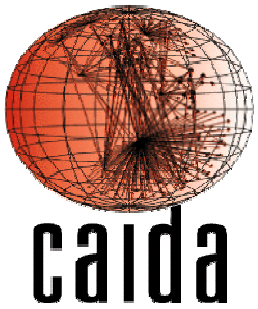
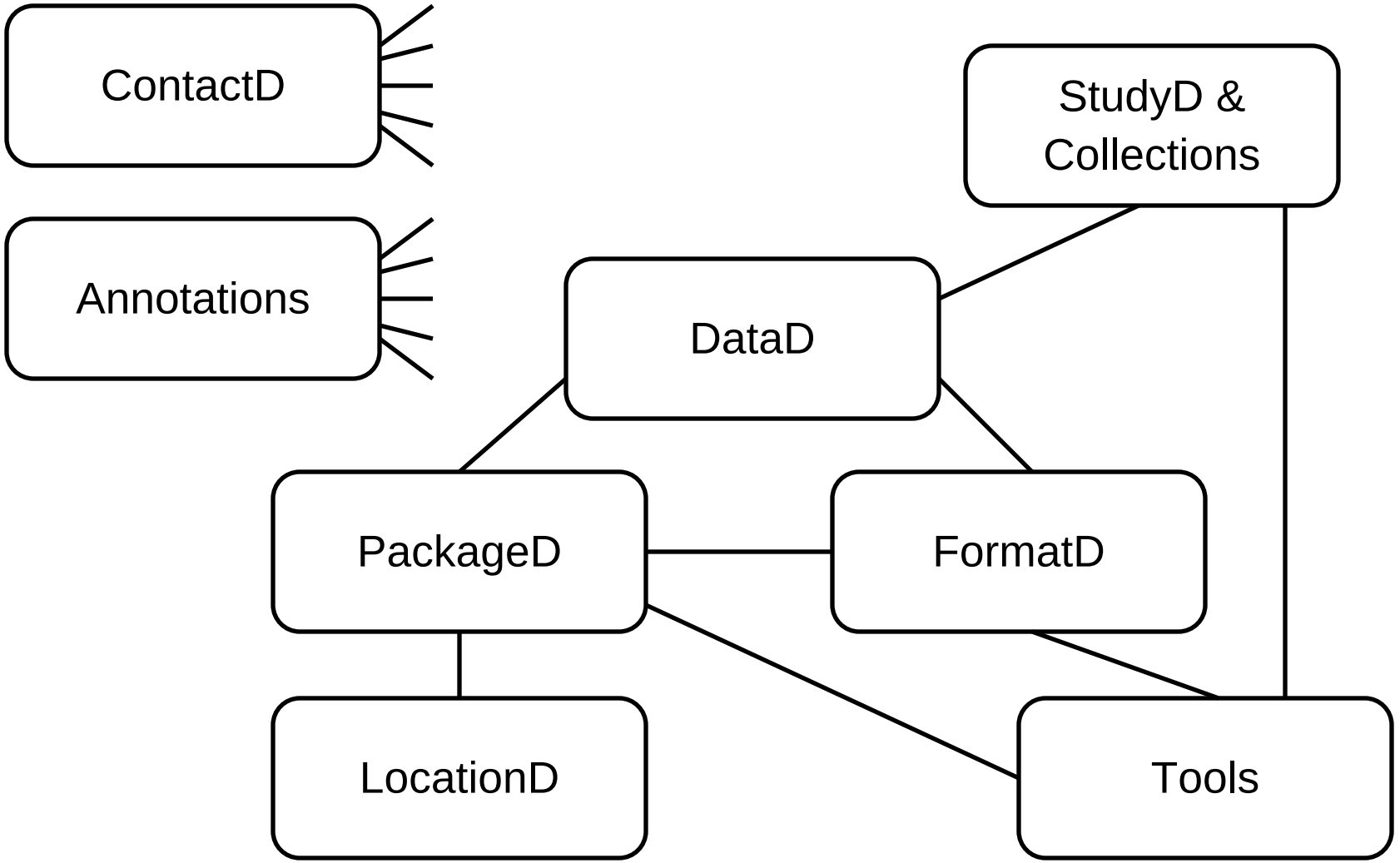
Internet Measurement Data Catalog

