Lessons Learned

cooperative association for internet data analysis

Bradley Huffaker <bradley@caida.org>

CAIDA/WIDE/CASFI Workshop - 12 June 2006

Goal

DatCat was designed to provide a unified metadata database for Internet data.

Data in DataCat

caida

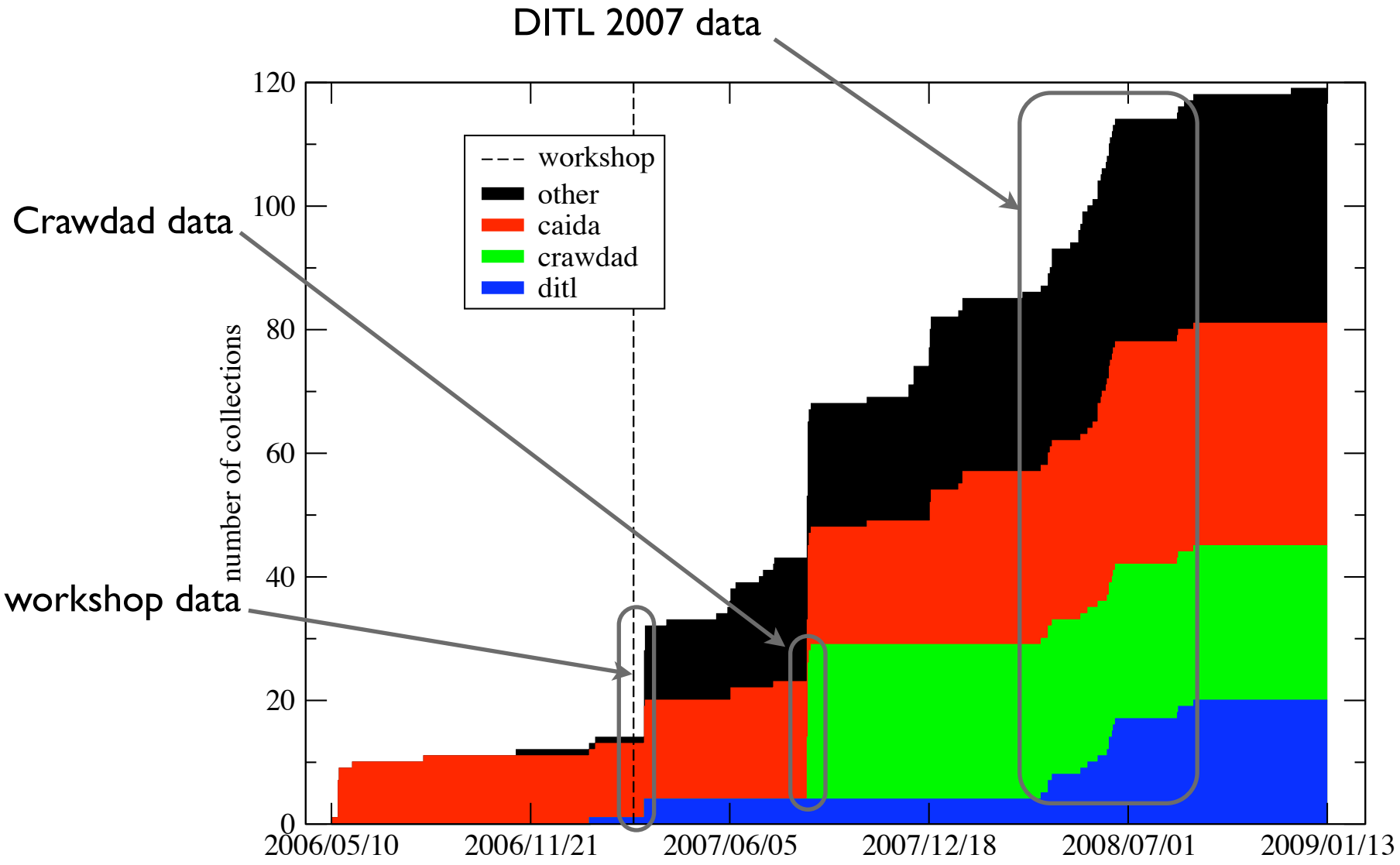
DatCat

http://www.datcat.org

- 104 collections of data + 13 publications
  - **All contributors needed help from CAIDA in one form of the other**
- 151k data objects, representing 20 TB of data
- 127k package objects, representing 11.5 TB of files

