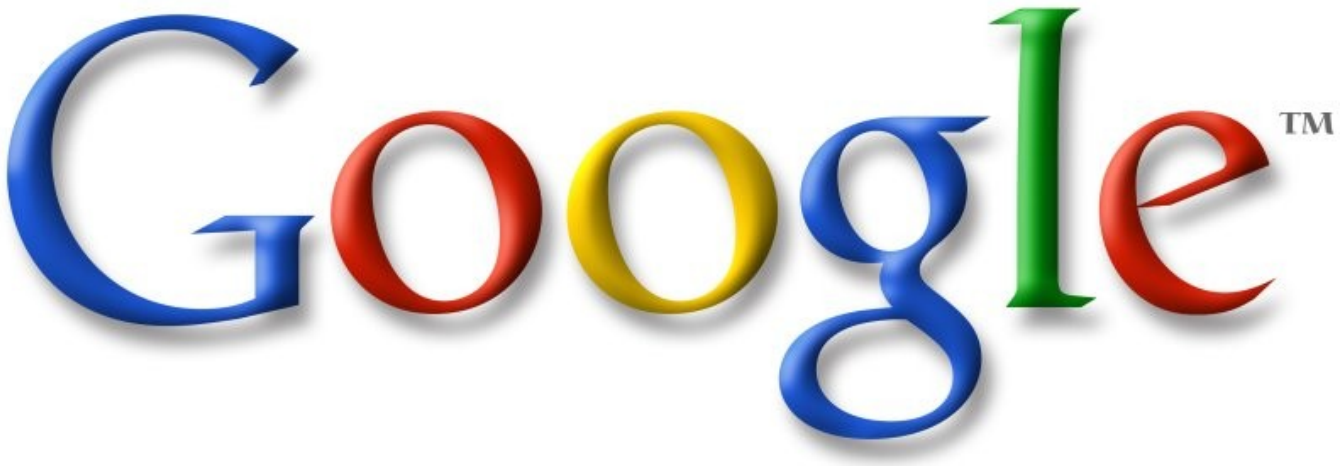
Alexandre Dulaunoy & Michael G. Noll

GooDiff

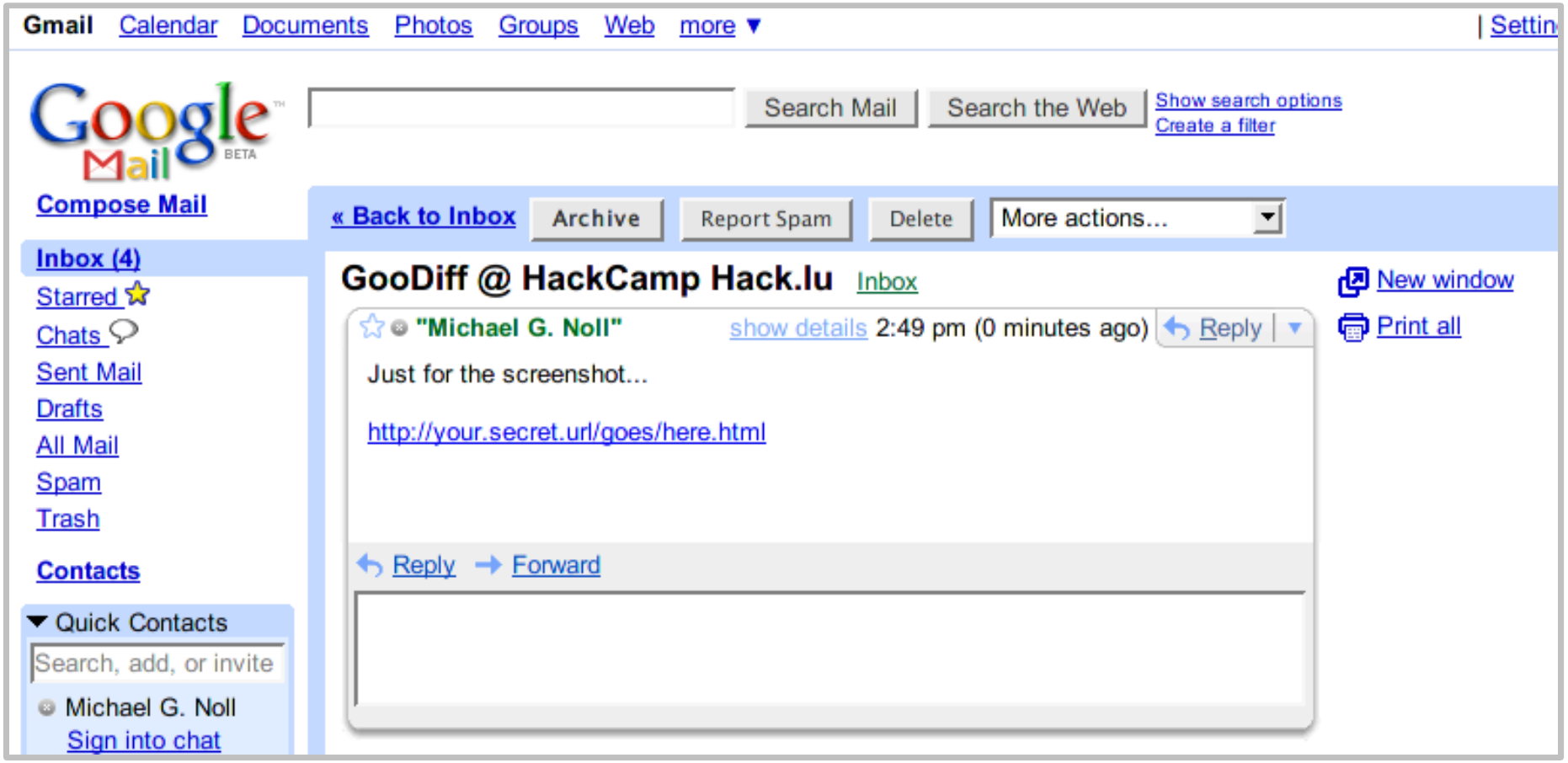
"GooDiff is a service for automated tracking of semantic changes in legal documents."

- http://www.goodiff.org/
- running since March 2006

Background


Google™

Background


Gmail Calendar Documents Photos Groups Web more ▼
Google Mail BETA
Search Mail
Search the Web
Show search options
Create a filter
Compose Mail
« Back to Inbox
Archive
Report Spam
Delete
More actions...
Inbox (4)
Starred
Chats
Sent Mail
Drafts
All Mail
Spam
Trash
Contacts
Quick Contacts
Search, add, or invite
Michael G. Noll
Sign into chat
GooDiff @ HackCamp Hack.lu
Inbox
New window
Print all
"Michael G. Noll"
show details 2:49 pm (0 minutes ago)
Reply
Just for the screenshot...
http://your.secret.url/goes/here.html
Reply
Forward

Background

- private hyperlink in GMail
- visited by GoogleBot some days later
- no answer from Google

Scientific questions:

- “Whazzup??” (what happened?)
- “Pwned??” (allowed to do that?)

Background

Google Groups, April 2006

http://groups.google.lu/group/Gmail-ABCs/browse_thread/thread/51faf9a3586d2314/2bf9d67b12f64506

Background

- Step #1:
  analyze Google privacy policy etc.
- Step #2:
  changes in the meantime?

= no changelogs

= no archives*

= “you must inform yourself”

Looking around...

- most legal documents (privacy, ToS, etc) have no timestamp, no version, ...
- end user forced to inform himself about changes
- no tool available to help out

