



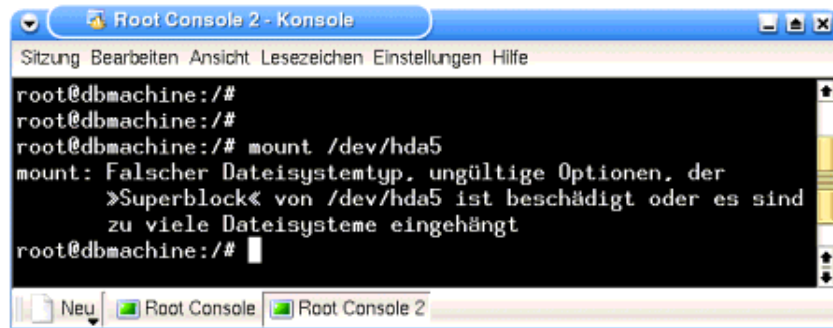
by Detlef Müller
<detlef_mue/at/web.de>

About the author:

They call me 'Linux' in the Internet cafe, even though I've only been working with Tux-OS for two years ... maybe it's time to get a BSD as well ... No 'job' at the moment, but I would like to get involved in Linux work sometime. So for me, Linux is both an ersatz job and a hobby. My other hobby is Attac, since the beginning of 2004. I'd like to contribute to implementing Linux usefully in that area. My first building site ... The vision: an e-democracy system that allows *all* participants to vote on the Internet - using free software, of course.

Translated to English by:
Orla Shanaghy
<orla(at)jostraca.org>

Data Loss: Worst Case Scenario



Abstract:

One of the best decisions I ever made about Linux was to only use journaling file systems. This decision was proven right yesterday in a very convincing way. A botched copy process ate all the data on the partition, including the entire data of a Linux project, and made the partition unmountable. On a ReiserFS journal file system ...

File systems with journaling are one of the goodies that make working with Linux secure. They ensure that you *can* use the reset button - usually (!) without any ill effects.

This report about a real-life data loss shows that it *can* sometimes have ill effects, and describes the bits and bytes' heroic rescue by a professionally operating Linux tool called 'reiserfsck'.

Linux introduction

Tux has been on my computer for about two years - three penguins now inhabit my computer. Two of the SuSE species, one of the Debian genus, Knoppix on the maternal line.

It all started with SuSE 7.3, a bargain snapped up on E-Bay. I'd already heard so much about Linux and wanted to become a Linux specialist myself, so this was my way of getting started.

Newbie problems ...

The first steps were definitely not easy. How often did I end up cursing the glut of new technical terms - especially since they are (usually) never explained.

When you read the first few sentences of the German distributors' manual, you are inundated with KDE, YaST, Bash, etc. ... and previously a big-name computer magazine had described it as the distribution with the best documentation ... Not a chance - nothing is simple or clear.

Sigh...but it passes. Back to the main point.

ReiserFS on EISA 486

This SuSE Linux 7.3 came on an old 486 that still had an EISA bus (...yes, those things do still exist.) The first hard reset (reset button) and subsequent restart caused problems. No more access to the file system, and only read-only mount (only read access).

"What's that supposed to mean?"

It means a lot of work. Repair attempts were fruitless...in the end, I just re-installed the whole SuSE. This happened 5 or 6 times. Each time I booted with the SuSE recovery system, used the repair tool e2fsck for ext2 file systems, and once I also edited the /etc/fstab file with the miserable vi editor. Then the file system was OK...or maybe not. Finally, I reinstalled Linux. By that stage, a whole day was gone. This kind of stuff takes longer for newbies...

Then I got the idea - inspired by an article in c't - to install a YaST journaling file system. No sooner said than done, and since then I have been spared from starting the recovery system etc.

If the PC hadn't been properly shut down beforehand, I got a factual 'replayed nnn transactions in ...' while Linux was being started, and the computer booted successfully.

"Halleluia!" I think. That's better. From now on, no more ext2 - journaling is the way to go from here!

'Journal replay' of a ReiserFS partition during system start ... (from log file) :

```
.....  
reiserfs: found format "3.6" with standard journal
```

```
reiserfs: checking transaction log (sd(8,4)) for (sd(8,4))
reiserfs: replayed 109 transactions in 10 seconds
reiserfs: using ordered data mode
.....
```

Stress tests

But I wanted to know for sure.

Once I was reasonably familiar with the JFs, I carried out stress tests. The file system was subjected to a hard reset with a fully-equipped interface.

Started KDE, with lots of programs, opened files with the editor, then pressed Reset button. The tests were successful. The file system *really* survived it.