System Design



Simplified System Diagram



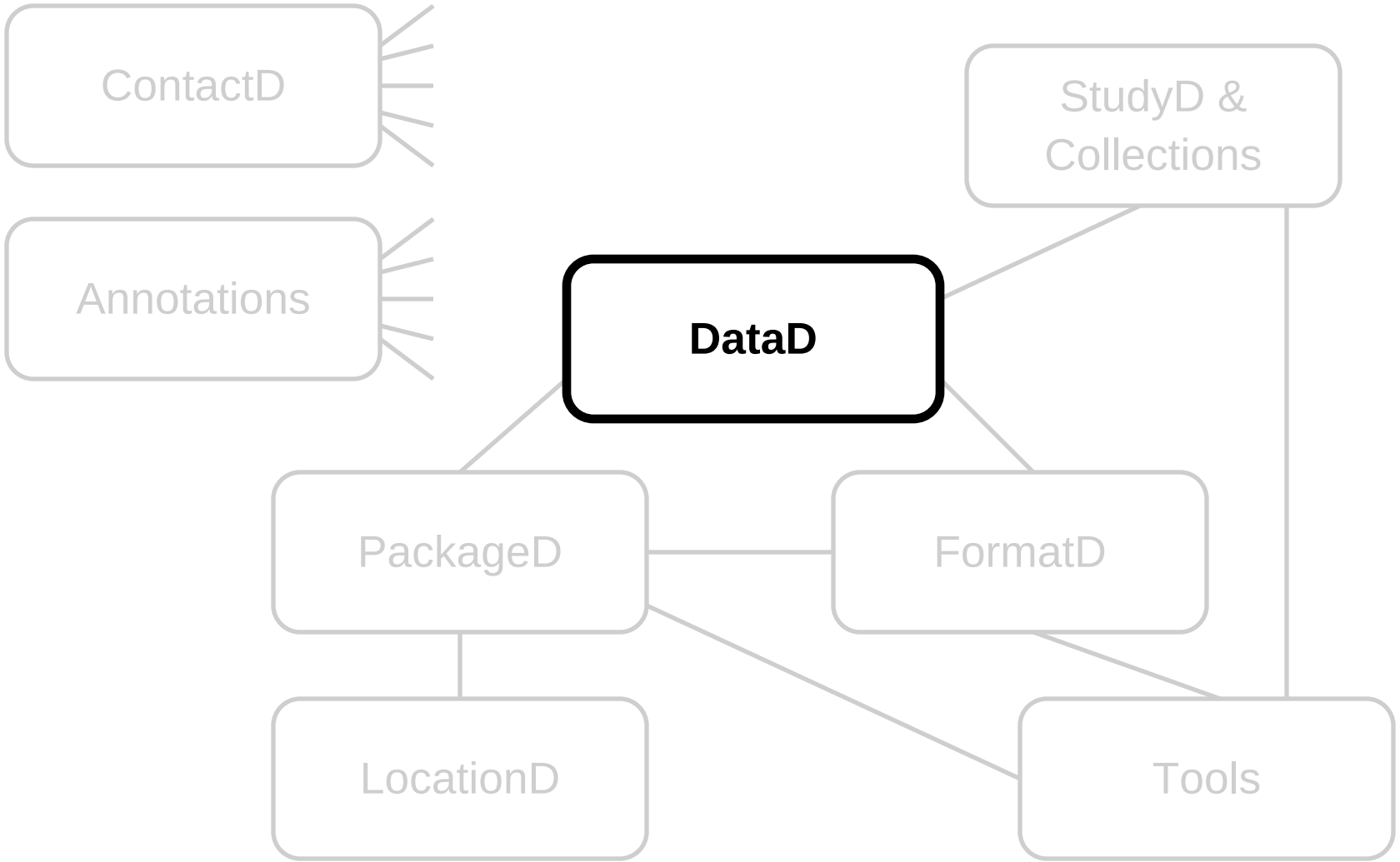
IMDC Focus

- Helping researchers find available network data.
- Central conceptual object: **Data Descriptor**.
- Everything else is to facilitate that, although some components are useful on their own.
- Note: IMDC does not store raw network data, just information about that data.

Data Descriptor (DataD)

- Maps conceptually to the level of a "file".
- Shared between all references to the same data item, even if available in multiple ways.
- Example:
  - "Route Views table dump on Dec 1, 2003 at 8am"
  - This same data item is available from:
    - routeviews.org compressed with bzip2
    - packetclearinghouse
    - with some skitter datasets

Simplified System Diagram



Data Descriptor Fields

- Name
- Description
  - long, short, URL
- Keywords
- File Size
- Format
- Location
  - geographic, network, logistic
- Platform
- Time Period
  - start time, end time, TZ offset, TZ name
- Creation Process
- md5 Hash

Common Fields

- Fields which appear throughout the system:

- Creator
- Contributor
- Creation Time
- Modification Time

Excluded DataD Fields

- Supporting tools – handled via format (next slide)
- Notes – handled by annotations (later slide)
- Format-specific information – annotations
  - Ex: pcap snaplen is 72