History

http://www.datcat.org



Lessons Learned

DatCat

http://www.datcat.org

- file level metadata hard
  - hard to fix errors across thousands of files
  - hard to display thousands of files
  - hard to generate

- process is too cumbersome for most users
  - majority of meta data is shared between files
  - many researchers are not programmers
  - researchers have limited time and motivation

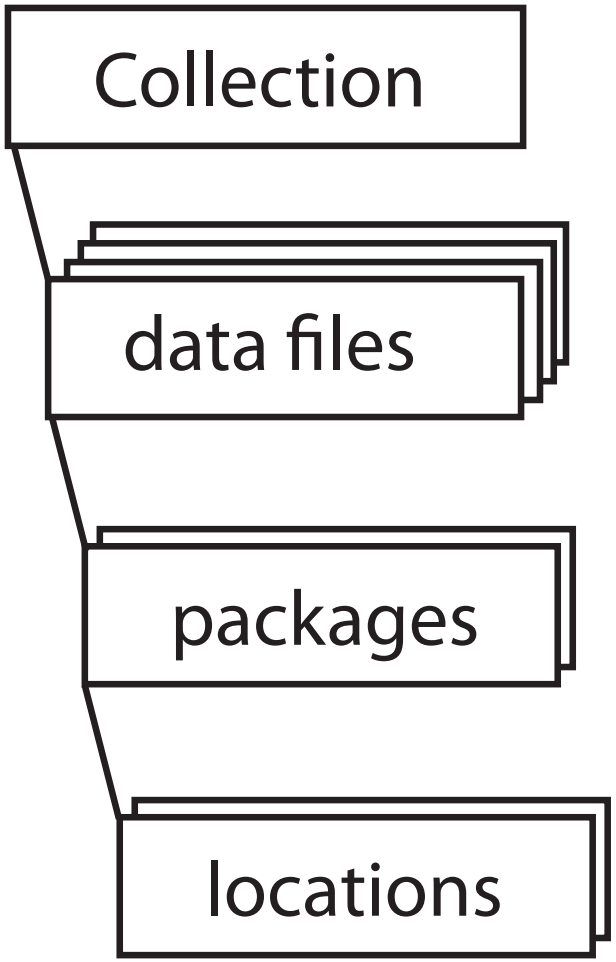
Future

From file to collection focus.

- stand alone collections
  - users will be able to add collections without the collection's individual files
- focus on helping researchers find contact for collections rather then individual files
- single transaction contribution
  - no tools to download
  - should take less then 15 minutes
  - funding has runs out

