

# THE ATLAS DATA MANAGEMENT SOFTWARE ENGINEERING PROCESS

Mario Lassnig (CERN PH-ADP)  
on behalf of the ATLAS collaboration

[mario.lassnig@cern.ch](mailto:mario.lassnig@cern.ch)  
[ph-adp-ddm-lab@cern.ch](mailto:ph-adp-ddm-lab@cern.ch)

CHEP, 2013-10-17, Amsterdam, NL

# Overview

2

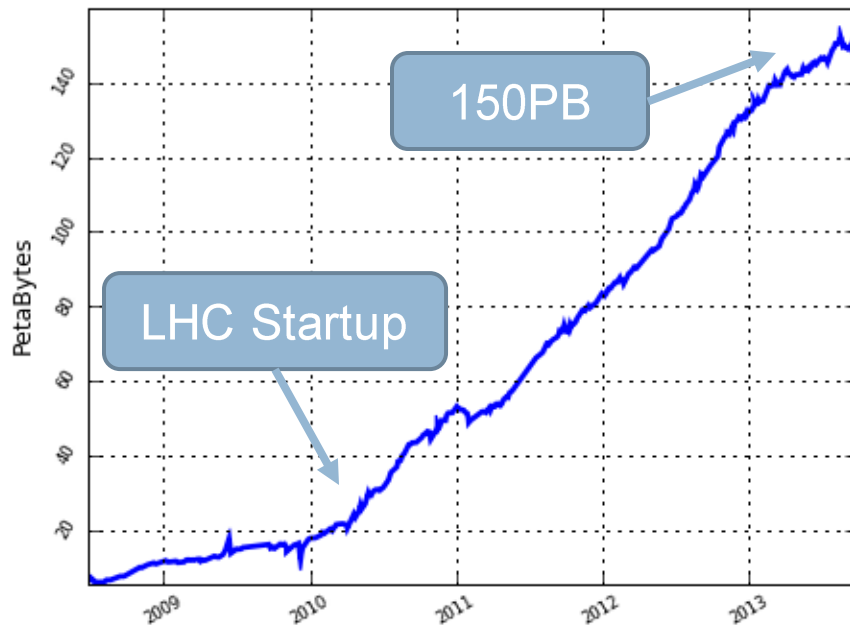
- ATLAS Data Management
- Requirements engineering
- Implementation and test
- The human factor
- Conclusion

# ATLAS Data Management

3

- Currently accumulating 40PB of data annually
  - ▣ Data is written to files, and aggregated into datasets
  - ▣ Distributed to data centres in the WLCG

Total GRID space usage according to DQ2



Total GRID files according to DQ2



# A new data management system?

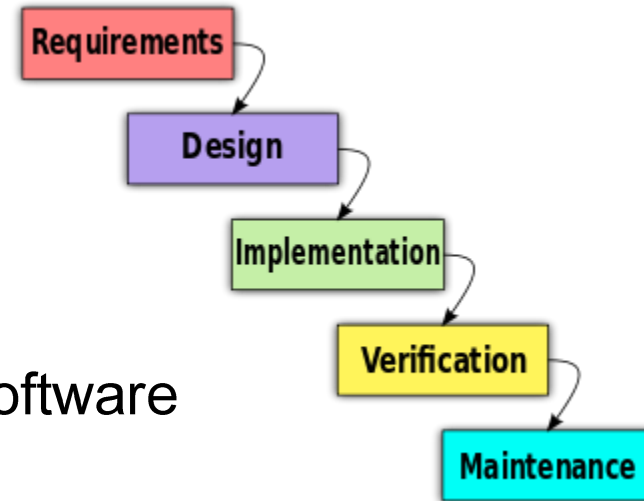
4

- Next-generation data management system
  - ▣ Rucio: cf. Vincent Garonne's talk from Tuesday
- Ensure system scalability beyond LHC Run-2
- Reduce operational overhead
- Support new ATLAS use cases
- New technologies, paradigms, middlewares
  
- Building Rucio required a new approach to software engineering
  - ▣ Existing Domino/Jenga approach not feasible

# Software process

5

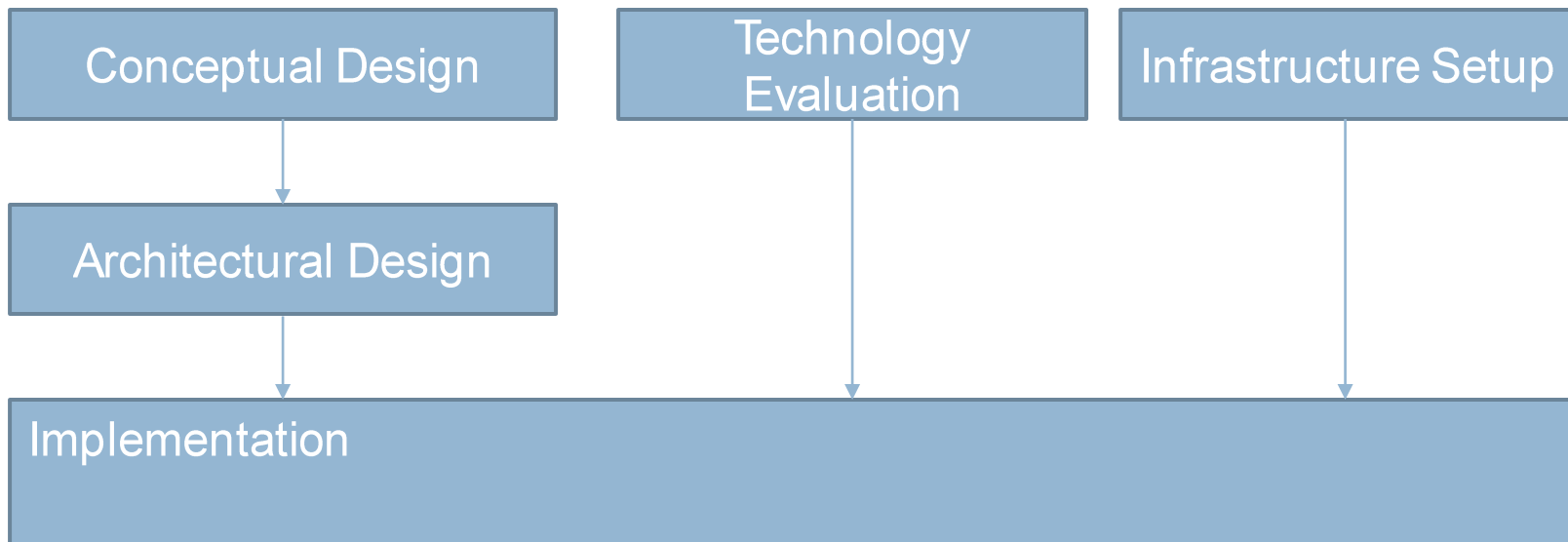
- Classic model: Waterfall
  - ▣ Fully complete each step
- However, not really applicable
  - ▣ Clients are required to state explicitly what they want upfront!
  - ▣ Critical dependencies on third party software e.g., a file transfer service
- Alternatives, such as Agile, Spiral, or Prototyping also have their own problems
  - ▣ Too consumer/business-focused, etc..
- Instead, we opted for our own modified waterfall model