Even activating 'emergency exit' during a running copy process caused no problems. The 486 SCSI system did cause a few problems, but **ReiserFS** 'does what it says on the tin'. It *always* returned the file system to a consistent, usable state. The opened files were likewise returned to their original state. Tests I carried out later on under the same conditions with ext3, the journal variety of ext2, were also successful.

This is how it all looks in the log file with ext3 during system bootup:

```
.....
Journalled Block Device driver loaded
(recovery.c, 256): journal_recover: JBD: recovery, exit status 0,
recovered transactions 450798 to 451415
(recovery.c, 258): journal_recover: JBD: Replayed 3756 and revoked 6/15 blocks
kjournald starting. Commit interval 5 seconds
EXT3 FS 2.4-0.9.19, 19 August 2002 on sd(8,1), internal journal
ext3_orphan_cleanup: deleting unreferenced inode 355953
ext3_orphan_cleanup: deleting unreferenced inode 355952
EXT3-fs: sd(8,1): 2 orphan inodes deleted
EXT3-fs: recovery complete.
EXT3-fs: mounted filesystem with ordered data mode.
.....
```

Other journal file systems

That was the preamble ...

Since then, I have also used **ext3** and **XFS**.

I've stayed away from **JFS**, as it's not supposed to be fully secure yet. I'm not saying anything negative about it, I just haven't tried it yet.

XFS has disappeared. I don't mind; I didn't have any problems with it, but I just hadn't used it in a long

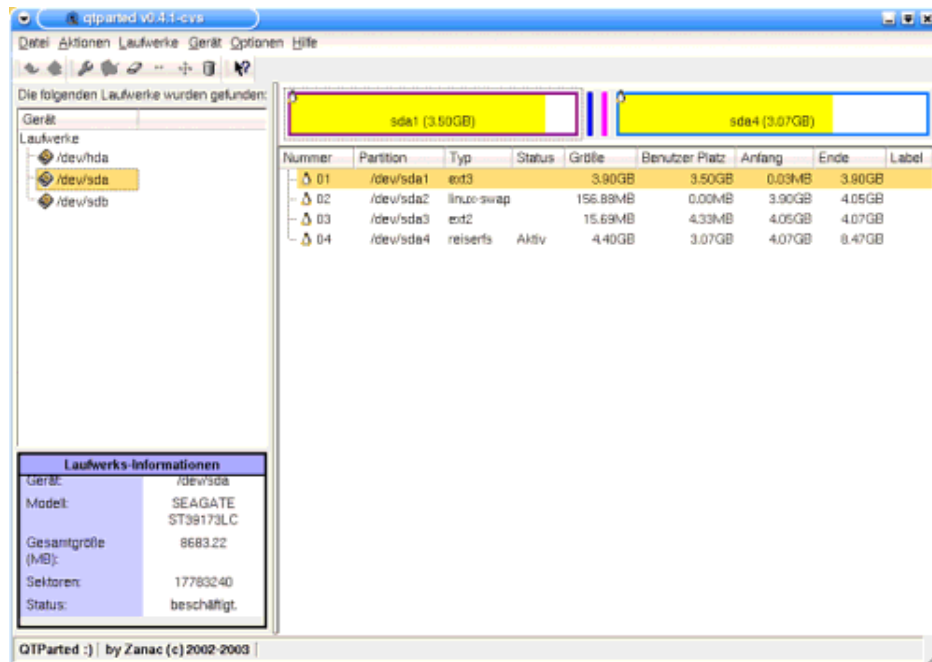
time.

I have kept using the ext3 file system. It's now running on the 486 alongside a Debian / unstable. With ext 2, it's even possible to change an existing partition to ext 3 *with data - on the live system*. I've tried it - it works!

