PARIS2011 CONFERENCE

12th-15th October

Understanding the source code via analysis of similarity

- Takeshi Abe

  tabe@fixedpoint.jp

Introduction

- Contributing to LibreOffice since Oct 2010
  - As a developer / TDF member
  - As a member of LibreOffice Japanese Team ←New!

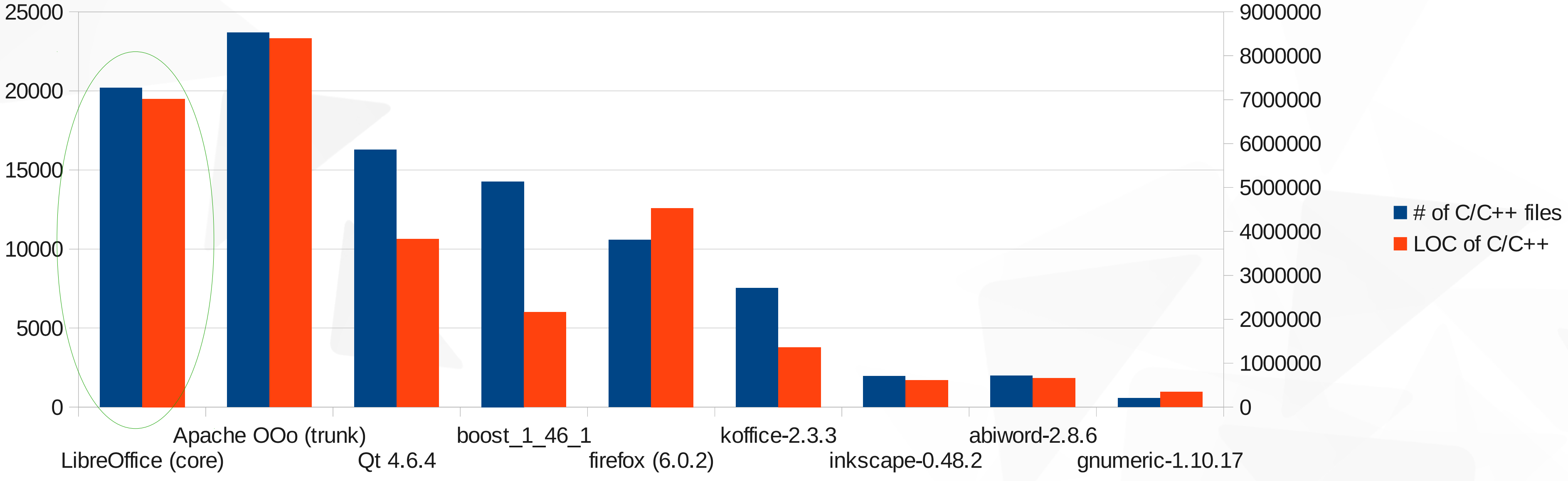
- A little bit of prehistory
  - Using OOo since 2.x
  - Not fully satisfied with it, but ...

Introduction: prehistory

Derived from http://9gag.com/gag/18811

Newbie wonders with the large code base

- Happy with EasyHack™ to start hacking
  - Code cleanup
  - Fix small bugs e.g. segv
- But wait, awful, how large is the C++ source code?



… any effective way to accustom them?

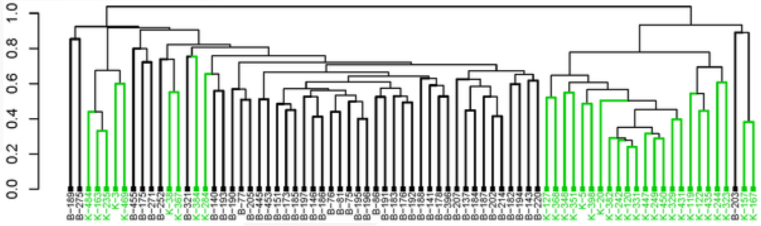
We have a bunch of cute tools

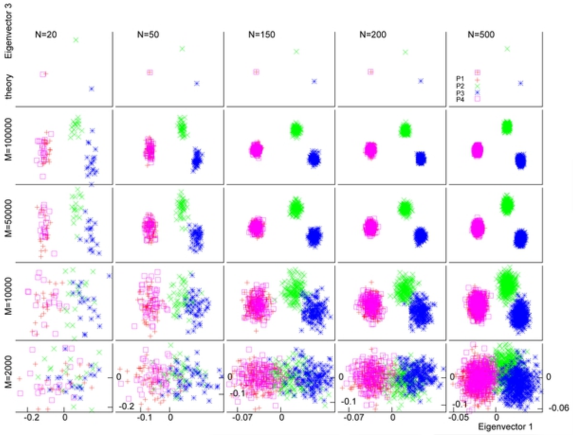
- Existing development tools helps us dealing with the large code base:
  - git
  - g++
  - gdb
  - Valgrind
  - Cppckeck
  - callcatcher
  - zzuf
  - …
- The point is: using them makes it possible to *analyze specific aspects of code without human interfering further*

Cluster analysis is the next one

Cluster analysis:

- Recognizing patterns without prior knowledge of supervisor, but only with similarity metrics





Chae M, Chen JJ, 2011 Reordering Hierarchical Tree Based on Bilateral Symmetric Distance. PLoS ONE 6(8): e22546. doi:10.1371/journal.pone.0022546

Ma J, Amos CI, 2010 Theoretical Formulation of Principal Components Analysis to Detect and Correct for Population Stratification. PLoS ONE 5(9): e12510. doi:10.1371/journal.pone.0012510

- Employing a similarity(distance) between compile units
  - The idea itself is rather old, and some of software exist
  - But mainly for code duplication detection only
  - We can expect more

What kind of similarity?

- You may define suitable metrics between C++ source files, but it would tend to be slow or hard-to-debug.
  - “*Compare tokens after parsing files*” sounds attractive (and popular in the research area), but impractical for our use
  - “*Compare files as is*” is a way to go
    - Many methods for string comparison available

Photo by quatre mains (http://www.flickr.com/photos/titrans/)

Hamming distance?

Count indices at which corresponding chars are different

- Can compute it fast with a small memory
- But it is weak against insertion of irrelevant code:

```
struct ScLookupCacheMapImpl
{
    ScLookupCacheMap aCacheMap;
    ~ScLookupCacheMapImpl()
    {
        freeCaches();
    }
    void clear()
    {
        freeCaches();
        // Zap map.
        ScLookupCacheMap aTmp;
        aCacheMap.swap( aTmp);
    }
private:
    void freeCaches()
    {
        for (ScLookupCacheMap::iterator
it( aCacheMap.begin()); it != aCacheMap.end();
++it)
            delete (*it).second;
    }
};
```

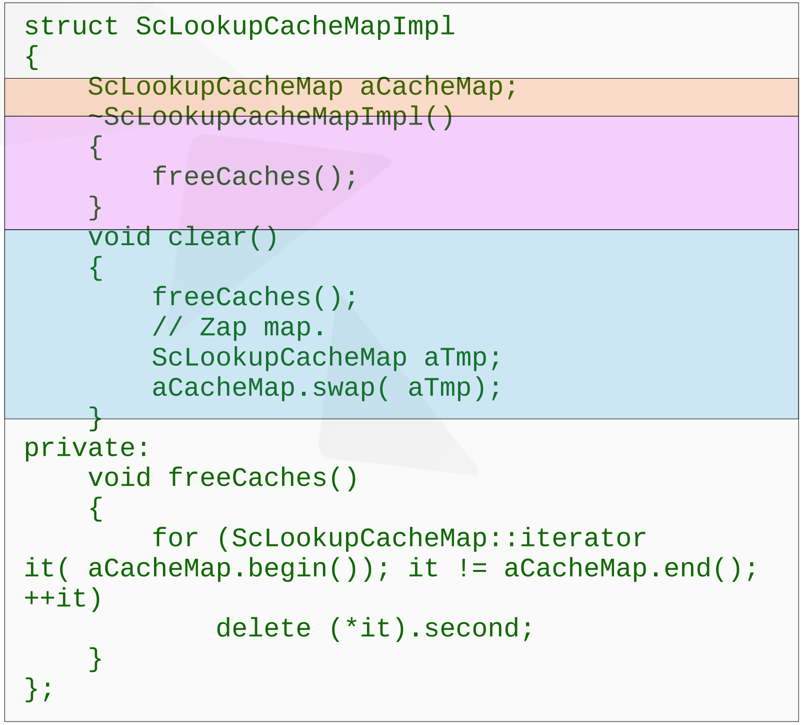
One empty line

```
struct ScLookupCacheMapImpl
{
    ScLookupCacheMap aCacheMap;
    ~ScLookupCacheMapImpl()
    {
        freeCaches();
    }

    void clear()
    {
        freeCaches();
        // Zap map.
        ScLookupCacheMap aTmp;
        aCacheMap.swap( aTmp);
    }
private:
    void freeCaches()
    {
        for (ScLookupCacheMap::iterator
it( aCacheMap.begin()); it != aCacheMap.end();
++it)
            delete (*it).second;
    }
};
```

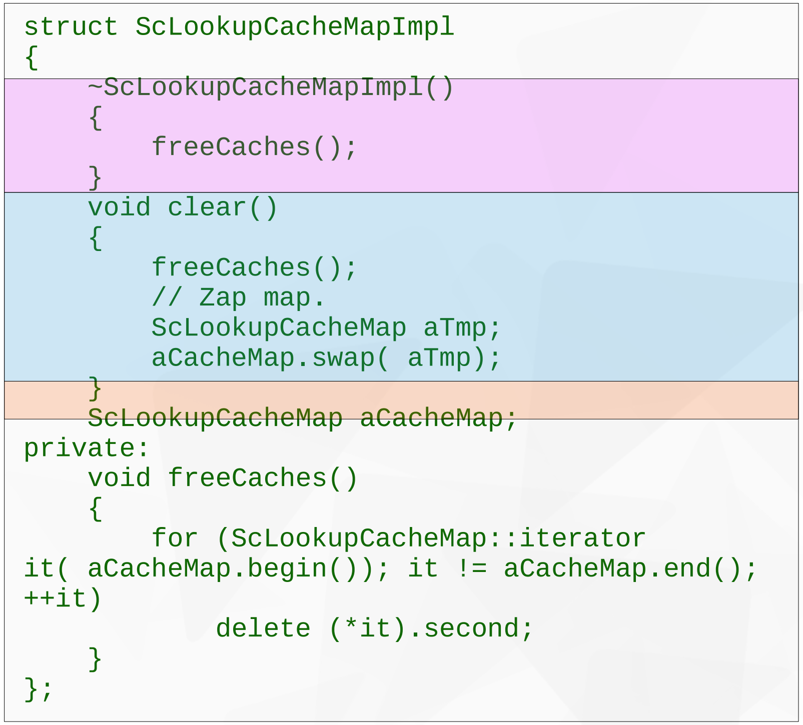
Edit distance (Levenshtein distance)?