# Modified software process

6

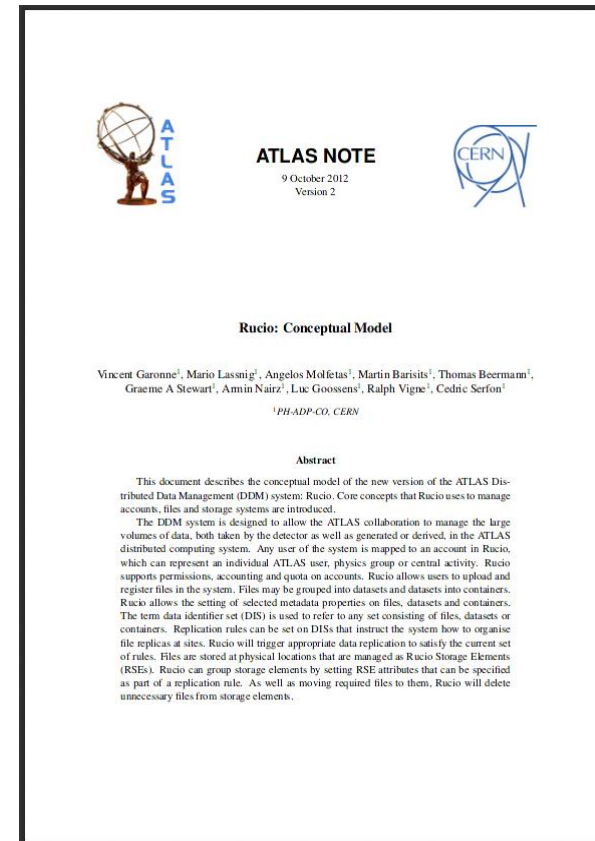
- Serialise specification and design
- However, in parallel, construct the foundation for the later stages of the process
  - ▣ Evaluation of technology – proof-of-concept
  - ▣ Setup of the infrastructure
- Couple the implementation and verification
  - ▣ Maintenance follows naturally



# Requirements engineering

7

- Started with a whitepaper
- Sent out surveys to our clients
  - ▣ Tier-0, AMI, data preparation, PanDA, endusers...
- Focused sessions with each client separately
- Technical meetings with other LHC experiments
- Workshop on the evolution of ATLAS Data Handling
- Explicit use case descriptions
- Resulting in the Rucio conceptual design
  - ▣ Signed off sentence-by-sentence



# Meanwhile...

8

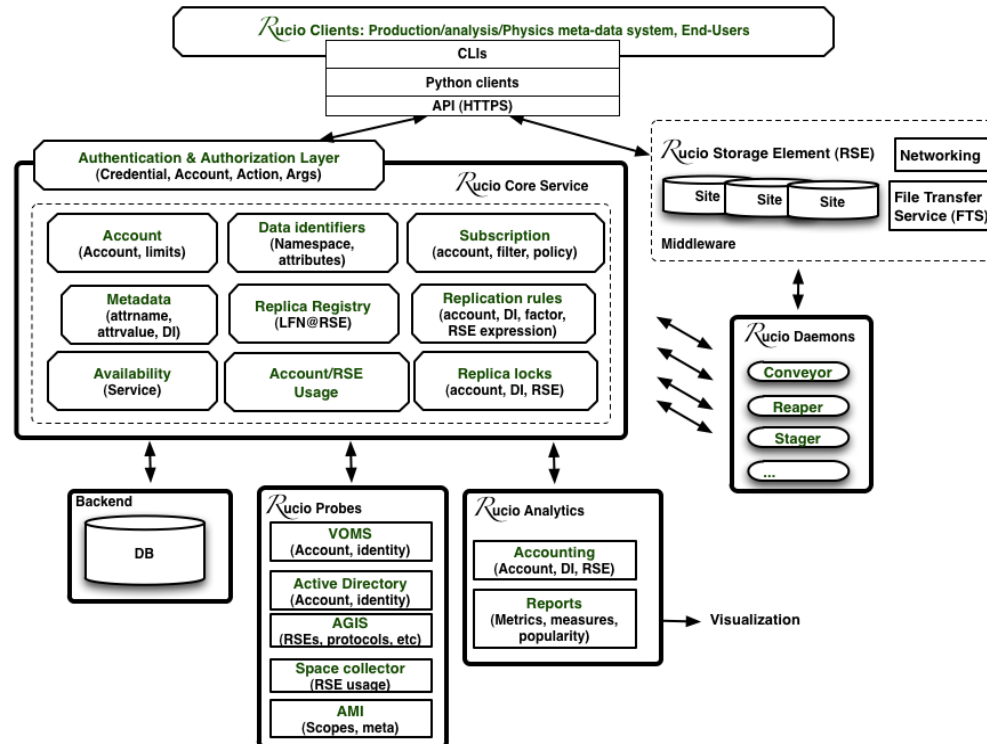
- Several technologies were evaluated
  - ▣ No specifications yet as the conceptual model not finished
  - ▣ However, experience from DQ2 leads to educated guesses
- Database management systems
- Database abstraction layers
- Communication protocols
- Source code management
- Queuing and notification
- Integration with external systems
- Testing and analysis
- Deployment



