Open Source Community Progress and Roadmap Strawman

Andrew_McMurry(@)hms.harvard.edu

THANKS!

- ✧ Sponsor — Harvard CBMI
- ✧ Co-Organizer — *Justin Quan (chip)*
- ✧ Infrastructure — *Mike Cherry (EagleI), Rick Agrella(chip)*
- ✧ Web — *Seth Paine (EagleI)*
- ✧ Videographer — *Kerry Folley (cbmi)*

- ✧ Project Presenters
- ✧ Invited Guest Speakers

- ✧ Distinguished tech savy programmers

Agenda

① Improve Open Source process and partnerships

- Keynote
- Panel

② Improve your programming prowess

- Whitebox Topics
- DIY Natural Language Processing

③ Learn about what is going on next door

- Slam Poetry 3 X 15

Then and Now

- 2006 : CHIP developers retreat
  - ~15 attendees
- 2009 : Harvard affiliated teaching hospitals
  - 30 attendees
- 2011 : Strengthening our Open Source Partnerships
  - 50 strong

Preparing the 2006 CHIP Retreat…

- Code sharing and reuse was low
- Sparse documentation
- Issue tracking system infrequently used
- Low automated test coverage
- Zak & Ken → *"prepare for hockey stick growth"*

Litmus Test

"IF and only IF" …

1 Process survives grant rush hour

2 Obviously saves programmer time

2006 proposed shift

1. *Document* Use Cases

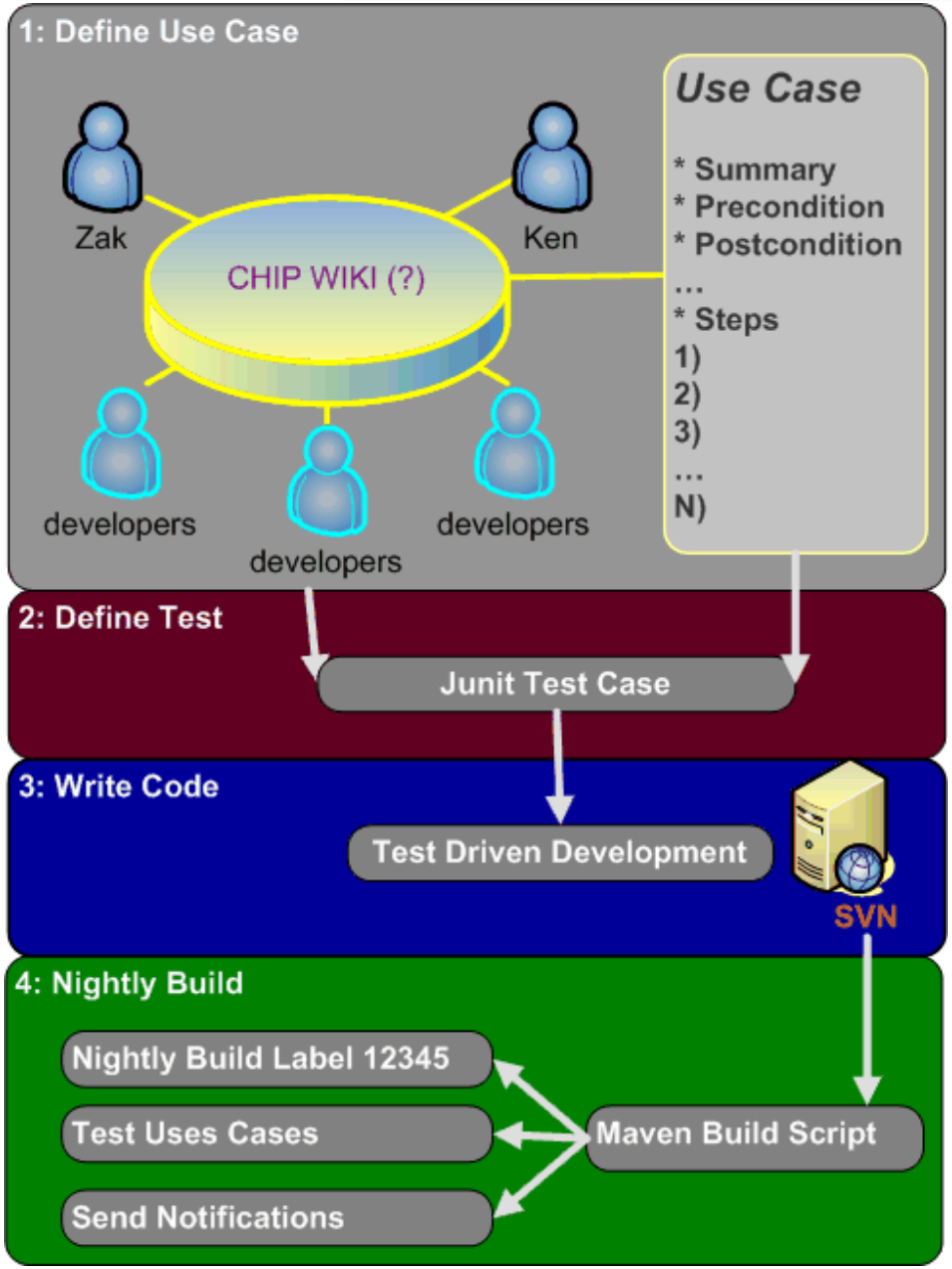
- Preconditions +
- Steps (Algorithm)+
- Post-conditions = Use Cases or "Scenarios"

