Cataloging DITL data for research use

Emile Aben <emile@caida.org>

WIDE - Jan 2008, Honolulu, HI, US

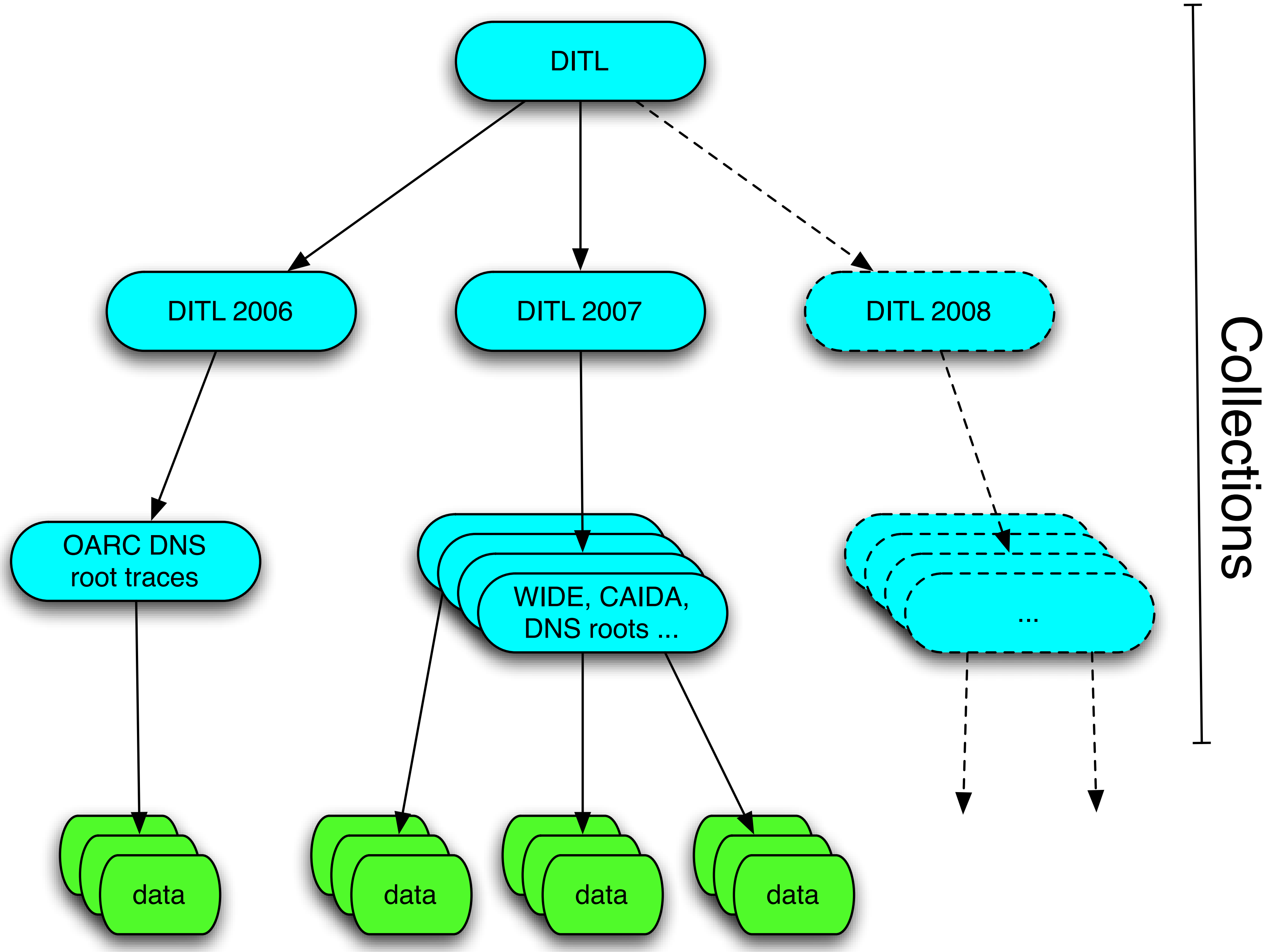
DatCat Catalog

- Internet Measurement Data Catalog
  - Searchable registry of information about network measurement datasets
  - Doesn’t store data itself
  - http://imdc.datcat.org

DatCat highlights 2007

- Datasets indexed in 2007 include:
  - Datasets from DITL 2007 (some still in progress)
  - Datasets from CRAWDAD (wireless)
  - Datasets from DCC1 workshop
- 284 registered users
- > 14k searches performed
- 14.5 TB data indexed

Structure of DITL in DatCat



Tour of DITL 2007 in DatCat

- *http://imdc.datcat.org/collection/1-031B-Q*
- find DITL 2007 through:
  - search for ‘DITL’
  - browse by keyword ‘DITL’ (or ‘DITL-2007-01-09’)
  - browse featured collections
  - ...

Describing datasets in DatCat

- contributing takes time and thought
  - how to best describe your dataset
  - collecting meta-data (some in advance)
  - possibly processing large amounts of data

Describing datasets in DatCat

- contributing takes time and thought
  - how to best describe your dataset
  - collecting meta-data (some in advance)
  - possibly processing large amounts of data
- worth the time and thought!
  - structurally enhances documentation
  - lets people know about your data

Describing datasets in DatCat

- contributing takes time and thought
  - how to best describe your dataset
  - collecting meta-data (some in advance)
  - possibly processing large amounts of data
- worth the time and thought!
  - structurally enhances documentation
  - lets people know about your data
- there are tools to help

Collecting Meta-data

- Meta-data to be recorded at collection time
  - generally by a human, some can be automated
  - examples: creation process (vlan), platform
- Meta-data that can be obtained by processing data
  - can be automated
  - example: IPv4 packet count in pcap trace
- How to document a data collection:
  - http://www.caida.org/data/how-to/how-to_document_data.xml

Meta-data fields in DatCat

- defined set of meta-data fields per object
  - for a collection:
    - name
    - contents
    - summary
    - motivation
    - creators/primary contact/contributor
    - start/end time
    - keywords
    - short description/description/description URL
- annotations allow for defining additional meta-data fields

Submitting to DatCat

- contribution tools
  - Perl API
    - useful for integration in existing data management system
    - flexible, but need to write code:

```
$submission = new IMDC::Submission;
$data1 = $submission->newData({name=>‘z-root pcap’});
$data1->short_desc(‘z-root pcap trace’);
...
```

  - *subcat*
    - very different approach (declarative)
    - preferred interface (we use it ourselves)
    - available since DCC1 workshop, and improved since

(more on next slide)

Contribution with *subcat*

- describe meta-data in human-friendly text files (YAML)

- use tools to extract additional meta-data (*data-to-yaml*)
  - pcap, gz, zip, tgz, dag, ...
  - write your own extractor

- *subcat* intuitively joins information together
  - templating
    - defaults
    - categories (e.g. pcap and snmp category)

Syntax example

```
---
.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
```

Conclusion

- some thoughts for next DITLs
  - import sooner rather than later
    - timetable for contributions to DatCat
  - more help in contributing / extracting meta-data?
  - provide people with example templates for meta-data
    - pcap for all traffic on a link
    - DNS pcap
  - what extra meta-data to capture
    - DNS stats?
- suggestions / questions?

Links

- DatCat: http://imdc.datcat.org
- DITL 2007 in DatCat:
  - http://imdc.datcat.org/collection/1-031B-Q
- contributing:
  - contribute@datcat.org
  - http://imdc.datcat.org/help/contributing
- http://www.caida.org/data/how-to/how-to_document_data.xml