# Design and architecture

9

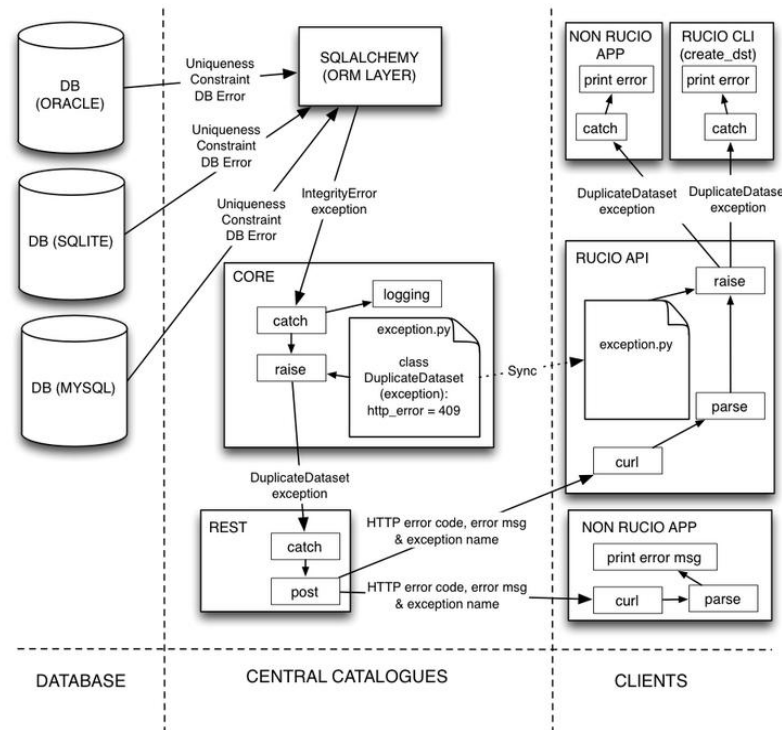
- Only after the conceptual model was finished
- First results from the technology evaluation
- Component interaction based design
  - ▣ Sequence/Flow diagrams for the most critical parts



# Design and architecture

10

- Only after the conceptual model was finished
- First results from the technology evaluation
- Component interaction based design
  - ▣ Sequence/Flow diagrams for the most critical parts



# Implementation

11

- We decided to follow a single rule:

Rucio must be *always releasable*

- This poses several requirements
  - ▣ We must be confident that the software works
  - ▣ No one is allowed to modify the software unsupervised
  - ▣ New features are implemented separately

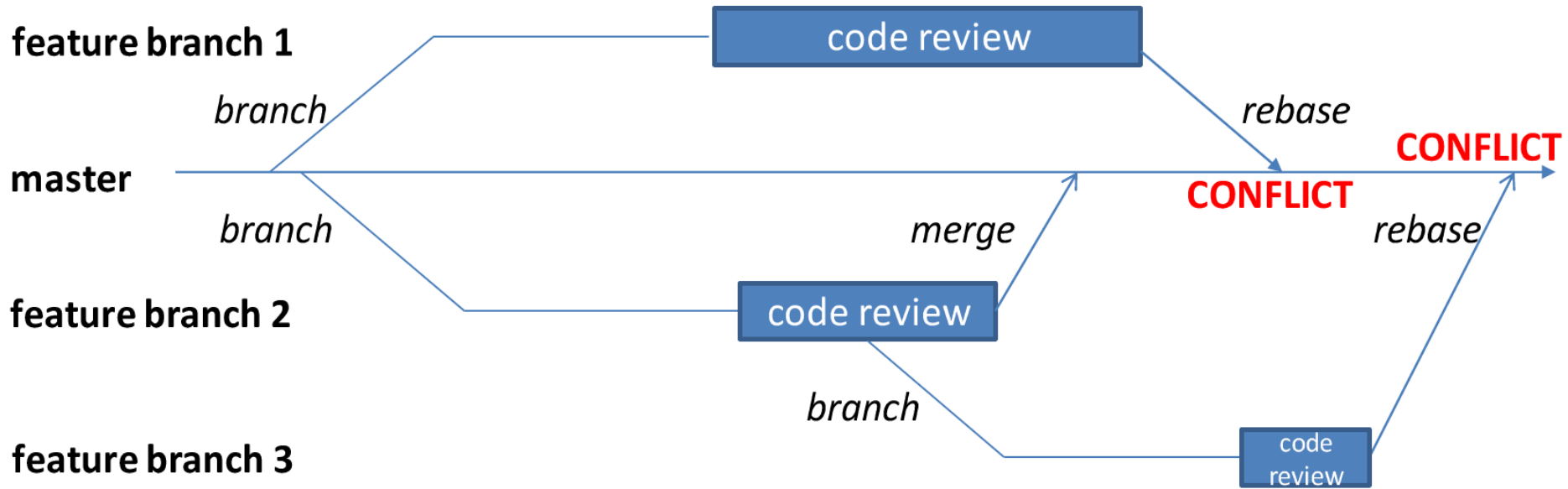
# Implementation

12

- Test-driven development
- Approval of features
  - Visual scanning of the source code
  - Local test of the source code
  - Explicit vote
- Feature branches are rejected without review, in case of
  - Missing test cases
  - Missing documentation or comments
  - Wrong coding style
- We use *git* (distributed version control) and *gerrit* (code review) to enforce this workflow
  - <http://git-scm.com/>
  - <http://code.google.com/p/gerrit/>

# The branching strategy

13



# Gerrit – Project overview

14

All **My** Projects People Plugins Documentation

status:open

Search

Mario Lassnig ▾

[Changes](#) [Drafts](#) [Draft Comments](#) [Watched Changes](#) [Starred Changes](#)