Count insertion/deletion/replacement of chars

- Robust, especially strong against insertion of irrelevant code
- Known clever algorithm like Myers' O(dm) in time
  - but still infeasible in our case
- Weak against transposition of chunks of code:



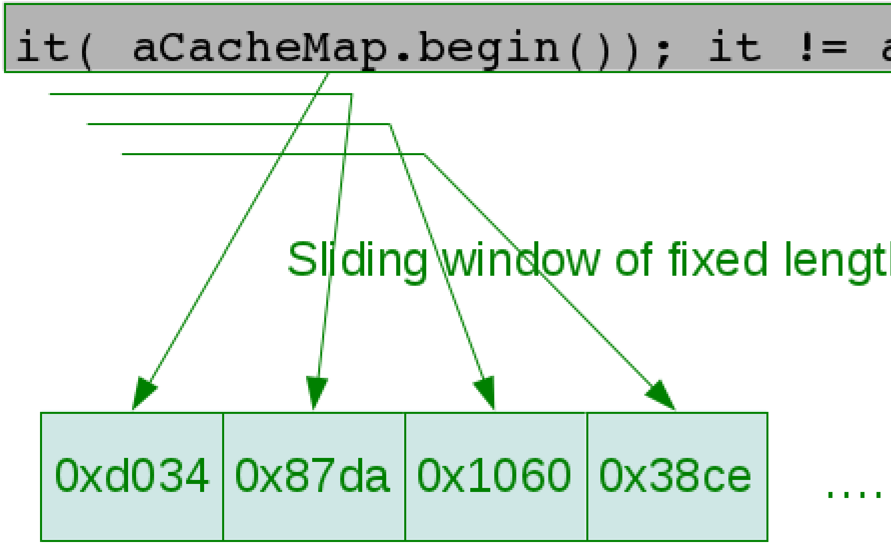
```
struct ScLookupCacheMapImpl
{
    ScLookupCacheMap aCacheMap;
    ~ScLookupCacheMapImpl()
    {
        freeCaches();
    }
    void clear()
    {
        freeCaches();
        // Zap map.
        ScLookupCacheMap aTmp;
        aCacheMap.swap( aTmp);
    }
private:
    void freeCaches()
    {
        for (ScLookupCacheMap::iterator
it( aCacheMap.begin()); it != aCacheMap.end();
++it)
            delete (*it).second;
    }
};
```



```
struct ScLookupCacheMapImpl
{
    ~ScLookupCacheMapImpl()
    {
        freeCaches();
    }
    void clear()
    {
        freeCaches();
        // Zap map.
        ScLookupCacheMap aTmp;
        aCacheMap.swap( aTmp);
    }
    ScLookupCacheMap aCacheMap;
private:
    void freeCaches()
    {
        for (ScLookupCacheMap::iterator
it( aCacheMap.begin()); it != aCacheMap.end();
++it)
            delete (*it).second;
    }
};
```

A solution: hash-based metrics

- “Mapping a source file to a set of hash values of its substrings of fixed length” seems a good compromise
  - Hashing may cause false positives of match
  - Rolling hash works effectively

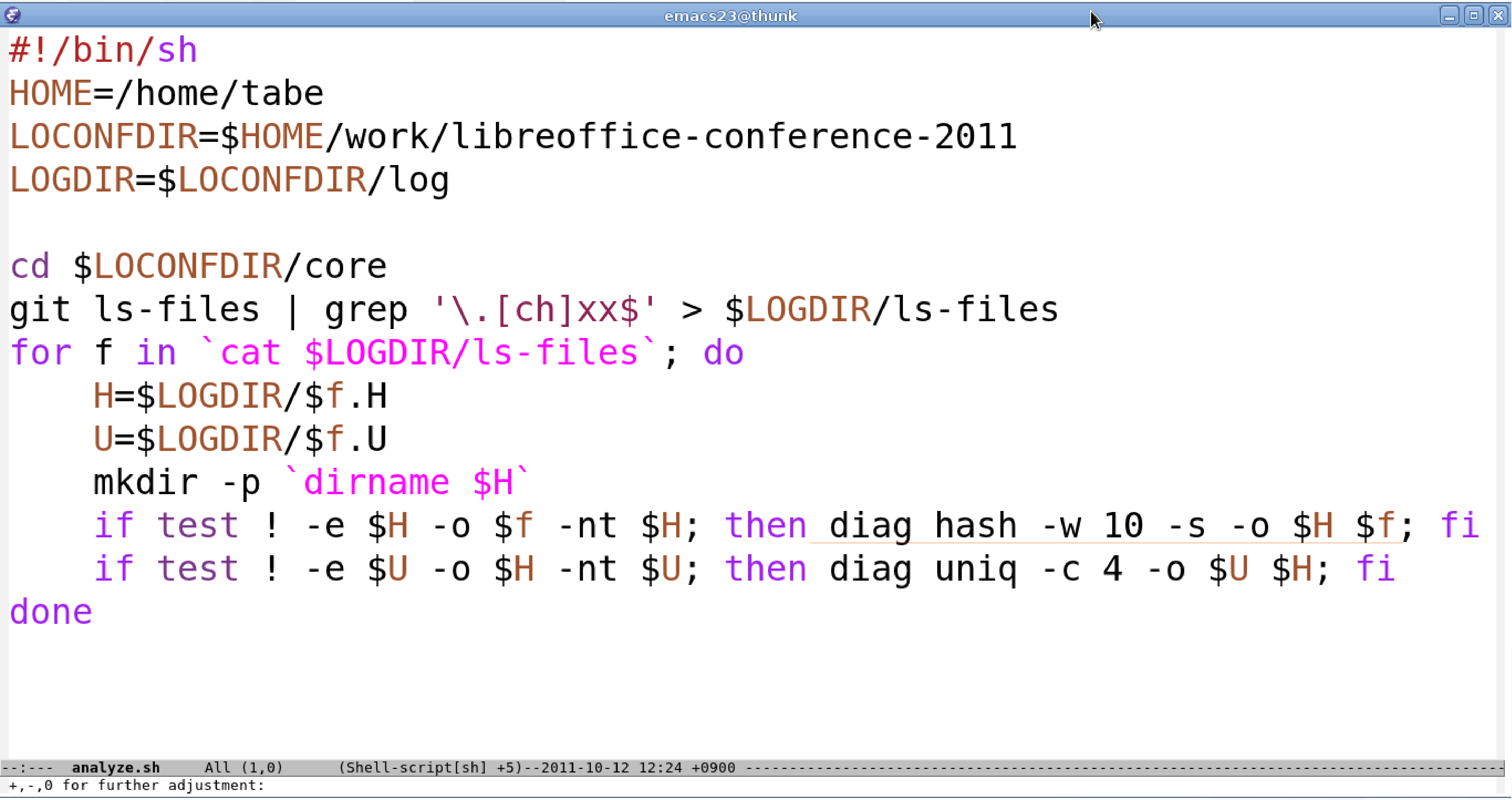


```
it( aCacheMap.begin()); it != aCacheMap.end(); ++it)
```

Diagonal for clustering

- Diagonal: http://diagonal.sourceforge.net/
  - Providing CLI utilities
  - implemented in C99 + POSIX
  - New BSD License
  - still unstable

