PCORnet CDM Forum

Tuesday, May 11, 2016, 1–2 PM Eastern
Hosted by Michael Matheny, MD, MS, MPH
Facilitated by Shelley Rusincovitch and Michelle Smerek

pcornet℠ | The National Patient-Centered Clinical Research Network

Agenda

- Welcome, announcements, and brief review of issue tracker
- Source control
- Repository Resources
- Wrap up

Announcements

pcornet℠

Recap: CDM Forum from April 19

- Why it may be of interest to this group:
- Discussion included:
  - Recent work with lab result reference table
  - The CONDITION table and “PCORnet-defined cohort algorithm” category
- Slides: https://github.com/CDMFORUM/CDM-GUIDANCE/wiki/CDM-Forum-Materials
- Recording: https://pcornet.imeetcentral.com/p/ZgAAAAAc3eR

pcornet

Upcoming data characterization office hours

Why it may be of interest to this group:

- Offers an opportunity for in-depth discussion with the team about your questions

Thursday, May 12, 2016
2:00 pm | Eastern Daylight Time (New York, GMT-04:00) | 1 hr
1-650-479-3207 Call-in number
Access code: 738 550 410

Monday, May 16, 2016
1:00 pm | Eastern Daylight Time (New York, GMT-04:00) | 1 hr
1-650-479-3207 Call-in number
Access code: 732 916 579

pcornet

Review of Issue Tracker (live)

CDM errata issue tracker:
https://github.com/CDMFORUM/CDM-ERRATA/issues

CDM guidance issue tracker:
https://github.com/CDMFORUM/CDM-GUIDANCE/issues

Version Control Software

Michael E. Matheny, MD, MS, MPH
Vanderbilt University / Tennessee Valley HS VA
pSCANNER and MidSouth CDRNs
Member of the PCORnet Data Committee

pcornet

Presentation Objectives

- To give the audience a working understanding of source control software:
  - the reason for its use
  - key features in choosing which solution to use
  - providing information about what existing PCORNet initiatives are using

- Primary Audience: Statistical Programmers, Database Analysts, and Software Developers (and their managers) not currently using VCS

pcornet

What this isn't

- Strong recommendations to change an existing solution or adopt any single solution
  - VCS infrastructure in a group is expensive to change (learning curve , work culture, and installation costs)
  - Each group must weight features of software and existing collaborations to come to optimized decision for themselves

pcornet

Brief Survey Question #1

Do you (or your technical team) have experience with version control software?

- High level of experience
- Moderate level of experience
- Low level of experience
- Not sure/not applicable/I don't belong to a technical team

Brief Survey Question #2

What is your (or your technical team's) *primary* development focus?

- Statistical programming (SAS, R, Stata, etc)
- Database development (for example, Oracle, SQL Server,MySQL, Postgres)
- Application development (for example, C#, VB, Java)
- Web development (for example, HTML, JavaScript, Ruby, PHP)
- Other not listed here
- Not sure/not applicable/I don't belong to a technical team

pcornet

Brief Survey Question #3

How important is source version control to you (or your technical team)?

- Very important
- Somewhat important
- Not important
- Not sure/not applicable/I don't belong to a technical team

PCORnet Data Committee: Data Infrastructure Software Development Environment Workgroup

- Leads:
  - Michael Matheny, MD, MS, MPH (TVHS VA, Vanderbilt)
  - Daniella Meeker, PhD (University of Southern California)
  - Shawn Murphy, MD, PhD (Partners – MGH)
- Objectives:
  - To review the state of version control software
  - Educate the PCORNet community on the strengths and weaknesses of different tools
  - Recommend that all PCORNet community members adopt **A** version control solution for statistical, database, and software programming projects

pcornet

What is Version Control?

- Software that allows a group of users to manage changes to programming code over time.
  - Maintains the history so bugs and errors can be rolled back and prior working versions restored
  - Allows users to contribute to a code base together
  - Provides an organization system for curation of code
  - Allows ease of access to code for users and developers to see and modify code (“check-out”)

For what types of development is this useful?