Search for status:open

Subject	Owner	Project	Branch	Updated	CR
★ [RUCIO-340] Added rule interaction to command line client	Martin Barisits	rucio	master	Oct 15	✗ Martin Barisits
★ Elastic Search	楚中 潘	rucio	master	Aug 22	-1 Martin Barisits

Press '?' to view keyboard shortcuts  
Powered by [Gerrit Code Review \(2.7\)](#) | [Report Bug](#)

All **My** Projects People Plugins Documentation

Change #, SHA-1, tr:tid or owner:email

Search

Mario Lassnig ▾

[Changes](#) [Drafts](#) [Draft Comments](#) [Watched Changes](#) [Starred Changes](#)

Change-Id:	I45d9b69f444af54d32f7add85748066e7a70efcf
Owner	Martin Barisits
Project	rucio
Branch	master
Topic	
Uploaded	Oct 8, 2013 15:03
Updated	Oct 15, 2013 14:04
Submit Type	Rebase if Necessary
Status	Review in Progress

★ [Commit Message](#) [Permalink](#)

[RUCIO-340] Added rule interaction to command line client

Change-Id: I45d9b69f444af54d32f7add85748066e7a70efcf

Reviewer	Code-Review
Martin Barisits	✗
Mario Lassnig	
Cedric Serfon	+1

Name or Email or Group

Add Reviewer

## ▼ Dependencies

Subject	Owner	Project	Branch	Updated
<b>Depends On</b>				
★ attach_dids_to_dids instead of add_replica (MERGED)	Ralph Vigne	rucio	master	Oct 8
<b>Needed By</b>				
(None)				

# Gerrit – Patch overview

15

Reference Version: Base ▾

▼ **Patch Set 2** 5c4173eb8cd6d54a203c93029be2827736f71541

Author	Martin Barisits <martin.barisits@cern.ch> Oct 8, 2013 15:03
Committer	Martin Barisits <martin.barisits@cern.ch> Oct 9, 2013 16:58
Parent(s)	9776fc31ec166d95c6827a9670686f7eff1d9ae4 attach_dids_to_dids instead of add_replica
Download	<a href="#">checkout</a>   <a href="#">pull</a>   <a href="#">cherry-pick</a>   <a href="#">patch</a>   <a href="#">Anonymous HTTP</a>   <a href="#">SSH</a>   <a href="#">HTTP</a>   <code>git fetch ssh://mario@atlas-gerrit.cern.ch:29418/rucio refs/changes/30/830/2 &amp;&amp; git cherry-pick FETCH_HEAD</code> 📄

[Review](#) [Abandon Change](#) [Rebase Change](#)

	File Path	Comments	Size	Diff	Reviewed
▶	Commit Message			Side-by-Side Unified	
M	bin/rucio	<b>2 comments</b>	+155, -1	Side-by-Side Unified	
M	lib/rucio/api/rule.py		+4, -0	Side-by-Side Unified	
M	lib/rucio/client/subscriptionclient.py		+1, -1	Side-by-Side Unified	
M	lib/rucio/core/rule.py		+24, -18	Side-by-Side Unified	
M	lib/rucio/tests/test_rule.py		+0, -1	Side-by-Side Unified	
M	lib/rucio/web/rest/did.py		+1, -1	Side-by-Side Unified	
M	lib/rucio/web/rest/rule.py		+2, -0	Side-by-Side Unified	
M	lib/rucio/web/rest/subscription.py		+1, -1	Side-by-Side Unified	
			+188, -23	All Side-by-Side All Unified	

▶ **Patch Set 1** 20925691d9dbca76197c91eda2c9055267d6184b

Comments	<a href="#">Expand Recent</a>   <a href="#">Expand All</a>   <a href="#">Collapse All</a>
<b>Mario Lassnig</b> Patch Set 1: (2 comments) typo Oct 8 15:13	
<b>Martin Barisits</b> Uploaded patch set 2. Oct 9 16:58	
<b>Cedric Serfon</b> Patch Set 2: Code-Review+1 (1 comment) Oct 10 16:06	
<b>Martin Barisits</b> Oct 15 14:04	
Patch Set 2: Code-Review-2 (1 comment)	

[Add Comment](#)

# Gerrit – Patch comparison

16

All My **Differences** Projects People Plugins Documentation  
Side-by-Side Unified Commit Message Preferences Patch Sets Files

Change #, SHA-1, tr:id or owner:email

Search Mario Lassnig ▾

bin/rucio

Reviewed & next⇒

Commit Message

Up to change

rule.py⇒

Patch Set Base 1 2		Patch Set 1 2	
1	#!/usr/bin/env python	1	#!/usr/bin/env python
2		2	
3	# Copyright European Organization for Nuclear Research (CERN)	3	# Copyright European Organization for Nuclear Research (CERN)
4	#	4	#
5	# Licensed under the Apache License, Version 2.0 (the "License");	5	# Licensed under the Apache License, Version 2.0 (the "License");
6	# You may not use this file except in compliance with the License.	6	# You may not use this file except in compliance with the License.
7	# You may obtain a copy of the License at http://www.apache.org/licenses/LICENSE-2.0	7	# You may obtain a copy of the License at http://www.apache.org/licenses/LICENSE-2.0
8	#	8	#
9	# Authors:	9	# Authors:
10	# - Mario Lassnig, <mario.lassnig@cern.ch>, 2012-2013	10	# - Mario Lassnig, <mario.lassnig@cern.ch>, 2012-2013
11	# - Vincent Garonne, <vincent.garonne@cern.ch>, 2012-2013	11	# - Vincent Garonne, <vincent.garonne@cern.ch>, 2012-2013
12	# - Thomas Beermann, <thomas.beermann@cern.ch>, 2012	12	# - Thomas Beermann, <thomas.beermann@cern.ch>, 2012
13	# - Yun-Pin Sun, <yun-pin.sun@cern.ch>, 2013	13	# - Yun-Pin Sun, <yun-pin.sun@cern.ch>, 2013
14		14	
15	"""	15	"""
16	Rucio CLI.	16	Rucio CLI.
17	"""	17	"""
18		18	
19	import argcomplete	19	import argcomplete
20	import argparse	20	import argparse
21	import os	21	import os
22	import sys	22	import sys
23	import time	23	import time
24		24	
25		25	
26	from rucio import client	26	from rucio import client
27	from rucio import version	27	from rucio import version
28	from rucio.client.accountclient import AccountClient	28	from rucio.client.accountclient import AccountClient
29	from rucio.client.didclient import DIDClient	29	from rucio.client.didclient import DIDClient
30	from rucio.client.metaclient import MetaClient	30	from rucio.client.metaclient import MetaClient
31	from rucio.client.pingclient import PingClient	31	from rucio.client.pingclient import PingClient
32	from rucio.client.rseclient import RSEClient	32	from rucio.client.rseclient import RSEClient
33	from rucio.client.ruleclient import RuleClient	33	from rucio.client.ruleclient import RuleClient
34	from rucio.client.scopeclient import ScopeClient	34	from rucio.client.scopeclient import ScopeClient
35		35	from rucio.client.subscriptionclient import SubscriptionClient
35	from rucio.common.utils import Adler32	36	from rucio.common.utils import Adler32
36	from rucio.rse import rsemanager	36	from rucio.rse import rsemanager
		37	