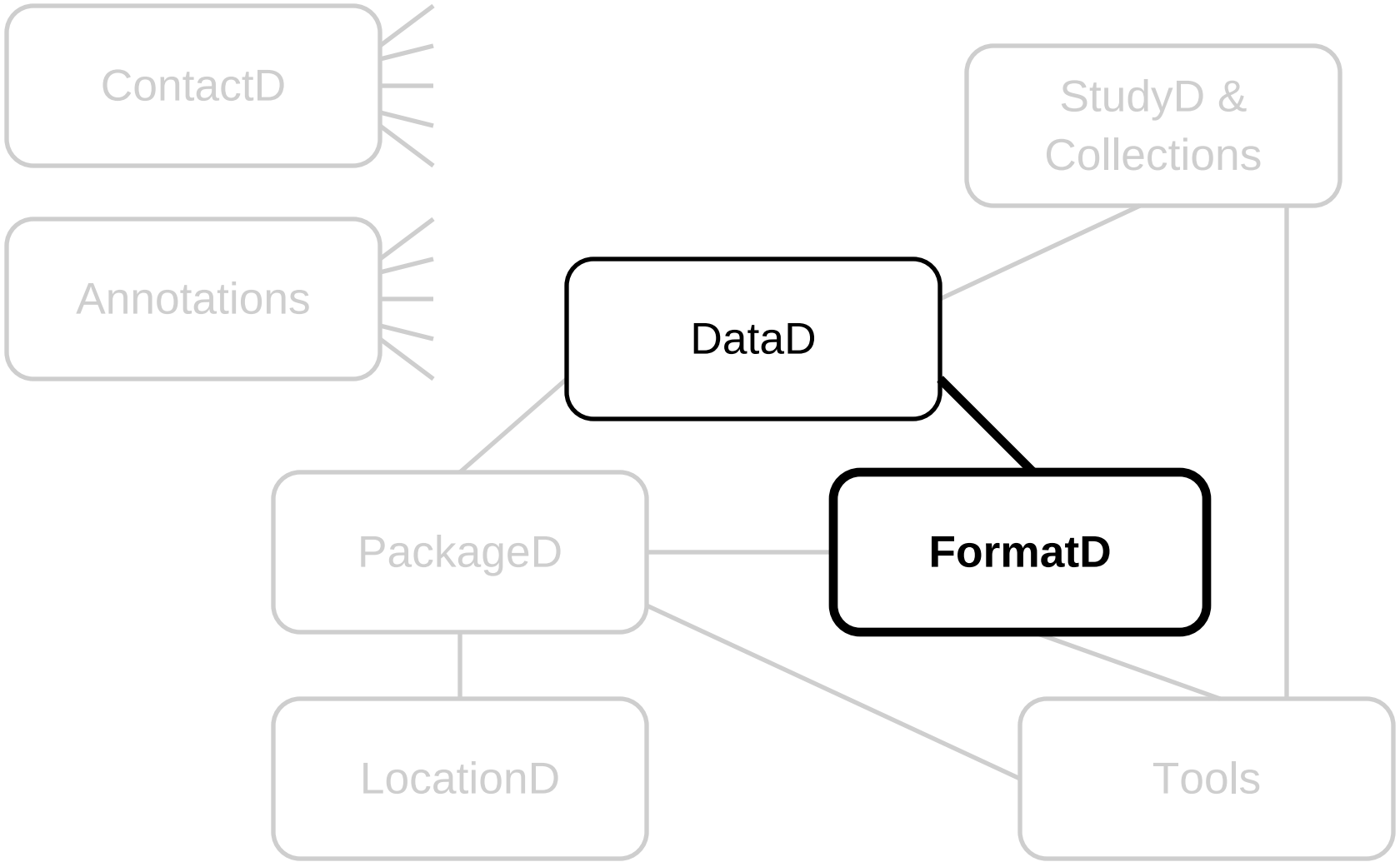
Format Descriptor (FormatD)

- Once a researcher gets some data, how can they process it? Or, one may only want to limit searches to particular formats, etc.

- Pointers to information about file formats.

- Mappings to tools (later slide) which read/write this format.

Simplified System Diagram



Format Descriptor Fields

- Name
- Description
  - long, short, URL
- Keywords
- Package or Data Format
- Type – ascii/binary/mixed
- File Suffix
  - list of suffices, need not be unique

