The Internet Measurement Data Catalog

Colleen Shannon, David Moore, Ken Keys,
Marina Fomenkov, Bradley Huffaker, k claffy *
CAIDA
San Diego Supercomputer Center
University of California, San Diego. †

Abstract

Internet data remains one of the basic components of computer science network research. Despite its necessity, available data is limited by legal, social, and technical constraints on its collection and distribution. Thus, optimal distribution of knowledge about available data is a valuable service to the research community. To this end, CAIDA has developed the Internet Measurement Data Catalog to:

- provide a searchable index of available data
- enhance documentation of datasets via a public annotation system
- advance network science by promoting reproducible research

This paper describes the impetus, design, and planned deployment of the Internet Measurement Data Catalog.

Categories and Subject Descriptors

H.3.1 [**Online Information Services**]: Data sharing. C.2.3 [**Network Operations**]: Network Monitoring. H.2.8 [**Databases**]: Database Application

Keywords

database, data sharing, measurement, monitoring, Internet, metadata, annotations

1. MOTIVATION

In early 2001, as CAIDA's collection of actively and passively collected Internet data [1] continued to grow rapidly, we began to encounter difficulties in effectively curating our data. Simply tracking data across storage machines as hardware failures and upgrades caused datasets to migrate over time composed one challenge. Another more complex problem lay in tracking the metadata for each dataset – everything from the details about how the collection was performed to problems discovered over subsequent years as the data is used. Moreover, we did not want to solve this problem only within our group; a significant component of our organization involves delivery of relevant Internet datasets to the research community. We wanted a way to effectively communicate information about the datasets we had available, including all of the operational details with the potential to influence research done on that data, to other researchers. Moreover, we wanted other researchers to be able to annotate datasets and give feedback about both the problems and the features they discovered in datasets, thereby increasing the utility of the datasets.

In addition to better documentation for datasets, we hoped that as the number of catalogued datasets increased, standard annotations would allow research questions to be answered directly using the available metadata. For example, some believe that the utility of a network grows as a square of the number of participants [5][1]. This naturally raises questions about how the amount and nature of traffic on the Internet grows as a function of the number of interconnected end hosts. Using just the annotations in a data catalog, one could easily examine the correlation between the number of packets traversing a link and the number of end hosts transmitting those packets at hundreds of measurement points across several years. Many other interesting questions could be answered using only annotations in the catalog.

In August 2001, we submitted a proposal to the National Science Foundation [2] to build a public system for dataset registry and annotation that could incorporate both CAIDA data and any other data available in the networking community. In early summer 2002, we began to build the *Internet Measurement Data Catalog* (IMDC) in earnest.

2. RELATED WORK

2.1 Scalable Internet Measurement Repository

As we began the design process, a recently published paper by Mark Allman et al, "A Scalable System for Sharing Internet Measurements" [3], proved to be an invaluable starting point. The authors detail a possible Scalable Internet Measurement Repository (SIMR) designed to distribute network data and the associated metadata, including details on user information, measurement tools, the dataset collection platform and location, dataset features, experiment information, and relationships between one dataset and others. SIMR also includes type-specific information for datasets, with suggested fields to include relevant information for packet traces. While CAIDA's Internet Measurement Data Catalog differs from the

* {cshannon, dmoore, kkeys, marina, bhuffake, kc}@caida.org
† This work was supported by grant ANI-0137121 from the Advanced Networking Infrastructure program of the National Science Foundation.

[1] In general, Metcalfe's law states that the value of a communication network is proportional to the square of the number of users.

proposed SIMR in many ways, the SIMR paper helped us to frame the system we were trying to build and provided a substantial jumping-off point for our development efforts.

2.2 IST-MOME Database

A similar data indexing and distribution effort by the MOME Project Consortium [7] has been available at `www.ist-mome.org` since April 2004. The MOME Measurement Data Database [4] includes packet, flow, and application traces, as well as routing, HTTP, and QoS data sources. Some data have analysis results available as well, including average traffic rate, packet sizes, arrival rate, and inter-arrival time. Graphical results include packet size histograms and bandwidth use by protocol and by application.

