Harvest FAQ

Kang-Jin Lee lee@arco.de

2003-11-08

Harvest frequently asked questions (FAQ) with answers

Contents

Chapter 1

Harvest

1.1 What is Harvest?

Harvest is a system to collect information and make them searchable using a web interface. Harvest can collect information on inter- and intranet using http, ftp, nntp as well as local files like data on harddisk, CDROM and file servers. Current list of supported formats in addition to HTML include TeX, DVI, PS, full text, mail, man pages, news, troff, WordPerfect, RTF, Microsoft Word/Excel, SGML, C sources and many more. Stubs for PDF support is included in Harvest and will use Xpdf or Acroread to process PDF files. Adding support for new format is easy due to Harvest's modular design.

1.2 Where can I get more information about Harvest?

See *Harvest homepage http://harvest.sourceforge.net/* for informations about Harvest.

1.3 Where can I download Harvest?

Harvest is available for download at *Harvest download page http://prdownloads.sourceforge.net/harvest/*.

1.4 Are there any information about Harvest in Russian?

Andrei Malashevich has translated the Harvest User's Manual to Russian. It is available at his *Harvest User's Manual page at http://baby.chg.ru/manual_harvest/*.

1.5 What is Harvest-ng?

Harvest-ng is a reimplementation of Harvest's gatherer by Simon Wilkinson. You can get more info about Harvest-ng at *Harvest-ng homepage http://webharvest.sourceforge.net/ng/*.

1.6 What is the copyright status of Harvest?

The core of Harvest located in *src* directory is under GPL. Additional components, located in *components* directory are under GPL or similar copyright.

1.7 Which Operating System do I need to run Harvest?

Harvest should run on any *nix like platforms including FreeBSD, Linux and Solaris.

1.8 Does Harvest run under Windows NT/2000/XP?

Michael Schlenker has ported Harvest to Windows platforms using *Cygwin* *http://sources.redhat.com/cygwin/*.

1.9 What Hardware do I need to use Harvest?

A Pentium 120MHz with 64MB RAM should achieve reasonable performance for around 350 MB of fulltext data in ca. 20.000 objects. A Pentium 650MHz with 256MB RAM should be able to handle around 1.5 GB of fulltext data in ca. 100.000 objects.

1.10 Which version of Harvest should I use?

- If you want to help developing Harvest, use the most recent version of Harvest.
- If you are cautious, a version older than a week should reasonably be safe to use.
- If you don't want to use development versions of Harvest, use the last version marked as stable.

1.11 What are "harvest-modified-by-RL-Stajsic", "harvest-MathNet", and "harvest-1.5.20-kj"?

After the original authors ceased working on Harvest, there were some periods where Harvest was unmaintained. During this time there were following forked versions of Harvest:

- "harvest-modified-by-RL-Stajsic" was released by R.L. Stajsic and Tim Samshuijzen with some bugfixes.
- "harvest-MathNet" is a modified version of Harvest-1.5.20 to improve the handling of German specical characters ("Umlaute", "scharfes S").
- "harvest-1.5.20-kj" series were released by me with bugfixes to Harvest 1.5.20.

All these forked trees were merged into Harvest 1.6.

1.12 What are the limits of Harvest?

- Harvest's Gatherer uses GDBM database to store the summarized data. On some architecture/OS, the maximum file size is 2 GB, so you can't have a database larger than 2 GB per Gatherer on those systems. To collect more data, you have to set up multiple Gatherers.
- The Broker stores the data as single files. On most OS, performance degrades noticeably with increasing number of files in a directory. Since the Broker uses finite number of directories defined in *src/broker/stor_man.c* to store the files, the broker will slow down with increasing number files.

1.13 Do I need root access to install and run Harvest?

For initial setup, you must be able to modify the webserver configuration and to schedule cron jobs. After the initial setup, it is recommended to run Harvest as a different user for security reasons.

