F/OSS Research Repositories & Research Infrastructures
Experiences with the Notre Dame OSS Archive and the VectorBase BRC

Greg Madey & Scott Christley
Computer Science & Engineering
University of Notre Dame

FOSSRRI 2008
University of California, Irvine
February, 2008
Overview

• SourceForge F/OSS Research Data Archive at Notre Dame
  • Data
  • Archive design
  • Limitations
• Lessons to be learned from the bioinformatics community
Background

- SourceForge Research Archive @ Notre Dame
- Evolved out of an NSF/DST funded project
- Current planning grant: NSF CRI grant
- VectorBase - an NIH/NIAIDS Bioinformatics Resource Center (BRC) @ Notre Dame
  - Resource center on insect vectors that transmit diseases
  - Genomic data, metadata, community and tools
F/OSS Research Data

- SourceForge.net
  - A large F/OSS development community
    - 168,000+ registered projects
    - 1,786,000+ registered users
  - Project data
    - Downloads, bug reports, forum activity, developers, project characteristics, etc.
  - Developer data
    - Activity data
    - Project membership
SourceForge Research Data Description

• Data warehouse
  – 600G total and growing at 12G/month.
  – Every dump has 80-120 tables.
  – Tables have up to 30 million records.
  – CVS & SVN metadata to be integrated

http://zerlot.cse.nd.edu/
Design

- Presentation Tier
  - Browser interface
  - Wiki-based portal
- Logic Tier
  - Authentication
  - Schema browser
  - Queries & download
- Data Tier
  - PostgreSQL
  - Monthly schema
Data Tier

- PostgreSQL
- Database - “Timeline”
- Monthly schema: one for each dump
- Mirrors the SourceForge.net backend
- Connection pool
- Persistent connections for improved performance
- CVS & SVN schema in development
Presentation Tier

- Various access methods
- Documentation and references
- Community support - FAQ, schema browser, table definitions
- Wiki interface
Schema Browser

Sf0407

Schema sf0407 has 83 tables.

- pg_stat_database_historical
- trove_egg
- user_perms
- groups_registration
- pg_autovac_skip
- ref_timezones
- stats_outage_log
- stats_rank_oldformula_byday
- stats_fileid_alltime_agg
- stats_group_rank_alltime
- stats_group_rank_byday
- stats_groupid_alltime_agg
- stats_toplist_week
- trove_cat_activity
- trove_ref_translation_to_iso8859
- trove_egg_minix
- artifact
- artifact_canned_responses
- artifact_category
- artifact_counts_egg
- artifact_file

Schemas

- January 2003 - sf0103
- November 2004 - sf1104
- December 2004 - sf1204
- February 2005 - sf0205
- March 2005 - sf0305
- April 2005 - sf0405
- May 2005 - sf0505
- June 2005 - sf0605
- July 2005 - sf0705
- August 2005 - sf0805
- September 2005 - sf0905
- October 2005 - sf1005
- November 2005 - sf1105
- December 2005 - sf1205
- January 2006 - sf0106
- February 2006 - sf0206
- March 2006 - sf0306
Stats project all

Appears in the following schemas:

- sfo103
- sfo104
- sfo1204
- sfo205
- sfo305
- sfo405
- sfo505
- sfo605
- sfo705
- sfo805
- sfo905
- sfo1005
- sfo1105
- sfo1205
- sfo1016
- sfo208
- sfo306
- sfo406
- sfo506
- sfo606
- sfo706
- sfo806
- sfo906
- sfo1006
- sfo1106
- sfo1206
- sfo1017
- sfo207
- sfo307

Most Recent Description

<table>
<thead>
<tr>
<th>Column</th>
<th>Type</th>
<th>Modifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>group_id</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>developers</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>group_ranking</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>group_metric</td>
<td>double precision</td>
<td></td>
</tr>
<tr>
<td>logo_showings</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>downloads</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>site_views</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>subdomain_views</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>page_views</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>msg_posted</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>msg_uniq_auth</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>bugs_opened</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>bugs_closed</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>support_opened</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>support_closed</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>patches_opened</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>patches_closed</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>artifacts_opened</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>artifacts_closed</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>tasks_opened</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>tasks_closed</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>help_requests</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>cvs_checkouts</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>cvs_commits</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>cvs_acids</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>svn_checkouts</td>
<td>integer</td>
<td></td>
</tr>
</tbody>
</table>
Researcher Must Know SQL

SourceForge.net Research Archive Query Form

Examples

```
SELECT *
FROM sf0305.users
WHERE user_id < 100

SELECT user_name
FROM sf1104.users a,
    sf1104.artifact b
WHERE a.user_id =
b.submitted_by AND
b.artifact_id = 304727
```

SELECT:

FROM:

WHERE:

Separator
- ;
- #
- ,
- XML

Add SQL query to result file?
- yes
- no

Submit Query  Clear

News
- The database version has been upgraded. If you notice any errors, please let us know (oss at nd dot edu)
- SQL query option added (as an attribute of the root element in XML output, as the first line in text file output).
- April schema (sf0407) now loaded.
Logic Tier