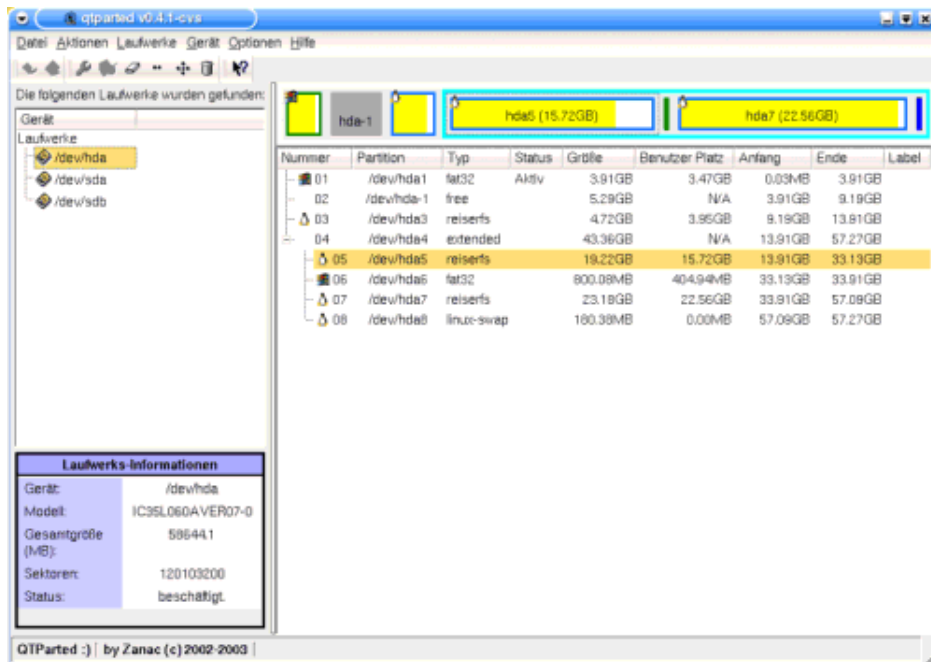
I used ext 3 again when I last installed Knoppix version 3.4 on the hard disk.

Most file systems on my work PC - *only* a PIII/500 - are currently ReiserFS.

How the two disks on my work PC are divided up:



Graphic 2, sda partitioning (SCSI disk)



Graphic 3, hda partitioning

D-day

I've been working on a documentation CD for Linux for the past 3/4 year. This involves large data volumes: howtos, tutorials, FAQs, plus different formats and archives in each case, and the same volume again for updates. I am also writing additional HTML files so that it's easy to get an overview of the CD-ROM.

There's been lots to do in the past few weeks. A free version of this CD is supposed to be available soon. So - put together an image, write a few command line burn scripts - it's just quicker than using a KDE program.

And put everything onto my hard disk. My data store is /dev/hda5 on a 60-Gig IDE disk. The partition is 20 Gigs (of which over 80% is already full). All important bits and bytes, lots of work involved. If anything were ever to happen to it ... oh, surely that's not very likely, it's not Windows with FATxx, after all.

I've often thought about backups, but never done one so far. I just have a few copies on a separate hard disk, and leave it at that.

Yesterday evening I left the Internet cafe, where I had downloaded packages from the SuSE website. They were all original SuSE documentation from 7.3 to 9.0 on CD 2. At home, I booted up the PC with SuSE 8.1. I usually use Debian, but because the packages were SuSE RPMs, I used 8.1 this time. And I was able to install the first 9.0 doc package. So it's no problem to install a newer package on version 8*. So I installed the 9.0 RPMs, copied them to the aforementioned hda5 partition, and then de-installed the RPMs. Then I did the same thing with 8.0.

Without closing down KDE, I changed to another console and entered <CTRL ALT> to shut

down the PC. I got an error message on the command line - I've forgotten exactly what it was - all I remember is...the PC has given up the ghost. Can't do anything...
OK, so I pressed the hard Reset button - I'm not afraid of doing that on Linux at all any more.

Worst case scenario

When I booted up Debian, I didn't notice anything at first. Then on the KDE level:
No directories are being shown on my work partition.
But it's full to bursting ... ?
It probably hasn't been mounted (no, rubbish - it's automatically mounted at startup).

And then, after a 'mount /dev/hda5' error message - too many file systems - or wrong superblock. The penny is dropping...

What I'm experiencing here is a real-life **data Worst Case Scenario**.

And now? Erm ... maybe try mounting again? No point - if it didn't mount the first time, it's not going to happen the second time.
But I tried it anyway ... nada! The partition containing the results of months of research, lots of HTML pages I had written myself, scripts for burning CDs, collections of DEBs and RPMs from the Internet, and lots of other stuff - all gone, disappeared, in Nirvana or wherever else.
Of course, some data is still on the disk, but will I be able to access it again?

Lean back, light a cigarette ...

As a DIY man at heart, the first thing that came to mind was data recovery. The partition is a ReiserFS. There are tools for that. I once read something in an article about c't Knoppix; I originally installed Debian as Knoppix. So the tools should be there.
And they are there.