# Gerrit – Patch comments

17

```
try:
    rules = client.list_subscription_rules(subscription_id=args.subscription_id)
except Exception, e:
    print 'Failed to list rules'
    print e
    return FAILURE
```

809  
810  
811  
812  
813  
814

**Cedric Serfon** Oct 10 16:06

Would be useful to have a query by subscription name + account. The subscription\_id is an internal thing and probably don't need to be exposed.

[Reply ...](#) [Reply 'Done'](#)

**Martin Barisits** Oct 15 14:04

I agree, I will add this to the next patch.

[Reply ...](#) [Reply 'Done'](#)

else:

815

**(Draft)**

lorem ipsum

[Save](#) [Discard](#)

```
print 'At least one option has to be given. Use -h to list the options.'
```

```
return FAILURE
```

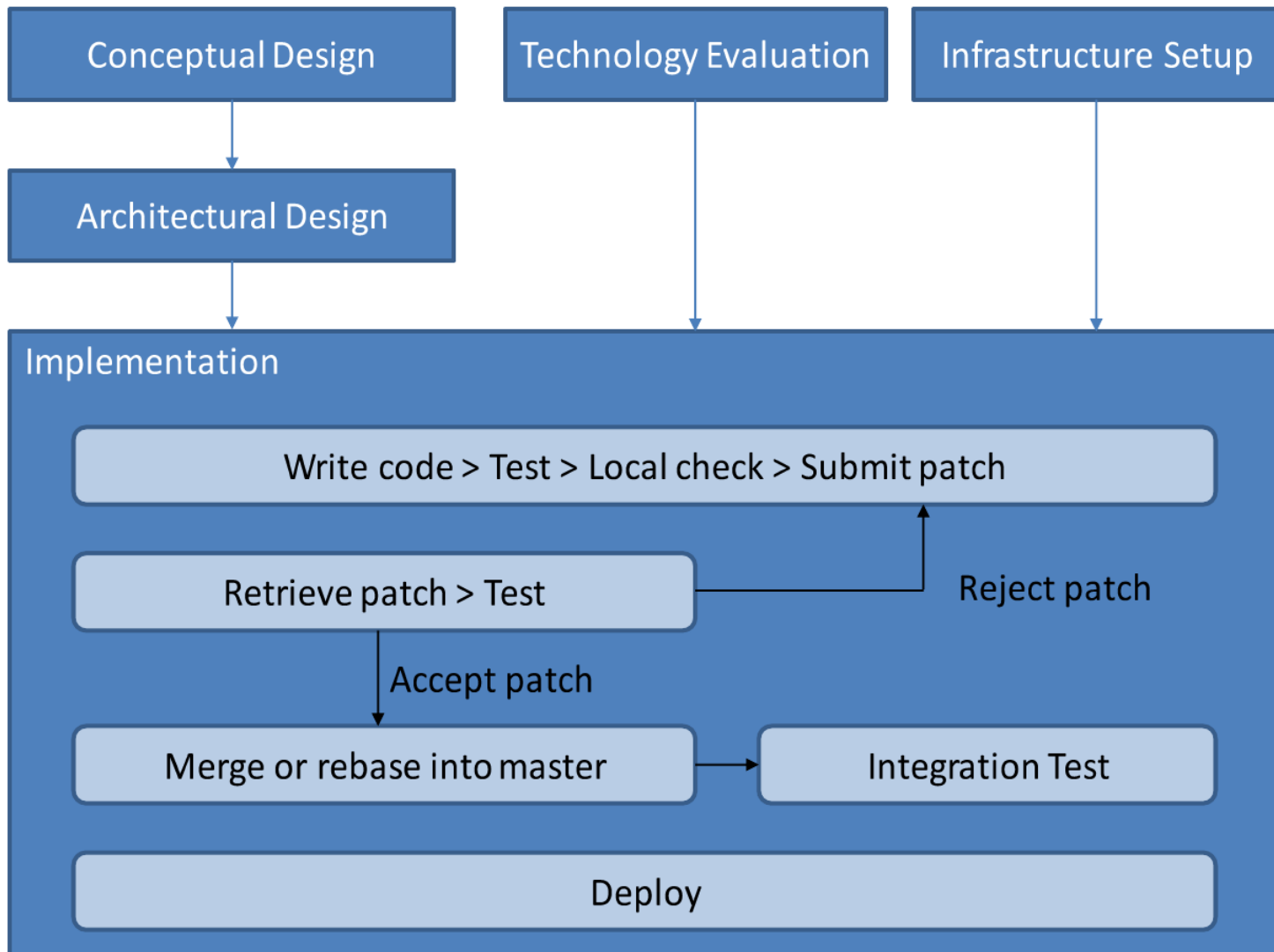
```
for rule in rules:
    print rule
```

```
return SUCCESS
```