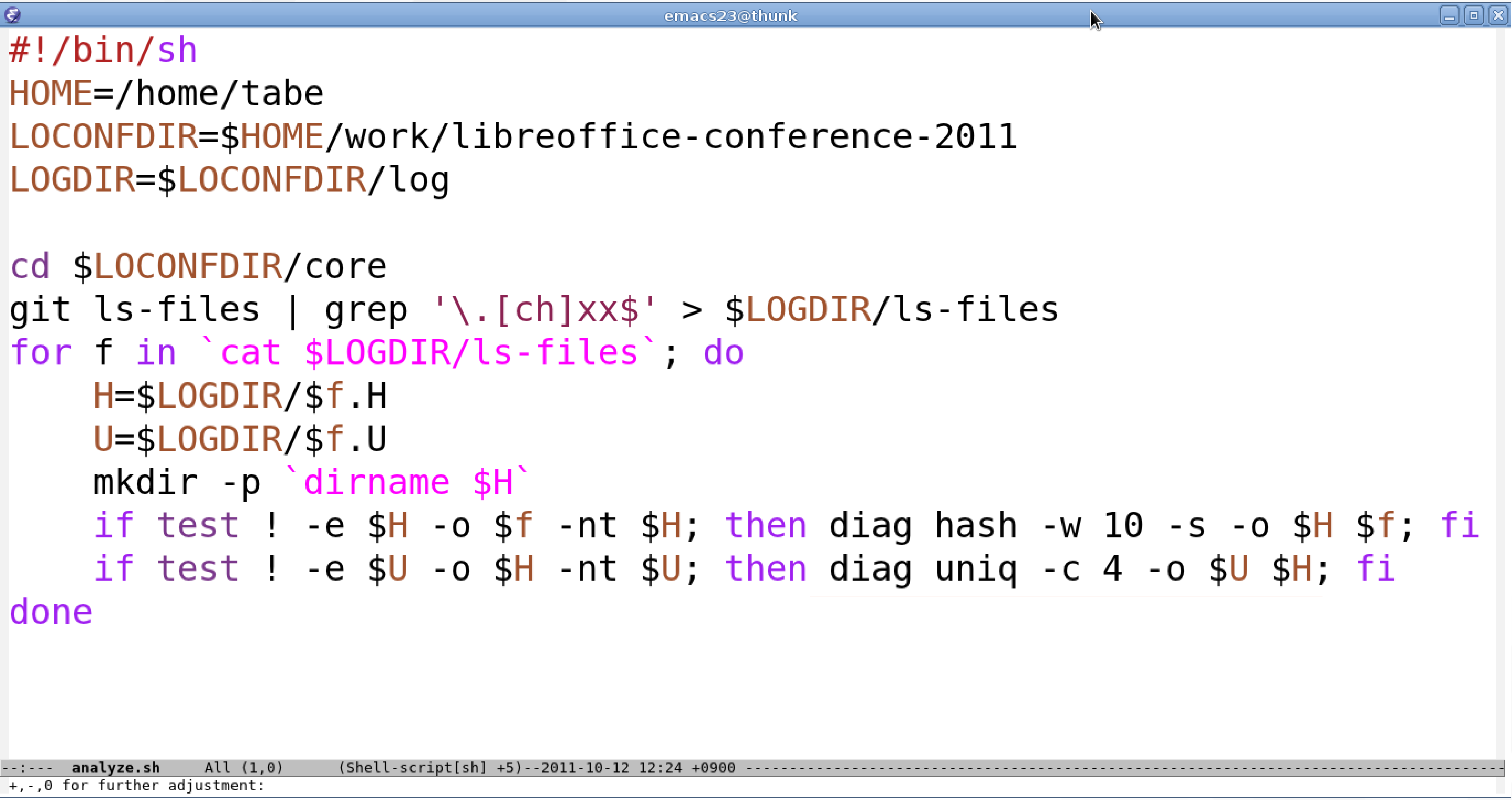
Preliminary setup: generate sorted hash files



```
#!/bin/sh
HOME=/home/tabe
LOCONFDIR=$HOME/work/libreoffice-conference-2011
LOGDIR=$LOCONFDIR/log

cd $LOCONFDIR/core
git ls-files | grep '\.[ch]xx$' > $LOGDIR/ls-files
for f in `cat $LOGDIR/ls-files`; do
    H=$LOGDIR/$f.H
    U=$LOGDIR/$f.U
    mkdir -p `dirname $H`
    if test ! -e $H -o $f -nt $H; then diag hash -w 10 -s -o $H $f; fi
    if test ! -e $U -o $H -nt $U; then diag uniq -c 4 -o $U $H; fi
done
```

```
--:---  analyze.sh      All (1,0)       (Shell-script[sh] +5)--2011-10-12 12:24 +0900 ----------------------------------------------------------------
+,-,0 for further adjustment:
```

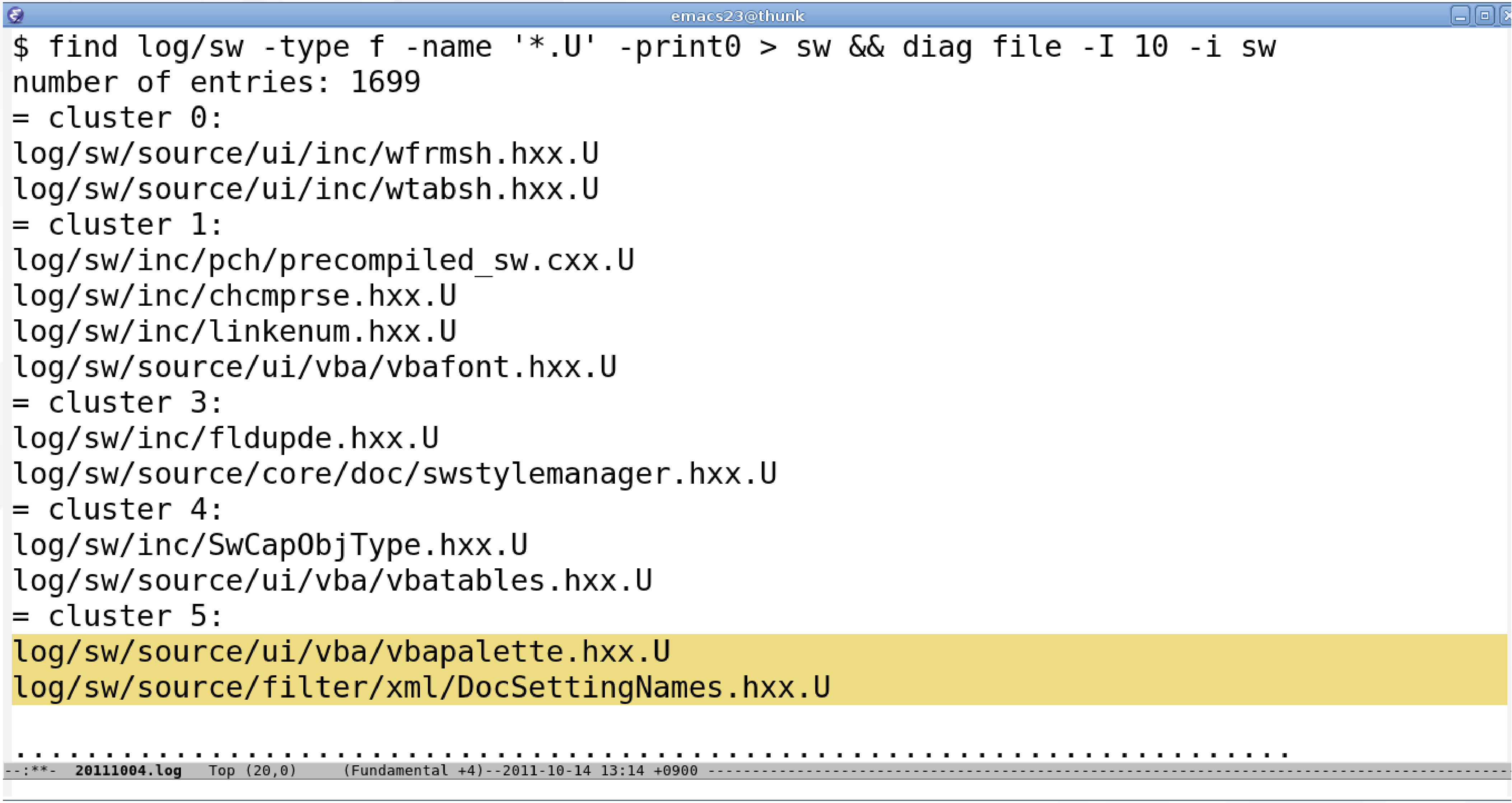
Preliminary setup: generate sorted hash files



```sh
#!/bin/sh
HOME=/home/tabe
LOCONFDIR=$HOME/work/libreoffice-conference-2011
LOGDIR=$LOCONFDIR/log

cd $LOCONFDIR/core
git ls-files | grep '\.[ch]xx$' > $LOGDIR/ls-files
for f in `cat $LOGDIR/ls-files`; do
    H=$LOGDIR/$f.H
    U=$LOGDIR/$f.U
    mkdir -p `dirname $H`
    if test ! -e $H -o $f -nt $H; then diag hash -w 10 -s -o $H $f; fi
    if test ! -e $U -o $H -nt $U; then diag uniq -c 4 -o $U $H; fi
done
```

