# Overview/Lessons Learned

cooperative association for internet data analysis

Bradley Huffaker <bradley@caida.org>

GEC7 - 17 March 2010

# Goal

DatCat was designed to improve data sharing by providing a unified metadata database for Internet data.

# Goal

**Makes the following processes easy for users.**

- finding data sets of interest
  - Many researchers lack access and/or expertise to collect data needed for their research.
- adding new data sets to the catalog
  - Contributors (who are generally underfunded and providing data out of dedication to the general good) want to minimize time lost to their own research.
- annotating data sets in the catalog
  - Provides a flexible way for contributors and users to mark up interesting facts about data sets. Such as the number of packets or that a given file is corrupted.
- does not store data

# overview

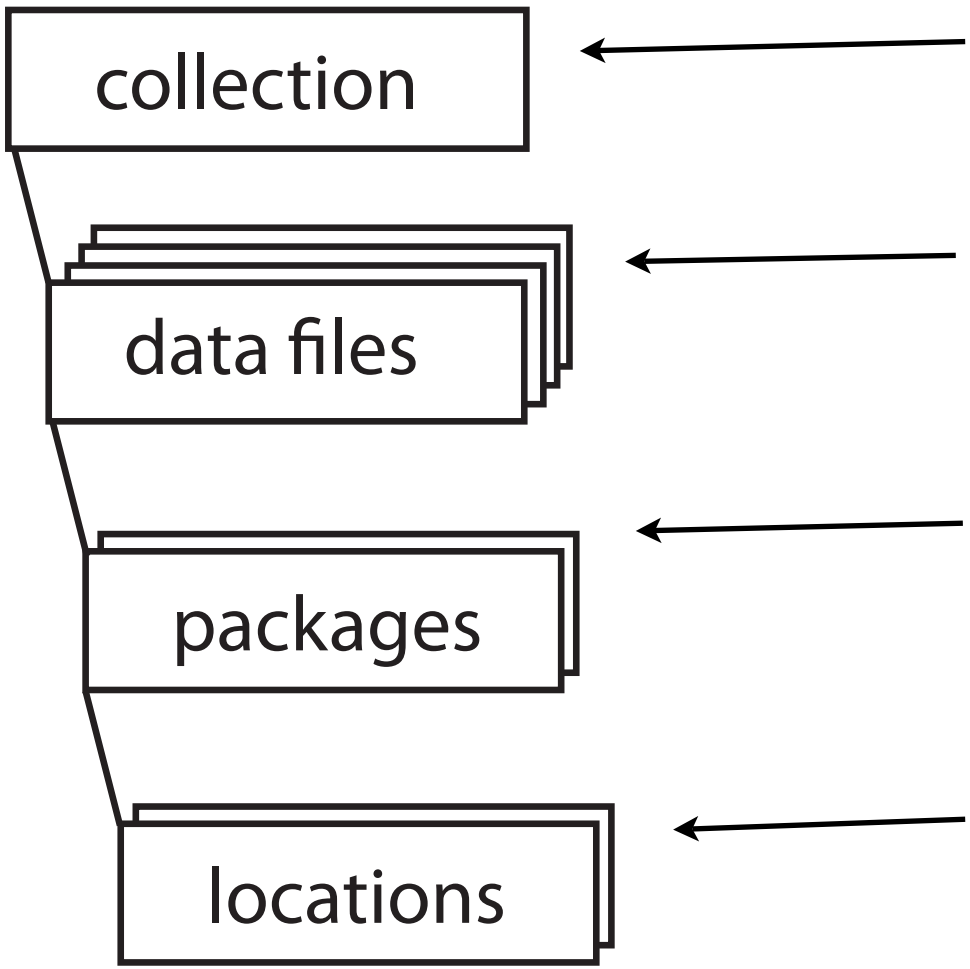
- data scheme
- submission
- web portal
- status/limitations
- future work

# Database Scheme

caida

DatCat

data scheme




logical group of files
(paper, project,...)

raw data files
(traceroutes, logs dumps, ...)

downloadable files
(single file, tarball, ...)

how to get the packages
(URL, contact address, ...)

http://www.datcat.org/help/contributing

# Annotation Key

caida

DatCat

data scheme

Annotations provide an extensible naming space for assigning domain specific values to files.