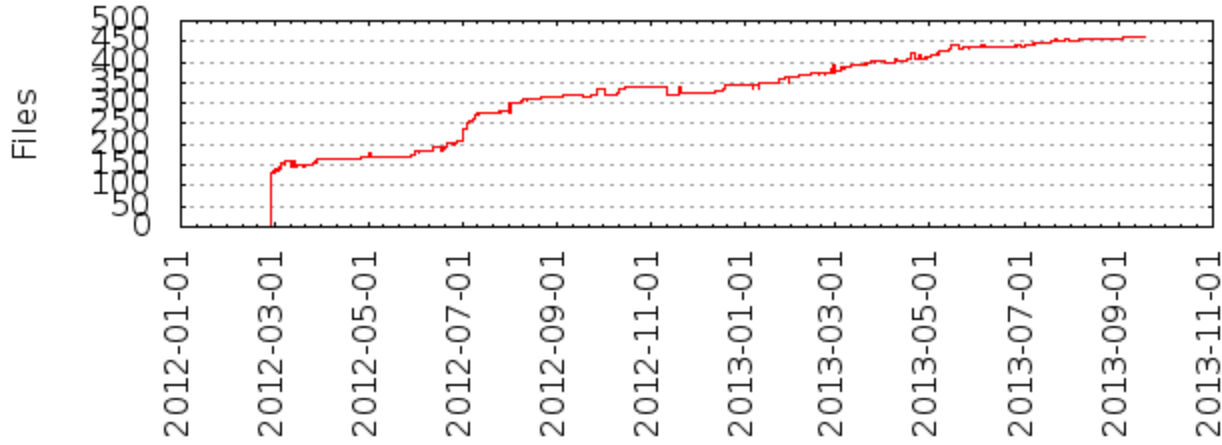
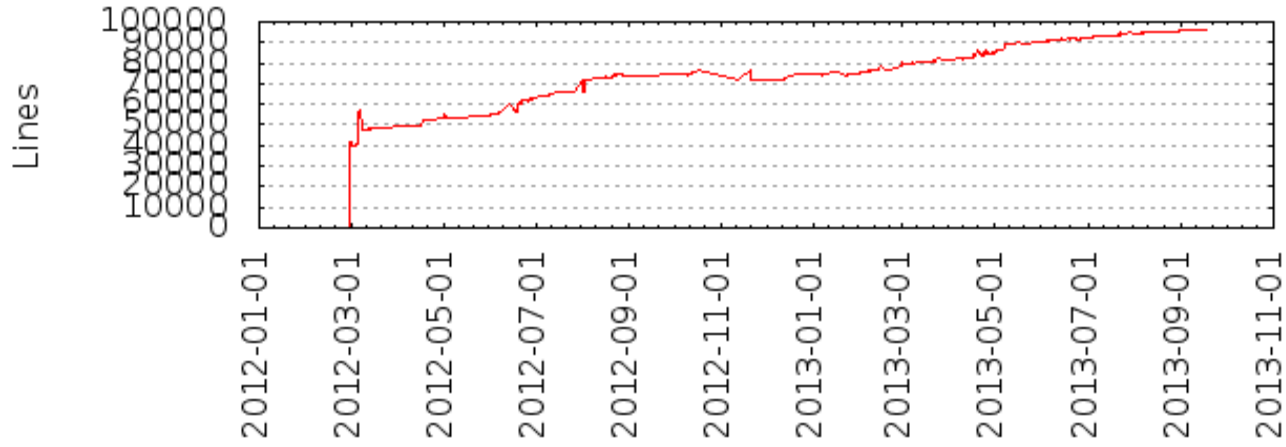
816  
817  
818  
819  
820  
821  
822

# The full process

18

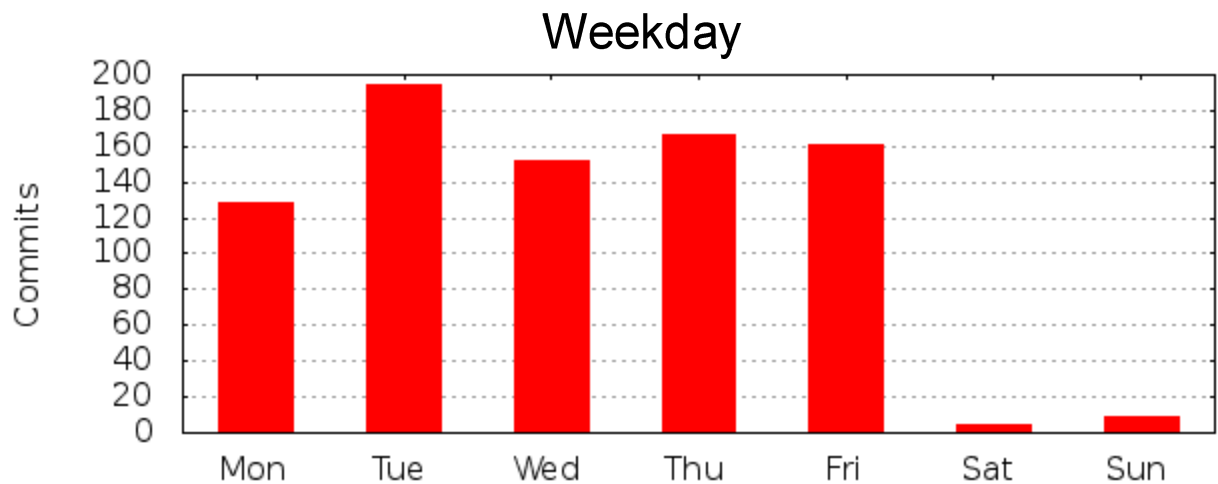
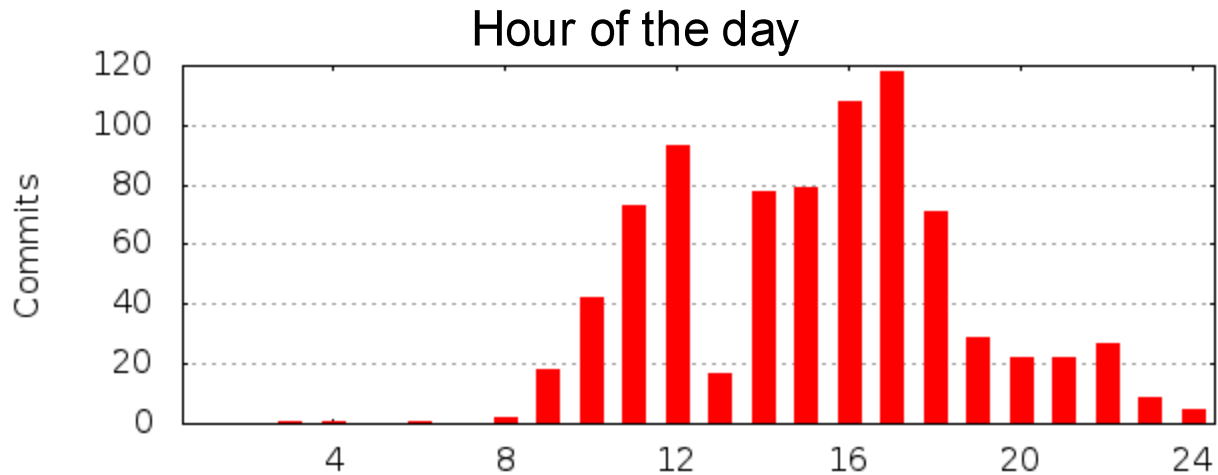