2. *Test* driven development

- Regression test everything all the time

3. *Integrate* Continuously

- Build, test, integrate after *every* code change



"It would be really nice"

Lessons Learned

✗ ***Document* by Use Cases**
- ✗ Not surviving rush hour

✓ ***Test* driven development**
- ✓ Junit

✓ ***Integrate* Continuously**
- ✓ Maven
- ✓ Bamboo

} = Saves programmer time

2009 "It Would Be Nice If…"

1. I want to make a FOO
2. Scribble what FOO does
3. Set a release date "release early, release often"
4. Write unit test for FOO
5. Write the code
6. Check-in changes with automated tests
7. Release Early Release Often
8. Create and deploy the Distribution
9. Announce!

Ideal lifecyle 2009++

1. **I want to make a FOO, does it already exist?**

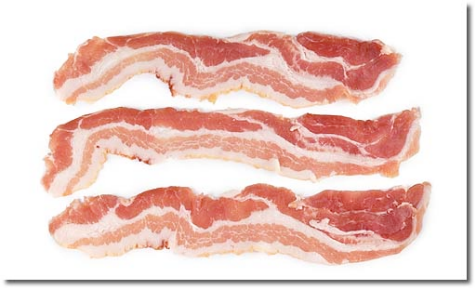
   - Increase of sharing, especially LIBS and SUBSYSTEMS
   - Knowledge of community projects mostly by word of mouth

   How can someone find your cool project?

Ideal lifecyle 2009++

1. I want to make a FOO
2. **Scribble what FOO does**
   - Documentation too often an afterthought
   - Lots of little word docs in various hard to find places

how to we motivate ourselves to write docs?