- each user has their own hierarchical name space
  - passive.IPv4.packet_count
  - active.RTT_95th_percentile
- both data contributors and general DatCat users may attach annotations
- any user may assign “note” annotations to any object

# Metadata Fields

data scheme

- collection
  - **fields:** name, contents, summary, motivation, creators/primary contact/contributor, start/end time, keywords, short description/description/description URL
  - **annotations:** note
- data
  - **fields:** name, creators/primary contact/contributor, keywords, format, file size, start/end time, duration, geographic/network location, time zone, MD5, description, creation process
  - **annotations:** passive.IPv4.packet_count, passive.IPv4.TCP.dst.port_count, cfg.passive.capture_len, AS_count, active.trace_count, active.RTT_10th_percentile, .....
- location
  - **fields:** package, creators, primary contact, status, download procedure, download URL, geographic/logistic location, availability

http://www.datcat.org/help/contributing

# Example

DatCat

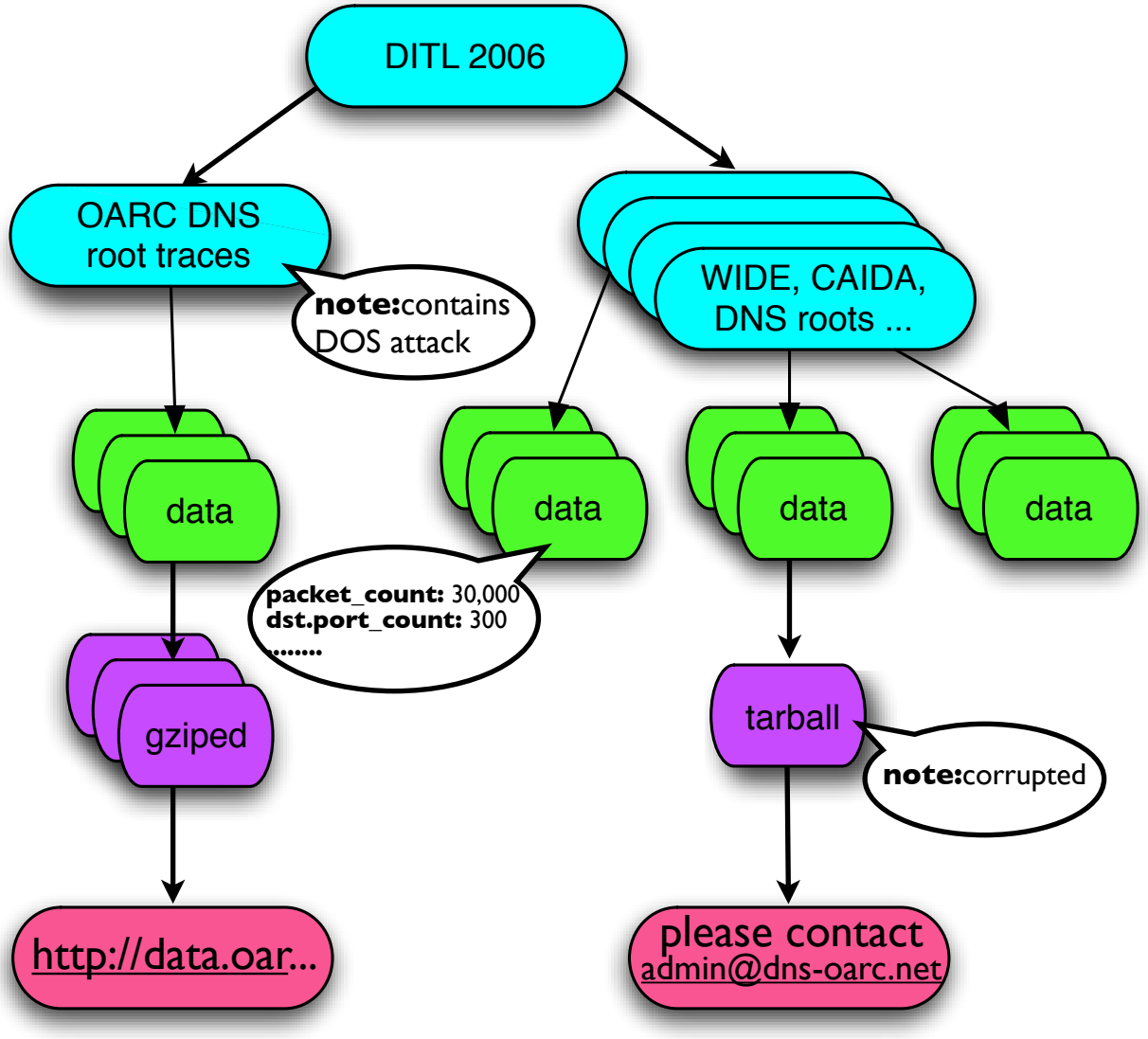
data scheme

collections

data files

packages

locations




http://www.datcat.org/help/contributing

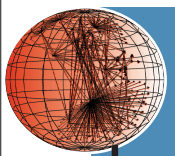
# DatCat Submission

DatCat submission

## contribution tools

- Perl API
  - useful for integrating into existing data management systems
  - flexible, but need to write code:
- *subcat*
  - different approach (declarative)
  - preferred interface (we use it ourselves)
  - available since DCC1 workshop, and improved since