# Some development statistics



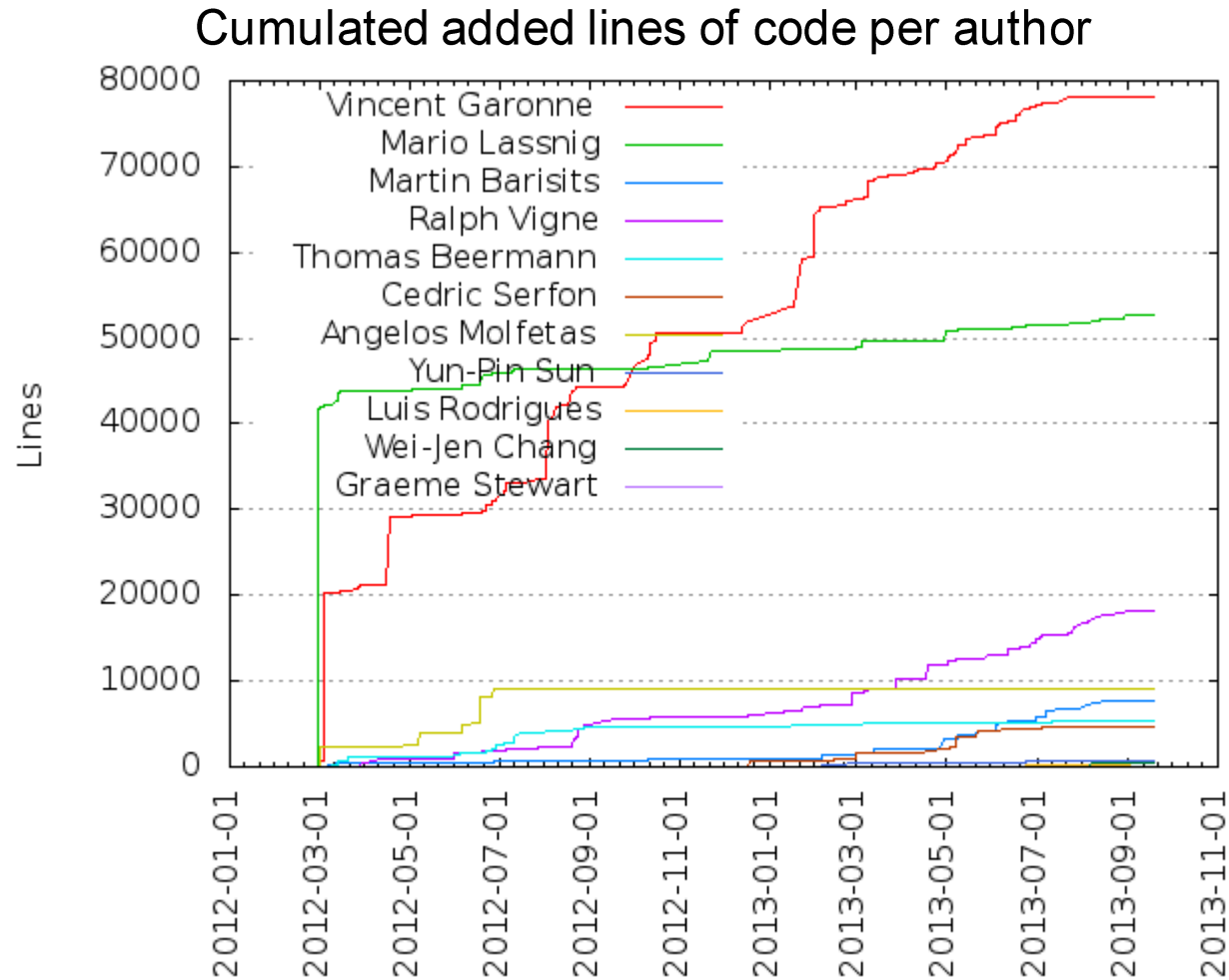
# Some development statistics

20



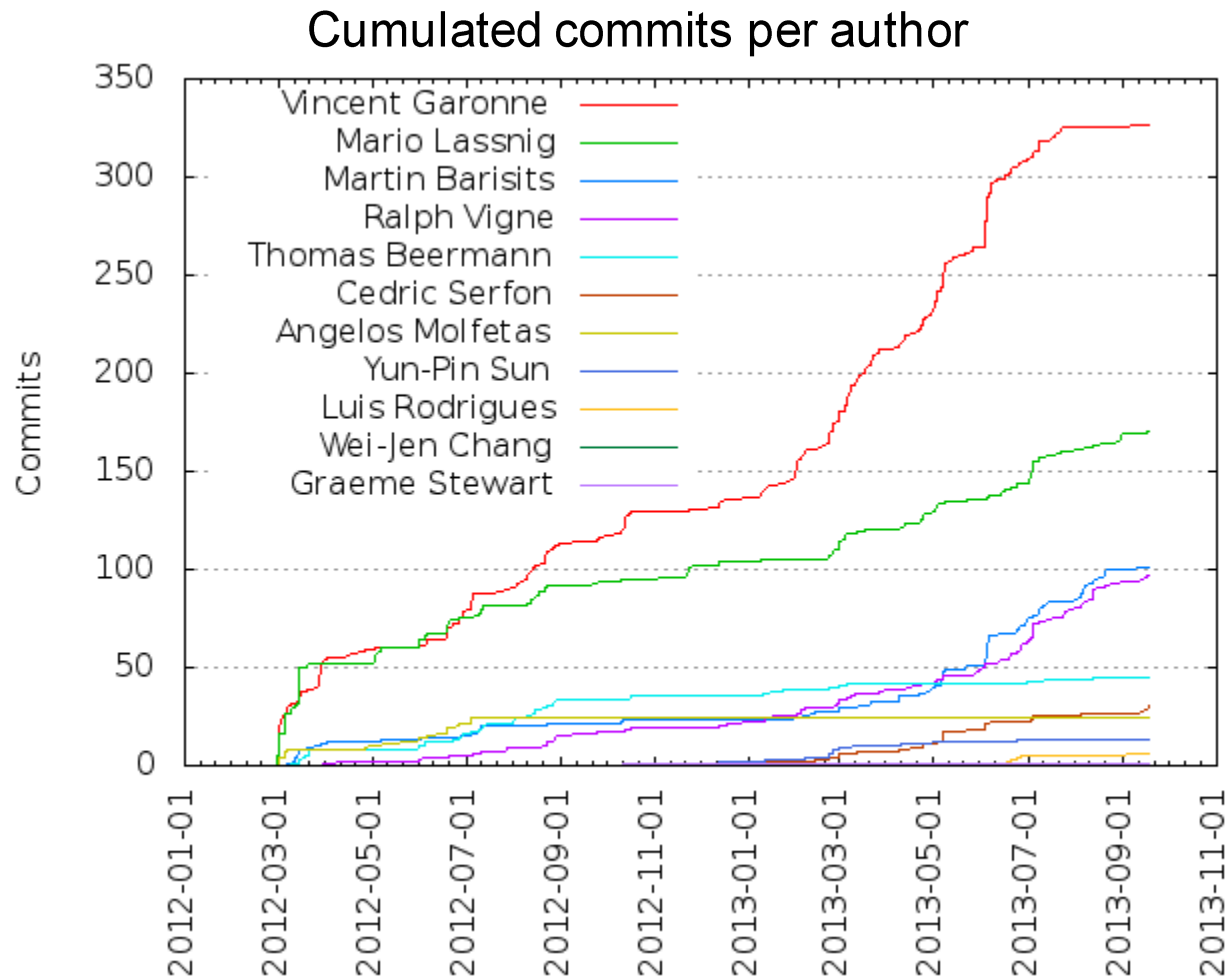
# Some development statistics

21



# Some development statistics

22



# What others think

23

- <http://ohloh.net/>
  - Open source software directory
    - Including Firefox, Apache HTTP, Linux, ...
  - Automatically collects code and analyses projects

In a Nutshell, Rucio...

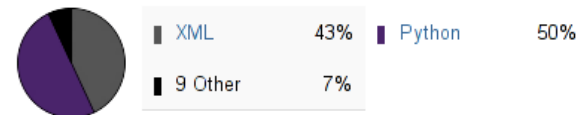
... has had 835 commits made by 14 contributors representing 50,906 lines of code

... is mostly written in Python with an average number of source code comments

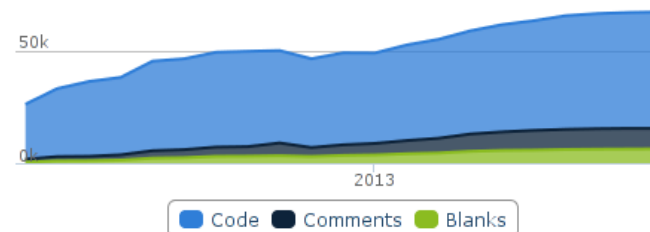
... has a young, but established codebase maintained by a large development team with stable Y-O-Y commits

... took an estimated 12 years of effort (COCOMO model) starting with its first commit in February, 2012 ending with its most recent commit 6 days ago

Languages



Lines of Code



# The human factor 1/2

24

- Won't this take a long time to get my changes into the source code?
  - ▣ Only ramp-up in the beginning
- What happens if I need to quickly fix something?
  - ▣ Self-approval of patches (public embarrassment)
- I really don't like that someone else is judging my work.