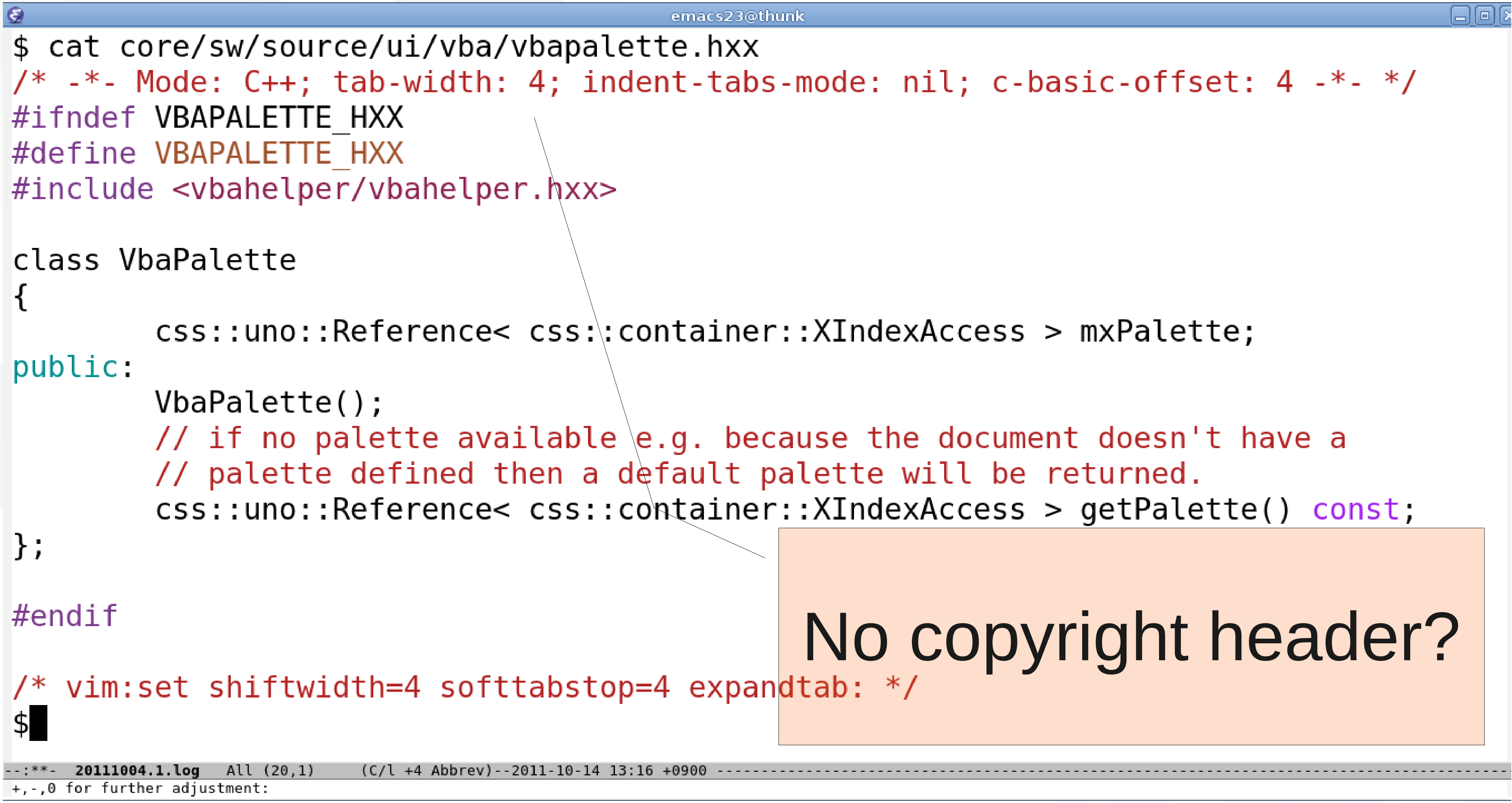
```
--:---  analyze.sh      All (1,0)      (Shell-script[sh] +5)--2011-10-12 12:24 +0900 ------------------------------------------------------------
+,-,0 for further adjustment:
```

Use case 1: 2011-10-04 sw



```
$ find log/sw -type f -name '*.U' -print0 > sw && diag file -I 10 -i sw
number of entries: 1699
= cluster 0:
log/sw/source/ui/inc/wfrmsh.hxx.U
log/sw/source/ui/inc/wtabsh.hxx.U
= cluster 1:
log/sw/inc/pch/precompiled_sw.cxx.U
log/sw/inc/chcmprse.hxx.U
log/sw/inc/linkenum.hxx.U
log/sw/source/ui/vba/vbafont.hxx.U
= cluster 3:
log/sw/inc/fldupde.hxx.U
log/sw/source/core/doc/swstylemanager.hxx.U
= cluster 4:
log/sw/inc/SwCapObjType.hxx.U
log/sw/source/ui/vba/vbatables.hxx.U
= cluster 5:
log/sw/source/ui/vba/vbapalette.hxx.U
log/sw/source/filter/xml/DocSettingNames.hxx.U

.........................................................................
--:**-  20111004.log    Top (20,0)     (Fundamental +4)--2011-10-14 13:14 +0900 ---------------
```

Use case 1: 2011-10-04 sw

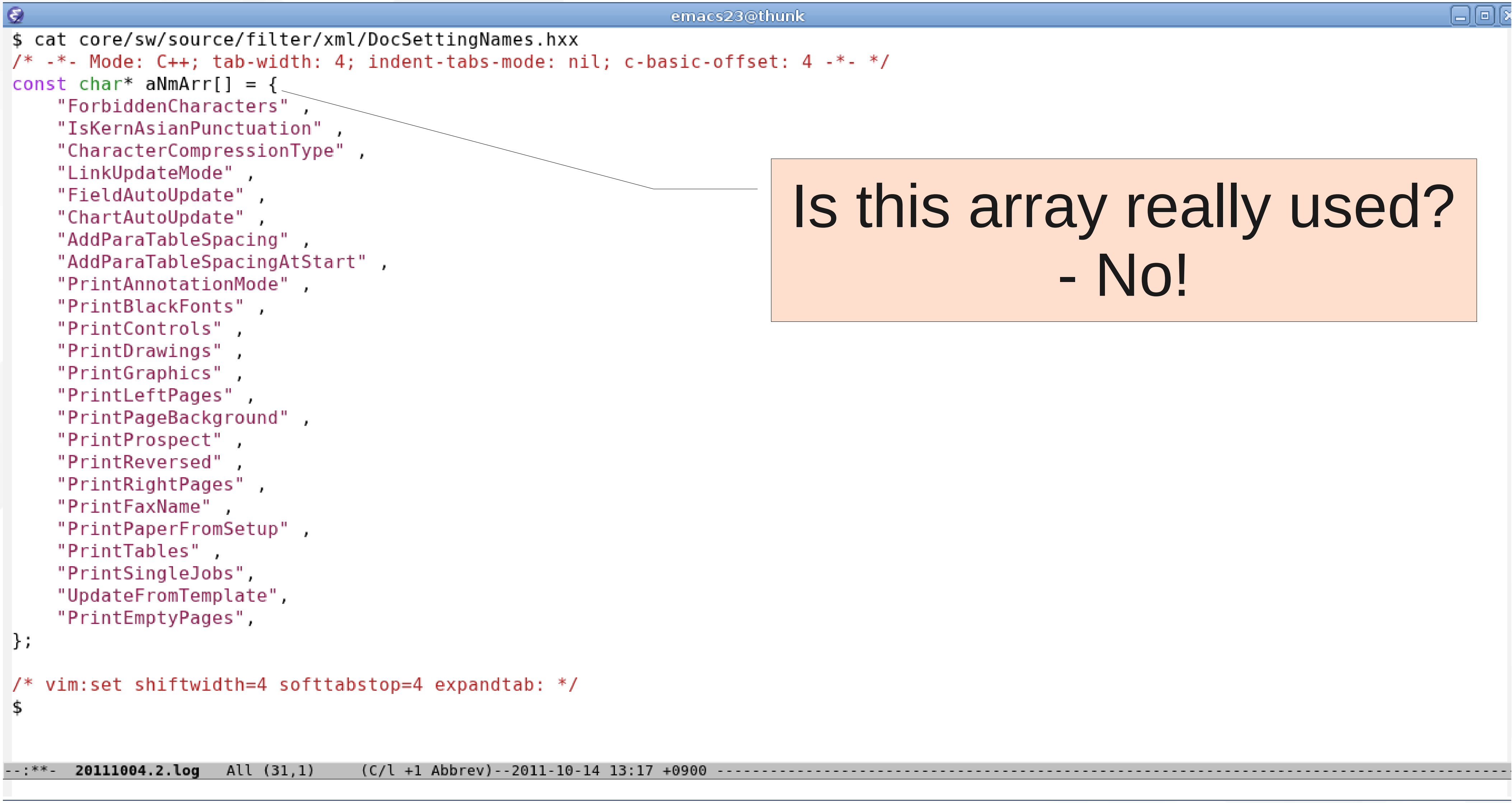


```
$ cat core/sw/source/ui/vba/vbapalette.hxx
/* -*- Mode: C++; tab-width: 4; indent-tabs-mode: nil; c-basic-offset: 4 -*- */
#ifndef VBAPALETTE_HXX
#define VBAPALETTE_HXX
#include <vbahelper/vbahelper.hxx>

class VbaPalette
{
        css::uno::Reference< css::container::XIndexAccess > mxPalette;
public:
        VbaPalette();
        // if no palette available e.g. because the document doesn't have a
        // palette defined then a default palette will be returned.
        css::uno::Reference< css::container::XIndexAccess > getPalette() const;
};

#endif

/* vim:set shiftwidth=4 softtabstop=4 expandtab: */
$
```

```
--:**-  20111004.1.log    All (20,1)     (C/l +4 Abbrev)--2011-10-14 13:16 +0900 ----------------------------------------
+,-,0 for further adjustment:
```

No copyright header?