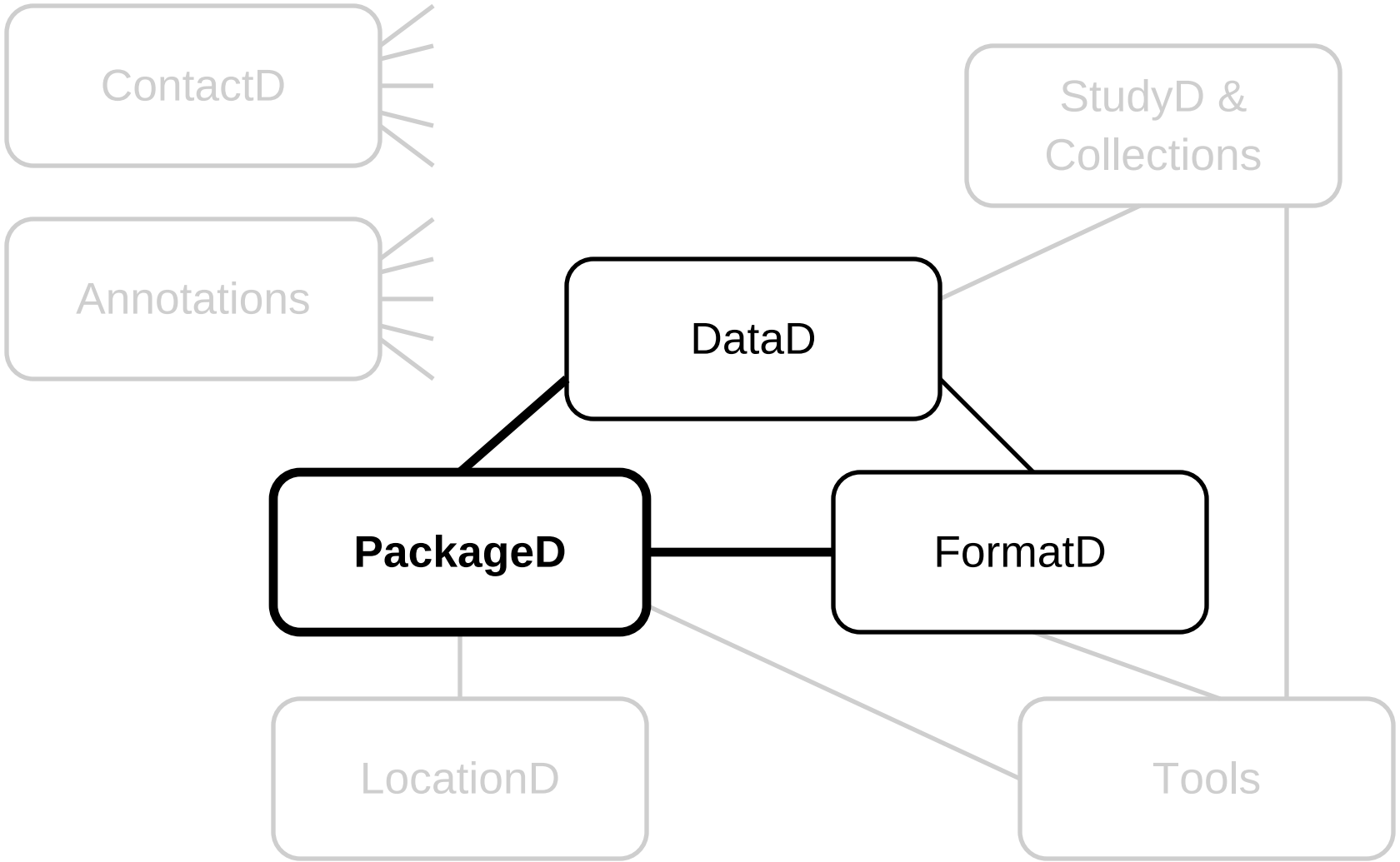
Package Descriptor (PackageD)

- How does a researcher actually fetch some data that they want?

- A package is a physical grouping of one or more data files. It can be thought of as a "downloadable unit".

- A package may have multiple data files in it.
- A particular data file may be in multiple packages.

Simplified System Diagram



Package Descriptor Fields

- Name
- Description
  - long, short, URL
- Keywords
- File Size
- Format ID
- md5 Hash

- Linkage to contained DataD/PackageD via a path.

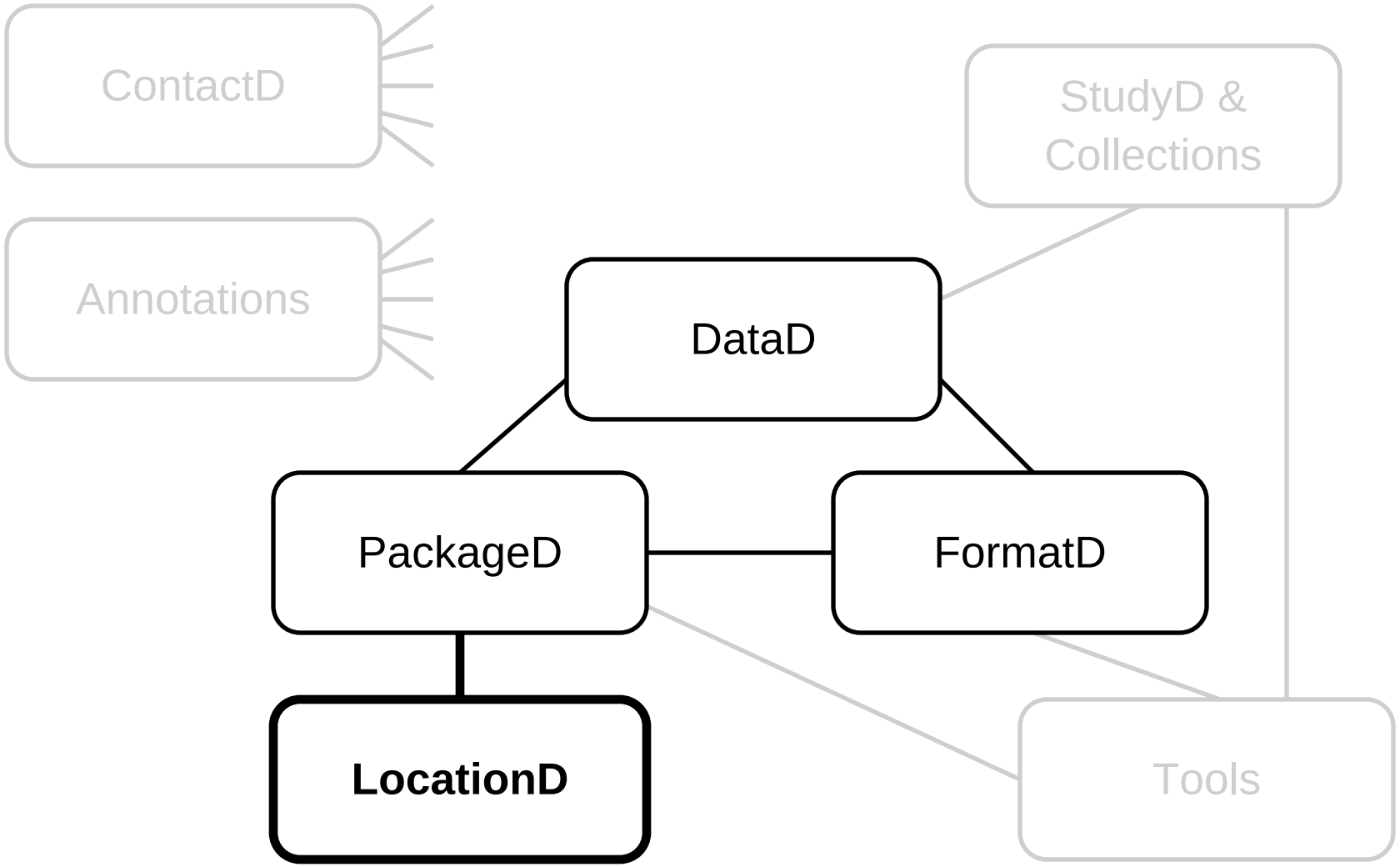
Location Descriptor (LocationD)