- Primarily for any text based programming
- This is a much broader scope than what most people consider:
  - Database programming
  - Statistical programming
  - Software development
  - While document management and content management systems have incorporated versioning, some users use these solutions for document versioning as well
- Usable but limited for binary and imaging versioning because the tools cannot detect fractional changes
- Example Gotcha: “This version has a bug, what did we change, and can we roll the change back?”

pcornet

Version Control Software – Key Features

- Repository Model - Client-Server versus Distributed
- Concurrency Model – Merge Versus Lock
- Licensing Model – Open Source (Free) versus Proprietary
- Operating System Compatibility
- VCS Client Tool Support

pcornet

Repository Model: Client – Server

- Centralized control – reading, writing, editing privileges can be managed
- One source of truth (Server)
- Backups are important – complete version history is only maintained on Server, risk for data loss if not properly backed up
- Can result in conflicts when users work on the same code at the same time (merge / conflict resolution)

Repository Model: Distributed

- Distributed control – “owner” of repository cannot prevent users from branching, modifying code, developing diverging code bases
- Multiple Redundancy – each person checks out a full version of the code, including all histories
  - Can work off line but checkout on mobile high data use
  - Data bloat on client because has all versions locally
  - Can recover from any copy of a repository if data loss
- Multiple Values of Truth
  - Determining the primary release and primary code is sometimes difficult with lots of branching occurring
- Merging only on demand: Since each user has a full copy of code, edit conflicts only have to be resolved during a merge request rather than immediately

pcornet

Concurrency Model

- When two users want to change the same file….
- Merge – the tool attempts to merge the changes between the users.
  - When there are no conflicts in that section of code, usually handled well, although multiple fixes in separate sections of code can break each other
  - When changes happen in same section, user has to resolve edits in a user interface
- Lock
  - A users “checks out” a file and only that user can modify the file until it is checked back in. Generally problematic for larger groups of developers (>2-3). Makes code edit management simple.

pcornet

Operating Systems

- Some VCS are only available on windows or unix-line (linux, etx)
- Macintosh compatibility varies for some tools

Client Tools

- At their heart, version control tools are a file communication and versioning protocol, so tools to access the repositories vary.
  - Desktop file folder integration (TortoiseSVN)
  - Software development environment integration
    - Visual Studio, Eclipse, etc
  - Stand-Alone User Interface (sometimes with visual graphing of branches and versions)
  - Web-Based

- Which tools support which methods of file manipulation impact productivity of development

pcornet

Licensing Model

- Open Source Tools
  - Free to use
  - Sometimes have security issues that proprietary tools do not have (SVN – VA)
  - Most popular tools have wide use
  - Generally lower funding support but larger developer community than proprietary
- Proprietary Tools
  - Sometimes tightly bound to a project set (Visual Studio – TFS) with high level of function
  - Better integration with project management tools and document management tools (suite of programs integrated)
  - Highly variable costs

pcornet

Proprietary Version Control Software

Software	Maintainer	Repository model		Concurrency model		Platforms supported				Cost
		Client-server	Distributed	Merge	Lock	Windows	Unix-Like	OS X	Others	
AccuRev SCM	Micro Focus	x		x	x	x	x	x		Non-free; $350 /seat
CA Harvest Software Change Manager	CA Technologies	x		x	x	x	x		i5/OS	Non-free
ClearCase	IBM Rational	x		x	x	x	x		Many	Non-free $4600 per floating license
Code Co-op	Reliable Software		x	x		x				Non-free $150 per seat
Dimensions CM	Serena Software	x		x	x	x	x		Many	Non-free
Endevor	CA Technologies	x		x	x				z/OS	Non-free
IC Manage	IC Manage Inc.	x		x	x	x	x	x		Non-free Commercial
MKS Integrity	Integrity, a PTC Company	x		x	x	x	x			Non-free
Perforce	Perforce Software Inc.	x		x	x	x	x	x		Cost free license, available on application, for OSS or educational use; Also free for up to 20 users, 20 workspaces, and unlimited files; Else $740–$900 per seat in perpetuity, or $144–$300 per seat per year
Plastic SCM	Codice Software	x		x	x	x	x	x		Free for up to 15 users; else starting at $595 per seat, or $3,500 per 25 developers per year
PVCS	Serena Software	x			x	x	x			Non-free
Rational Team Concert	IBM Rational	x		x	x	x	x	x	Many	Free for up to 10 users; else non-free
SCM Anywhere	Dynamsoft	x		x	x	x	x	x		Non-free Single user free; $299 per user
SourceanywhereStandalone	Dynamsoft	x		x	x	x	x	x		Non-free Single user free; $299 per user
StarTeam	Borland	x		x	x	x	x	x	*	Non-free Quoted on an individual basis.
Surround SCM	Seapine Software	x		x	x	x	x	x		Non-free $595 per named user; $29/month subscription
Team Foundation Server	Microsoft	x	x	x	x	x	*	*	*	Free <= 5 users in Visual Studio Online, else Non-Free
Synergy	IBM Rational	x	x	x	x	x	x			Non-free
Vault	SourceGear	x		x	x	x	x			Non-free $300 per user