  - ▣ Tough luck.



# The human factor 2/2

25

- Writing test cases is bothersome.
  - ▣ If you don't have time now, then you will not have time later anyway.
- We will never agree on a common coding style.
  - ▣ We did and software now enforces it.
- Branching? Merging? What is this?
  - ▣ Git tutorials
- How do we deal with new members in the team?
  - ▣ Up to date documentation
- How can we deal with outside contributions?
  - ▣ Submit patches to gerrit

# Conclusion

26

- Rucio is the new data management system of ATLAS
- Its development follows a modified waterfall process
  - ▣ From conceptual design
  - ▣ Via architectural design
  - ▣ To test-driven and peer-reviewed implementation
- Initial inhibition threshold well overcome
  - ▣ Social uncertainties almost negligible
  - ▣ Conduct enforced where possible by software
- Resulted in
  - ▣ High throughput of essentially error free code
  - ▣ Easy injection of new engineers into the team

# THE ATLAS DATA MANAGEMENT SOFTWARE ENGINEERING PROCESS

Mario Lassnig (CERN PH-ADP)  
on behalf of the ATLAS collaboration

[mario.lassnig@cern.ch](mailto:mario.lassnig@cern.ch)  
[ph-adp-ddm-lab@cern.ch](mailto:ph-adp-ddm-lab@cern.ch)

CHEP, 2013-10-17, Amsterdam, NL