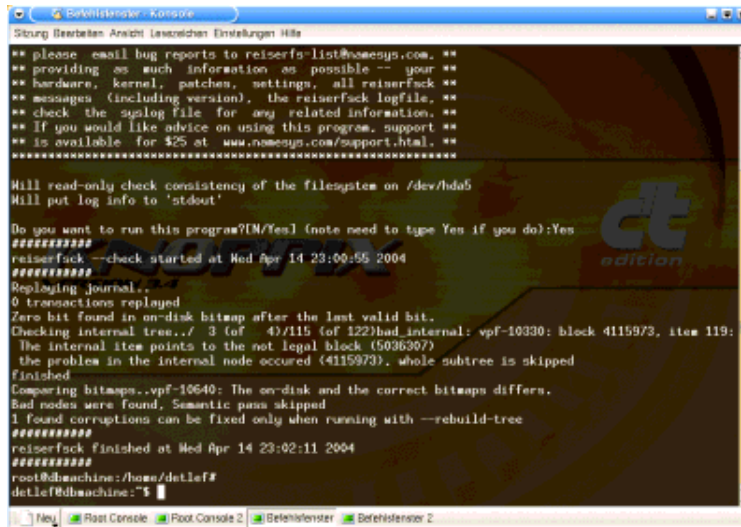
reiserfsck in an emergency operation

So: first, look for the doc directory. Must be under /usr/share/doc/reiser-something. Under Something (...should be called reiserfsprogs) I find a few English files, one for each tool, converted from the manpages.

A quick look through the operation tools for data recovery reveals that reiserfsck must be the 'scalpel'.
OK, let's get started...

First, I called it up without changing anything. -check seems to be the right thing to do at the beginning.
First the diagnosis, then the operation...

```
# reiserfsck -check
```



```
Sitzung Bearbeiten Ansicht Lesezeichen Einstellungen Hilfe

** please email bug reports to reiserfs-list@namesys.com. **
** providing as much information as possible -- your **
** hardware, kernel, patches, settings, all reiserfsck **
** messages (including version), the reiserfsck logfile, **
** check the syslog file for any related information. **
** If you would like advice on using this program, support **
** is available for $25 at www.namesys.com/support.html. **
*****

Will read-only check consistency of the filesystem on /dev/hda5
Will put log info to 'stdout'

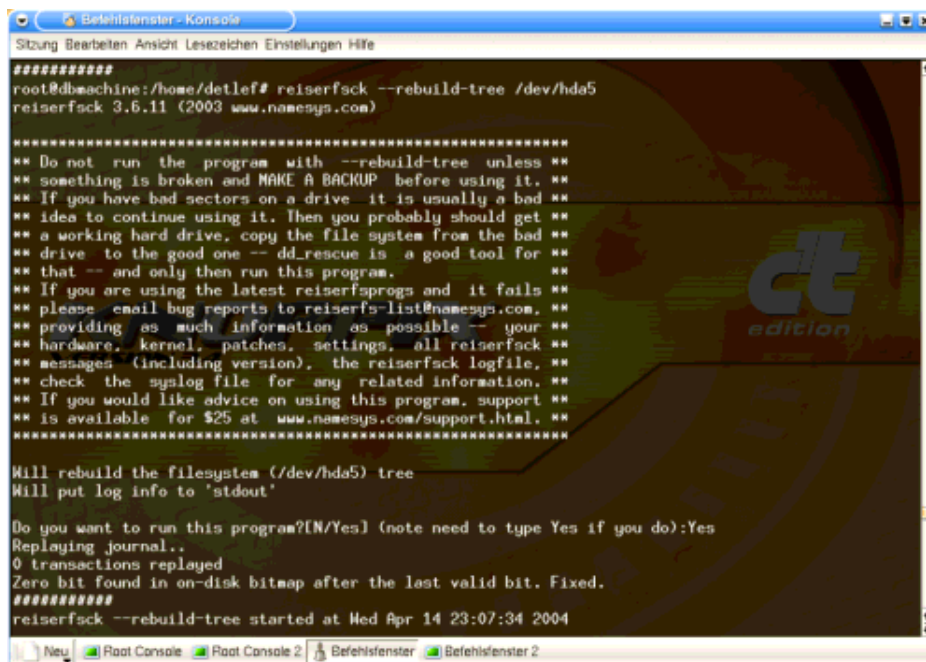
Do you want to run this program?[N/Yes] (note need to type Yes if you do):Yes
*****
reiserfsck --check started at Wed Apr 14 23:00:55 2004
*****
Replaying journal..
0 transactions replayed
Zero bit found in on-disk bitmap after the last valid bit.
Checking internal tree.. / 3 (of 4)/115 (of 122)bad internal: vpf-10330: block 4115973, item 119:
The internal item points to the not legal block (5036307)
the problem in the internal node occurred (4115973), whole subtree is skipped
Finished
Comparing bitmaps.. vpf-10640: The on-disk and the correct bitmaps differs.
Bad nodes were found, Semantic pass skipped
1 found corruptions can be fixed only when running with --rebuild-tree
*****
reiserfsck finished at Wed Apr 14 23:02:11 2004
*****
root@dbmachine:/home/detlef#
detlef@dbmachine:~$
```

Bild 4, reiserfsck -check

I don't understand it all, but I do understand that reiserfsck found errors and says it can fix them. Sounds good.