pcornet

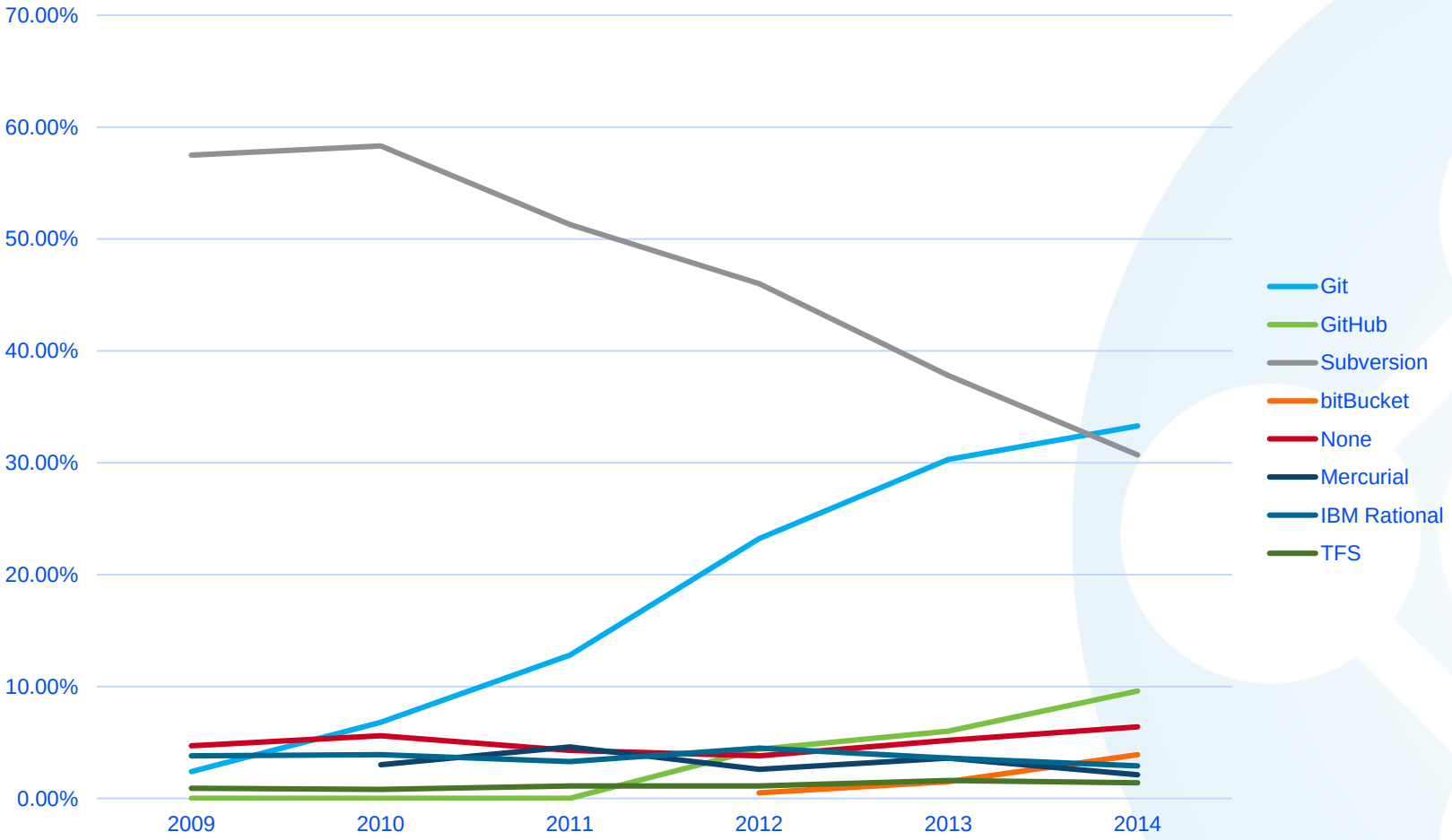
Source: Wikipedia & Laura Paoletti (Children's Hosp of LA)
https://en.wikipedia.org/wiki/Comparison_of_version_control_software