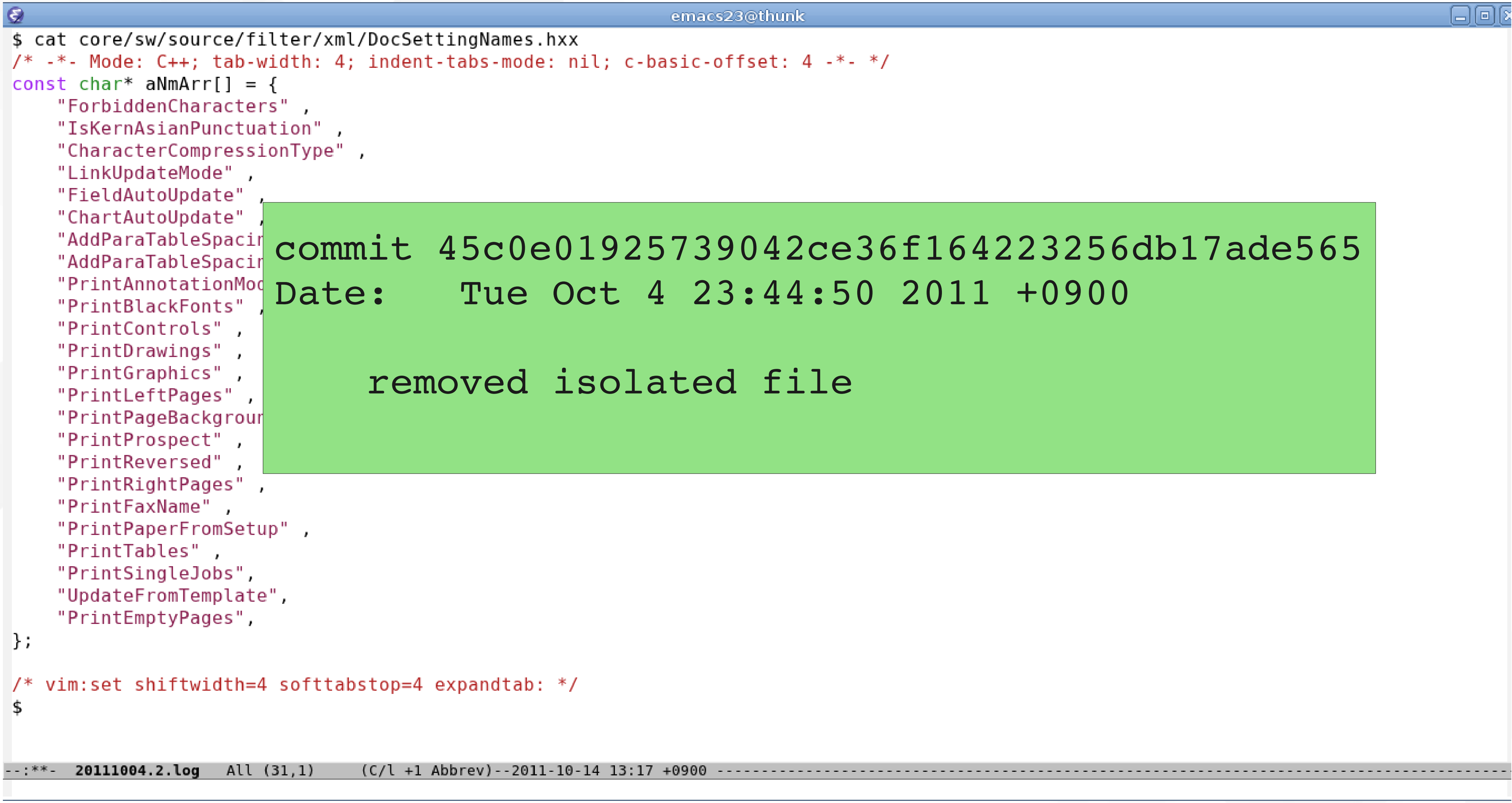
Use case 1: 2011-10-04 sw



```
$ cat core/sw/source/filter/xml/DocSettingNames.hxx
/* -*- Mode: C++; tab-width: 4; indent-tabs-mode: nil; c-basic-offset: 4 -*- */
const char* aNmArr[] = {
    "ForbiddenCharacters" ,
    "IsKernAsianPunctuation" ,
    "CharacterCompressionType" ,
    "LinkUpdateMode" ,
    "FieldAutoUpdate" ,
    "ChartAutoUpdate" ,
    "AddParaTableSpacing" ,
    "AddParaTableSpacingAtStart" ,
    "PrintAnnotationMode" ,
    "PrintBlackFonts" ,
    "PrintControls" ,
    "PrintDrawings" ,
    "PrintGraphics" ,
    "PrintLeftPages" ,
    "PrintPageBackground" ,
    "PrintProspect" ,
    "PrintReversed" ,
    "PrintRightPages" ,
    "PrintFaxName" ,
    "PrintPaperFromSetup" ,
    "PrintTables" ,
    "PrintSingleJobs",
    "UpdateFromTemplate",
    "PrintEmptyPages",
};

/* vim:set shiftwidth=4 softtabstop=4 expandtab: */
$

--:**-  20111004.2.log    All (31,1)     (C/l +1 Abbrev)--2011-10-14 13:17 +0900 ----------------------------------------
```

Is this array really used?
- No!

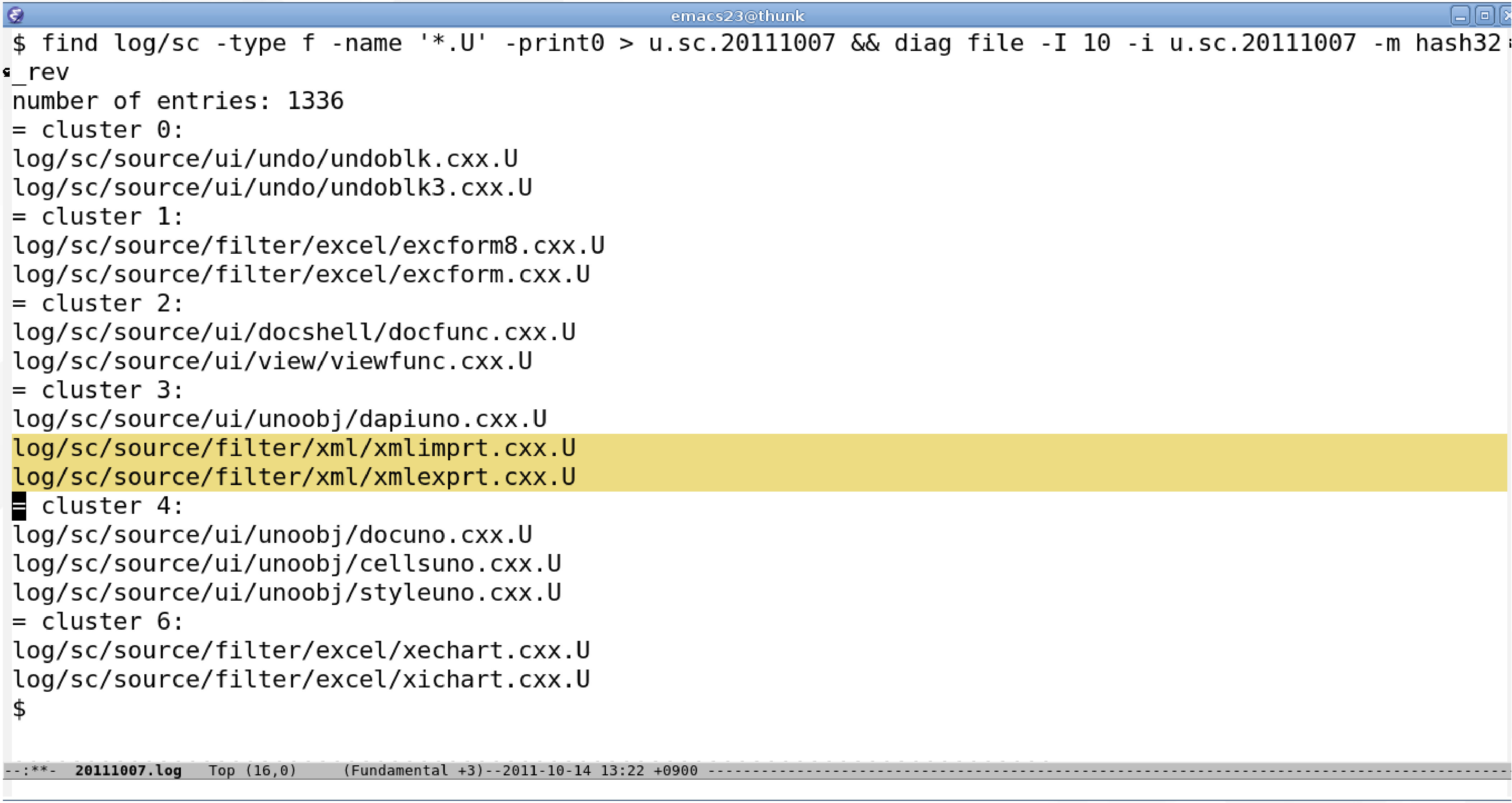
Use case 1: 2011-10-04 sw



```
$ cat core/sw/source/filter/xml/DocSettingNames.hxx
/* -*- Mode: C++; tab-width: 4; indent-tabs-mode: nil; c-basic-offset: 4 -*- */
const char* aNmArr[] = {
    "ForbiddenCharacters" ,
    "IsKernAsianPunctuation" ,
    "CharacterCompressionType" ,
    "LinkUpdateMode" ,
    "FieldAutoUpdate" ,
    "ChartAutoUpdate" ,
    "AddParaTableSpaci
    "AddParaTableSpaci
    "PrintAnnotationMo
    "PrintBlackFonts" ,
    "PrintControls" ,
    "PrintDrawings" ,
    "PrintGraphics" ,
    "PrintLeftPages" ,
    "PrintPageBackgrou
    "PrintProspect" ,
    "PrintReversed" ,
    "PrintRightPages" ,
    "PrintFaxName" ,
    "PrintPaperFromSetup" ,
    "PrintTables" ,
    "PrintSingleJobs",
    "UpdateFromTemplate",
    "PrintEmptyPages",
};

/* vim:set shiftwidth=4 softtabstop=4 expandtab: */
$

--:**-  20111004.2.log    All (31,1)     (C/l +1 Abbrev)--2011-10-14 13:17 +0900 ----------------
```

```
commit 45c0e01925739042ce36f164223256db17ade565
Date:   Tue Oct 4 23:44:50 2011 +0900

    removed isolated file
```