GooDiff
+++good
---evil

GooDiff to the rescue

- “Goo” + “Diff” = GooDiff

Components

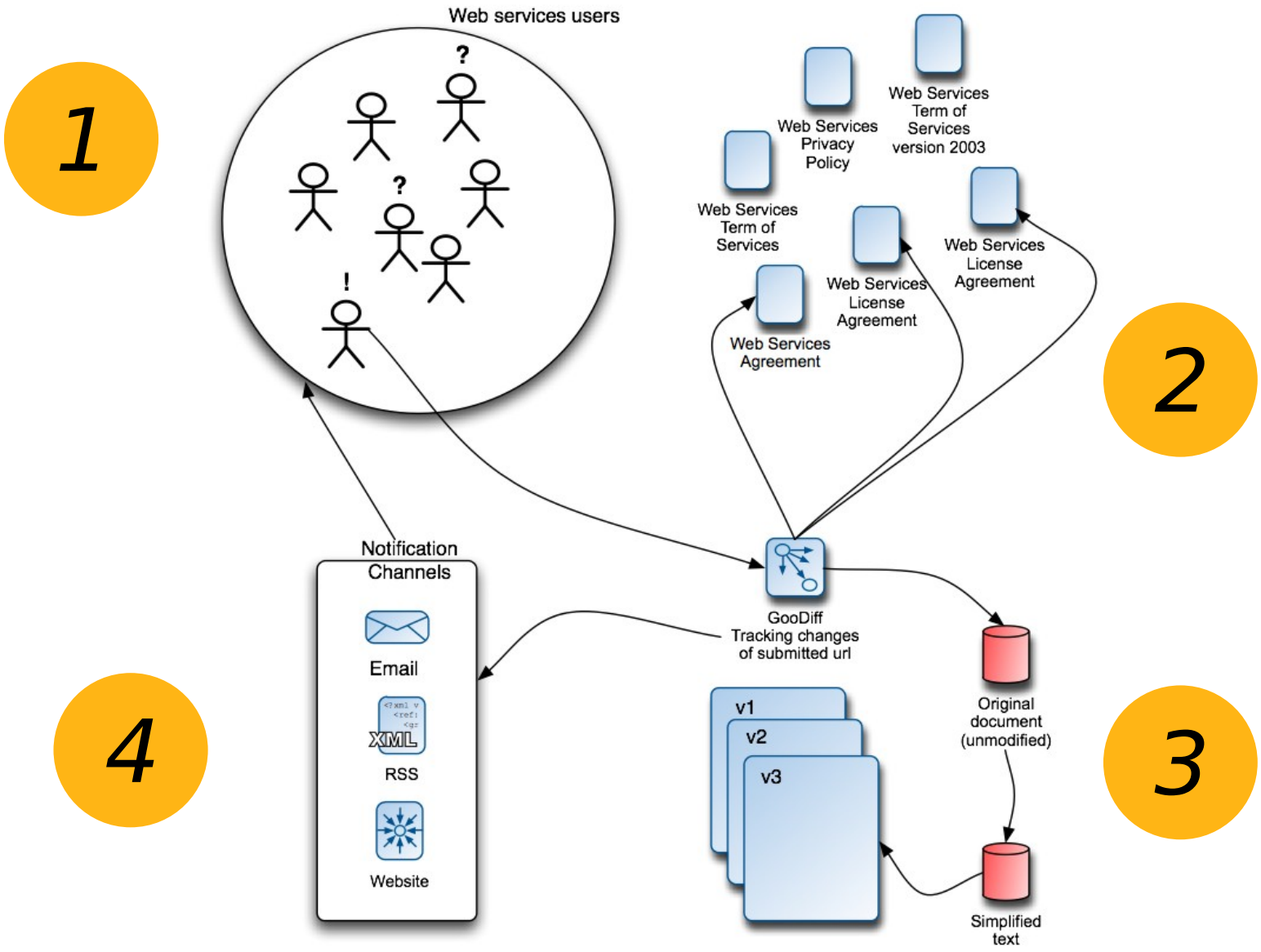
- monitor: fetch documents daily
- revision control: track changes
- UI: browsing, notifications

GooDiff to the rescue

- core code implemented in ~ 2 weeks (yes, full rewrite already done...)