- How does a researcher actually fetch some data that they want? (part2)

- Packages may be available from multiple locations (mirror sites, etc).

- Not all packages will be directly available for download, there may be a procedure to follow or AUP describing terms of data use.

Simplified System Diagram



Location Descriptor Fields

- Download URL
  - if available
- Download Procedure
  - AUP requirements, how to obtain if not publically available
- Geographic Location of Server
- Logistic Location of Server

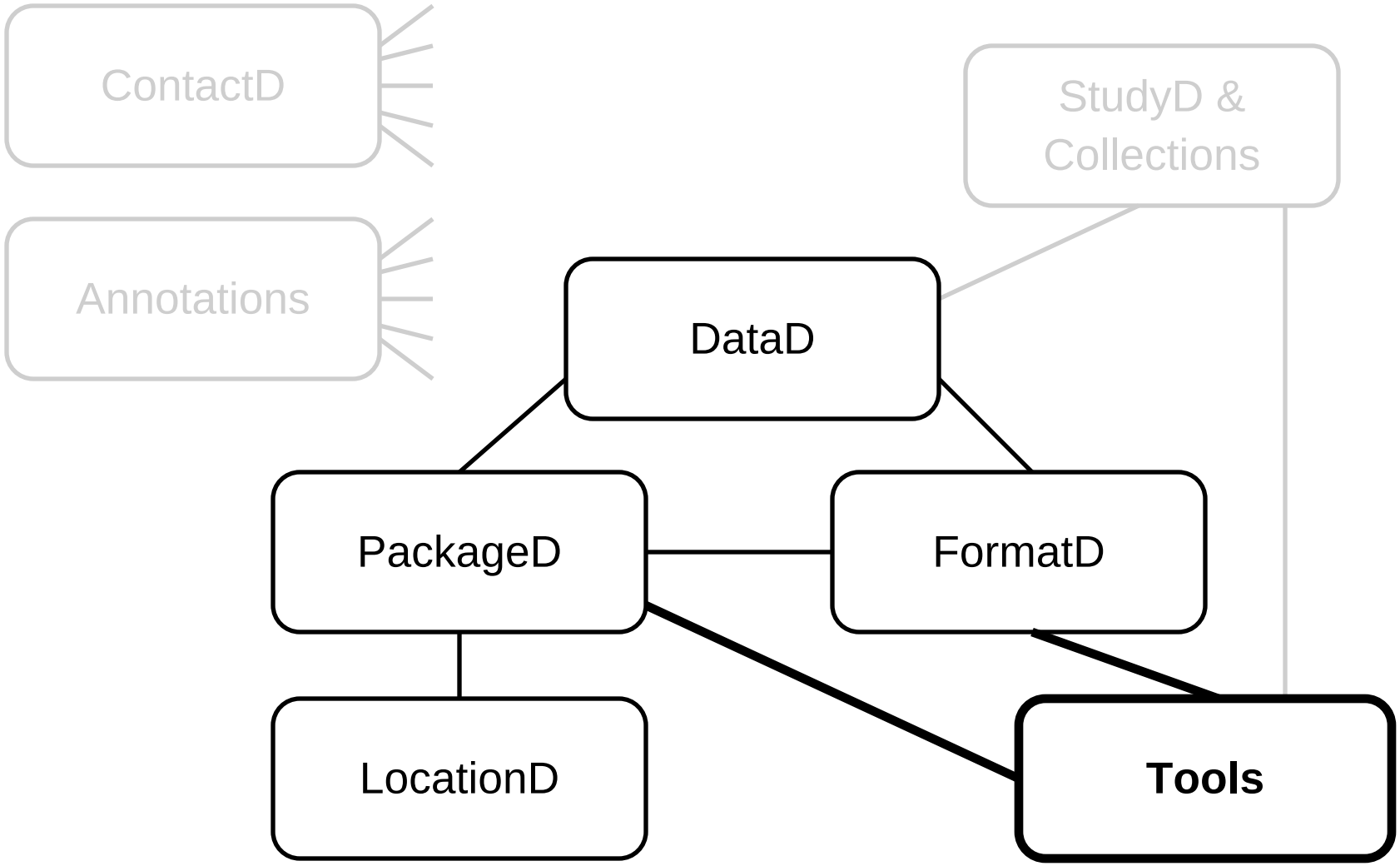
Tool/ToolSet Descriptors

- What tools are available to process some data?
- What tools were used to generate a data?