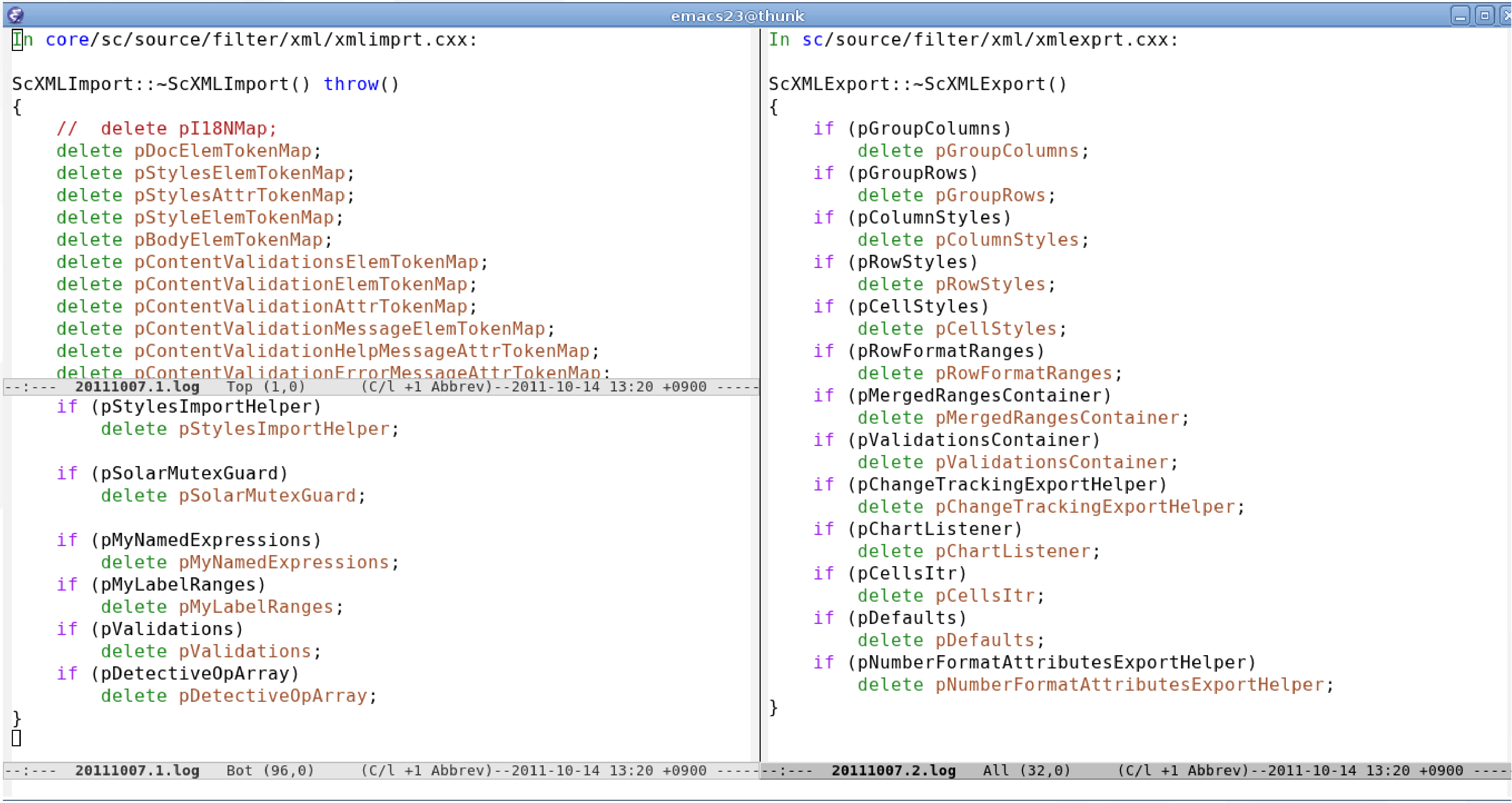
Use case 2: 2011-10-07 sc



```
$ find log/sc -type f -name '*.U' -print0 > u.sc.20111007 && diag file -I 10 -i u.sc.20111007 -m hash32
_rev
number of entries: 1336
= cluster 0:
log/sc/source/ui/undo/undoblk.cxx.U
log/sc/source/ui/undo/undoblk3.cxx.U
= cluster 1:
log/sc/source/filter/excel/excform8.cxx.U
log/sc/source/filter/excel/excform.cxx.U
= cluster 2:
log/sc/source/ui/docshell/docfunc.cxx.U
log/sc/source/ui/view/viewfunc.cxx.U
= cluster 3:
log/sc/source/ui/unoobj/dapiuno.cxx.U
log/sc/source/filter/xml/xmlimprt.cxx.U
log/sc/source/filter/xml/xmlexprt.cxx.U
= cluster 4:
log/sc/source/ui/unoobj/docuno.cxx.U
log/sc/source/ui/unoobj/cellsuno.cxx.U
log/sc/source/ui/unoobj/styleuno.cxx.U
= cluster 6:
log/sc/source/filter/excel/xechart.cxx.U
log/sc/source/filter/excel/xichart.cxx.U
$
```

emacs23@thunk
--:**- 20111007.log   Top (16,0)   (Fundamental +3)--2011-10-14 13:22 +0900 ----

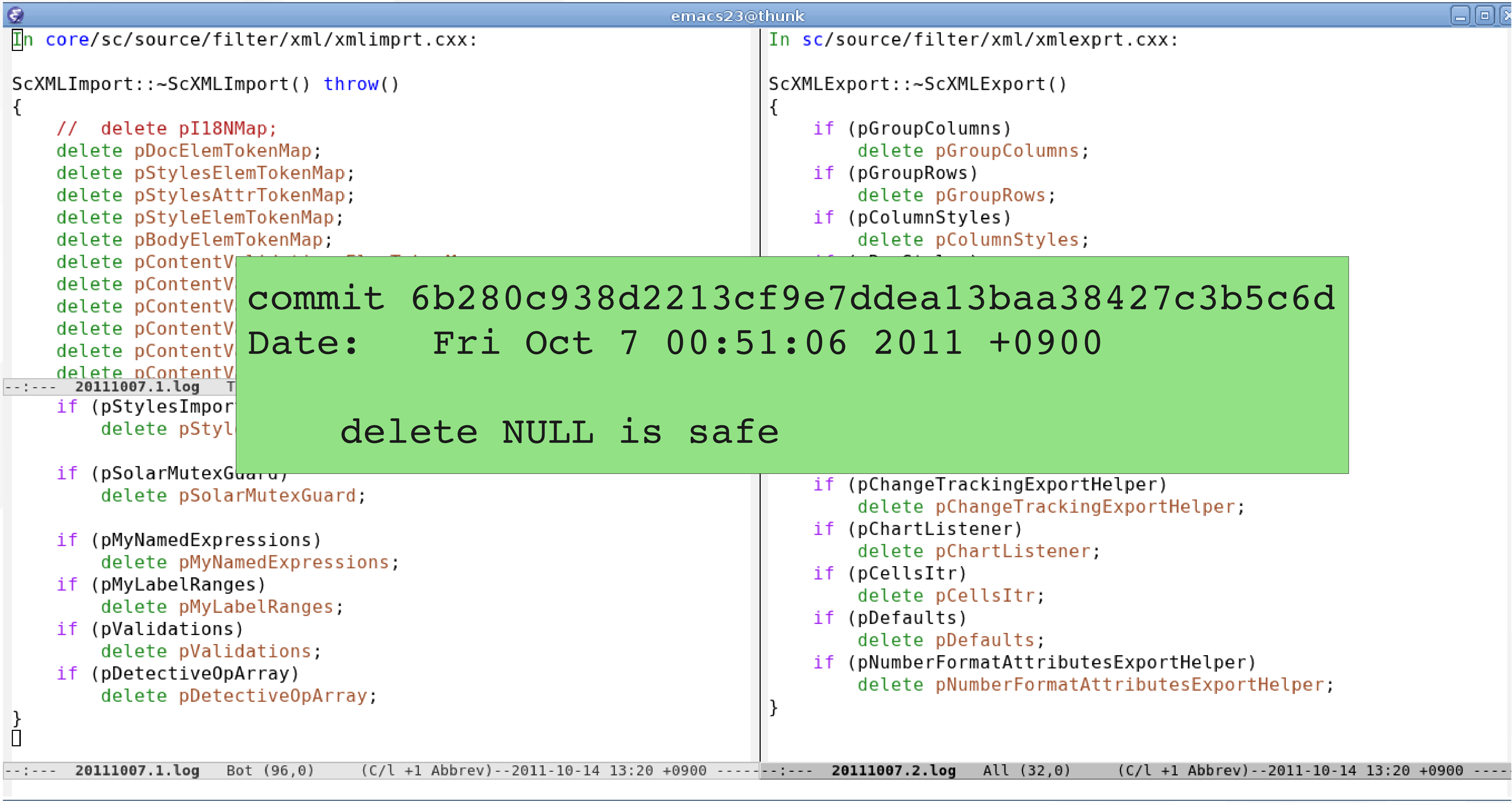
Use case 2: 2011-10-07 sc



```
In core/sc/source/filter/xml/xmlimprt.cxx:

ScXMLImport::~ScXMLImport() throw()
{
    //  delete pI18NMap;
    delete pDocElemTokenMap;
    delete pStylesElemTokenMap;
    delete pStylesAttrTokenMap;
    delete pStyleElemTokenMap;
    delete pBodyElemTokenMap;
    delete pContentValidationsElemTokenMap;
    delete pContentValidationElemTokenMap;
    delete pContentValidationAttrTokenMap;
    delete pContentValidationMessageElemTokenMap;
    delete pContentValidationHelpMessageAttrTokenMap;
    delete pContentValidationErrorMessageAttrTokenMap;
--:---  20111007.1.log    Top (1,0)      (C/l +1 Abbrev)--2011-10-14 13:20 +0900 -----
    if (pStylesImportHelper)
        delete pStylesImportHelper;

    if (pSolarMutexGuard)
        delete pSolarMutexGuard;

    if (pMyNamedExpressions)
        delete pMyNamedExpressions;
    if (pMyLabelRanges)
        delete pMyLabelRanges;
    if (pValidations)
        delete pValidations;
    if (pDetectiveOpArray)
        delete pDetectiveOpArray;
}

--:---  20111007.1.log    Bot (96,0)     (C/l +1 Abbrev)--2011-10-14 13:20 +0900 -----
```

```
In sc/source/filter/xml/xmlexprt.cxx:

ScXMLExport::~ScXMLExport()
{
    if (pGroupColumns)
        delete pGroupColumns;
    if (pGroupRows)
        delete pGroupRows;
    if (pColumnStyles)
        delete pColumnStyles;
    if (pRowStyles)
        delete pRowStyles;
    if (pCellStyles)
        delete pCellStyles;
    if (pRowFormatRanges)
        delete pRowFormatRanges;
    if (pMergedRangesContainer)
        delete pMergedRangesContainer;
    if (pValidationsContainer)
        delete pValidationsContainer;
    if (pChangeTrackingExportHelper)
        delete pChangeTrackingExportHelper;
    if (pChartListener)
        delete pChartListener;
    if (pCellsItr)
        delete pCellsItr;
    if (pDefaults)
        delete pDefaults;
    if (pNumberFormatAttributesExportHelper)
        delete pNumberFormatAttributesExportHelper;
}

--:---  20111007.2.log    All (32,0)     (C/l +1 Abbrev)--2011-10-14 13:20 +0900 ----
```