Open Source Version Control Software

Software	Maintainer	Repository model		Concurrency model		Platforms supported			
		Client-server	Distributed	Merge	Lock	Windows	Unix-Like	OS X	Others
GNU Bazaar	Canonical Ltd.	x	x	x		x	x	x	
BitKeeper	BitMover Inc.		x	x		x	x	x	
darcs	The Darcs team		x	x		x	x	x	
Fossil	D. Richard Hipp		x	x		x		x	POSIX
Git	Junio Hamano		x	x		x		x	
Mercurial	Matt Mackall		x	x		x	x	x	
Monotone	Nathaniel Smith, Graydon Hoare		x	x		x	x	x	
Subversion	Apache Software Foundation	x		x	x	x	x	x	

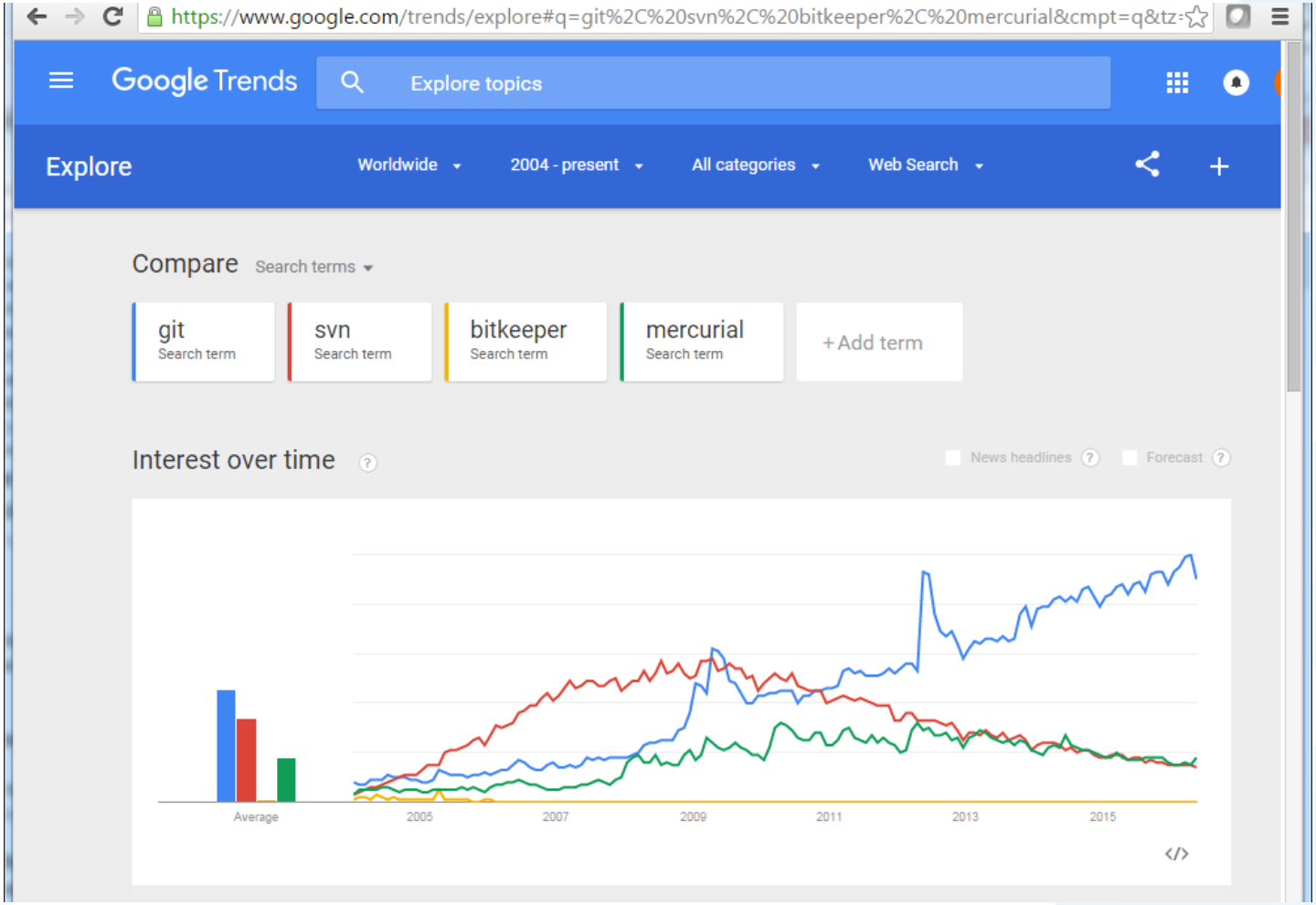
Source: Wikipedia & Laura Paoletti (Children's Hosp of LA)
https://en.wikipedia.org/wiki/Comparison_of_version_control_software