The MOME Database aims to provide data in common, standardized formats through a unified interface. This goal differs slightly from the CAIDA data catalog in that we focus on describing data as is, in whatever format it happens to have, rather than trying to convert all the data in a given category to a specific format. Both styles of data description are useful and valuable, and we expect that both CAIDA's Internet Measurement Data Catalog and the IST-MOME Database will complement each other.

Integration and inter-operation with the MOME Database and any other available repositories of network trace data is a significant part of our future plans for CAIDA's Internet Measurement Data Catalog.

3. GOALS AND BENEFITS OF A DATA CATALOG

Many challenges await those who work in Internet research, including keeping up with the conditions of the ever-changing operational environment, privacy concerns, legal complications, and resource access. One of the most fundamental problems remains access to current data. For many projects, the relevant datasets simply do not exist, and researchers must go through a laborious process of securing permission and deploying measurement infrastructure before they can begin to study the problem at hand. For others, though, the necessary data may exist and even be publicly available. Unfortunately, if word-of-mouth has insufficiently propagated the information about the data ownership and access procedures, researchers may waste time and effort creating a new dataset, use a dataset inappropriate for the problem at hand, or worst of all, abandon the project.

In addition, the dearth of centralized knowledge about available data results in the few datasets that do become widely known being used long after they are no longer an accurate reflection of the current network conditions. Correspondingly, lack of dataset publicity limits longitudinal study of network conditions since comparable datasets that span months or years are difficult to find.

While the resource, legal, and privacy concerns limiting new Internet data collection efforts remain largely intractable, significant research could be promoted through more widespread use of existing data. To that end, CAIDA began developing an Internet Measurement Data Catalog – an index of existing datasets possibly available for research.

In addition to the obvious utility of locating datasets relevant to research projects, a large data catalog provides many other benefits to Internet research. The focus on specifically indexing data provides a forum for robust documentation of data collection procedures. This has the potential to positively impact the scientific merit of studies performed, since the collection process and resulting artifacts in the data can significantly bias the results of data analysis. Vern Paxson's "Strategies for Sound Internet Measurement" paper [6] describes in detail both common pitfalls and best practices for data collection efforts.

Currently, critical experimental design details are largely exempt from the scientific review process, as paper length limits and the perception that data collection minutiae are boring and irrelevant cause data collection details to be elided from papers. In other cases, authors simply do not know the collection process or history behind the data they are using. An independent repository of dataset information allows researchers sufficient space to describe the data collection process, resulting in better documentation of a dataset's strengths and weaknesses that is accessible to paper reviewers and future users of that dataset. Enhanced ability to determine that the results of a paper reflect the system being studied, rather than an artifact of a data collection process, is a significant asset to the Internet measurement community as a whole.

With researchers able to find datasets that are relevant to the topic they wish to investigate, new scopes of research, including both comparison across many sites at a single point in time and trend analysis over long periods of time, become possible. Moreover, since these studies have clearly documented data sources, they represent the heretofore elusive holy grail of reproducible Internet research.

4. ARCHITECTURAL CHOICES AND CHALLENGES

A complete description of the architecture of the Internet Measurement Data Catalog is beyond the scope of this overview; rather, this section highlights some significant decisions made about the structure of the metadata repository.

One of the most prominent features of the Internet Measurement Data Catalog is the fact that external data is not stored by CAIDA or in the database. This eliminates many difficulties associated with the need to purchase and maintain storage systems to hold a mind-boggling volume of data (CAIDA's current data archive includes some 32 TB of uncompressed data; curating our own collection is challenging enough!). While time and resources can solve such technical problems, the ownership, privacy, and legal complications involved in storing data owned by others would prohibit inclusion of many valuable datasets in the repository. Finally, current data provision models span a wide range from free download by anyone, to requiring an internship, to restricting access to site visits with offsite results vetted. Storing only metadata, rather than the data itself, lets the IMDC index datasets with a wide variety of access controls – including those that are not publicly available. One might initially wonder what the point of indexing unavailable data might be, but realities of data provision demonstrate that data that is not available now may be available later. Also, many interesting papers have been published using "unavailable" data. A catalog entry documenting a dataset's existence provides a starting point to broker a relationship or collaboration that results in a significant scientific contribution.