Use case 2: 2011-10-07 sc

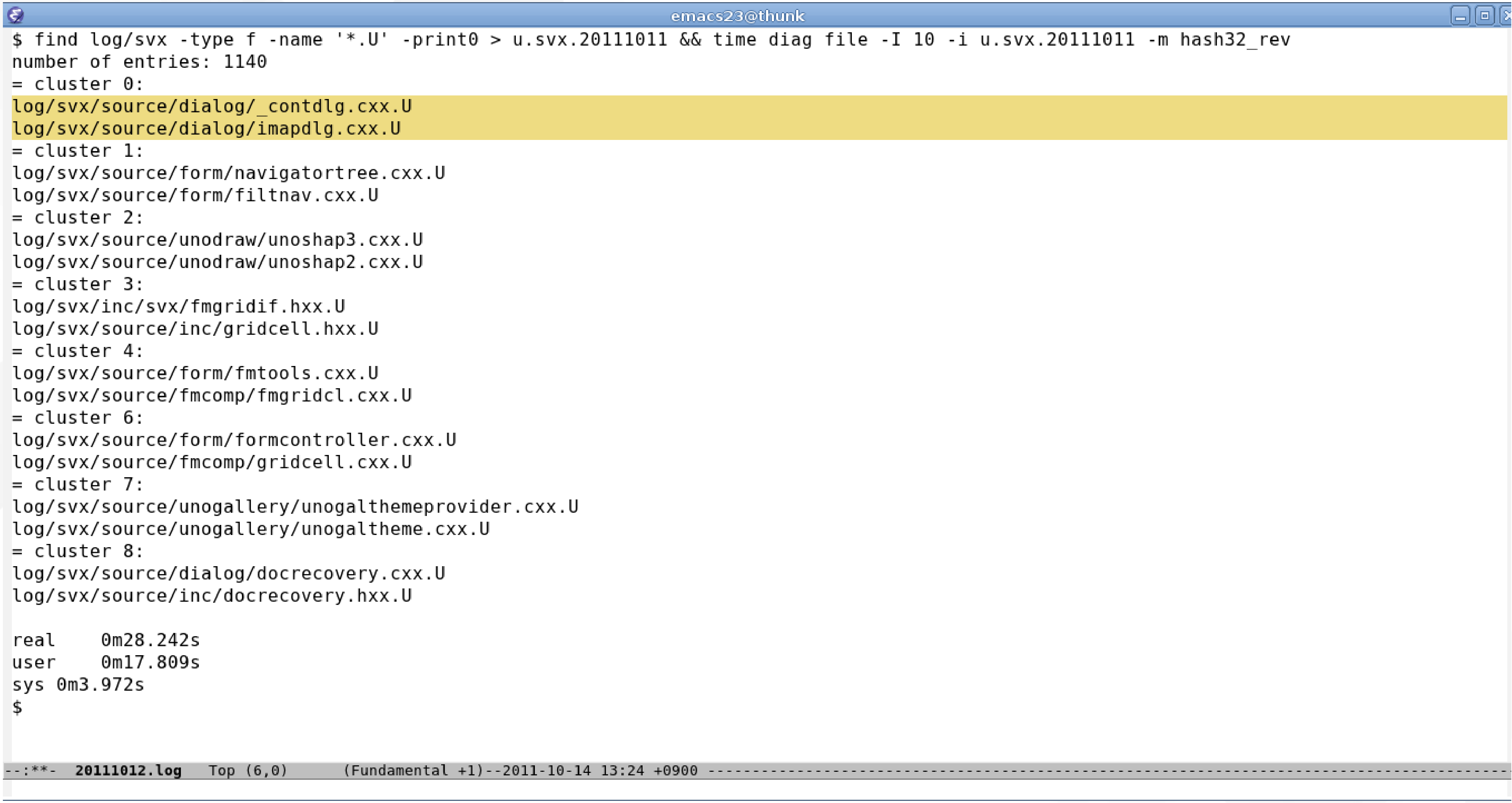


```
In core/sc/source/filter/xml/xmlimprt.cxx:

ScXMLImport::~ScXMLImport() throw()
{
    //  delete pI18NMap;
    delete pDocElemTokenMap;
    delete pStylesElemTokenMap;
    delete pStylesAttrTokenMap;
    delete pStyleElemTokenMap;
    delete pBodyElemTokenMap;
    delete pContentV
    delete pContentV
    delete pContentV
    delete pContentV
    delete pContentV
    delete pContentV
--:---  20111007.1.log   T
    if (pStylesImpor
        delete pStyl

    if (pSolarMutexG
        delete pSolarMutexGuard;

    if (pMyNamedExpressions)
        delete pMyNamedExpressions;
    if (pMyLabelRanges)
        delete pMyLabelRanges;
    if (pValidations)
        delete pValidations;
    if (pDetectiveOpArray)
        delete pDetectiveOpArray;
}

--:---  20111007.1.log   Bot (96,0)     (C/l +1 Abbrev)--2011-10-14 13:20 +0900 -----
```

```
In sc/source/filter/xml/xmlexprt.cxx:

ScXMLExport::~ScXMLExport()
{
        if (pGroupColumns)
            delete pGroupColumns;
        if (pGroupRows)
            delete pGroupRows;
        if (pColumnStyles)
            delete pColumnStyles;

        if (pChangeTrackingExportHelper)
            delete pChangeTrackingExportHelper;
        if (pChartListener)
            delete pChartListener;
        if (pCellsItr)
            delete pCellsItr;
        if (pDefaults)
            delete pDefaults;
        if (pNumberFormatAttributesExportHelper)
            delete pNumberFormatAttributesExportHelper;
}

--:---  20111007.2.log   All (32,0)     (C/l +1 Abbrev)--2011-10-14 13:20 +0900 ----
```

```
commit 6b280c938d2213cf9e7ddea13baa38427c3b5c6d
Date:   Fri Oct 7 00:51:06 2011 +0900

    delete NULL is safe
```

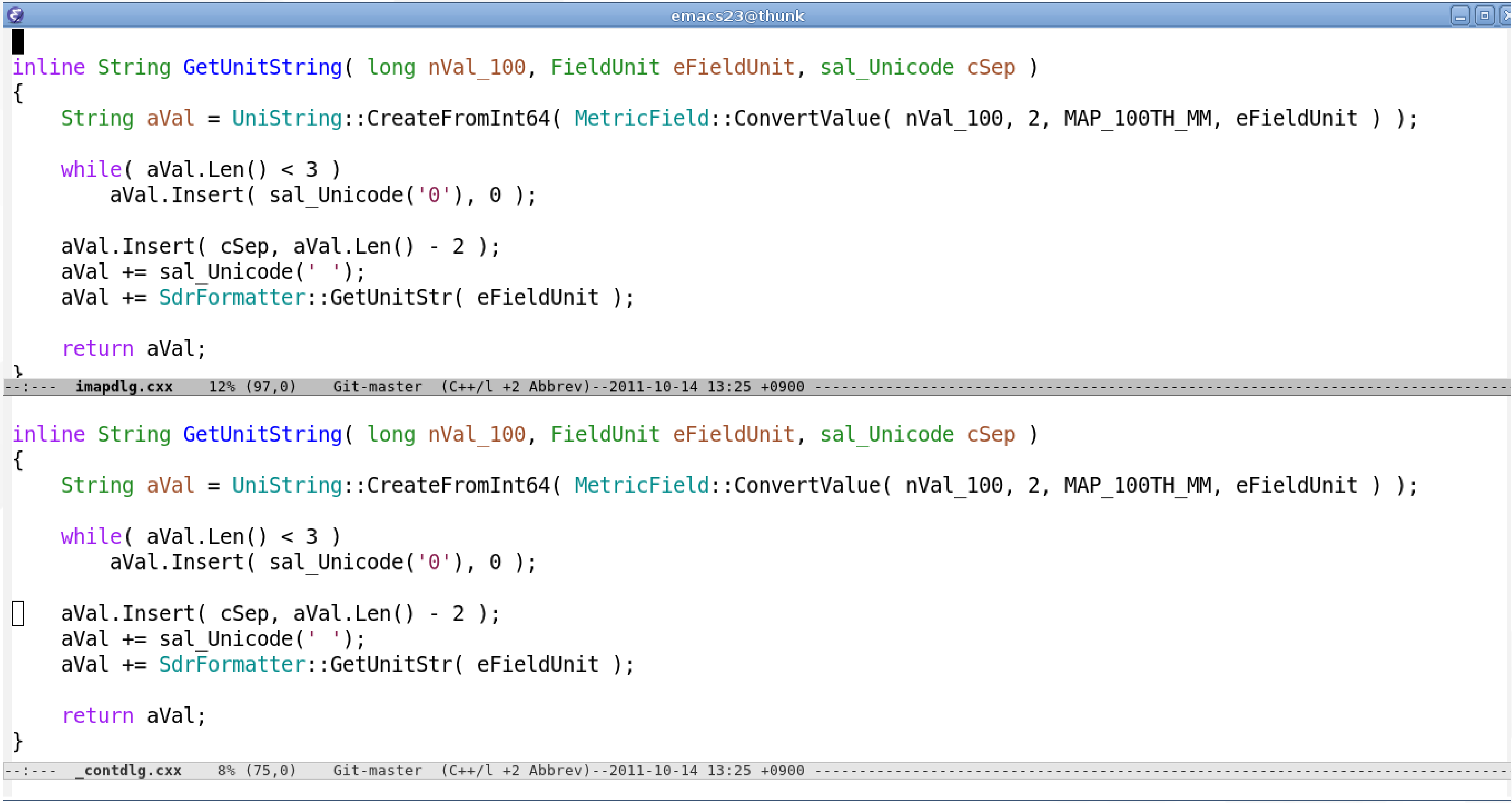
Use case 3: 2011-10-12 svx



```
$ find log/svx -type f -name '*.U' -print0 > u.svx.20111011 && time diag file -I 10 -i u.svx.20111011 -m hash32_rev
number of entries: 1140
= cluster 0:
log/svx/source/dialog/_contdlg.cxx.U
log/svx/source/dialog/imapdlg.cxx.U
= cluster 1:
log/svx/source/form/navigatortree.cxx.U
log/svx/source/form/filtnav.cxx.U
= cluster 2:
log/svx/source/unodraw/unoshap3.cxx.U
log/svx/source/unodraw/unoshap2.cxx.U
= cluster 3:
log/svx/inc/svx/fmgridif.hxx.U
log/svx/source/inc/gridcell.hxx.U
= cluster 4:
log/svx/source/form/fmtools.cxx.U
log/svx/source/fmcomp/fmgridcl.cxx.U
= cluster 6:
log/svx/source/form/formcontroller.cxx.U
log/svx/source/fmcomp/gridcell.cxx.U
= cluster 7:
log/svx/source/unogallery/unogalthemeprovider.cxx.U
log/svx/source/unogallery/unogaltheme.cxx.U
= cluster 8:
log/svx/source/dialog/docrecovery.cxx.U
log/svx/source/inc/docrecovery.hxx.U

real    0m28.242s
user    0m17.809s
sys 0m3.972s
$
```

```
--:**-  20111012.log    Top (6,0)     (Fundamental +1)--2011-10-14 13:24 +0900 ------------------------------------------
```