http://www.datcat.org/help/contributing



# subcat

DatCat

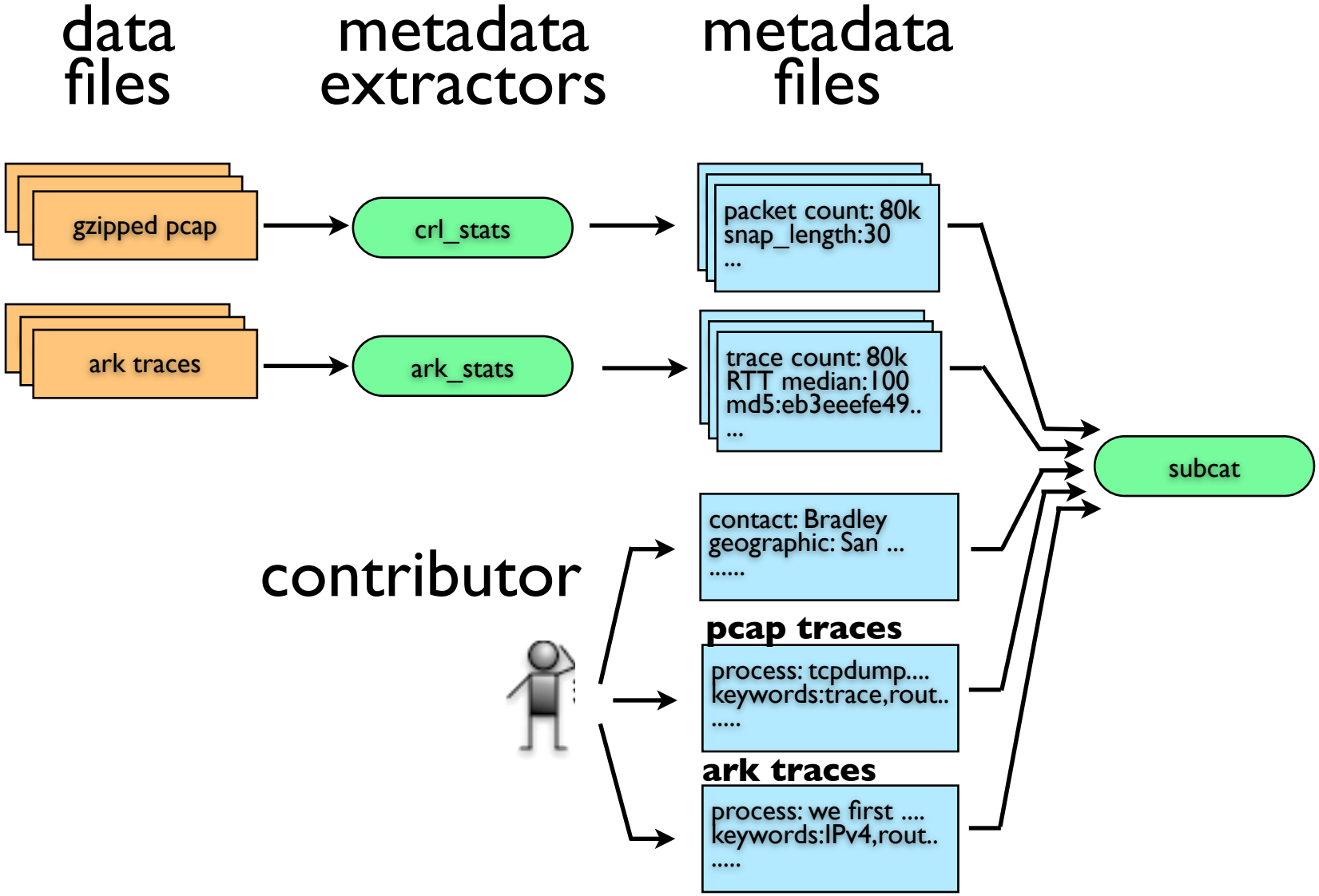
## submission

- describe metadata in human-friendly text files (YAML)
- CAIDA provides tools to extract additional metadata (data-to-yaml)
  - pcap, gz, zip, tgz, dag, ...
  - write your own extractor
- subcat intuitively joins information together
  - templating
    - defaults
    - categories (e.g. pcap and snmap category)