- ToolSet is an organized set of tools.
  - Ex: CoralReef, DAG Tools, OSU Flow Tools, etc.

- Version info – want to allow separation of versions when it is meaningful, but not require too much overhead from users

Simplified System Diagram



Tool/ToolSet Descriptor Fields

- Name
- Description
  - long, short, URL
- Keywords
- Release Date
- Operating Systems
- …

- We have a couple ways we've thought of to handle tools. We're waiting to see how things are initially used before making final decisions on this.
- Important to not require too much overhead from contributors.

Excluded Tool/ToolSet Fields

- Notes – annotation
- Bugs – annotation
  - we expect this to be very important

Data "Creation Process"

- Lots of ways that data might be created, especially for derived data.
- We thought a lot about this, including various methods of describing derivation chains.
- However, this seemed complicated and might not be sufficiently broad to cover things we hadn't thought of.
- Also, often people don't remember or know all of what was done.
- So, currently text fields until we gain better understanding of what people might want to put there.

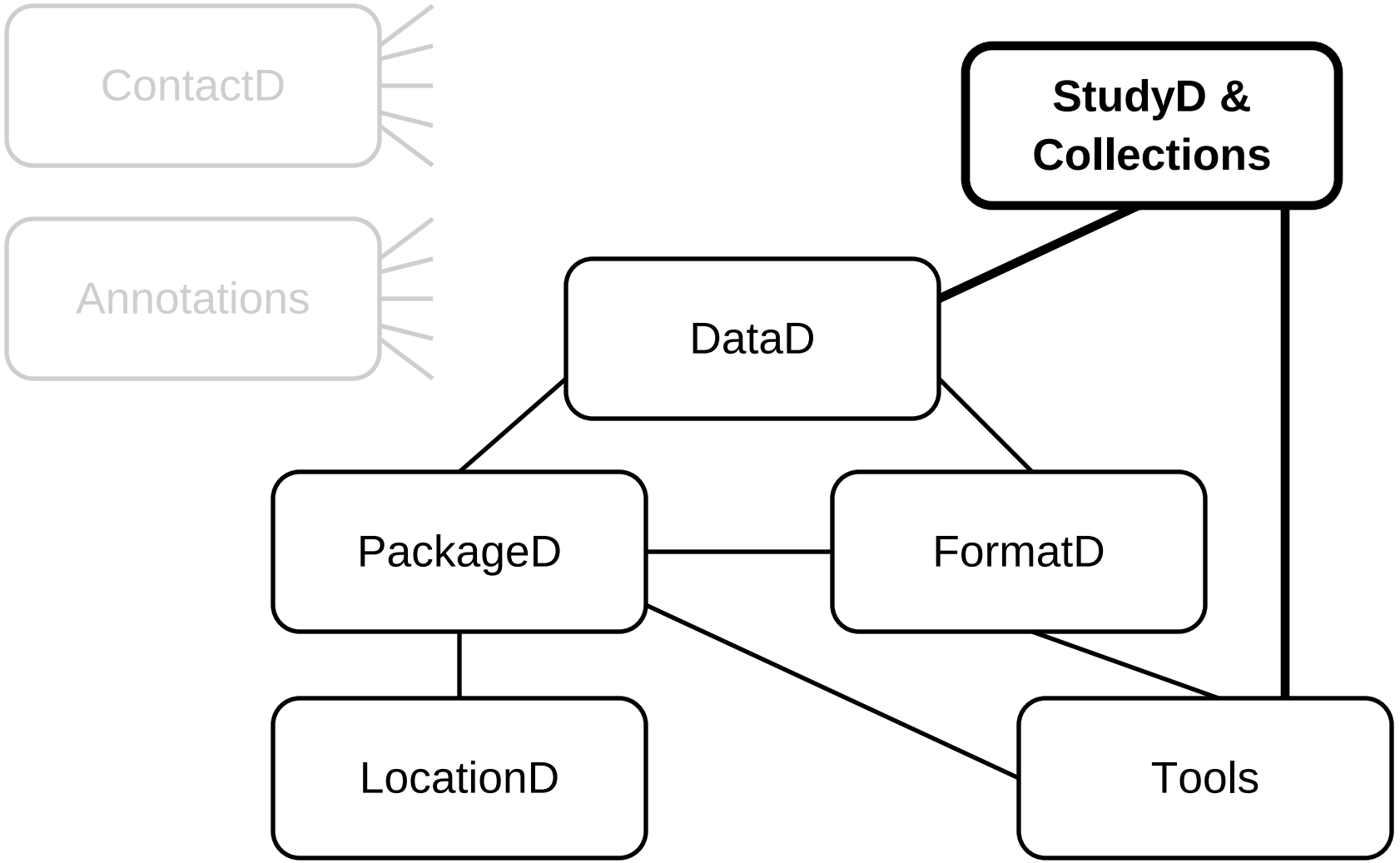
Study Descriptor (StudyD)

- What data were used in a paper/web writeup?
- What results are available about specific data?

