A lot of different sources of data in Debian

With different data formats:
- text files
- BerkeleyDB
- SQL databases
- JSON
- YAML

Need to combine them all:
Mainly for Quality Assurance, e.g:
- Packages of priority ≥ standard with RC bugs?
- Maintainers with lots of outdated/buggy packages?
Ultimate Debian Database

Idea:

- **Import all the data in a single (Postgre)SQL DB**
- Easier to query (relatively well-known interface)
- *The* proper way of joining data together
- No need to write problem-specific scripts
History

Started as a Google Summer of Code project in 2008

Student: Christian von Essen (Neronus)

Mentors:
- Lucas Nussbaum (lucas)
- Marc Brockschmidt (HE)
- Stefano Zacchiolli (zack)
Design choices

Not problem-specific, no typical queries (not projectb!)

Schema:
- Typical user == human
- Make it easy to write/run queries
- Performance? important, but not a critical goal
- No surrogate keys
Surrogate key

- Unique identifier (usually integer)
- Used as primary key
- Not derived from any application data packages (**package_id**, package_name, ...)

MySQL: AUTO_INCREMENT
PostgreSQL: serial

Has both advantages and disadvantages

Details: http://en.wikipedia.org/wiki/Surrogate_key
Design choices (2)

Data:
- Correctness is critical
- Partial updates? Often difficult/risky
- Solution: complete data reloads (for most importers)
  ♦ Using transactions to avoid temporary unavailability
Design choices (3)

Debian is inconsistent
  ▶ What does "package" mean?

Inconsistency can be interesting for QA
→ Keep inconsistency in UDD
→ No foreign keys between data sources
→ Some VIEWs that hide inconsistency
Current status

- Running as udd.debian.org
- Uses PostgreSQL 9.4
- You can connect from {qa,alioth}.d.o using: psql service=udd
- Even non-DDs can connect!

More info:
http://wiki.debian.org/UltimateDebianDatabase
What’s in it?

▶ Main stuff:
♦ Sources and Packages
♦ Bugs (including archived bugs)
▶ Identities:
♦ Carnivore
♦ Debian LDAP (restricted to DDs)
♦ Debian Maintainers
♦ PTS subscriptions (OLD)
▶ Derivatives:
♦ Sources/Packages for Ubuntu, Aptosid
♦ Ubuntu bugs
♦ Ubuntu upload history
♦ Ubuntu popcon
♦ Ubuntu lintian
▶ Additional stuff:
♦ Migrations to testing
♦ History of uploads
♦ FTPmasters NEW, removals, autorejects, deferred queue
♦ Debtags
♦ DDTP (translation status)
♦ Popularity contest
♦ Orphaned packages
♦ Upstream status (uscan)
♦ Screenshots
♦ Lintian
♦ Reproducible builds
♦ security tracker
♦ VCSwatch
♦ Mentors
♦ CI
♦ Duck
♦ Key packages
♦ Release team hints
♦ wanna-build
♦ testing auto-removals
So, what can we find out about Debian using UDD?
Number of different lintian errors or warnings

```
select package, count(distinct tag) as cnt
from lintian
where tag_type in('error','warning')
group by package
order by cnt desc limit 15;
```
Number of different lintian errors or warnings

```sql
select package, count(distinct tag) as cnt
from lintian
where tag_type in('error','warning')
group by package
order by cnt desc limit 15;
```

<table>
<thead>
<tr>
<th>package</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>muddleftpd</td>
<td>19</td>
</tr>
<tr>
<td>cyclades-serial-client</td>
<td>18</td>
</tr>
<tr>
<td>lpr</td>
<td>18</td>
</tr>
<tr>
<td>spellcast</td>
<td>17</td>
</tr>
<tr>
<td>bonnie++</td>
<td>16</td>
</tr>
<tr>
<td>vflib3</td>
<td>16</td>
</tr>
<tr>
<td>scalapack-doc</td>
<td>15</td>
</tr>
<tr>
<td>jbofihe</td>
<td>15</td>
</tr>
<tr>
<td>netmaze</td>
<td>15</td>
</tr>
<tr>
<td>bookview</td>
<td>15</td>
</tr>
<tr>
<td>jailer</td>
<td>15</td>
</tr>
<tr>
<td>debbugs</td>
<td>15</td>
</tr>
</tbody>
</table>
Who uploaded sid’s packages?

```sql
select changed_by, count(*) from sources s, upload_history uh
where s.source = uh.source and s.version = uh.version
and s.distribution='debian' and s.release = 'sid'
group by changed_by order by count desc limit 8;
```
Who uploaded sid’s packages?

```
select changed_by, count(*) from sources s, upload_history uh
where s.source = uh.source and s.version = uh.version
and s.distribution='debian' and s.release = 'sid'
group by changed_by order by count desc limit 8;
```

<table>
<thead>
<tr>
<th>changed_by</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>gregor herrmann <a href="mailto:gregoa@debian.org">gregoa@debian.org</a></td>
<td>1324</td>
</tr>
<tr>
<td>Clint Adams <a href="mailto:clint@debian.org">clint@debian.org</a></td>
<td>527</td>
</tr>
<tr>
<td>Andreas Tille <a href="mailto:tille@debian.org">tille@debian.org</a></td>
<td>512</td>
</tr>
<tr>
<td>Emmanuel Bourg <a href="mailto:ebourg@apache.org">ebourg@apache.org</a></td>
<td>462</td>
</tr>
<tr>
<td>Maximiliano Curia <a href="mailto:maxy@debian.org">maxy@debian.org</a></td>
<td>383</td>
</tr>
<tr>
<td>Thomas Goirand <a href="mailto:zigo@debian.org">zigo@debian.org</a></td>
<td>382</td>
</tr>
<tr>
<td>Jonas Smedegaard <a href="mailto:dr@jones.dk">dr@jones.dk</a></td>
<td>374</td>
</tr>
<tr>
<td>Joachim Breitner <a href="mailto:nomeata@debian.org">nomeata@debian.org</a></td>
<td>296</td>
</tr>
</tbody>
</table>
UDD Dashboards

- UDD bugs search
  https://udd.debian.org/bugs/ (example)

- UDD Maintainer Dashboard
  https://udd.debian.org/dmd/ (example)

- UDD Bapase (search for "interesting" packages)
Things to improve

- Mostly a one-person project
  - Good starting point: debug importers failures
    https://udd.debian.org/udd-status.cgi

- No real good development environment
  - Vagrant environment now available

- More collaboration with DDPO and Tracker
  - Hard to share data inside Debian infrastructure

Contact: #debian-qa or debian-qa@l.d.o

http://wiki.debian.org/UltimateDebianDatabase