• Interactive web query system
  – Authorized user can submit query to the backend repository through a web query
  – Results are provided as text with various formats, delimiters, and XML

• Dynamic web schema browser
  – Users can browse the schema using a dynamic schema browser
Utilization - Sample

• Monthly activity (June 2006)
  – Total queries submitted: 16,947
  – Total data files retrieved: 13,343
  – Total bytes of query data downloaded: 26,684,556,278

• Monthly activity (Feb 2007)
  – Total queries submitted: 38,659
  – Total data files retrieved: 24,422
  – Total bytes of query data downloaded: 13,048,335,165

• Typical number of monthly users: 5 – 10
• Total number of users with significant activity: ~75

Queries and Downloads by Month (2006-2007)

- Number of Queries
- Gigabytes Downloaded

Month:
- Jan 2006
- Feb 2006
- Mar 2006
- Apr 2006
- May 2006
- Jun 2006
- Jul 2006
- Aug 2006
- Sep 2006
- Oct 2006
- Nov 2006
- Dec 2006
- Jan 2007
- Feb 2007
- Mar 2007
- Apr 2007
- May 2007
- Jun 2007
- Jul 2007
- Aug 2007
- Sep 2007
- Oct 2007
- Nov 2007
- Dec 2007
Limitations

- Data is production oriented - not primarily collected with research as a goal
- Data is project oriented
- Limited to SourceForge.net, e.g., how to link data to other repositories (Conklin, 2007)
- No active community around the data archive
- Limited services, other than the raw data, some metadata, a FAQ, schema browser, and SQL query tool

Lessons to be learned from Bioinformatics?
Bioinformatics

- Starts with lot’s of raw data
- Added value
  - Assembly (organization of data)
  - Annotation
  - Data browsers
  - Search tools
  - Computational tools
  - Programmatic interfaces (DAS, Web Services)
- Educational components
- Cybercommunities
Annotation

- Genome features: genes, exons, introns, gene products, promoter regions, target sites, transposable elements, repeats
  - Structural elements
- Methods: automatic, expert curators, community
  - Sequencing centers - automatic gene prediction
  - Flybase.org - expert curators (expense!)
  - Vectorbase.org - community annotators
- Ontology/Controlled Vocabulary is important
- F/OSS objects of study:
  - Artifacts, processes, projects, users, community, knowledge
  - Simple data type (programming language, operating system)
  - Knowledge constructs (architecture, norms, roles)
Annotation

- Data integration
  - Links between different annotation types
  - Provenance, governance

- Example: F/OSS Project
  - Artifacts developed/maintained
  - Processes utilized
  - Communication norms
  - Incentive structures
Data Browser

- Interactive view of data
  - DNA sequence has directionality
  - Annotations integrated as tracks side-by-side with DNA sequence (DAS)

- F/OSS: what is the data?
  - Raw data
  - Knowledge constructs (annotations)
    - Social Network? Connections between users, projects, tasks, actions, involvement in processes

- How to visualize these knowledge constructs?
  - Information scientists, HCI experts
Search

• What is being searched/indexed?
  • Raw data, meta data
  • Keyword searches
  • Field-specific searches

• Entrez (NCBI Search)
  • Federated search across broad range of databases
  • Meta data search

• Pubmed: literature search
  • Raw and meta data search

• Use of ontology/controlled vocabulary
Computer Tools

- **BLAST**
  - Specialized search process for specific types of raw data (sequences).
- **SQL queries**
- **CLUSTALW**
  - Comparison (alignment) of sequence data
- **F/OSS Project Alignment?**
  - Given set of projects, “align” them according to artifacts, processes, knowledge constructs, etc. (annotations).
- **HMMER**
  - Probabilistic search of sequence data
- **F/OSS search?**
  - Learn/Define a prototype project, search for projects that probabilistically match prototype.
Summary

• F/OSS annotations are more complex
  • Biology annotations primarily structural, functional annotation is descriptive not mechanistic.
  • Biology community just starting to think about process representation (pathways, networks), temporal, spatial.
  • Much is hidden in literature (descriptive) or in modeling/simulation.
  • F/OSS has mechanistic data (user actions, message posts).

• Biology use of ontologies is more for controlled vocabulary, less for knowledge representation.
  • F/OSS researchers could have same problem; disparate groups agree on common terms.
  • Beware: annotation standard ---> path dependence

• Complete the cycle!
  • Data --> new knowledge (article) --> new annotation (new data!)
  • Allow more complex and integrated studies
Thank You!

GNU
MySQL
OpenOffice.org
FreeBSD
PHP
Eclipse
KDE
Perl
Python
Mozilla
Firefox
Ubuntu
Apache
Cathedral & Bazaar
Creative Commons
Sendmail
The Apache Software Foundation