

# The National Cancer Informatics Program Open Development Initiative

Juli Klemm, Ph.D.  
Bioinformatics Open Source Conference  
July 19, 2013

# NCIP is a cross-NCI program to support biomedical informatics in cancer research



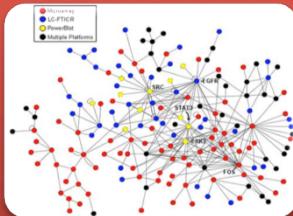
## National Cancer Informatics Program

*A cross-NCI program supporting biomedical informatics in cancer research*

- Foster community development of informatics capabilities
- Provide access to well-described data collections
- Maintain an interoperability infrastructure and standards
- Reinforce collaborative relationships among researchers
- Train biomedical investigators to use informatics capabilities

<http://cbiit.nci.nih.gov/ncip>

# Why is the NCI interested in open source?



# Biology has become a data intensive science

- Rapidly evolving, unique needs of cancer researchers
  - Big data



## Sustainability

- Budget realities
  - Innovation



# Open science movement

- Open source, open data, open access
  - Open government

# The NCIP Open Development Initiative

- ***Support*** the rapid pace of innovation in cancer research
- ***Enable*** the creation of better tools by crowdsourcing innovations
- ***Empower*** the research community to govern project priorities and roadmaps
- ***Insulate*** software development from the ebbs and flows of the federal funding cycle
- ***Engage*** a broader community of talented developers to contribute to a critical public health concern

# Progress to date

2011

- RFI: Models of open development for biomedical research
- Research...

2012

- Open source think tank
- Decision to move to BSD 3-Clause license
- NCIP GitHub Channel established <https://github.com/ncip>