- Python + SVN + Trac
  - Python: all cool stuff
  - SVN: 2x, for storing documents
  - Trac: GUI (we're laaaazy)

GooDiff :Raising public awareness of web services policies
by automatically tracking semantic changes



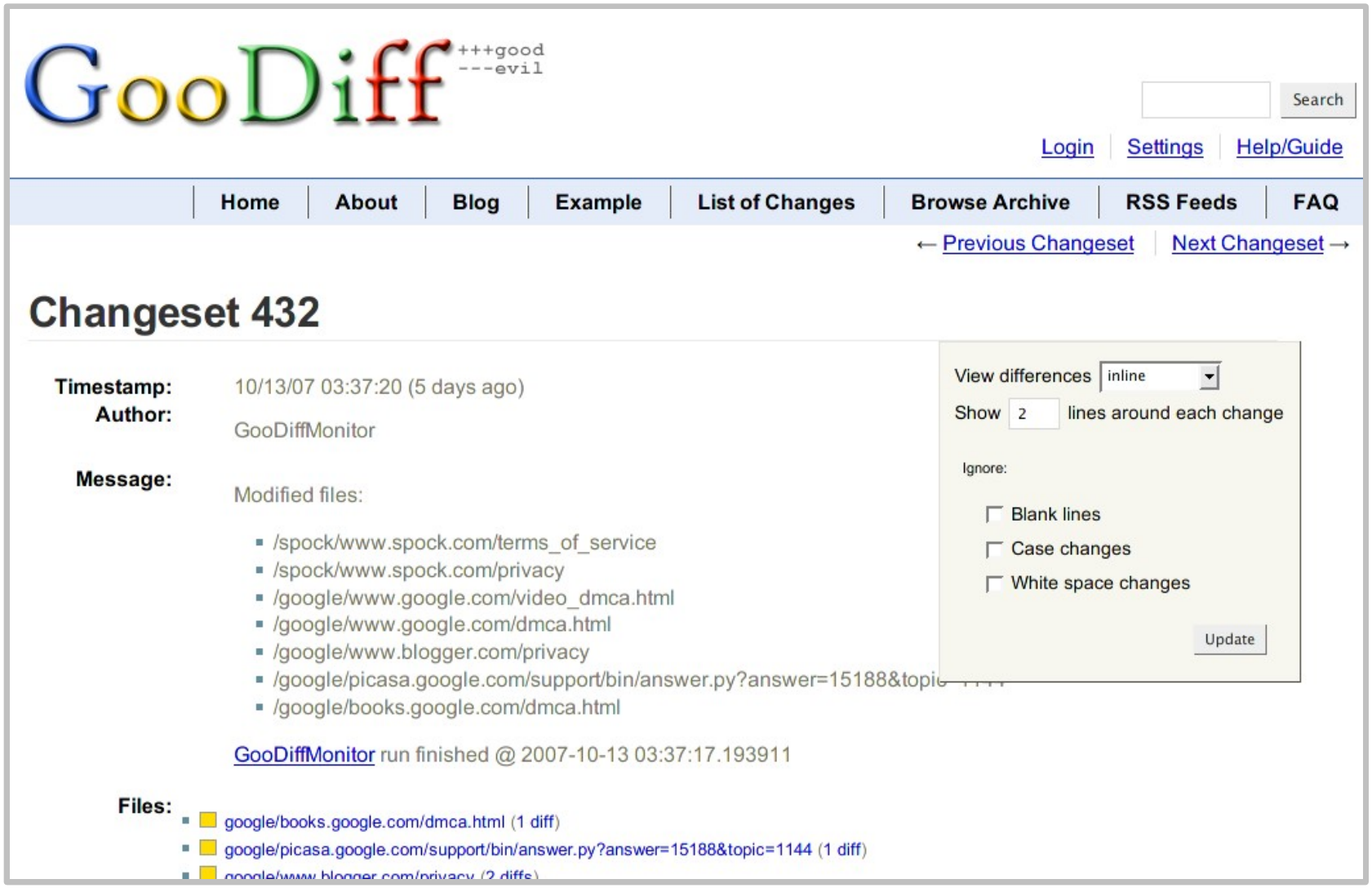
Configuration (sample)

```
<provider name="facebook">
        <service name="facebook">
            <document url="http://www.facebook.com/terms.php">
            <replace pattern='\?pwstdfy=[0-9a-fA-F]{32}' with="?pwstdfy=|REMOVED|" />
          </document>
        </service>
</provider>
```

What we're tracking

+ others...

