection, analysis, dissemination, and early-warning organization specifically designed to support the unique environment and needs of organizations connected to and served by higher education and research networks. It has 315 partner institutions with Doug Pearson as director.

Both workshops will solicit and incorporate survey feedback on DatCat and data needs to improve and promote DatCat to a wider audience of users. We will summarize and publish workshop reports.

3.3.4 Public Forum for Discussion of Data Sharing Issues

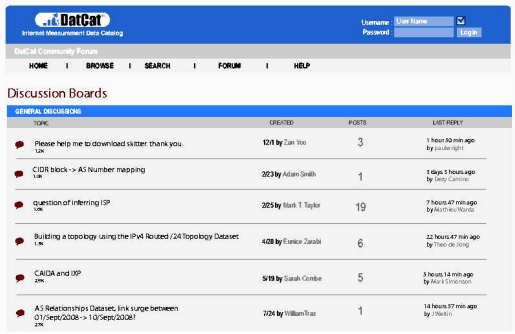

Figure 4: *Proposed datCat forum interface (mockup)*

Ultimately, the goal of this project is to enable more empirically-grounded cybersecurity and cyberinfrastructure research, with expected broader impacts on a range of network-related scientific disciplines. A strong metric of success for metadata sharing is feedback from users on whether they found the data they were looking for, or other data enabling their investigations.

To promote evaluation, we will modify existing tools to support a public Internet forum [25], where users post messages to initiate or respond to topical discussions The threaded forum structure facilitates and archives historical discussion among data users, providers, and consumers, which can then be searched internally or via search engines. We will enhance the DatCat public forum functionality to allow posts to be tied directly to DatCat objects, allowing users to easily find which forum posts relate to a *Collection* or *Publication* of interest. We will introduce three categories to facilitate discussions: **Dataset Request**, **Dataset Discussion**, and **General Discussion**.

The **Dataset Request** category will help researchers who are not familiar with the catalog to find data sets. DatCat users, including CAIDA staff, can post responses in the forum to help users find data relevant to their questions. We will experiment with moderator mechanisms and archiving frequently asked questions to a FAQ. The **Dataset Discussion** category supports inquiries and comments about specific datasets, including problems with file format, documentation, access, or other errors. Topics may be requests, suggestions, open questions, or statements. The **General Discussion** forum will support more open discussion of topics of interest to the Internet research community, spanning from capture hardware to writing applications to Institutional Review Boards [26].

Public DatCat forums will help formalize and expand the current methods of knowledge exchange, with a stable resource for the community. We will attract users to this forum via discussions at workshops and conferences, and references on research and operational mailing lists.

4 Broader Impacts

4.1 Integrating Research and Education

Some of the most valuable training for future researchers is not found in carefully controlled classroom experiences – real world data has unexpected problems. CAIDA frequently hosts graduate students from other institutions, to acquaint them with our data sets and advance collaborations. DatCat offers a unique opportunity for those students to gain access to massive (and messy) datasets while taking advantage of experienced CAIDA researchers providing guidance and advice. Day-to-day experience with processing and documenting research data exposes students to

the range of problems that datasets, and networks, can have and at the same time teaches them how to apply scrutiny and a healthy skepticism of unusual results. We expect to have at least one student doing his/her summer internship at CAIDA in each year of this project.

CAIDA has a highly successful track record of regularly participating in the REU program: in the last six years we supported and trained 20 undergraduates working on various NSF-sponsored projects. We will apply for REU funds for this project as well.

Our software tools will facilitate cybersecurity research but also help annotate various Internet measurement data, and enable the creation of specialized "educational data kits" focusing on specific cybersecurity topics.

PI Kc Claffy is an Adjunct Professor at the Computer Science and Engineering Department of UCSD. She guest lectures for graduate and undergraduate classes and regularly gives seminars on empirical and theoretical underpinnings of the Internet. She will use data and metadata from DatCat as educational aids for teaching and mentoring, both in and out of the classroom.

PI Gupta will offer course projects using DatCat data in the advanced security course she teaches at Indiana University. Further, she will contribute to outreach activities, including advertising DatCat and soliciting cybersecurity data from researchers attending prominent security conferences.

Both PIs will also collaborate with other faculty at their respective institutions to develop hands-on class projects using DatCat.

4.2 Supporting Diversity

Based at UC, San Diego, CAIDA has a strong record of integrating diversity into our research activities. Since 1999, the composition of our 90 paid interns has included 25 females, 21 Asians and 4 Hispanics.

PI Gupta has a track record in recruiting minorities and women, and she will continue this trend. Her School's Assistant Dean for Diversity and Education is the Program Manager for the Alliance for the Advancement of African-American Researchers in Computing (A4RC), so Dr. Gupta has a direct engagement with that initiative and talent pool from 15 Historically Black Colleges and Universities (HBCU's). She has leveraged this channel to supervise African American students during summers.