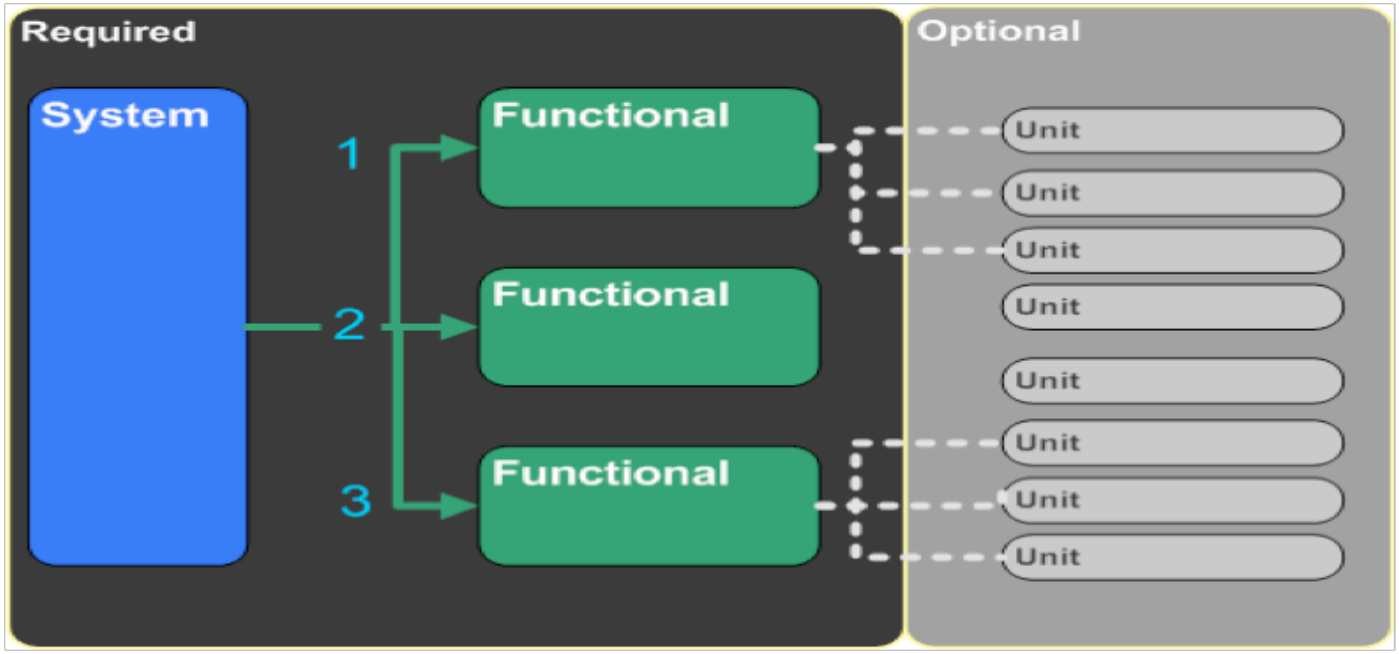
Ideal lifecyle 2009++

1. I want to make a FOO
2. Scribble what FOO does
3. **Set a release date "release early, release often"**
   - We SPRINT well in 1 month cycles
   - Release Candidates for external adopters

What is planned for the next release?

Ideal lifecyle 2009++

1. I want to make a FOO
2. Scribble what FOO does
3. Set a release date “release early, release often”
4. **Write a unit test for FOO**



Ideal lifecyle 2009++

1. I want to make a FOO
2. Scribble what FOO does
3. Set a release date “release early, release often”
4. Write unit test for FOO
5. **Write the code**
   - Pair Programming
     - Integration Scenarios
     - Really hard problems
   - “Committers”
     - Done well to keep down the number of committers
     - TOO well

Ideal lifecyle 2009++

1. I want to make a FOO
2. Scribble what FOO does
3. Set a release date “release early, release often”
4. Write unit test for FOO
5. Write the darn code
6. **Check-in changes with automated tests**
   - Publicly accessible source repositories
   - Build automatically (mostly bamboo)

Ideal Lifecyle 2009++

1. I want to make a FOO
2. Scribble what FOO does
3. Set a release date "release early, release often"
4. Write unit test for FOO
5. Write the darn code
6. Check-in changes with automated tests
7. Release Early Release Often
8. **Create Distribution**
   - Packaging distributions
   - Deploy to public test environment

How long would it take someone else to install / extend your software?

Today

- Local development lifecycle improved
- Local code sharing and reuse getting better
- External adoption is HIGH
- Open Source partnerships slowly improving
- "External" development is LOW

Why so little "external" development?

Lower the Hacktivation energy to 15 minutes or less!

--By order of John Resig (JQuery)

What *is* necessary, however, is that enough investment be put into presentation that newcomers can get past the initial obstacle of unfamiliarity. Think of it as the first step in a bootstrapping process, to bring the project to a kind of minimum activation energy. I've heard this threshold called the *hacktivation energy*: the amount of energy a newcomer must put in before she starts getting something back. The lower a project's hacktivation energy, the better. Your first task is bring the hacktivation energy down to a level that encourages people to get involved.