http://www.datcat.org/help/contributing

# subcat Input Files

DatCat submission




http://www.datcat.org/help/contributing

# subcat File

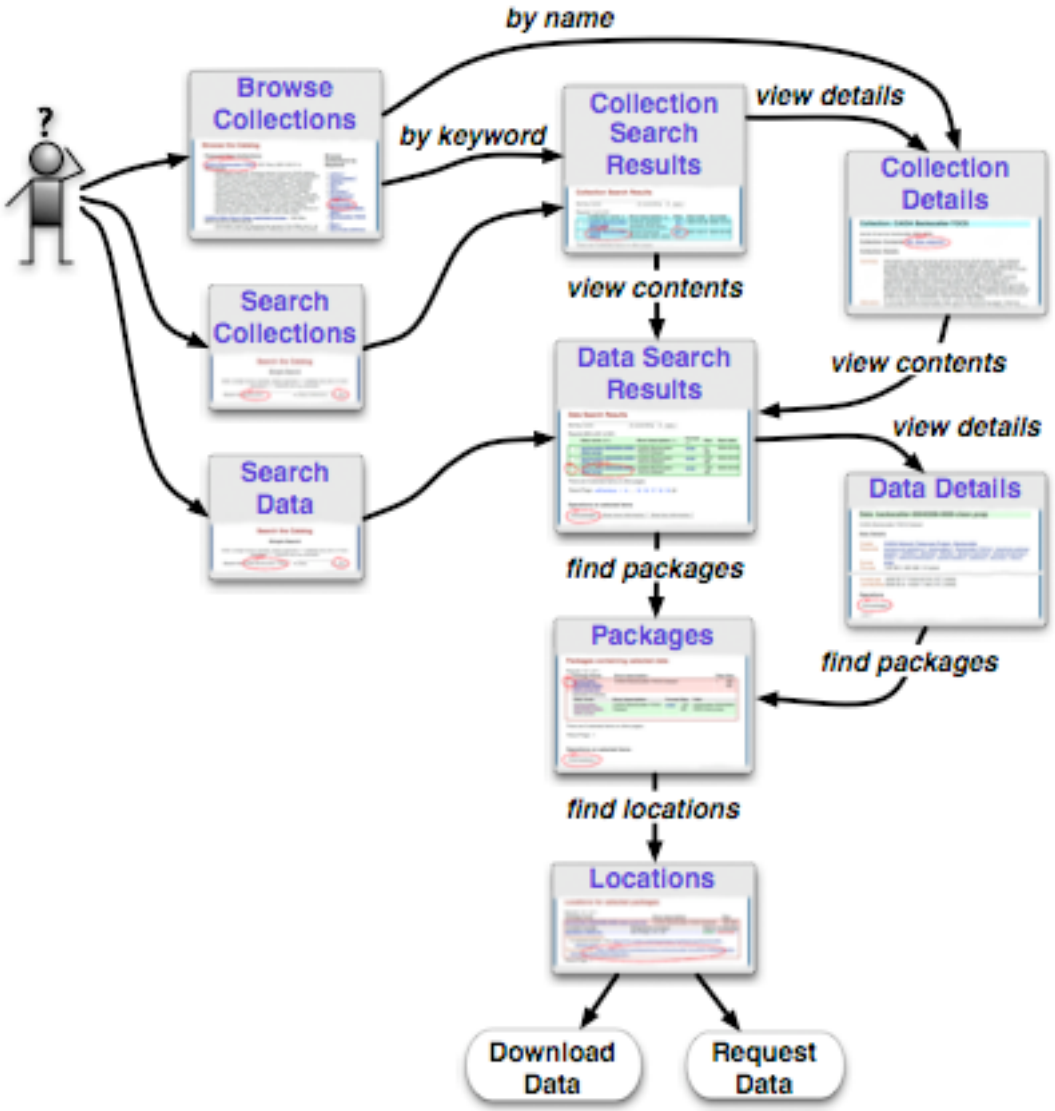
DatCat submission

```
.object: collection
name: Day in the Life of the Internet (DITL)
creators: contact.caida_ditl
primary_contact: contact.caida_ditl
short_description: simultaneous Internet measurement events
keywords: DITL, synchronized, DNS, DNS roots
motivation: This collection groups all Day in the Life of the Internet measurements.
summary: >-
    The Day in the Life of the Internet (DITL) measurement project aims to provide
    simultaneous capture of a variety of worldwide Internet measurements
    for further analysis by research scientists.
description_markup: html
description: >-
    The Day in the Life of the Internet (DITL) measurement project aims to provide
    simultaneous capture of a variety of measurements from and across many
    strategic links around the globe for further analysis by research scientists.
    <p>
    Examples of possible measurements are:
    <ul>
    <li>Packet traces from the DNS root nameservers and AS112 servers</li>
    <li>Packet traces from backbone links</li>
    <li>Netflow data</li>
    <li>Topology data</li>
    <li>Logs and traces from critical infrastructure, such as DNS</li>
    </ul>
description_url: 'http://www.caida.org/projects/ditl/'
start_time: 2006-01-10 00:00:00 UTC
duration: ongoing
```