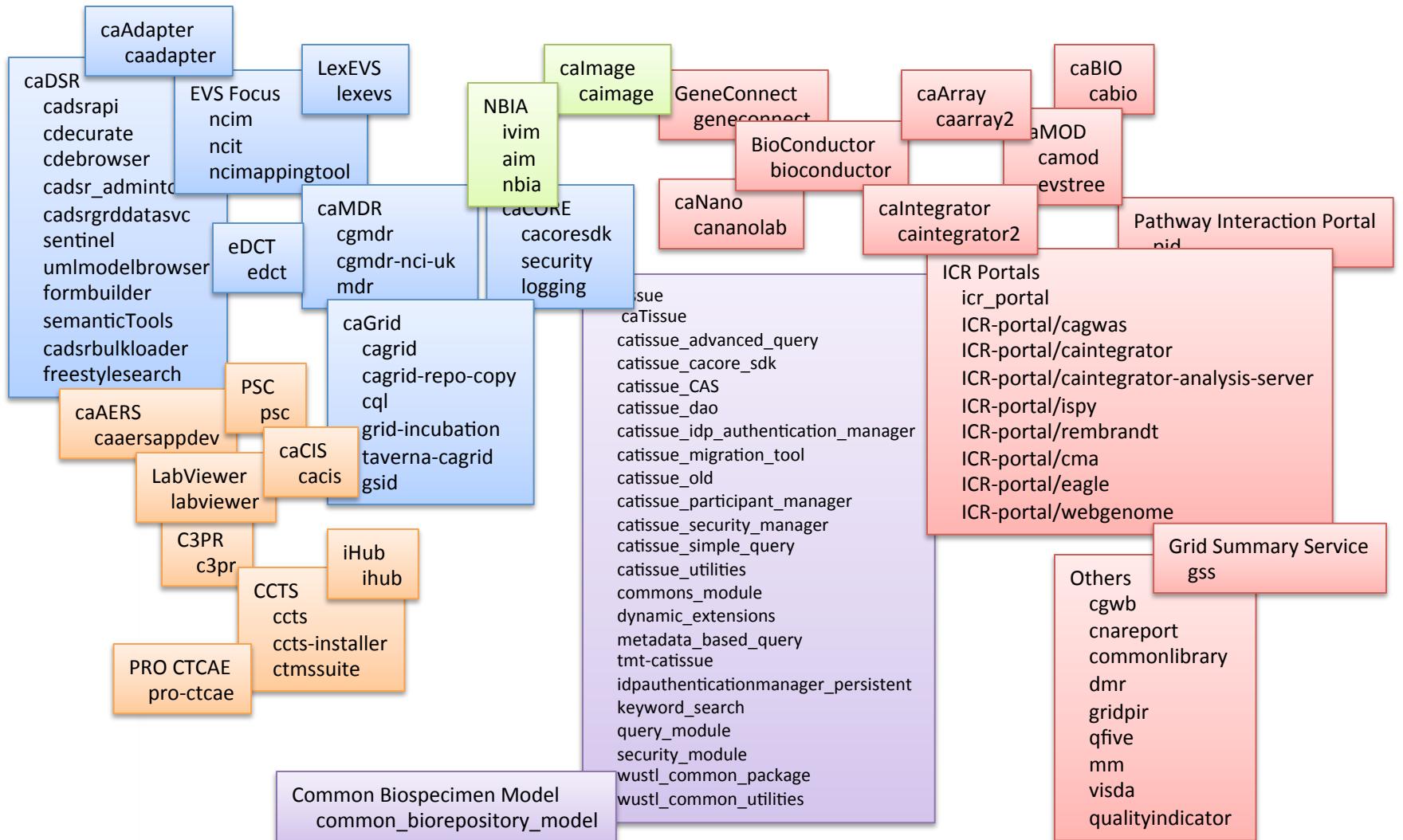
2013

- Begin code and license migration
- Today: 112 repositories representing 47 projects migrated

# Lowering the barriers: New license and new repository

- BSD 3-Clause License
  - OSI-approved, simple, well-known
  - Attractive to both academic and commercial users
  - Copyright held by the non-govt developers
- GitHub
  - DVCS required for distributed development
  - Government has pre-negotiated terms and conditions
  - Fork and pull model

# Migration Projects



# Migration resources and lessons learned

- Communication, process, SOPs, templates
  - How to Migrate From SVN to GitHub
  - Templates
    - README, NOTICE, LICENSE, CONTRIBUTING
  - Access control methodology
  - Learning resources, best practices
- Repository clean-up
  - Size, content, structure
- Security considerations for building from a remote repository
  - Local mirroring



# Next steps: Building Community

“May all of your problems be technical.” – Jim Gray



Photo: Chris Ware/Keystone Features/Getty Images

# Diversity is a big challenge

- Some community development
- Gov't development only
- Dormant

Activity

- Cancer Biology & Genomics
- Clinical & Translational Research
- Semantics & Interoperability

Domains

Technology

Readiness

- Database applications
- Toolkits
- Data models
- Analytical scripts
- ...

- Code modularity
- Documentation
- Repository structure

# Early steps at building communities

| Project Name | Initiated by | Current “Owner” | Activities                                       |
|--------------|--------------|-----------------|--|
| caLIMS       | Government   | Community       | Community-led management committee, low activity |
| caTissue     | Government   | Government      | Two “Code Jamborees” have been held              |
| XIP/AVT      | Community    | Community       | “Hack-fest” recently held – off and running      |

# Join us!

- Fork the code
- File an issue
- Fix a typo
- Watch a repo
- Become a member

**NCIP Open Source**



National Cancer Informatics Program

**at NCI**

**Supporting Cancer Research with Open Source**

Visit [ncip.nci.nih.gov](http://ncip.nci.nih.gov)  
[Wiki Pages](#)

**Recently updated** [View All on GitHub](#)

**cagrid2** Jul 18, 2013 - 1 stargazers - 0 forks  
**caarray** Jul 18, 2013 - 1 stargazers - 0 forks  
**cadrs-api** Jul 18, 2013 - 0 stargazers - 0 forks

**Statistics**

**112** public repos  
**10** members  
[opensource@nci.nih.gov](mailto:opensource@nci.nih.gov)

**Cancer Biology and Genomics**

**caGrid**

caGrid is a service-oriented platform that provides tools to integrate data, securely share data, and compose analysis pipelines

**caCORE**

caCORE contains tools and APIs that provide the building blocks for the development of interoperable information management systems

**EVS**

EVS provides tools and services to accurately code, analyze and share cancer and biomedical research, clinical, and public health information

**caDSR**

caDSR is a collection of tools for documenting and sharing human- and machine-readable data descriptions

**calIntegrator Portals**

The calIntegrator portals bring together diverse data types for integrative query and analysis

**caTissue**

caTissue is a web-based tool for managing biospecimens

**cgMDR**

cgMDR (Cancer Grid Metadata Repository) provides a framework for supporting a light weight ISO 11179 standards metadata repository

New! NCIP Channel landing page: <http://ncip.github.io>

Follow us on Twitter @NCI\_NCIP

Connect with us on LinkedIn

NCIP Blog: <http://ncip.nci.nih.gov/blog>



# Acknowledgments

## NCI CBIIT

- Robert Shirley
- Anthony Kerlavage
- George Komatsoulis

Juli Klemm  
@jklemm  
juli.klemm@nih.gov

## SAIC-Frederick, Inc.

- Sreenath Nampally
- Larry Brem

## Kitware

- Brad King
- Luis Ibanez