I thought about it for a minute, then started the operation. Called up the scalpel using ...

```
# reiserfsck --rebuild tree /dev/hda5
```



```
Sitzung Bearbeiten Ansicht Lesezeichen Einstellungen Hilfe

*****
root@dbmachine:/home/detlef# reiserfsck --rebuild-tree /dev/hda5
reiserfsck 3.6.11 (2003 www.namesys.com)

*****
** Do not run the program with --rebuild-tree unless **
** something is broken and MAKE A BACKUP before using it. **
** If you have bad sectors on a drive it is usually a bad **
** idea to continue using it. Then you probably should get **
** a working hard drive, copy the file system from the bad **
** drive to the good one -- dd_rescue is a good tool for **
** that -- and only then run this program. **
** If you are using the latest reiserfsprogs and it fails **
** please email bug reports to reiserfs-list@namesys.com. **
** providing as much information as possible -- your **
** hardware, kernel, patches, settings, all reiserfsck **
** messages (including version), the reiserfsck logfile, **
** check the syslog file for any related information. **
** If you would like advice on using this program, support **
** is available for $25 at www.namesys.com/support.html. **
*****

Will rebuild the filesystem (/dev/hda5) tree
Will put log info to 'stdout'

Do you want to run this program?[N/Yes] (note need to type Yes if you do):Yes
Replaying journal..
0 transactions replayed
Zero bit found in on-disk bitmap after the last valid bit. Fixed.
*****
reiserfsck --rebuild-tree started at Wed Apr 14 23:07:34 2004
*****
```

Graphic 5, reiserfsck --rebuild-tree

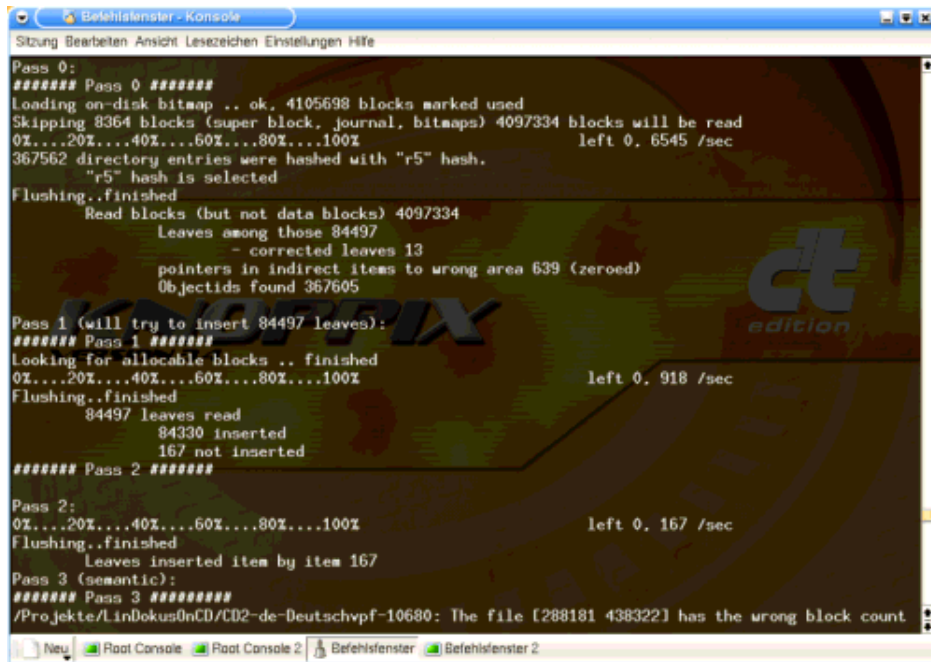
This makes me nervous. No wonder - I'm about to find out what I'll have to do in the next few weeks. "Should the file system be restored now?" ... Yes, it should.

I get the good old 'replaying journal' message. It's this Good Samaritan that makes the restore possible -

a kind of table of contents for all sub-partitions. Two lines later, reiserfsck comes across a faulty null bit and ... corrects it.

Next comes **Pass 0** of the restore, visually separated on the console. This process takes ca. 15 minutes for my 20 GB ... a percentage display allows the user to keep an eye on progress.