http://www.datcat.org/help/contributing

# Web Portal

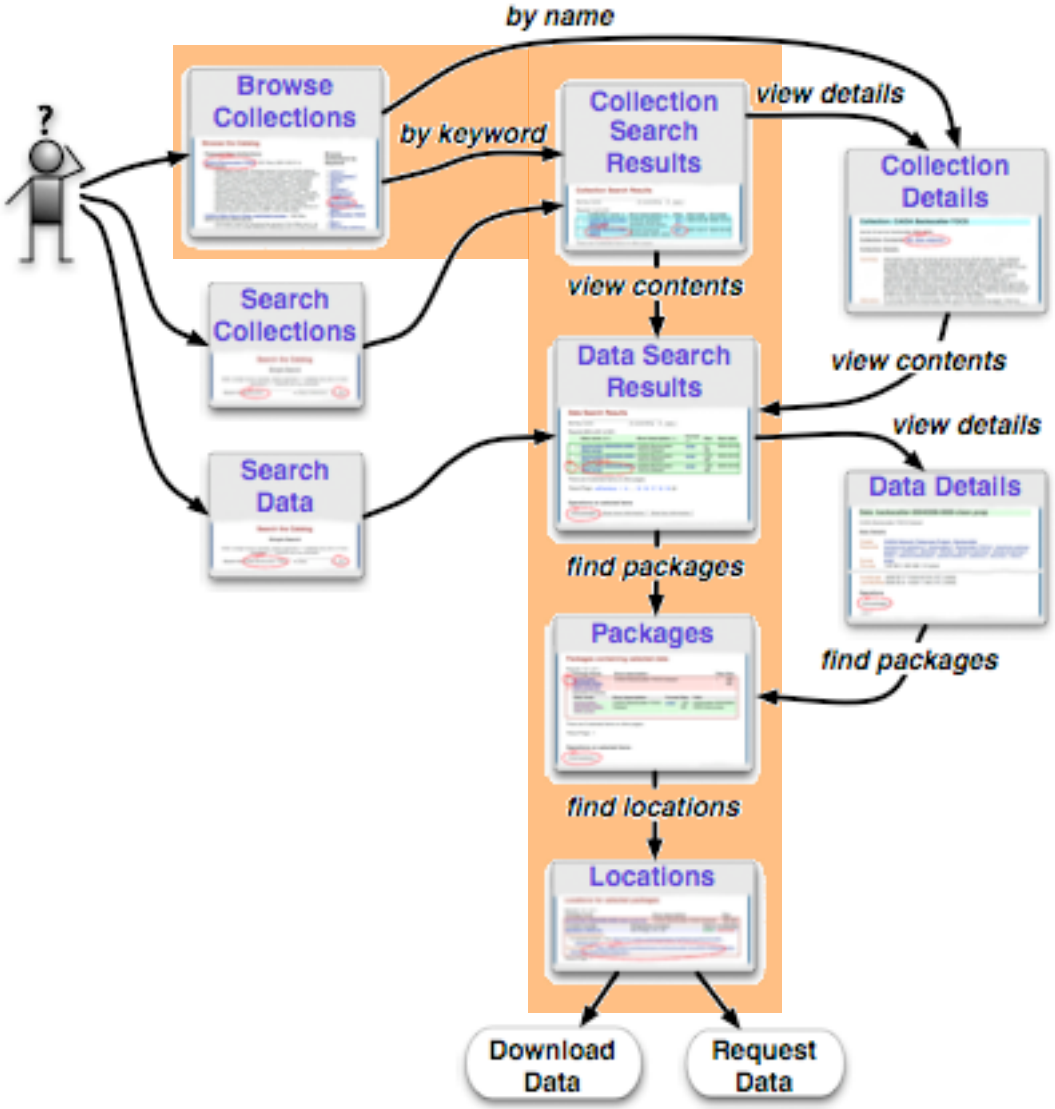
web portal




http://www.datcat.org/browse

# Web Portal

web portal




http://www.datcat.org/browse

# Status of DatCat

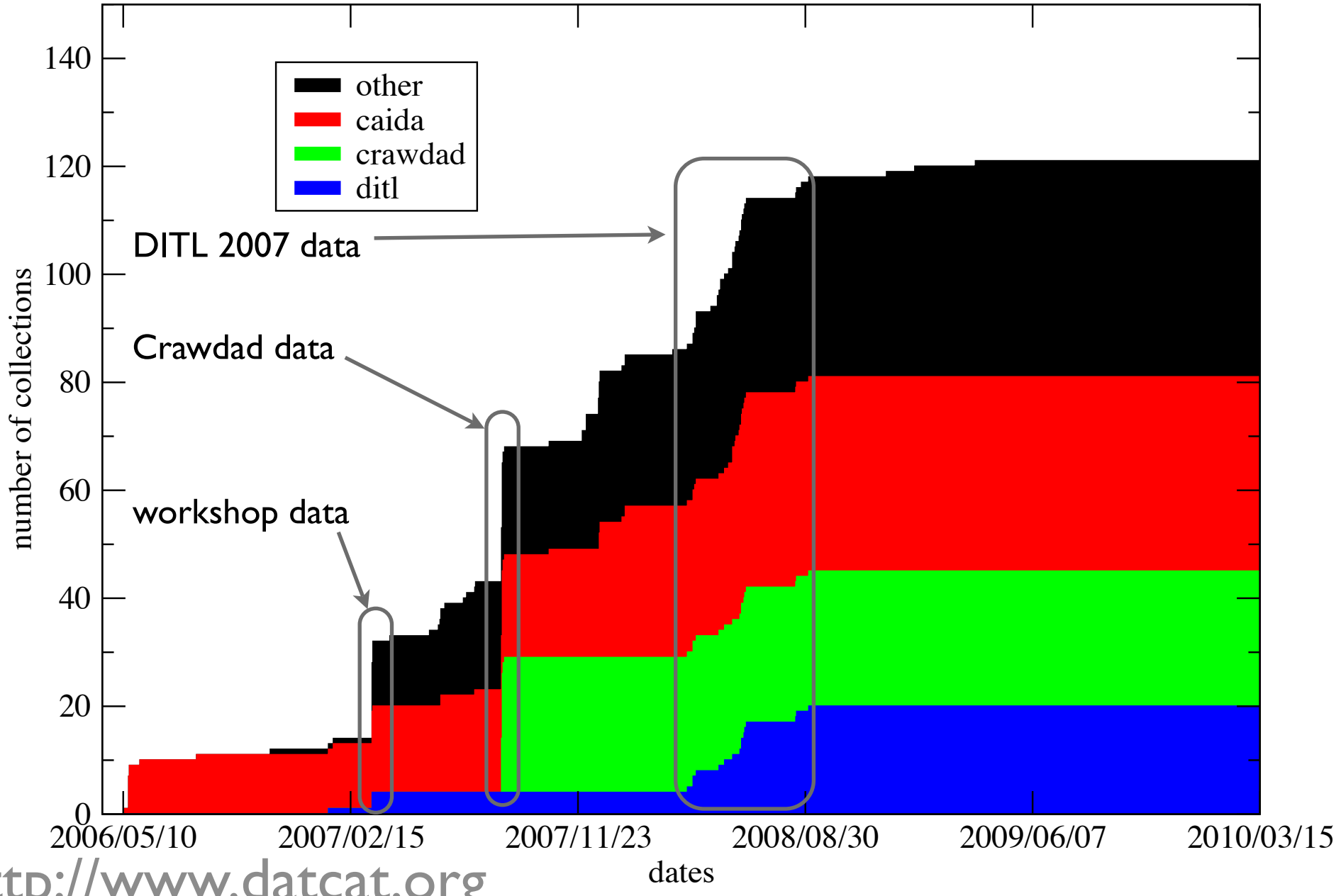


106 collections of data + 12 publications

- **All contributors needed help from CAIDA**
- **DatCat contains**
  - 121 packages
  - 154k data objects
- process blocked on funding (ended in 2008)

# History

## status/limitations



# Lessons Learned

caida

DatCat

status/limitations

file-level metadata hard

- hard to fix errors across thousands of files
- hard to display thousands of files
- hard to generate

submission process too cumbersome for most users

- majority of metadata is shared between files
  - creator, creation process, location, etc
- many researchers are not programmers
- researchers have limited time and motivation

# Lots of Redundant Information

caida

DatCat

status/limitations

For a single contribution, a majority of data objects have identical metadata shared across a large number of data objects.