- Non-goal: citeseer
  - we want to link studies to data, not track bibliography

- Example:
  - "I've invented the Mauve queue management system, but want to test it against the trace data from the Blue paper."

Simplified System Diagram



Study Descriptor Fields

- Name
- Description
  - long, short, URL
- Keywords

- Linkage to DataDs, ToolDs
- Linkage to StudyWriteup (i.e. text of paper)

Excluded Study Descriptor Fields

- Bibtex Entry – annotation
- Citeseer URL – annotation

- Reasons these are annotations, not fields:
  - Suspect there will be many similar things (e.g. DBLP)
  - Makes it similar to other organizational groupings of data, tools, packages, papers, etc (next slide)

Generalized Collections

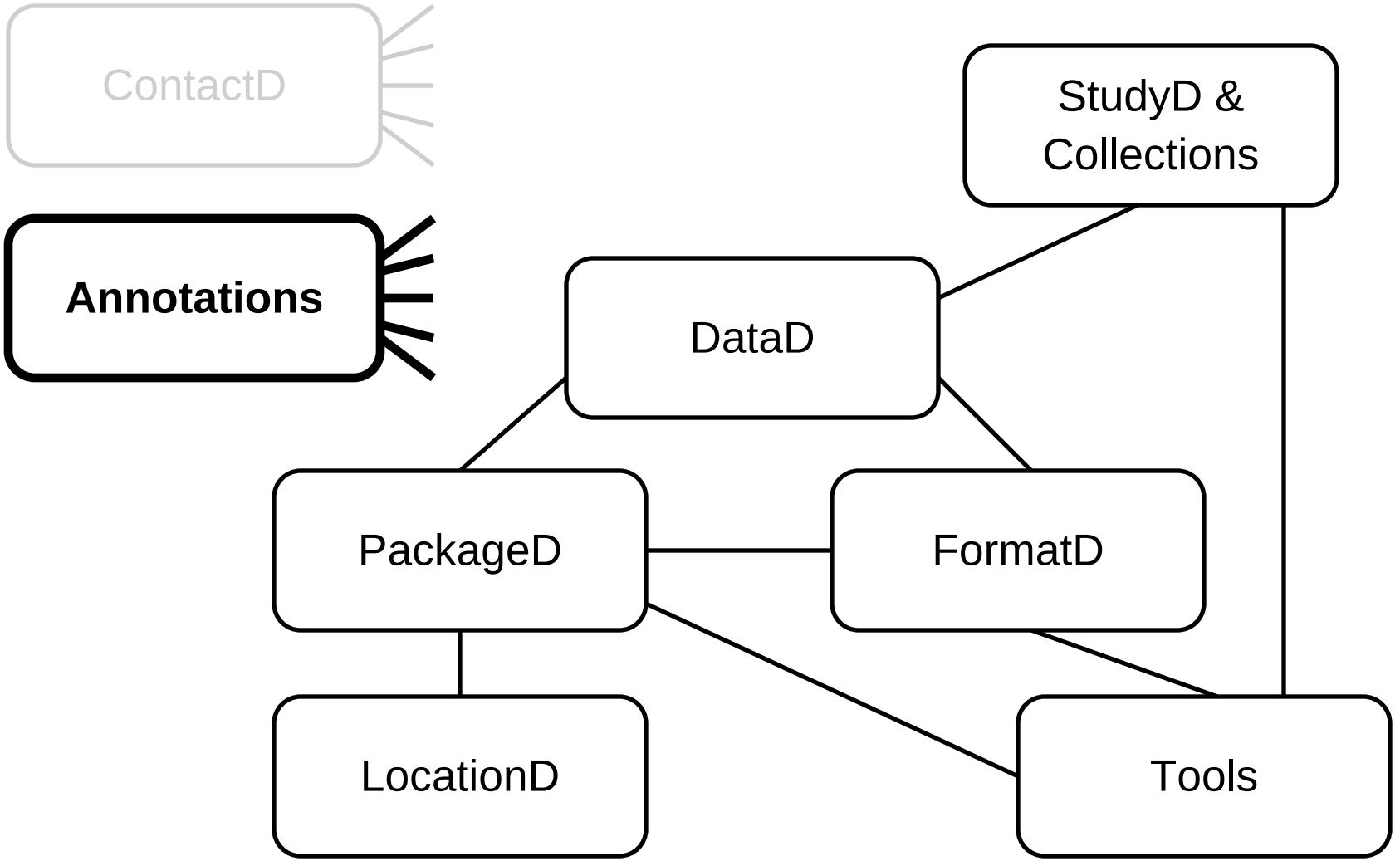
- Organizational groupings of data (etc) with a specific purpose.
- Similar to a StudyD, but there may not be a writeup available.
- These groupings may not exist physically.
- Examples:
  - "AS Topology Collection #27"
    - all skitter, surveyor traceroute data for Dec 1, 2003 (60 DataDs)
    - RouteViews and Ripe RIS Table Dumps (11 DataDs)
    - ASFinder tool (1 ToolD)
  - "Witty Worm"
    - various packet header and flow files from different locations
  - "Suggested IDS Test Suite #1"

Annotations

- What other information is there about this data (or tool or study or package or …)?
- How do I let other people know something important I learned about this data (or …)?