Examples


GooDiff
+++good
---evil
Search
Login
Settings
Help/Guide
Home
About
Blog
Example
List of Changes
Browse Archive
RSS Feeds
FAQ
← Previous Changeset
Next Changeset →
Changeset 432
Timestamp: 10/13/07 03:37:20 (5 days ago)
Author: GooDiffMonitor
Message:
Modified files:
/spock/www.spock.com/terms_of_service
/spock/www.spock.com/privacy
/google/www.google.com/video_dmca.html
/google/www.google.com/dmca.html
/google/www.blogger.com/privacy
/google/picasa.google.com/support/bin/answer.py?answer=15188&topic
/google/books.google.com/dmca.html
GooDiffMonitor run finished @ 2007-10-13 03:37:17.193911
Files:
google/books.google.com/dmca.html (1 diff)
google/picasa.google.com/support/bin/answer.py?answer=15188&topic=1144 (1 diff)
View differences inline
Show 2 lines around each change
Ignore:
Blank lines
Case changes
White space changes
Update

Examples

- Google Picasa (changeset 28, 30-Mar-06)
- If you send a request to Google's servers, we record standard log information, including Internet Protocol addresses and information related to your request. We will ask before collecting personally identifying information from you and give you an opportunity to opt out.

Examples

- Google Picasa (changeset 28, 30-Mar-06)
- If you send a request to Google's servers, we record standard log information, including Internet Protocol addresses and information related to your request. We will ask before collecting personally identifying information from you and give you an opportunity to opt out.
- ...and information related to your request. **We also log information about the installation process when you download Picasa. We also log information about the installation process and your system and settings when you download Picasa.** We will ask before ...

(and yes, there have been two sentences added)

Examples

- Blogger.com (changeset 432, 13-Oct-07)
- Google-Server zeichnen automatisch Daten zu Ihrer Verwendung des Service auf, z. B. dazu, wann Sie Blogger verwenden und die Haufigkeit sowie die Grosse von Datentransfers. Informationen, die auf der Blogger-Oberflache angezeigt werden oder auf die geklickt wird (z. B. UI-Elemente, Einstellungen und andere Informationen), werden ebenfalls aufgezeichnet.

Examples

- Blogger.com (changeset 432, 13-Oct-07)
- Google-Server zeichnen automatisch Daten zu Ihrer Verwendung des Service auf, z. B. dazu, wann Sie Blogger verwenden und die Haufigkeit sowie die Grösse von Datentransfers. Informationen, die auf der Blogger-Oberflache angezeigt werden oder auf die geklickt wird (z. B. UI-Elemente, Einstellungen und andere Informationen), werden ebenfalls aufgezeichnet.
- ...werden ebenfalls aufgezeichnet. **If you are logged in we may associate that information with your account.**

(and yes, the new sentence is really in English)

Examples

- Apple
  - addition of iPhone to the portfolio
  - “+ ringtones”
- Google Finance
  - real-time courses instead of X mins delay

Yes, we also get notifications of product updates and other stuff...

Statistics & comments

- tracking stuff since ~ 18 months
- #changesets: 435 (raw), ~300 (real)
  - changeset: >=1 changed documents
- documents change more often than we would have expected
- timestamps IN documents (if present) are often not matching the time when we track a change

Theory vs. Practice

- how to handle 404, 301, ...?
  - “YouTube will be undergoing scheduled maintenance, starting around 7:00 pm PDT.“
- dynamic content:
  ads, web bugs, session IDs, ...
- how to rebuild DB from scratch while preserving change time etc.?
- <add more here>

Final remarks

- “We're not there yet!”
- can't spot things already in the docs
- user feedback for finding semantic changes
  - Web 2.0 with rounded corners!!11!1!!
- +awareness, +transparency
- feel free to send us your track requests
  - info@goodiff.org