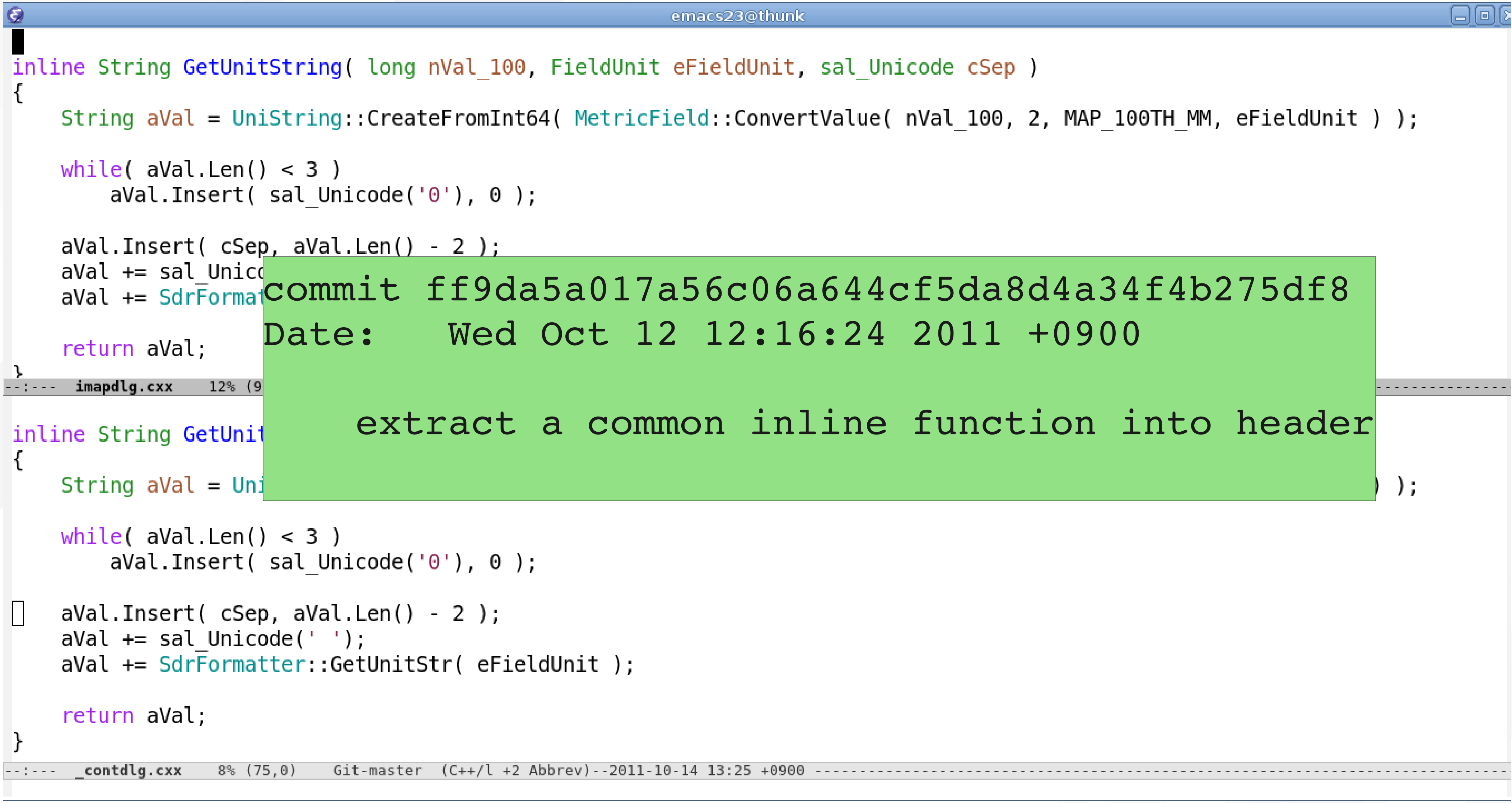
Use case 3: 2011-10-12 svx



```
inline String GetUnitString( long nVal_100, FieldUnit eFieldUnit, sal_Unicode cSep )
{
    String aVal = UniString::CreateFromInt64( MetricField::ConvertValue( nVal_100, 2, MAP_100TH_MM, eFieldUnit ) );

    while( aVal.Len() < 3 )
        aVal.Insert( sal_Unicode('0'), 0 );

    aVal.Insert( cSep, aVal.Len() - 2 );
    aVal += sal_Unicode(' ');
    aVal += SdrFormatter::GetUnitStr( eFieldUnit );

    return aVal;
}
--:---  imapdlg.cxx     12% (97,0)     Git-master   (C++/l +2 Abbrev)--2011-10-14 13:25 +0900 ----------------------------------------------------------------
```

```
inline String GetUnitString( long nVal_100, FieldUnit eFieldUnit, sal_Unicode cSep )
{
    String aVal = UniString::CreateFromInt64( MetricField::ConvertValue( nVal_100, 2, MAP_100TH_MM, eFieldUnit ) );

    while( aVal.Len() < 3 )
        aVal.Insert( sal_Unicode('0'), 0 );

    aVal.Insert( cSep, aVal.Len() - 2 );
    aVal += sal_Unicode(' ');
    aVal += SdrFormatter::GetUnitStr( eFieldUnit );

    return aVal;
}
--:---  _contdlg.cxx     8% (75,0)     Git-master   (C++/l +2 Abbrev)--2011-10-14 13:25 +0900 ----------------------------------------------------------------
```

Use case 3: 2011-10-12 svx



```
inline String GetUnitString( long nVal_100, FieldUnit eFieldUnit, sal_Unicode cSep )
{
    String aVal = UniString::CreateFromInt64( MetricField::ConvertValue( nVal_100, 2, MAP_100TH_MM, eFieldUnit ) );

    while( aVal.Len() < 3 )
        aVal.Insert( sal_Unicode('0'), 0 );

    aVal.Insert( cSep, aVal.Len() - 2 );
    aVal += sal_Unic
    aVal += SdrFormat

    return aVal;
}
--:---  imapdlg.cxx    12% (9

inline String GetUnit
{
    String aVal = Uni                                                                                    ) );

    while( aVal.Len() < 3 )
        aVal.Insert( sal_Unicode('0'), 0 );

    aVal.Insert( cSep, aVal.Len() - 2 );
    aVal += sal_Unicode(' ');
    aVal += SdrFormatter::GetUnitStr( eFieldUnit );

    return aVal;
}
--:---  _contdlg.cxx    8% (75,0)    Git-master  (C++/l +2 Abbrev)--2011-10-14 13:25 +0900 ----------------
```

```
commit ff9da5a017a56c06a644cf5da8d4a34f4b275df8
Date:   Wed Oct 12 12:16:24 2011 +0900

    extract a common inline function into header
```

Summary

- Cluster analysis helps you understanding the code base
  - Get diagonal by

```
git clone \
git://diagonal.git.sourceforge.net/gitroot/diagonal/diagonal
```

- Related works
  - Detecting code duplication / plagiarism
    - Code clone literature
    - How to detect code duplication during development?
    - Plagiarism Prevention and Detection

PARIS2011 CONFERENCE

12th-15th October

Thank you!

- Aknowledgement