

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

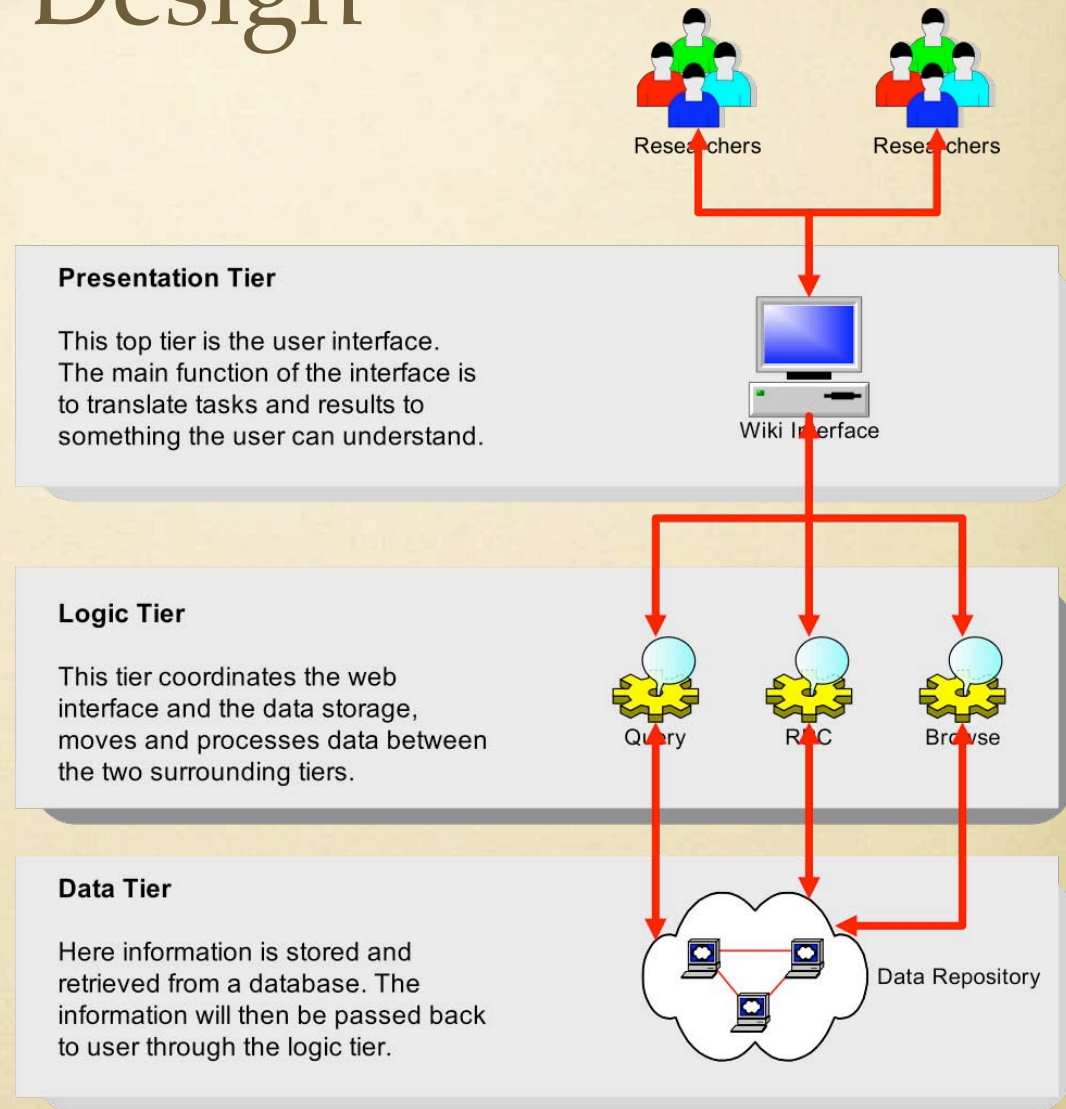
# SourceForge Research Data Description

- Data warehouse
  - 38 monthly dumps between January 2003 - January 2008.
  - 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/>