--Producing Free and Open Source Software

HARVARD CATALYST | THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

cbmi Center for Biomedical Informatics

where's the stuff?

cbmi
Center for Biomedical Informatics

Home » Research » Roundup

- ASD Phenotype-Genotype Correlation
  Collaborative Research
- Autworks
  Autism Gene Network
- Co-Location Collaboration
  Academic Ecosystem Modeling
- Medvane
  Forcasting Research Trends
- Rodeo
  Rapid BLAST for Researchers
- Roundup
  Comprehensive Ortholog Detection
- SPIN
  Real-time Aggregated Query
- Harvard CTSC Informatics Program
  Clinical Translational Science: Informatics Program

SPiN

Software

Logon to the Pat

- Concept & Research
- Roadmap and Status
- Software
- Community
- About Us
- FAQ
- Distributed IRB
- Publications

i2b2
Informatics for Integrating

About Us | Driving Biology Projects

Software

- i2b2 Software
- i2b2 Community Wiki
- i2b2 JIRA Bug Tracker
- i2b2 Subversion Repository
- Archived Source Code
- Contributed
- Tutorial
- Guestbook *
- Statistics *

eagle-i consortium

dip Children's Hospital Informatics Program

INDIVO HOME | RESEARCH & PUBLICATIONS | INDIVO COLLABORATORS | LICENSE & TRADEMARK | DEVELOPER COMMUNITY | TEAM & CONTACT INFO

INDIVO
THE PERSONALLY CONTROLLED HEALTH RECORD

dip HL

Developer Community

Indivo X is the new, web-platform version of Indivo. It is extensible via a standard API, of which significant portions are being developed in collaboration with Dossia. Indivo X will be released in

INDIVO X PUBLIC BETA 1 RELEASED
Indivo X Public Beta 1, the first stable version of Indivo X, is now available.

HARVARD CATALYST | THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

About Harvard Catalyst | National CTSA Consortium | Contact Us | News | Events | Spotlights | Search Harvard Catalyst

People & Collaboration | Consulting & Advice | Education & Training | Funding | Research Resources | Programs

Tools & Services
- Core Facilities
- Harvard Catalyst Central Laboratory (HCCL)
- HCCRC Resource Request
- IRB Ceded Review Request
- Laboratory for Innovative Translational Technologies
- Pathology Specimen Locator
- REDCap (Research Electronic Data Capture)
- SHRINE

Information & Support
- Atlas
- Community Connect to Research
- Countway Library Research Services
- HarvardTrials
- Medvane
- Regulatory Atlas
- Regulatory Binder

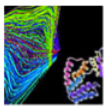
Research Resources

Harvard Catalyst offers clinical and translational investigators what they need to conduct clinical experiments, including clinical research nursing, experimental design assistance, sample processing, data management, and information about core facilities.

How do I:

- Request research support for my clinical study?
- Get advice on what resources I might need for my project?
- Find facilities or services that can process my samples or experiments?
- Find data or samples in support of my work?
- Access tools to help me organize my data and my research?
- Ensure that I have met all of the regulatory requirements for my research?
- Find funding to support my research?
- Find researchers with whom to collaborate?
- Explore research trends within the Harvard Catalyst community?

Spotlight



Core Facilities
Resources You Need: From sequencing to imaging to samples and more.

See Also
- Consulting & Advice
- Funding
- Harvard Catalyst Profiles
- Semantic Search

New Litmus Test : solve the matrix

	Dev Page	Demo	Software	Docs	Community
Scrubber					
Indivo					
SHRINE					
I2B2					
Eagle-I					
Medvane					
SHRIMP					
SPIN					
cTAKES					
SMART					
Flow Language					
YOUR PROJECT					

Open.Med

- Proposal to Solve the Matrix
- Common infrastructure = more code sharing
- Lower “Hacktivation” energy
- Local Project Hosting
- For “External” developers, even across the street

Open.MED

Welcome to Open.MED | Host Project | Services | Who uses Open.MED | FAQ

Open.med wiki > Home > ... > How to Write Use Cases > Host Project

Host Project

updated Feb 24, 2011 by Andrew McMurry

Open Source informatics projects at Harvard affiliated teaching hospitals can use the following services

* Confluence Wiki
* Jira Issue Tracker
* Bamboo Continuous Build Integration
* Mailman Mailing Lists
* Subversion Source Repository
* Nexus Artifact & Download Repository

To use any of the above services, contact Andrew McMurry.

As much or as little as you like,
completely autonomous,
have it your way

Scrubber

Research | Demo | Status | Get Software | Community | Docs | FAQ

Open.med wiki > Scrubber > Research

Research

Motivation

"Free Text" medical notes contain information which can be used to **locate human biospecimens** and even predict patient
Because medical notes often contain Protected Health Information, it is necessary to "scrub" notes of sensitive information
investigator. Towards this goal, we have developed Open Source software that removes PHI from raw text, XML, or database
The software has been approved for use by numerous hospital IRBs, and has been manually reviewed by physician experts.

