Recovering the OpenOffice.org Code History
Why Code History?

- first-hand reference on how code evolved
- when the developer knew most about it
- detailed references to external resources
- developers have limited memories
- original developers leave projects
OpenOffice Repositories History

- 1988-2000 Proprietary
- 2000-2003 CVS trunk-only
- 2003-2009 CVS with branches
- 2008-2009 Subversion
- 2009-2011 Mercurial
- 2011-2014 Subversion
- 2014-20XX Git (read-only now)
OpenOffice VCS Transition Losses

- 1988-2000 All history lost
- 2000-2003 CVS trunk preserved
- 2003-2009 CVS branches lost
- 2008-2009 SVN branches lost
- 2009-2011 HG mostly preserved
- 2011-2014 SVN still available
- 2014-20XX GIT look great
The Lost Heritage

- OOo repository changes dropped branches
- From 2003-2011 all development work was done on branches
  ➔ about 5000 CVS branches lost
  ➔ about 1000 SVN branches lost
  ➔ the Mercurial branches are not easily available
Why worry about lost branches?

Branch before merge

History-Preserving Merge
A small excursion

Branch before merge

History-Preserving Rebased Merge
OOo-Style Merging

Branch before merge

Branch-Crushing OOo-Style Merge
What was lost?

Commits from branches were squashed:

- most commit messages were lost
- file-level change relationships was lost
- commit message ↔ changeset was lost
- authorship was lost / re-attributed
Chances to get the history back

- The CVS sub-repositories once were available as one rsync'able tarball
- the OOo SVN repository was available via svnsync
- the HG repositories were available unless they were integrated
Making them Usable

OOo-CVS Tarball

- dba
- framework
- graphics

CVS

- converted to SVN

OOo-SVN

- converted to GIT

OOo-HG

- converted to GIT

All-OOo GIT with grafts and a unified mailmap

HG A

HG B

HG C
Problems of the CVS-History

• squashed branch-accumulated commits

• codebase only partially tagged
  ➔ branches have many missing files
  ➔ the conversion has to introduce “glue” commits

• many partial merges (for each file)
  – no proper merge commits
History Losing Partial Merges

Branch before merge

History-Crushing File-Based Merge

Apache OpenOffice™
Problems of the CVS-History

• “resyncs” messed up branch histories
• originated from multiple CVS-Repos
  e.g. framework, graphics, gsl, ...
• some branch names were deleted
  ➔ there are “unnamed branches”
Problems of the SVN-History

- squashed accumulation of commits
- no proper merge commits
- “resyncs” messed up branch histories
- Most SVN branches are not yet connected to their CVS counterparts
Minor Problems of the HG-History

• many wrong author names
  ➔ can be solved with mail-mapping

• HG-Commit-Hashes were lost
  ➔ can be solved by a re-import
The Repository Histories

CVS

SVN

HG
The HistOOory in GIT

- all former repositories were converted to GIT
- they have been merged into one archive
  (at http://people.apache.org/~hdu/HistOOory.zip)
  - all the code history is compressed into 2GB
  - it contains all branches, commits and files
  - except binary artifacts like GIFs, Templates, Fonts
What can be done with it?

- All former repositories are preserved
- All non-empty branches are preserved
- All commits can be researched individually
- Historical sources can be recreated
- Bad merging means “blame” doesn't work
Questions?