Eclipse (Java) User Survey – Source Control Use



Source: Eclipse Yearly Community User Surveys

Google Search Trends



Source: https://www.google.com/trends/

Example of SVN File Structure

GIT vs SVN - CodeForest | GitSvnComparison - Git S | sha1 - Google Search | Programming_SQL_Comm

https://subversion.matheny.info/!/#Programming_SQL_Common/view/head/trunk

VISUALSVN SERVER

Programming_SQL_Common / trunk

Revision: HEAD

..			
Download		r90	michael.matheny
Reference Code		r75	michael.matheny
Reference Table Views		r155	michael.matheny
051 - Make [Med].[usp_Process_RxOutPat_X].sql	62 KB	r14	jason.denton
[dbo].[ilfCalculateAge].sql	3 KB	r195	jason.denton
[dbo].[SetDefaultSchema].sql	1 KB	r109	michael.matheny
[Dim].[usp_LocalDrug_Updated].sql	60 KB	r70	michael.matheny
[Intermed].[usp_Condition_Era].sql	34 KB	r224	jason.denton
[Intermed].[usp_Inpatient_Filtered].sql	67 KB	r232	michael.matheny
[Intermed].[usp_Patient_Demographics].sql	19 KB	r232	michael.matheny
[Intermed].[usp_RxOutPatFill_Med_Era].sql	23 KB	r242	michael.matheny
[Intermed].[usp_RxOutPatFill_MedClass_Era].sql	20 KB	r199	jejo.koola
[Intermed].[usp_RxOutPatFill_MorphineEquivDaily_Era].sql	12 KB	r241	michael.matheny
[Intermed].[usp_Vitals_Inpatient_Relative].sql	63 KB	r99	michael.matheny
[Map].[CCS_X].sql	979 KB	r64	michael.matheny
[Map].[ConditionID_to_DimSID].sql	5 KB	r152	michael.matheny
[Map].[ConvertLabTextToValue].sql	7 KB	r192	jason.denton
[Map].[MedSchedule].sql	5 KB	r105	michael.matheny
[Map].[NarcoticEquivalent].sql	9 KB	r105	michael.matheny

https://subversion.matheny.info/!/#Programming_SQL_Comm...

GIT vs SVN - CodeForest | GitSvnComparison - Git S | sha1 - Google Search | Programming_SQL_Comm