Challenge

"Free Text" medical notes can be "messy", often lacking even complete sentences. Assuring that Free Text data is "cleaned"
challenging. Furthermore, differences between hospital coding styles make it difficult to reuse NLP technology at other inst
medical notes are used in the research setting despite the wealth of data notes provide.

Approach

De-Identify Patients

1 pager, what is the problem you are solving?

Scrubber

Research | Demo | Status | Get Software | Communi

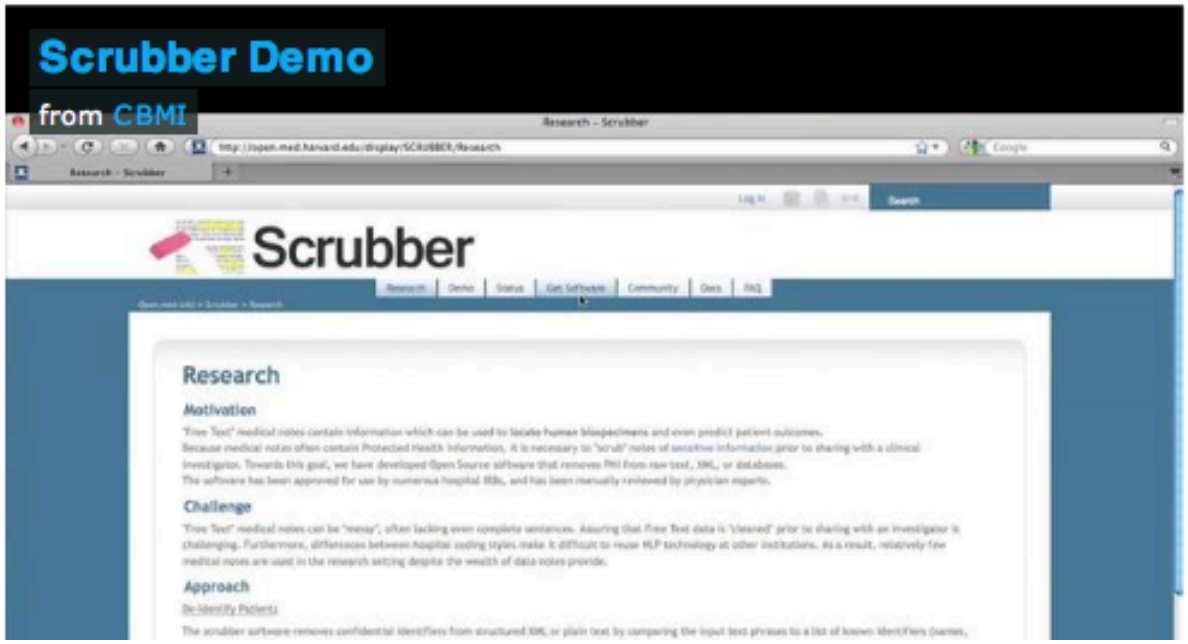
Open.med wiki > Scrubber > Research > Demo

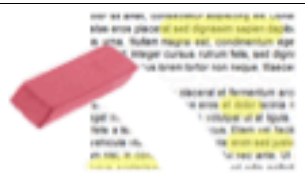
Demo

Assume low attention span

updated Mar 09, 2011 by britt fitch

2 minutes: install & demo video





Scrubber

Research | Demo | Status | Get Software | Community | Docs | FAQ

Open.med wiki > Scrubber > ... > Demo > Status

Status

is this project active?

What is scheduled for the next release? (Jira plugin)

updated Feb 23, 2011 by Andrew McMurry

This project is **actively sponsored** by Cancer.gov to aid **sharing human biospecimens** with select diagnostic and treatment

Status:

2005: Approved for use at four Harvard affiliated teaching hospitals
2006: Initial open source release for Pathology Diagnoses (Linux.com article)
2007: Completely rewritten API to improve performance, reproducibility, and hospital-specific customizations.
2008: Extended to support scrubbing other kinds of notes such as patient discharge summaries.
2009: Approved for use at two large HMO sites.
2010: Machine Learning work begins using millions of peer-reviewed publications to train "ham" (medical concepts) from "sp

2011 Roadmap

- Currently statistical evaluation of the scrubber performance is underway for upcoming publications.
- Active development on De-ID improvements using corpus data.
- Active development on new Concept Extraction module for Scrubber.