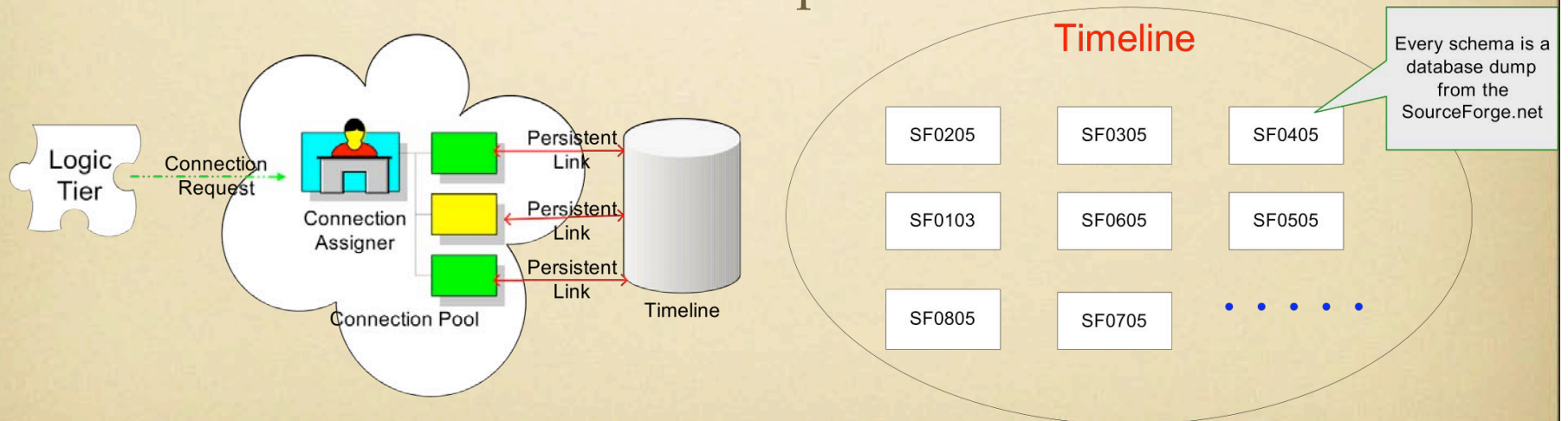
# Design

- 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

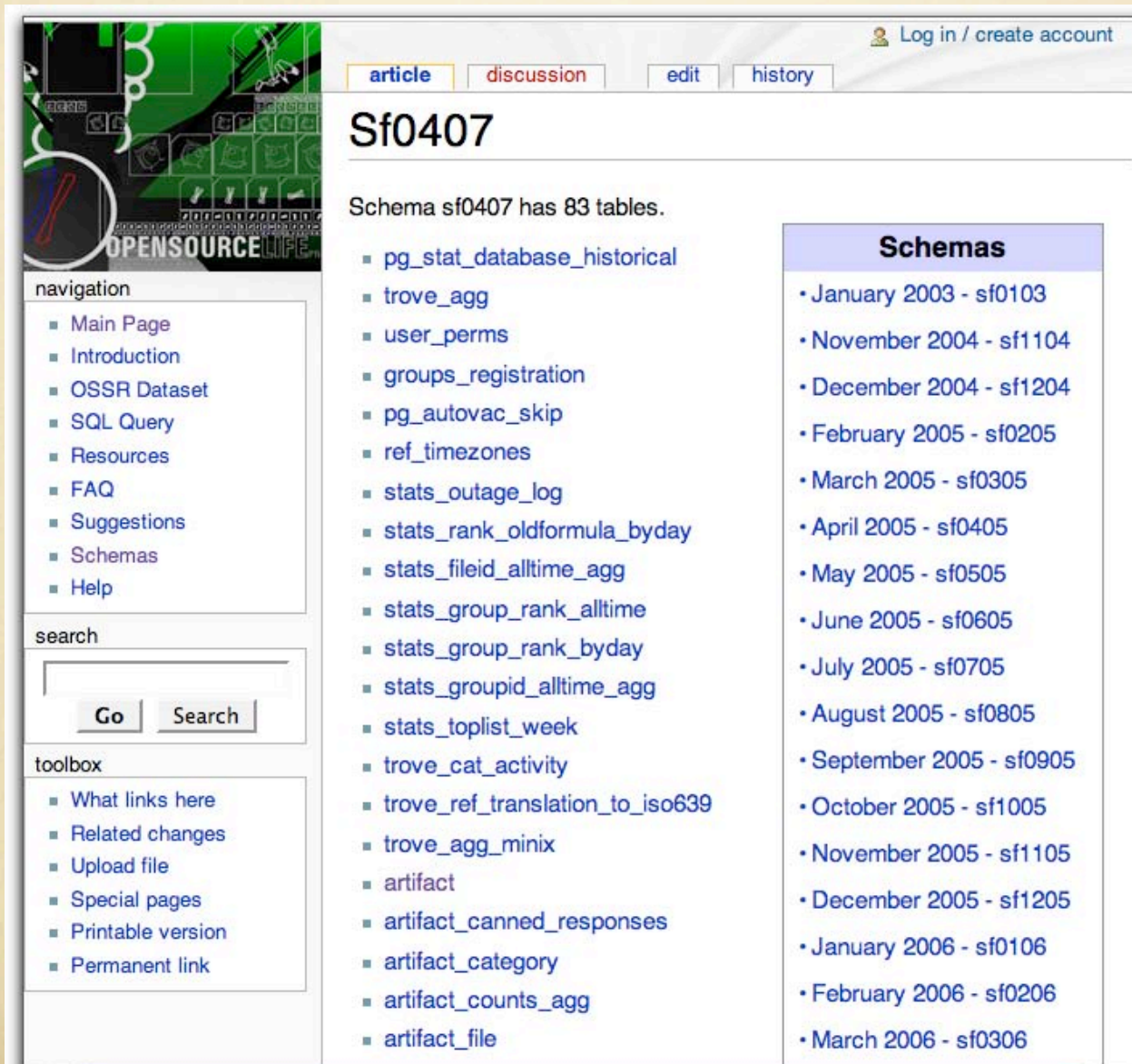
- Various access methods
- Documentation and references
- Community support - FAQ, schema browser, table definitions
- Wiki interface

The screenshot displays the SourceForge Research Data Archive website. At the top right, there is a "Log in / create account" link. Below this, navigation tabs for "article", "discussion", "view source", and "history" are visible. The main heading is "Main Page". The central content area is titled "SourceForge Research Data Archive: A Repository of FLOSS Research Data". It contains several bullet points: the first describes the wiki's purpose and access; the second mentions the project's funding by NSF and provides a link to the introduction; the third states the repository is hosted by the Department of Computer Science & Engineering at the University of Notre Dame; the fourth notes the material is based on work supported by the National Science Foundation; and the fifth provides contact information for questions. To the right of the main content is a "Top Links" sidebar with links to Schema Browser, Query Form, Research Data, Making Queries, Resources, Papers, Contact, FAQ, Schemas, and All tables. On the left side of the page, there is a "navigation" menu with links to Main Page, Introduction, OSSR Dataset, SQL Query, Resources, FAQ, Suggestions, Schemas, and Help. Below this is a "search" box with "Go" and "Search" buttons. At the bottom left is a "toolbox" with links for What links here, Related changes, Upload file, Special pages, Printable version, and Permanent link.

