Debian Data Export

A standard for publishing Debian information.

Feb 7, 2009
15 slides
Enrico Zini (enrico@debian.org)

Enrico Zini enrico@debian.org

Debian: the data hell

- Package files, for binary packages
  - Format: rfc822-like
  - Split per distribution, then per architecture
- Package files, for source packages
  - Same as before
- Maintainer <-> Source package mapping
  - Available at DDPO, deprecated
  - Available at BTS, somehow
  - Extractable from source packages
  - Available at projectb
- Debtags information
  - From /var/lib/debtags or from Alioth

Enrico Zini enrico@debian.org

Debian: the data hell

- Debtags vocabulary
- Extra debtags sources
- Popcon rankings
- Bug information
- Changelogs
- .desktop files of packages not installed
- New queue
- Screenshots
- Apt-file information
- http://ftp-master.debian.org/~joerg/pkg-nums
- http://ftp-master.debian.org/~joerg/arch-space
- License information

Enrico Zini enrico@debian.org

Debian: the data hell

- Localisation information
- uscan status
- Buildd logs
- BTS data
- sloccount run results
- Debian Weather
- Debian Pure Blend specific information
- UDD!

Can you think of more?

Debian: the data formats hell

- rfc822-like files
  - Description subformat
  - Tag subformat
  - Dependency subformat
- SOAP interfaces
- LDAP interfaces
- SQL interfaces
- Lots of ad-hoc formats
- HTML scraping
- post-processing occasionally needed

Can you think of more?

Enrico Zini enrico@debian.org

Debian: the data access hell

- Something in mirrors
- Something on people.debian.org
- Something on specific Debian machines
- Something other machines elsewhere
- Something can only be accessed FROM specific machines
- Something can only be computed on the user's system

Enrico Zini enrico@debian.org

My general goals

- Producing data should be easy. The major task should be computing it: all the rest should be a no brainer
- Finding data should be easy
- Getting data should be easy
  - In terms of protocol to download it
  - In terms of format to parse it

More wishes?

Enrico Zini enrico@debian.org

My specific goals

- debtags.debian.net
  - Must (ideally) index information for all packages in Debian, Ubuntu, Pure Blends, other derivatives, all distros, all arches.
  - Only one version per package. If a package is in more than one distribution, I want to use the data in *testing*.
- Autocompletion in web form fields
  - Of binary package names
  - Of source package names
  - Of maintainer names
  - Of ...
- Machine readable interface to all the data that I produce

The solution (so far)

A demo should happen now.

If you are reading the slides after the presentation, you may find the video in the Debian video archives:

`http://meetings-archive.debian.net/pub/debian-meetings/2009/`

Enrico Zini enrico@debian.org

Scope of the DDE data space

- Export views corresponding to common use cases
  - Do not reimplement SQL, or LDAP
  - For special needs, people can craft a SQL or LDAP query.
  - If the need becomes more general, we turn the query into a DDE plugin

Enrico Zini enrico@debian.org

Uses of DDE

Current

- debtags.debian.net, screenshots.debian.net
- Completion in web forms
- Some example mashups
- apt-file without local database

Future

- Extra features in package managers
- apt fetcher for extra data
- More external sites to feed (Blends?)
- Switch existing tools to use blend-specific data sources

More ideas?

Enrico Zini enrico@debian.org

Deployment

In theory

It is a WSGI application, it is trivial to deploy it in any way you like

In practice

- CGI does not scale
- Cherrypy 2.x does not allow WSGI apps to stream
- Cherrypy 3.x has a bad chain of conflicts
- Paste won't reload without killing running streams
- Fastcgi needs careful tuning, or you are killed if you run for long
- mod_wsgi runs in apache's process space

Help?

Scalability

In theory

All data is read only, there is no state: it's a cache wonderland!

- Put varnish in front of it
- Aggressive cache headers
- Can be replicated, can use DNS round robins

In practice

- Demand may reach insane levels
  - All web forms making lots of small queries!
  - All package managers!
- It's not static data, cannot use the mirror network

Ideas?

New possibilities opening

- Javascript mashups
  - We need to wait for FireFox 3.1 to fix multi-source XMLHTTPRequest
  - Currently DDE supports JSONP
  - JSONPP is easy to add, but this path makes me sick

Enrico Zini enrico@debian.org

Who will make it happen?

Not me alone.

I will chase a couple of itches of mine, but I won't reach DDE's full potential just on myself.

If DDE can scratch an itch of yours, I'll be happy to show you the ropes.

Enrico Zini enrico@debian.org