Scrubber

Research | Demo | Status | Get Software | Community | Docs | FAQ

Open.med wiki > Scrubber > Software

Software

updated Feb 23, 2011 by Andrew McMurry

Latest Release

- **>> Download v2.8 and Install Now <<**
- User Guide
- Release Notes
- Build from Source Code
  - $ svn co http://scm.open.med.harvard.edu/svn/repos/spin/scrubber/releases/2.8/ scrubber-2.8
  - $ mvn install

Development Trunk

$ svn co http://scm.open.med.harvard.edu/svn/repos/spin/scrubber/trunk/ scrubber-trunk
$ mvn install

See also Developer Guide

Scrubber

Research | Demo | Status | Get Software | Community | Docs | FAQ

Open.med wiki > Scrubber > ... > Status > Community

- Bugs and Feature Requests
- Development Process
- Mail
- Contact

Community

updated Feb 14, 2011 by britt fitch

The Scrubber software has been approved by numerous hospital review boards (IRB) with deployments across 4 Harvard hospitals and 2 large
If you would like to try out the Scrubber software in your hospital, download and use "out of the box".

Evaluate **Scrubber**

* Project Status
* Software Download
* Scrubber-User-Guide
* Publications using real data

Report **issues and request new features**

* JIRA Issues Tracker
* Software Roadmap

Contribute **code**

* Development Process
* Job Opportunities

Contact

* Contact Us
* Mailing List

Scrubber

Research | Demo | Status | Get Software | Community | Docs | FAQ

Issues

updated Feb 22, 2011 by Andrew McMurry

When submitting a new issue or feature request, please follow these steps:

1. Visit http://open.med.harvard.edu/jira/scrubber
2. Click 'create new issue' in the top navigation
3. Required fields:
    1. **Summary**: Give a brief high level summary of your issue or request.
    2. **Affects Version**: Select the version of Scrubber that you are running.
    3. **Environment**: Please let us know what kind of environment you are working in. For example, operating system
    4. **Description**: Here is where you get to give us a detailed description of your issue or request. For issues, PLEASE
    5. **Attachment**: A picture is worth a thousand words. If you have a supporting document or a screen shot of an issu
4. Click the 'create' button at the bottom of the screen and the development team will be notified of your request.

JIRA +
release plugin

How to submit bugs &
feature requests

Scrubber

Research | Demo | Status | Get Software | Community | Docs | FAQ

Open.med wiki > Scrubber > Development Process

Development Process

updated Feb 24, 2011 by Andrew McMurry

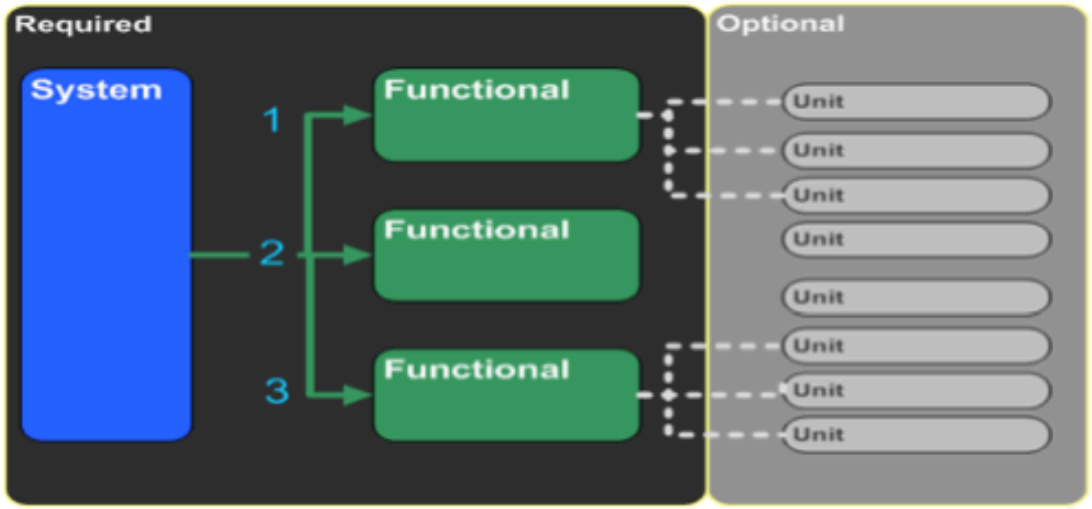
How do you want others to contribute?