https://subversion.matheny.info/!/#Programming_SQL_Common/history/head/trunk

VISUALSVN SERVER

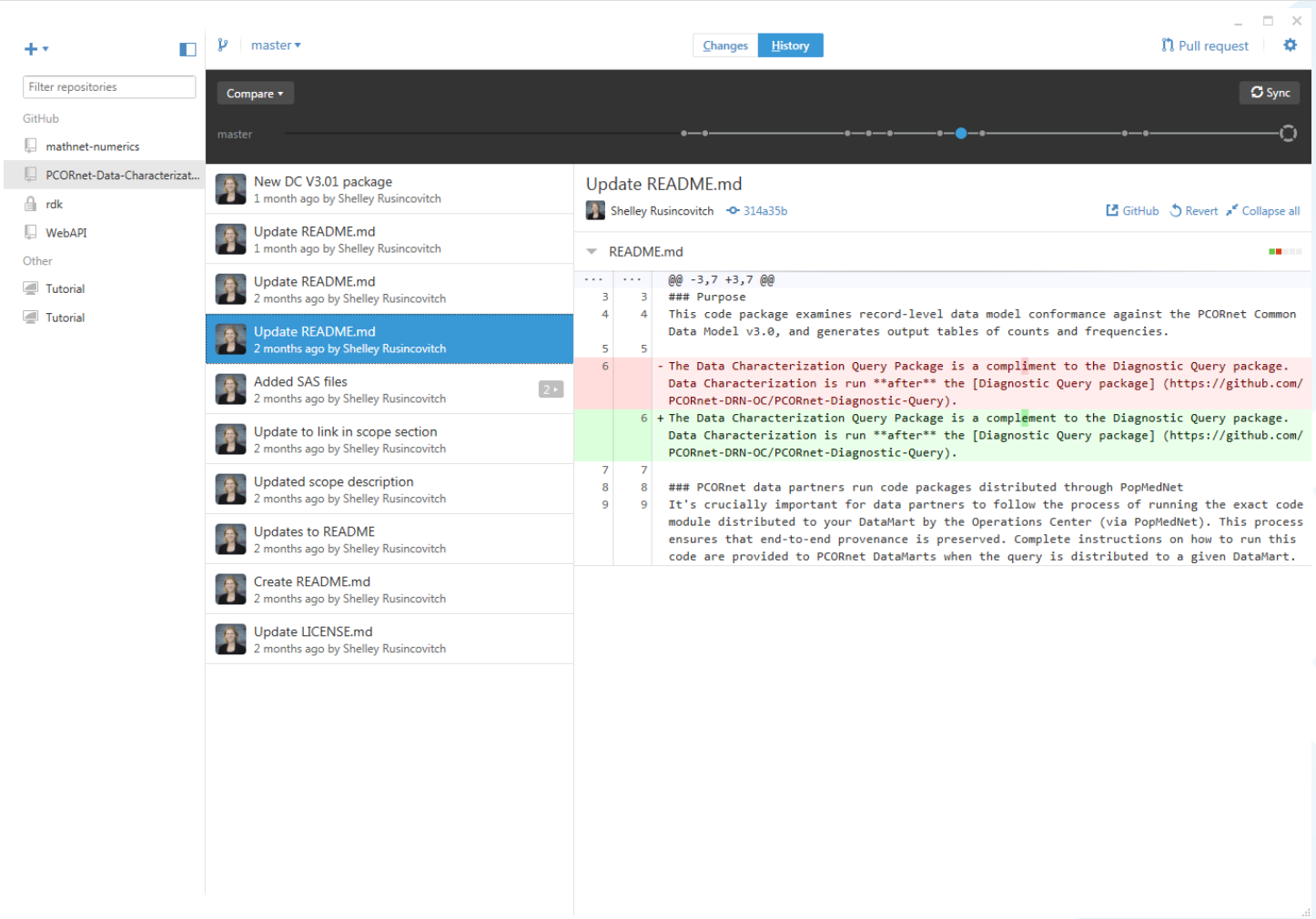
Programming_SQL_Common / trunk

Revision: HEAD

Rev	Author	Date	Message
r248	jason.denton	5/4/2016, 1:46:20 PM	
r247	jason.denton	4/6/2016, 4:28:29 PM	
r245	jason.denton	3/31/2016, 7:24:03 PM	
r244	michael.matheny	3/28/2016, 2:49:12 PM	
r243	michael.matheny	3/28/2016, 2:48:01 PM	
r242	michael.matheny	3/28/2016, 2:45:31 PM	
r241	michael.matheny	3/28/2016, 2:41:49 PM	
r240	michael.matheny	3/28/2016, 2:39:54 PM	
r239	michael.matheny	3/28/2016, 2:19:45 PM	Massive Map.Medication refactoring
r238	jason.denton	3/25/2016, 5:09:40 PM	Corrected definition of Tramadol
r234	michael.matheny	3/22/2016, 7:48:51 AM	
r233	michael.matheny	3/22/2016, 7:39:28 AM	
r232	michael.matheny	3/22/2016, 7:21:04 AM	
r230	michael.matheny	3/22/2016, 7:04:26 AM	
r228	michael.matheny	3/22/2016, 6:52:54 AM	branched to vinci mirror, purpose is to re-merge modded code
r227	jason.denton	3/14/2016, 8:09:57 AM	Replaced inner-most cursor for era generation with Itzik Ben-Gan interval packing algorithm.
r226	jejo.koola	3/5/2016, 1:05:29 PM	
r225	jejo.koola	3/5/2016, 12:55:25 PM	
r224	jason.denton	3/3/2016, 8:08:20 PM	Added @CMS flag for carry through to Temp.usp_Condition_Events
r223	jason.denton	3/3/2016, 2:14:31 PM	Completed the CMS injection code.
r222	jason.denton	3/3/2016, 2:11:27 PM	Started to add Hypoglycemia meds.
r221	jason.denton	3/3/2016, 2:09:00 PM	Removed CMS logic as CMS data is not integrate at the Temp level.
r220	jejo.koola	3/1/2016, 2:49:12 PM	
r219	jejo.koola	2/26/2016, 4:03:29 PM	