The content of the catalog must somehow accommodate two irreconcilable requirements of its target audience: researchers searching for data want all known information included in the entry for each dataset, while contributors to the catalog want

to do as little work as possible in entering their datasets in the repository. While this is clearly a no-win situation, we compromised by making a minimum subset of information about each dataset required to encourage participation, while relying on our annotation system (described below) to provide researchers with the ability to fill in any missing or otherwise useful information about the datasets they use.

Further simplifying the time and hassle costs to potential contributors, CAIDA has developed an API that will allow contributors with large and/or continuously updating datasets to automate insertion of catalog entries. For contributors with a small, one-time contribution to make, we also have a web-based input system that walks users through the creation of a dataset entry.

4.1 Identification

The ability to uniquely (and concisely) reference datasets is a prerequisite to creating citations. Because a major goal of the IMDC is to allow more complete dataset descriptions in conjunction with published papers, we have developed citable persistent identifiers for data and for all other catalog objects.

4.2 Annotations

The ability for all users to add annotations to catalog objects is a core feature of the IMDC. Annotations from both standard and user namespaces will be available for use. Within any namespace, annotations for various objects can be created. This allows great flexibility in documenting specific types or formats of data. For example, packet traces may include such standard namespace annotations as packet and byte counts, IPv4 packet, byte, and address counts, non-IP-protocol count, collection filter, capture length, link type, and link bandwidth. Annotations can document the presence and scope of data features, such as "collection failure from 15:04 PST to 15:34 PST" or "contains Slammer worm traffic." Thus data contributors will not shoulder the burden of dataset documentation alone, and errors and omissions in catalog entries can be corrected.

4.3 Beyond Data

While datasets are certainly the core of any data catalog, their utility is enhanced by the ability to organize data. The IMDC will include two structures that allow both contributors and users to group related data. A *Study* group is designed to point to the data used to produce a published result. For the IMDC, the definition of publication has limited requirements; the results must be available in some way, but they need not be accepted to a peer-reviewed conference or journal. Technical reports and simple web pages are regularly cited in more rigorously reviewed papers, and tracking the data used to produce such results remains valuable.

However, published results are not by any stretch of the imagination the only useful category by which to organize datasets. The IMDC will also include the ability to categorize data into *Functional Groups*. The defining characteristic of a Functional Group is a stated purpose – anything from "All of the data collected by this monitor in 2005" to "Datasets containing Denial-of-Service attacks" to "Datasets containing measurement card errors" to "My set of most valuable datasets." We expect functional groups to provide a significant method (beyond raw searching) to locate datasets with features in common.

Finally, data is only useful to the extent that one can read and process that data. The IMDC will include tool and tool version information that can be associated with file formats to help researchers best utilize the data available to them. We hope that documentation of bugs and surprising features of dataset collection and analysis tools will increase the integrity of Internet research.

The Internet Measurement Data Catalog has many other interesting features; we hope you will explore them as they become available.

5. THE FUTURE OF THE IMDC

The Internet Measurement Data Catalog will become available in several stages, beginning with the core functionality and adding features as researchers use the system. Initially, the catalog will contain all CAIDA datasets, as well as contributions from several other groups known to collect significant volumes of Internet data. These datasets will be accessible via both simple and advanced search interfaces, and complete information about data access will be available. The next release phase will add the annotation system, with future additions including data organization features (Study and Functional Group), Tool information, and open public contribution to the catalog. We expect exact order of new feature release will depend both on development progress and user input. Our overarching goal is to make access to a significant volume of catalog data available as soon as possible.

5.1 Community Support

The Internet Measurement Dataset Catalog project will not succeed without significant support from the research community. At the most basic level, we hope that users will use the IMDC as a tool to locate datasets for their research, and that they will contribute annotations to those datasets to describe any significant problems or features they discover. IMDC's utility also depends on researchers contributing catalog entries for the data they collect. While we encourage researchers to make the datasets entered in the catalog publicly available, we recognize that many barriers to sharing Internet datasets exist and that documentation of unavailable data remains a significant contribution to the catalog.