## Recent News

- April schema ([sf0407](#)) has been loaded.
- Major updates to the [Schema](#) pages. Here is a list of [all tables](#) that appear in at least one schema.
- Table pages now contain primitive framework for adding information about the data. Please make changes to these sections if you have surmised relevant information about a given table. Please do

# Schema Browser



The screenshot shows a web interface for a Schema Browser. At the top right, there is a user profile icon and the text "Log in / create account". Below this, there are four tabs: "article" (selected), "discussion", "edit", and "history". The main heading is "Sf0407". Below the heading, it states "Schema sf0407 has 83 tables." followed by a list of table names. On the right side, there is a box titled "Schemas" containing a list of month-year pairs with corresponding schema IDs. On the left side, there are three sections: "navigation" with a list of links, "search" with a search box and "Go" and "Search" buttons, and "toolbox" with a list of utility links.

Log in / create account

article discussion edit history

## Sf0407

Schema sf0407 has 83 tables.

- pg\_stat\_database\_historical
- trove\_agg
- 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\_iso639
- trove\_agg\_minix
- artifact
- artifact\_canned\_responses
- artifact\_category
- artifact\_counts\_agg
- 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

navigation

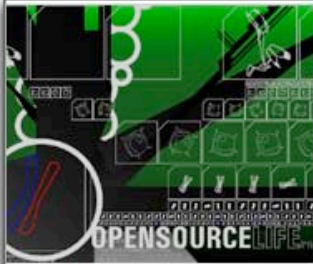
- Main Page
- Introduction
- OSSR Dataset
- SQL Query
- Resources
- FAQ
- Suggestions
- Schemas
- Help

search

Go Search

toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link



[Log in / create account](#)

[article](#) [discussion](#) [edit](#) [history](#)

# Stats project all

Appears in the following schemas:

- [sf0103](#)
- [sf1104](#)
- [sf1204](#)
- [sf0205](#)
- [sf0305](#)
- [sf0405](#)
- [sf0505](#)
- [sf0605](#)
- [sf0705](#)
- [sf0805](#)
- [sf0905](#)
- [sf1005](#)
- [sf1105](#)
- [sf1205](#)
- [sf0106](#)
- [sf0206](#)
- [sf0306](#)
- [sf0406](#)
- [sf0506](#)
- [sf0606](#)
- [sf0706](#)
- [sf0806](#)
- [sf0906](#)
- [sf1006](#)
- [sf1106](#)
- [sf1206](#)
- [sf0107](#)
- [sf0207](#)
- [sf0307](#)

## navigation

- [Main Page](#)
- [Introduction](#)
- [OSSR Dataset](#)
- [SQL Query](#)
- [Resources](#)
- [FAQ](#)
- [Suggestions](#)
- [Schemas](#)
- [Help](#)

## search

## toolbox

- [What links here](#)
- [Related changes](#)
- [Upload file](#)
- [Special pages](#)
- [Printable version](#)
- [Permanent link](#)

## Most Recent Description

[\[edit\]](#)

Table "sf0407.stats\_project\_all"

Column	Type	Modifiers
group_id	integer	
developers	integer	
group_ranking	integer	
group_metric	double precision	
logo_showings	integer	
downloads	integer	
site_views	integer	
subdomain_views	integer	
page_views	integer	
msg_posted	integer	
msg_uniq_auth	integer	
bugs_opened	integer	
bugs_closed	integer	
support_opened	integer	
support_closed	integer	
patches_opened	integer	
patches_closed	integer	
artifacts_opened	integer	
artifacts_closed	integer	
tasks_opened	integer	
tasks_closed	integer	
help_requests	integer	
cvs_checkouts	integer	
cvs_commits	integer	
cvs_adds	integer	
svn_checkouts	integer	

# Researcher Must Know SQL

[WIKI](#) | [QUERY](#) | [SCHEMAS](#) | [FAQ](#) | [RESULTS](#)