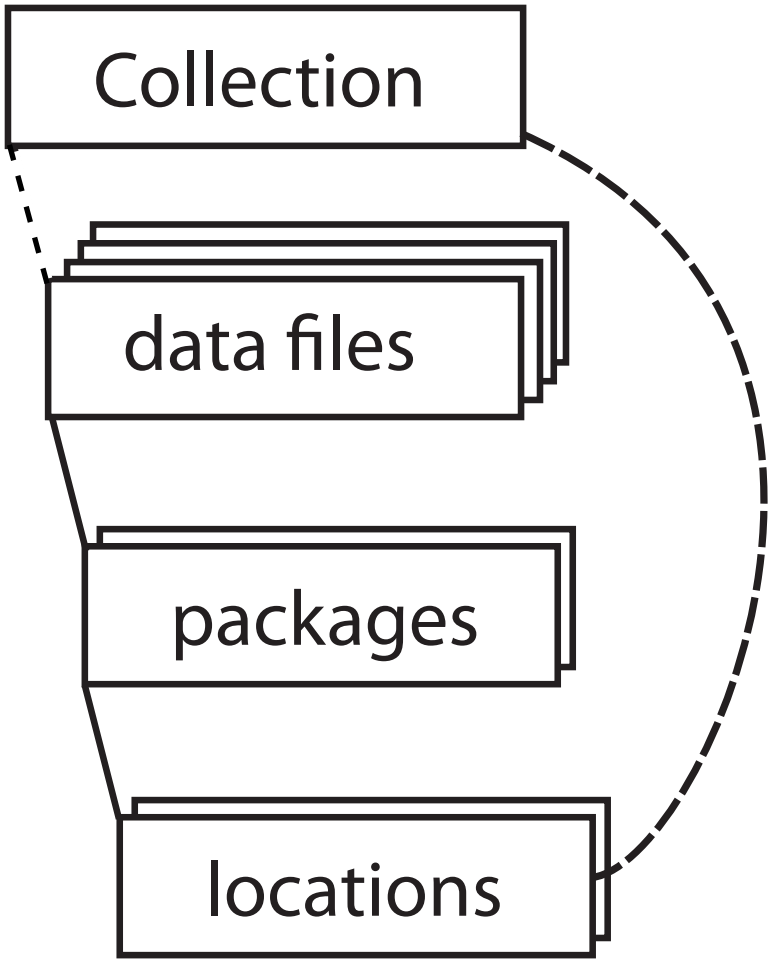
Stand Alone Collections

http://www.datcat.org



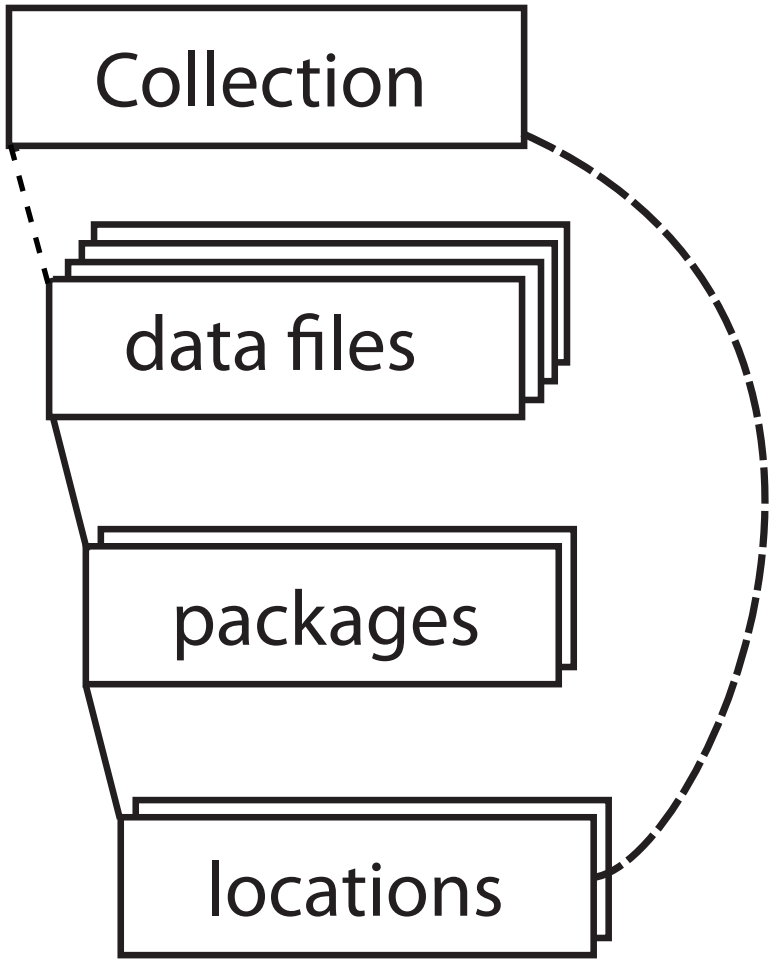
Stand Alone Collections

http://www.datcat.org



Stand Alone Collections

http://www.datcat.org



- users already search by collections
- contributors will only need to fill in the collection information
- shorten "clicks" from collection to location

Final Thoughts

DatCat

http://www.datcat.org

- contributions too complicated
  - contributions often required multiple attempts
  - contributions often block on moderator time
- researchers have limited time/skills
- make the “simple” first, the “better” later
- better to have lots of collections, then lots of files