However, opportunities for research community support do not end with individual use of the catalog. A significant problem limiting the public contribution of datasets is the lack of incentives for researchers to divert time, energy, and resources from their other work in order to provide data to others. Unfortunately, the performance evaluation for researchers in their jobs rarely values data provision, and the scientific value of reproducible results is not an influence on paper acceptance. There are many ways that the Internet research community could provide incentives for making data available, including:

- devoting one session (typically three papers) out of a multi-day conference to reproducible papers – those with the data available at the time of the submission process.
- providing a conference Best Paper using a Public Dataset award
- providing a conference award for the Best Public Dataset

While such additions might indirectly assist the IMDC, the majority of their impact would benefit the measurement community at large.

6. CONCLUSION

CAIDA's Internet Measurement Data Catalog will soon provide the research community with a persistent source of information about available Internet data. A flexible annotation system will allow the catalog to incorporate pertinent details about current and future data types, as well as recording user comments and corrections for data objects. We anticipate that the benefits of the IMDC will include encouraging users to make more datasets available, enhancing documentation of data collection methodologies and their impact on subsequent research, providing unique identifiers for data used in research studies, expanding the scope of research studies to include more comprehensive data across longer timescales, and perhaps most importantly, promoting reproducible research.

For future information about the Internet Measurement Data Catalog, including notification of our grand opening, subscribe to our `imdc-announce@caida.org` mailing list by visiting `http://mail.caida.org/mailman/listinfo/imdc-announce` or sending email to `imdc-announce-subscribe@caida.org`.

7. ACKNOWLEDGMENTS

Marina Fomenkov, Bradley Huffaker, Ken Keys, David Moore, and Colleen Shannon form the design team for CAIDA's Internet Measurement Data Catalog project. Ken Keys is responsible for the implementation of the data catalog, and Marina Fomenkov is the project manager. However the goals and design of this project are truly a group effort; the final project design is greater than the sum of the individual visions from whence it came.

The Mark Allman et al paper [3] described in section 2.1 exerted a seminal influence on the eventual structure of the Internet Measurement Data Catalog. We remain grateful to the authors for their clear treatment of a revolutionary idea.

We would also like to thank the many folks (they are too numerous to mention individually) who attended our ISMA Data Catalog Workshop, attended our presentation at the 2004 Internet Measurement Conference, or otherwise provided feedback on catalog design and features. Significant improvements would not have been possible without your insightful comments.

Development of the Internet Measurement Data Catalog would not have been possible without the support of the NSF Advanced Networking Infrastructure program, and specifically, without the patience and encouragement of our program manager, Kevin Thompson.

8. REFERENCES

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

[2] The National Science Foundation. http://www.nsf.gov/.

[3] Mark Allman, Ethan Blanton, and Wesley M. Eddy. A Scalable System for Sharing Internet Measurements. In *Proceedings of the 2002 Passive and Active Measurement Workshop*, Fort Collins, USA, March 2002. http://www.icir.org/mallman/papers/simr-pam2002.ps.

[4] Pedro A. Aranda Gutierrez, Antal Bulanza, Marek Dabrowski, Baiba Kaskina, Juergen Quittek, Carsten Schmoll, Felix Strohmeier, Attila Vidacs, and Kardos Sandor Zsolt. A Scalable System for Sharing Internet Measurements. In *Proceedings of the IPS-MoMe2005 Workshop*, Warsaw, Poland, March 2005. http://www.ist-mome.org/publications/mome_paper.pdf.

[5] Andrew Odlyzko and Benjamin Tilly. A refutation of Metcalfe's Law and a better estimate for the value of networks and network interconnections, March 2005. http://www.dtc.umn.edu/ odlyzko/doc/metcalfe.pdf.

[6] Vern Paxson. Strategies for Sound Internet Measurement. In *Proceedings of the 2004 ACM Internet Measurement Conference*, Taormina, Italy, October 2004. http://www.icir.org/vern/papers/meas-strategies-imc04.pdf.

[7] The MOME Project Consortium. Information Technologies Society - Cluster of European Projects aimed at MOnitoring and MEasurement. http://www.ist-mome.org/.