SOURCEFORGE.net

## 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

### 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.

Submit Query

Clear

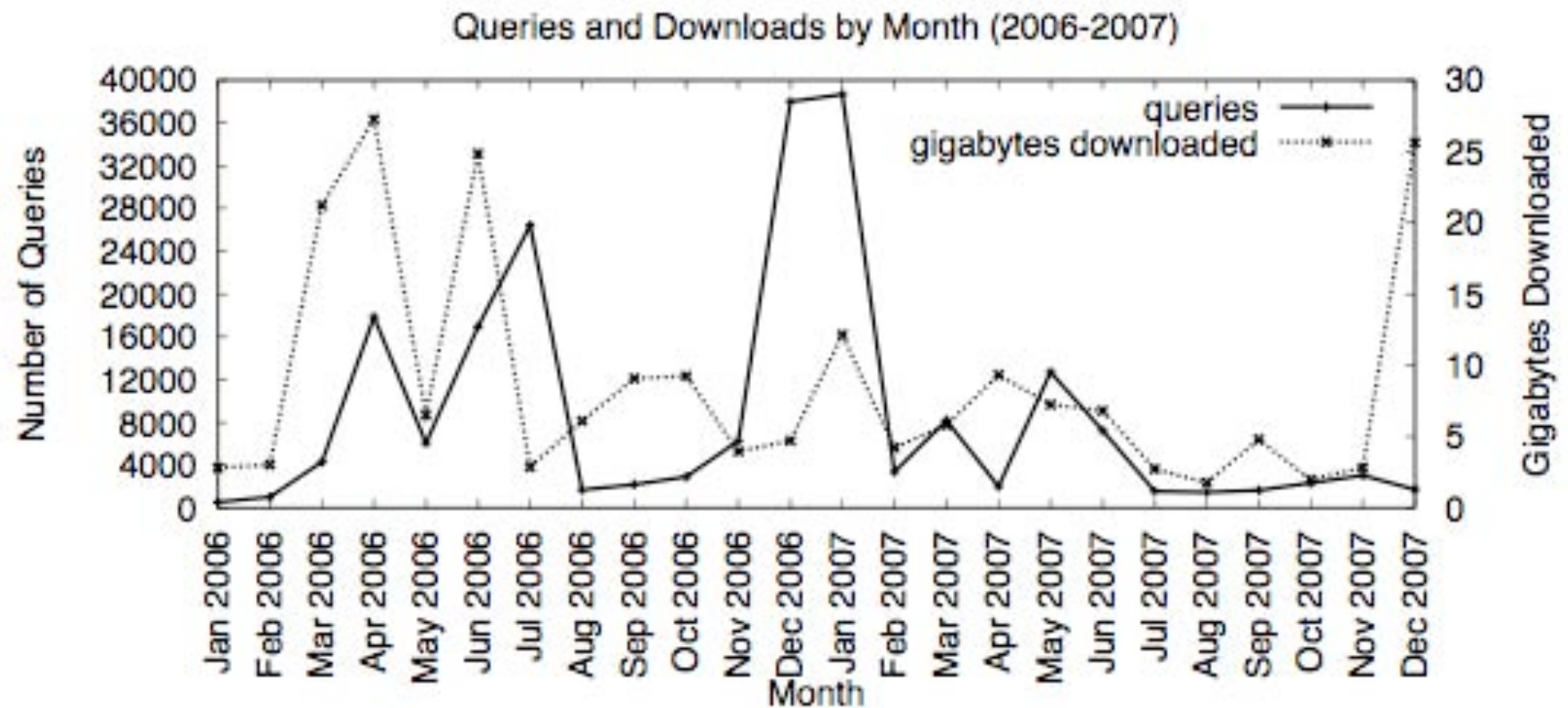
# Logic Tier

- Interactive web query system
  - Authorized user can submit query to the back end 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 of monthly users: **5 - 10**
- Total number of users with significant activity: **~75**

# Utilization Statistics: 2006-2007



# 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
    - Processes? Software development lifecycle. Communication patterns and norms.
- 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

# Computational 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.

ATCCGTT  
ATC--TT

# 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



debian



OpenOffice.org

eclipse



MySQL



K DESKTOP ENVIRONMENT

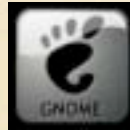


open source



GNU

# Thank You!



mozilla.org



creative commons



Python



The Apache Software Foundation

<http://www.apache.org/>