Graphic 2 shows details of an error. What does it mean exactly? Hmm ... ask me something else. :)

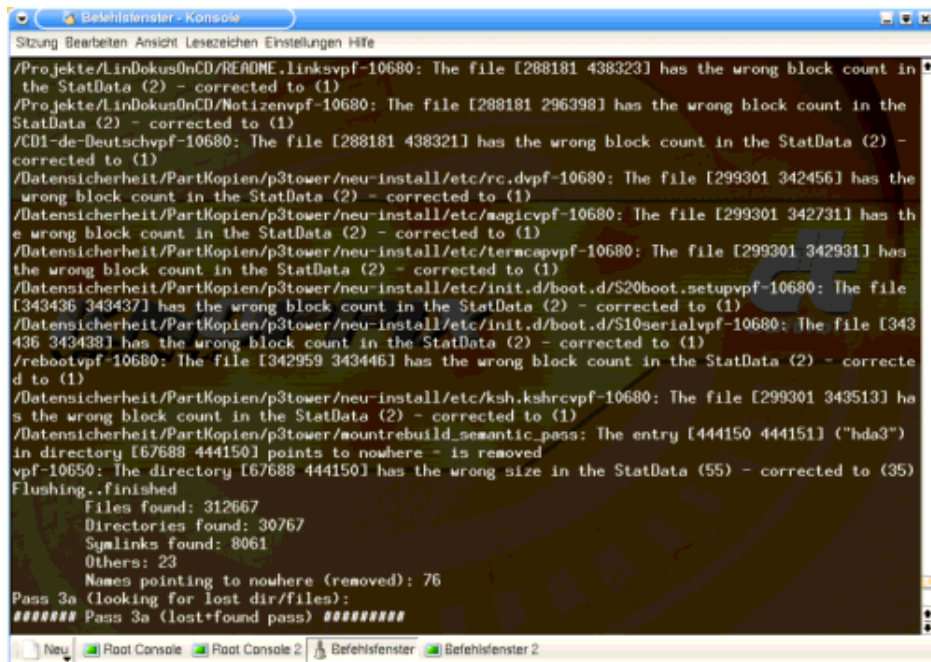


Graphic 6, Pass 0 up to 2, 3 (beginning only)

On it goes ... **Pass 1** is really fast. No data error messages.

Pass 2 is the same.

In **Pass 3** I get inundated with data error messages. I recognise the files, they are from the SuSE documentation copy process. That proves that something went wrong with that specific copy process. Was it the KDE 3 conqueror or a bug in ReiserFS?