Gupta's School of Informatics and Computing at Indiana University is also actively involved in the leadership of the Academic Alliance for the National Center for Women in Information Technology (NCWIT). She will use that resource as a significant base from which to recruit women students.

Community resources like DatCat are critical to the success of underrepresented groups in computer science and engineering, including women and minorities. Acquisition of data typically involves personal trust relationships and people-networking with engineering and management personnel outside the campus environment, which can leave underrepresented groups at a social disadvantage. DatCat offers universal access to information on available data sets, providing a leveling influence on the research playing field.

References

[1] N. R. C. Committee on Research Horizons in Networking, *Looking Over the Fence at Networks: A Neighbor's View of Networking Research*. National Academies Press, 2001.

[2] R. Hamming, "You and your research: transcription of the Bell Communications Research Colloquium Seminar," Mar. 1986. http://www.cs.virginia.edu/~robins/YouAndYourResearch.pdf.

[3] M. Allman and V. Paxson, "Issues and etiquette concerning use of shared measurement data," in *IMC*, 2007. http://www.icir.org/mallman/papers/etiquette-imc07.pdf.

[4] Rendon Group, "Conficker Working Group, Lessons Learned," 2011. http://www.confickerworkinggroup.org/wiki/uploads/Conficker_Working_Group_Lessons_Learned_17_June_2010_final.pdf.

[5] Keith Johnson, "Qualified Success Claimed Against Computer Worm," January 2011. http://online.wsj.com/article/SB10001424052748704279704576102433926728902.html?mod=WSJ_Tech_LEFTTopNews.

[6] N. S. Foundation, "Dissemination and Sharing of Research Results," 2010. http://www.nsf.gov/bfa/dias/policy/dmp.jsp.

[7] Ellen Zegura, "Network Science and Engineering (NetSE) Research Agenda," July 2009. http://www.cra.org/ccc/docs/NetSE-Research-Agenda.pdf.

[8] M. Allman, E. Blanton, and W. M. Eddy, "A Scalable System for Sharing Internet Measurements," in *Passive and Active Measurement 2002 (PAM2002)*, (Fort Collins, USA), Mar. 2002. http://www.icir.org/mallman/papers/simr-pam2002.ps.

[9] C. Shannon, D. Moore, K. Keys, M. Fomenkov, B. Huffaker, and K. Claffy, "The Internet Measurement Data Catalog," *SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 5, pp. 97–100, 2005. http://www.caida.org/publications/papers/2005/ccr_imdc/ccr_imdc.pdf.

[10] CAIDA, "ISMA Data Catalog Workshop," 2004. http://www.caida.org/workshops/isma/0406/.

[11] CAIDA, "DCC - DatCat Community Contribution Workshop Series," 2007. http://www.caida.org/workshops/dcc/.

[12] CAIDA, "8th CAIDA-WIDE Workshop," 2007. http://www.caida.org/workshops/wide/0707/.

[13] CAIDA, "1st CAIDA-WIDE-CASFI Workshop," 2008. http://www.caida.org/workshops/wide/0808/.

[14] CAIDA, "DatCat: How to Document a Data Collection," 2008. http://www.caida.org/data/how-to/how-to_document_data.xml.

[15] CAIDA, "Topostats topology statistics calculation tool," 2010. http://www.caida.org/tools/utilities/topostats/.

[16] Department of Homeland Security, "DHS Highlights Two Cybersecurity Initiatives to Enhance Coordination with State and Local Governments and Private Sector Partners," November 2010. http://www.dhs.gov/ynews/releases/pr_1290115887831.shtm.

[17] Government Accountability Office, "Critical Infrastructure Protection: Key Private and Public Cyber Expectations Need to Be Consistently Addressed," July 2010.

[18] "MySQL Community Server." http://www.mysql.com/downloads/mysql/.

[19] "PostgreSQL." http://www.postgresql.org/.

[20] "GNU General Public License 2.0." http://www.gnu.org/licenses/old-licenses/gpl-2.0.html.

[21] "Open Source Initiative: The PostgreSQL License (TPL)." http://www.opensource.org/licenses/postgresql.

[22] VeriSign, "Domain name industry brief," 2010. http://www.verisign.com/domain-name-services/domain-information-center/domain-name-resources/domain-name-report-feb10.pdf.

[23] Bojan Zdrnja, Nevil Brownlee, and Duane Wessels, "Passive Monitoring of DNS Anomalies," in *Fourth GI International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA)*, 2007. http://www.caida.org/publications/papers/2007/dns_anomalies/.

[24] "CAIDA Internet Data." http://www.caida.org/data/.

[25] "Internet forums." http://en.wikipedia.org/wiki/Internet_forum.

[26] CAIDA, "CAIDA's Application to the UCSD IRB," 2008. http://www.caida.org/home/about/irb/caida_irb_app_cover_17oct2008.xml.