- Examples:
  - "At 11:30am, primary router failed"
  - "Between time T1 and T2 there was a DoS attack on host X"
  - "pcap snaplen is 72 bytes for this file"

Simplified System Diagram



Annotations Dictionary

- Key Name
  - hierarchical namespace (e.g. "FORMAT-pcap-snaplen")
- Description
  - long, short, URL
- Value Type
  - String, Number
- "Position" Type
  - Time Range, "ALL", String

- Standardized namespaces and per-user namespaces. The per-user namespaces are globally visible, but allow users to invent their own categories.
- Specific annotations in user namespaces can be promoted when widely accepted by community. We don't know what will be important.

Annotation Fields

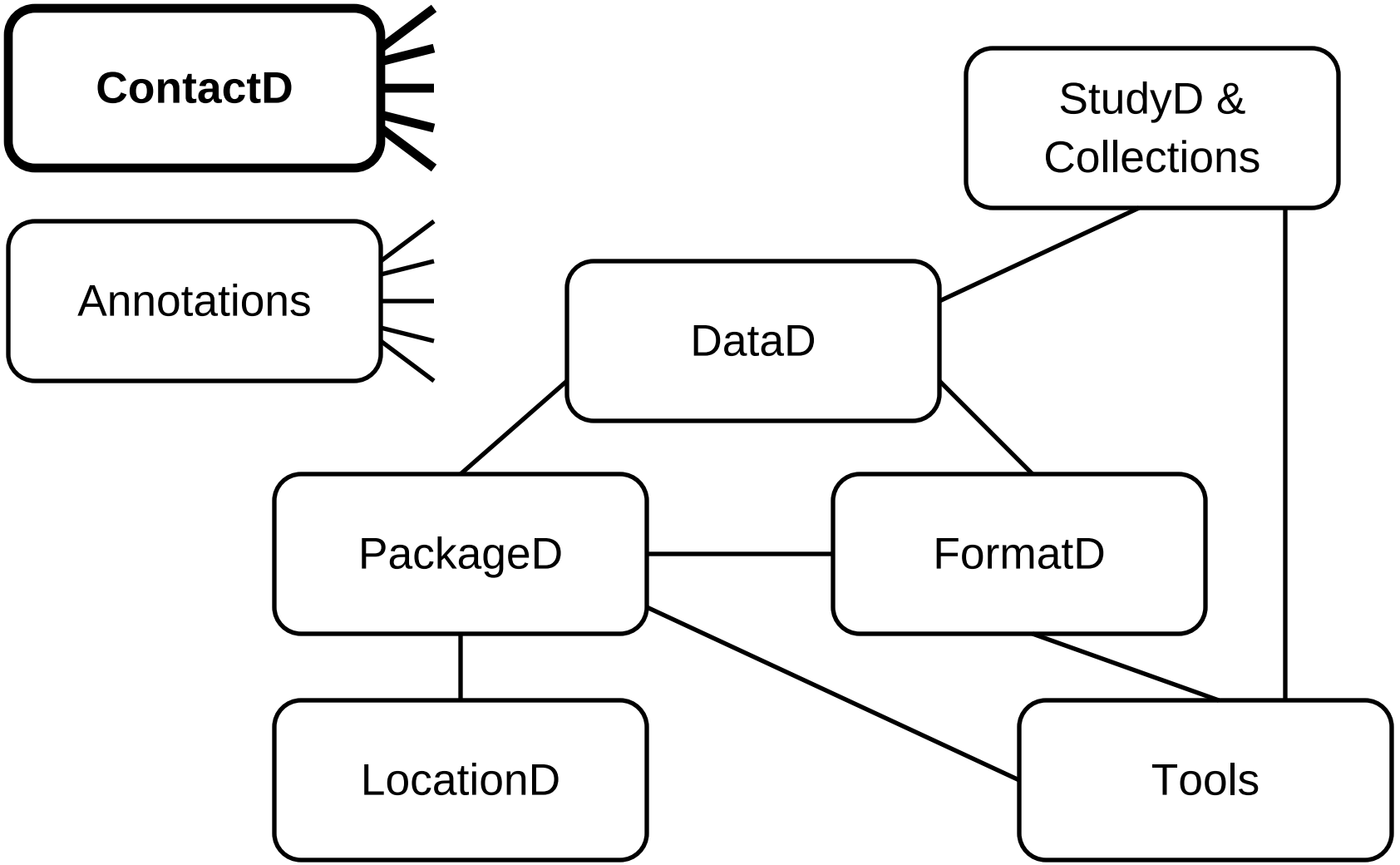
- Dictionary Key
- "Object" of annotation
  - specific DataD, PackageD, LocationD, …
- Value
- Position

Contact Descriptor (ContactD)

- Who submitted a piece of information to the database?

- Who was the original creator of some data, a tool, a package, etc?

Simplified System Diagram



Contact Descriptor Fields

- Login
- Password
- Name
- Description
  - long, short, URL
- email (hideable)
- Phone (hideable)
- Address (hideable)
- Country (hideable)
- Organization
- Research Interests

Current Status