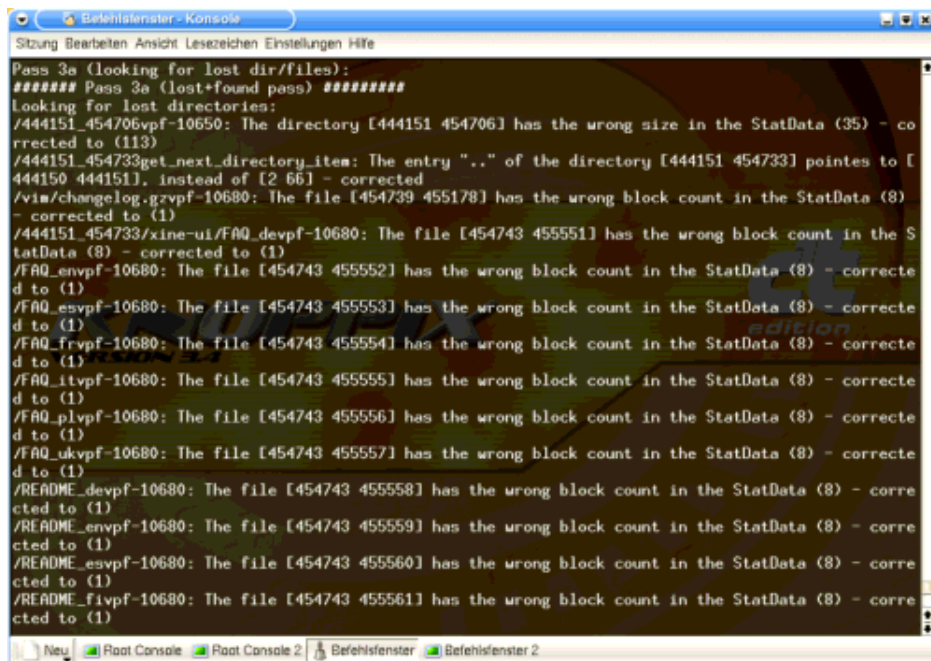


```
Befehlsfenster - Konsole
Sitzung Bearbeiten Ansicht Lesezeichen Einstellungen Hilfe

/Projekte/LinDokus0nCD/README.linksvpf-10680: The file [288181 438323] has the wrong block count in the StatData (2) - corrected to (1)
/Projekte/LinDokus0nCD/Notizenvpf-10680: The file [288181 296398] has the wrong block count in the StatData (2) - corrected to (1)
/CD1-de-Deutschvpf-10680: The file [288181 438321] has the wrong block count in the StatData (2) - corrected to (1)
/Datensicherheit/PartKopien/p3tower/neu-install/etc/rc.d/vpf-10680: The file [299301 342456] has the wrong block count in the StatData (2) - corrected to (1)
/Datensicherheit/PartKopien/p3tower/neu-install/etc/magicvpf-10680: The file [299301 342731] has the wrong block count in the StatData (2) - corrected to (1)
/Datensicherheit/PartKopien/p3tower/neu-install/etc/ternacvpf-10680: The file [299301 342931] has the wrong block count in the StatData (2) - corrected to (1)
/Datensicherheit/PartKopien/p3tower/neu-install/etc/init.d/boot.d/S20boot.setupvpf-10680: The file [343436 343437] has the wrong block count in the StatData (2) - corrected to (1)
/Datensicherheit/PartKopien/p3tower/neu-install/etc/init.d/boot.d/S10serialvpf-10680: The file [343436 343438] has the wrong block count in the StatData (2) - corrected to (1)
/rebootvpf-10680: The file [342959 343446] has the wrong block count in the StatData (2) - corrected to (1)
/Datensicherheit/PartKopien/p3tower/neu-install/etc/ksh.kshrcvpf-10680: The file [299301 343513] has the wrong block count in the StatData (2) - corrected to (1)
/Datensicherheit/PartKopien/p3tower/mountrebuild_semantic_pass: The entry [444150 444151] ("hda3") in directory [67688 444150] points to nowhere - is removed
vpf-10650: The directory [67688 444150] has the wrong size in the StatData (55) - corrected to (35)
Flushing..finished
Files found: 31267
Directories found: 30767
Symlinks found: 8061
Others: 23
Names pointing to nowhere (removed): 76
Pass 3a (looking for lost dir/files):
##### Pass 3a (lost+found pass) #####
```

Graphic 7, Pass 3 (end)

According to the description, a search is carried out in **Pass 3a** for lost files or directories.



```
Befehlsfenster - Konsole
Sitzung Bearbeiten Ansicht Lesezeichen Einstellungen Hilfe

Pass 3a (looking for lost dir/files):
##### Pass 3a (lost+found pass) #####
Looking for lost directories:
/444151_454706vpf-10650: The directory [444151 454706] has the wrong size in the StatData (35) - corrected to (113)
/444151_454733get_next_directory_item: The entry ".." of the directory [444151 454733] points to [444150 444151], instead of [2 66] - corrected
/via/changelog.gzvpf-10680: The file [454739 455178] has the wrong block count in the StatData (8) - corrected to (1)
/444151_454733/xine-ui/FAQ_devvpf-10680: The file [454743 455551] has the wrong block count in the StatData (8) - corrected to (1)
/FAQ_envvpf-10680: The file [454743 455552] has the wrong block count in the StatData (8) - corrected to (1)
/FAQ_esvpf-10680: The file [454743 455553] has the wrong block count in the StatData (8) - corrected to (1)
/FAQ_frvpf-10680: The file [454743 455554] has the wrong block count in the StatData (8) - corrected to (1)
/FAQ_itvpf-10680: The file [454743 455555] has the wrong block count in the StatData (8) - corrected to (1)
/FAQ_plvpf-10680: The file [454743 455556] has the wrong block count in the StatData (8) - corrected to (1)
/FAQ_ukvpf-10680: The file [454743 455557] has the wrong block count in the StatData (8) - corrected to (1)
/README_devvpf-10680: The file [454743 455558] has the wrong block count in the StatData (8) - corrected to (1)
/README_envvpf-10680: The file [454743 455559] has the wrong block count in the StatData (8) - corrected to (1)
/README_esvpf-10680: The file [454743 455560] has the wrong block count in the StatData (8) - corrected to (1)
/README_fivpf-10680: The file [454743 455561] has the wrong block count in the StatData (8) - corrected to (1)
```

Graphic 8, Pass 3a

The tool usually finds what it's looking for, specifies the error, and corrects the relevant entries, commenting them with a 'corrected to ...' at the end of the line.