pcornet

Example of GitHub Cloned Repository (with an Update)



Git/GitHub Versus Subversion

- Summary of Comparisons
  - If you want to use what is most commonly used: Either, both have high market penetrance, but Git is ascending (14M users, 35M projects)
  - If you need tighter control on access and editing: SVN
  - If you want to be able to just check out a sub-section of Code: SVN
  - Speed of Operations (important for larger code bases): Git
  - Shorter & Predictable Version Numbers: SVN (Git uses a hash)
  - Ability to Represent Richer Branching History: Git
  - Ability to develop code off-line: Git (complete version history local)

Sources: https://git.wiki.kernel.org/index.php/GitSvnComparison
https://en.wikipedia.org/wiki/Comparison_of_version_control_software
http://www.codeforest.net/git-vs-svn

Source Control Conversion & Mirroring Resources

- Subversion to Git
  - SubGit: migration AND mirroring
- Git to Subversion
  - http://stackoverflow.com/questions/661018/pushing-an-existing-git-repository-to-svn/1056817#1056817
  - Has trouble with large volumes of branching

Key Reference Links to Code Repositories

If helpful, my “shortlist” of links:

- ADAPTABLE base phenotype : https://github.com/ADAPTABLETRIAL/PHENOTYPE
- CDM errata issue tracker: https://github.com/CDMFORUM/CDM-ERRATA/issues
- CDM guidance issue tracker: https://github.com/CDMFORUM/CDM-GUIDANCE/issues
- PCORnet diagnostic query package: https://github.com/PCORnet-DRN-OC/PCORnet-Diagnostic-Query
- PCORnet data characterization query package: https://github.com/PCORnet-DRN-OC/PCORnet-Data-Characterization
- PCORnet Data Committee on GitHub: https://github.com/PCORnet/DataCommittee
- DRN OC home page: https://pcornet.imeetcentral.com/p/aQAAAAAB6T9b

pcornet

Source: Shelley Rusincovitch

Conclusions

- We highly recommend the use of a version control software solution for text based documents with frequent changes, even for a single user
  - Statistical programming
  - Database programming
  - Software development

- Because of low cost (free) and high utilization, use of subversion or Git would be recommended as a solution for those not already using a VCS solution

- PCORNet is generally hosting its code in GitHub, but other source control solutions can be used, and links posted to the PCORNet commons.
  - An excellent example is the PCORNet Data Committee's repositories list
    - https://github.com/PCORnet/DataCommittee/wiki/Community

pcornet

Brief Survey Question #4

After the presentation, how important is source version control to you (or your technical team)?

- Very important
- Somewhat important
- Not important
- Not sure/not applicable/I don't belong to a technical team

Wrap up

Next CDM Forum: June 8, 2016

PCORnet Common Data Model (CDM) Implementation Forum

Wednesday, June 8, 2016, 2:00 – 3:00 PM Eastern time
Hosted by Keith Marsolo, MD; facilitated by Shelley Rusincovitch and Michelle Smerek

(Calendar invites will be sent tomorrow)

Online:
https://dukemed.webex.com/dukemed/j.php?MTID=m13d7d4f519aed700a09592e1b68059e0
Call-in: 1-650-479-3207 / Access code: 735 621 006

pcornet