The Scrubber build process is *test driven* with continuous build integration.

Scrubber follows the open source mantra *release early release often*, and all features and schedules are available in the R

By *test driven* we intend that all new functionality is checked in with a corresponding functional test. The minimum test le

- Adding two integers does not require a unit test.
- Adding a new feature or fixing a bug requires a functional test.



Scrubber

Research | Demo | Status | Get Software | Community | Docs | FAQ

Open.med wiki > Scrubber > ... > Issues > Mail

Mail

updated Feb 22, 2011 by Andrew McMurry

Announcements

Stay tuned with scrubber releases, milestones, and news

Users

Discuss experiences, issues, and features

Development

Contribute to scrubber programming and discuss new features

Recommendation from John “JQuery” Resig

Scrubber

Research | Demo | Status | Get Software | Community | Docs | FAQ

Open.med wiki > Scrubber > Research > Developer Guide

- System Overview
- User Guide
- Developer Guide
- Publications

Crucial

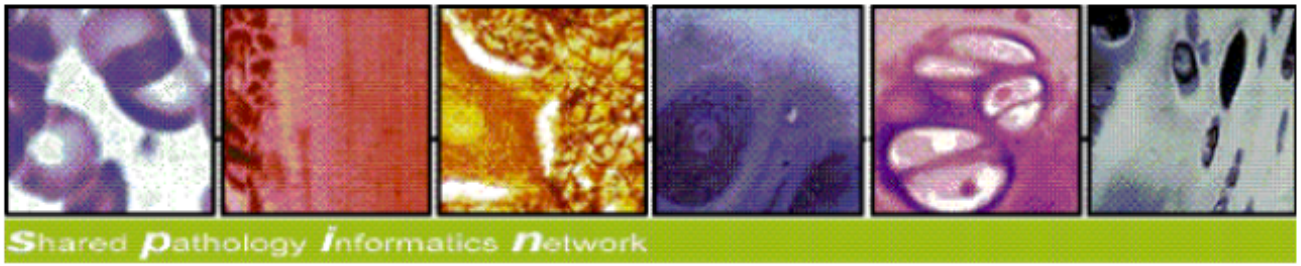
Developer Guide

updated Feb 10, 2011 by Andrew McMurry

Title: **SPIN Scrubber Developer Guide**

Author: Andrew McMurry

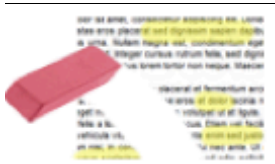
Contact: Andrew_McMurry@hms.harvard.edu



Intended Audience :

Developers looking to extend, customize, or contribute to this scrubber utility.

It is assumed the reader has already reviewed the *Scrubber User Guide*.



Scrubber

Research | Demo | Status | Get Software | Community | Docs | FAQ

- System Overview
- User Guide
- Developer Guide
- Publications

FAQ

updated Feb 10, 2011 by Andrew McMurry

Has the scrubber actually been used to share free-text patient data?

Yes. The scrubber was used to de-identify over 1 million pathology reports across Harvard affiliated teaching hospitals.

Will my hospital IRB approve using this Scrubber?

This will depend on the expected stringency and intended use of the de-identified data. At Harvard, the scrubber was IRB approved for fu
pathology reports. In the Harvard implementations, up to 400 reports could be shared with an approved investigator.

How well does the scrubber work?

See publications for an in depth manual review of cases.
Recent independent reviews at 2 large HMO sites suggest even better performance since the 2006 report.
Updated report forthcoming.

Does the scrubber also help find medical concepts? (autocoding)

This is the top priority for 2011, see project roadmap.
Since the scrubber can find phrases of any kind, this is definitely possible, even in the existing released code.

Summary

- 2006 Major Dev Lifecycle Improvements
- 2009 Increased Code Sharing and Reuse
- 2011 Solving the project matrix: Open.Med

Questions?

Andrew_McMurry(@)hms.harvard.edu