Then it gives a summary of its emergency rescue operation. In **Pass 4**, it gives the simple message that the synchronization (of the journal in its current state on the hard disk) has finished.

```
Sitzung Bearbeiten Ansicht Leesezeichen Einstellungen Hilfe
444151 444152], instead of [2 66] - corrected
/444152_444179get_next_directory_item: The entry "." of the directory [444152 444179] points to [
444151 444152], instead of [2 66] - corrected
/454706_454708get_next_directory_item: The entry "." of the directory [454706 454708] points to [
444151 454706], instead of [2 66] - corrected
/454706_454709get_next_directory_item: The entry "." of the directory [454706 454709] points to [
444151 454706], instead of [2 66] - corrected
/454706_454710get_next_directory_item: The entry "." of the directory [454706 454710] points to [
444151 454706], instead of [2 66] - corrected
/454706_454711get_next_directory_item: The entry "." of the directory [454706 454711] points to [
444151 454706], instead of [2 66] - corrected
/454706_454712get_next_directory_item: The entry "." of the directory [454706 454712] points to [
444151 454706], instead of [2 66] - corrected
/454706_454713get_next_directory_item: The entry "." of the directory [454706 454713] points to [
444151 454706], instead of [2 66] - corrected
/Topic-15/bak/index.html.devpf-10680: The file [454731 455033] has the wrong block count in the Sta
tData (8) - corrected to (1)
Flushing..finished4, 236 /sec
  Objects without names 26
  Empty lost dirs removed 12
  Dirs linked to /lost+found: 26
    Dirs without stat data found 1
Pass 4 - finished done 83101, 199 /sec
  Deleted unreachable items 25
Flushing..finished
Syncing..finished
*****
reiserfsck finished at Wed Apr 14 23:26:35 2004
*****
root@dbmachine:~/home/detlef#
detlef@dbmachine:~$
```

Graphic 9, Pass 4 and end

Now my data should be accessible again.

I get no messages during the mount - a sure sign on UNIX that a command has executed successfully. :-))

All's well that ends well?

And the conqueror is showing me my old familiar directories on partition hda5. Everything is back again ... or should I say, almost everything. A few of the copied files are missing - naturally. You can't expect a perfect result from a faulty process. I can copy them back over.

Today, the day after, I still haven't checked all the data on hda5. But it's very likely that everything was fully restored. The tool seemed *very professional* in active use!

It's now 16:30 hrs on D-Day+1. The alarm bell rang just 18 hours ago. The report (this one) is nearly finished - that's how successful the emergency rescue operation was.

I'm glad that I saved the progress of the 'console' after the recovery in a file yesterday. It meant I could include screenshots of the original 'accident photos' in this article.

P.S. (2 days later): no indication of any data loss. I work on the affected partition all the time.

Verdict

Data loss can also occur on a journal file system, but the chances of a full recovery are higher. JFs are

reliable and easy to maintain.

A **journal file system** is a *must* for every Linux user (you'll forgive such a strong opinion in the free software world).

Most distributors now offer the user a journal file system as a default setting during the installation procedure.

And ... that means backup sinners have struck lucky and can get by without backing up their data. However, this is not to advocate not backing things up. So, always make backups!

References

Journal file systems articles:

- Journaling file systems for Linux - Linux Gazette issue 68, July 2001 (de | en); .. with lots of details.
- Adventure ReiserFS - Linux Netmag 4/2000 (de | en)
- Doppelte Buchführung - Linux Magazine 1/2002 (de), comparison of journal file systems (German only).
- Darf es etwas mehr sein ? - Linux Magazine 6/2000 (de), extending a ReiserFS on LVM (German only).
- Buchführung für die Festplatte - Linux Magazine 4/2000 (de), a comparison of four journal file systems (German only).
- Crashfest - Linux Magazine 7/2001 (de) XFS on SuSE 7.1 (German only).

Journal file systems websites:

- ReiserFS - the homepage of ReiserFS.
- ext2 / ext3 - or try this website
- XFS - the SGI journal file system.
- JFS - ... an IBM Open Source project.

Backup articles:

- RSync: the best backup system - LinuxFocus March 2004.
- storeBackup, the unconventional backup tool - LinuxFocus January 2004.
- Arkeia, a commercial, professional backup solution for networks - LinuxFocus May 2000.

<p>Webpages maintained by the LinuxFocus Editor team © Detlef Müller "some rights reserved" see linuxfocus.org/license/ http://www.LinuxFocus.org</p>	<p>Translation information: de --> -- : Detlef Müller <detlef_mue/at/web.de> de --> en: Orla Shanaghy <orla(at)jostraca.org></p>
--	--