- data in the database

  name, creators/primary contact/contributor, keywords, format, file size, start/end time, duration, geographic/network location, time zone, MD5, description, creation process

- could be solved by pushing subcat-type categories into the database

# Future



## From file to collection focus.

stand-alone collections

- users will add collections without the collection's individual files

focus on helping researchers find contact for collections rather than individual files

single transaction contribution

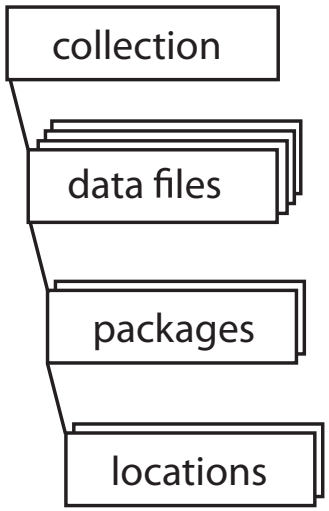
- no tools to download
- should take less then 15 minutes

# Stand Alone Collections

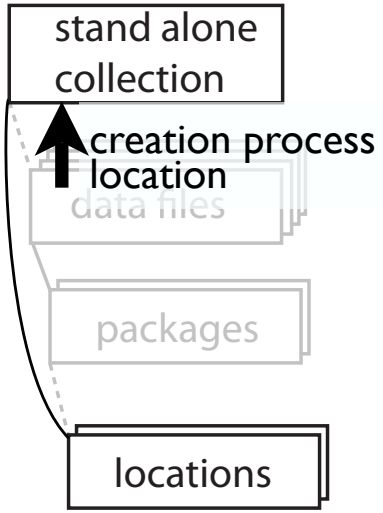
DatCat

future

current




additional option




- users already search by collections
- contributors will only need to fill in the collection information, plus
  - creation process
  - location
  - data files optional
- shorten search path from collection to locations

# DatCat Forum

caida

DatCat future

DatCat℠
Internet Measurement Data Catalog

Username : User Name
Password :
Log in

DatCat Community Forum

HOME | BROWSE | SEARCH | FORUM | HELP

## Discussion Boards

**GENERAL DISCUSSIONS**

| TOPIC | CREATED | POSTS | LAST REPLY |
|---|---|---|---|
| Please help me to download skitter. thank you. 1.2K | 12/1 by Zan Yoo | 3 | 1 hour 50 min ago by paulwright |
| CIDR block -> AS Number mapping 1.0K | 2/23 by Adam Smith | 1 | 3 days 3 hours ago by Dezy Camino |
| question of inferring ISP 1.6K | 2/25 by Mark T. Taylor | 19 | 7 hours 47 min ago by Mathieu Warda |
| Building a topology using the IPv4 Routed /24 Topology Dataset 1.3K | 4/28 by Eunice Zarabi | 6 | 22 hours 47 min ago by Theo de Jong |
| CAIDA and IXP 2.9K | 5/19 by Sarah Combe | 5 | 5 hours 14 min ago by Mark Simonson |
| AS Relationships Dataset, link surge between 01/Sept/2008 -> 10/Sept/2008? 27K | 7/24 by WilliamTrax | 1 | 14 hours 57 min ago by J Weltin |

# Final Thoughts

future

- contributions process too complicated for users
  - contributions often required multiple attempts
  - contributors no longer remember metadata
  - metadata difficult to create for large data sets
- researchers have limited time/skills/motivation
- make the "simple" first, the "better" later
- better to have lots of collections, then lots of files