1.14 How do I block Harvest from my site? How do I identify Harvest?

Put a line like this to your robots.txt:

```
User-agent: Harvest
Disallow: /
```

1.15 What can I do to help?

There are many ways to help depending your skills and time you want to contribute to improve Harvest:

- Use Harvest and let others know that you are using Harvest.
- Use Harvest and let me know why you are using Harvest.
- Submit ideas, feature requests and bug reports.
- Contribute localization.
- Contribute documentation.
- Contribute code.

Chapter 2

Building Harvest

2.1 How do I uninstall Harvest?

Harvest keeps all of its files in */usr/local/harvest* or whichever **prefix** you have assigned during `configure`. To uninstall Harvest, simply delete the Harvest directory.

If you did following when installing Harvest:

```
# ./configure --prefix=/home/data/harvest
```

then, this should uninstall Harvest:

```
# rm -fr /home/data/harvest
```

You might also want to check the start scripts which start Harvest daemons during system boot and remove `cron` jobs necessary for running Harvest.

2.2 Where can I get bison and flex?

Bison and flex are available at *GNU FTP Site* `<ftp://ftp.gnu.org/>` and its mirrors.

2.3 How can I install Harvest in "/my/directory/harvest" instead of "/usr/local/harvest"?

Do

```
# ./configure --prefix=/my/directory/harvest
# make
# make install
```

2.4 How can I avoid "syntax error before 'regoff_t'" error message when compiling Harvest?

On some systems, building Harvest may fail with following message:

```
Making all in util
gcc  -I../include -I./../include -c buffer.c
In file included from ../include/config.h:350,
                 from ../include/util.h:112,
                 from buffer.c:86:
/usr/include/regex.h:46: syntax error before ‘regoff_t’
/usr/include/regex.h:46: warning: data definition has no type or storage class
/usr/include/regex.h:56: syntax error before ‘regoff_t’
*** Error code 1
```

If you get this error, edit *src/common/include/autoconf.h* and add "#define USE_GNU_REGEX 1" before typing `make` to build Harvest.

2.5 Where can I get more information for building Harvest on FreeBSD?

See *FreshPorts Harvest page http://www.freshports.org/www/harvest/* for more informations about building Harvest on FreeBSD.

Chapter 3

Gatherer

3.1 Does the Gatherer support cookies?

No, Harvest's Gatherer doesn't support cookies.

3.2 Why doesn't Local-Mapping work?

In Harvest 1.7.7, the default HTML enumerator was switched from `httpenum-depth` to `httpenum-breadth`. The breadth first enumerator had a bug in **Local-Mapping**, which was fixed in Harvest 1.7.19. To make **Local-Mapping** work, use depth first enumerator or update to Harvest 1.7.19 or later.

Local mapping will fail if the file is not readable by the gatherer process, or the file is not a regular file, or the file has execute bits set, or the filename contains characters that have to be escaped (like tilde, space, curly brace, quote, etc). So, for directories, symbolic links and cgi scripts, the gatherer will always contact the server instead of using local file.

3.3 Does the Gatherer gather the Root- and LeafNode-URLs periodically?

No, the Gatherer gathers Root- and LeafNode URLs only once. To check the URLs periodically, you have to use cron (see "man 8 cron") to run `$HARVEST_HOME/gatherers/YOUR_GATHERER/RunGatherer`.

3.4 Can Harvest gather https URLs?

No, https is not supported by Harvest. To gather https URLs, use Harvest-ng from Simon Wilkinson. It is available at *Harvest-ng homepage http://webharvest.sourceforge.net/ng/*.

3.5 When will Harvest be able to gather https URLs?

This is not on top of my to-do list and may take some time.

3.6 Does Harvest support client based scripting/plugin like Javascript, Flash?

No, Harvest's gatherer does not support Javascript, Flash, etc., and there are no plans to add support for them.

3.7 Why does the gatherer stop after gathering few pages?

Harvest's gatherer doesn't support Javascript, Flash, etc. Check the site you want to gather and make sure that the site is browsable without any plugins, Javascript, etc.

3.8 How can I index local newsgroups? How can I put hostname into News URL?

You will find a News URL hostname patch by Collin Smith in the *contrib* directory.

NOTE: Even though most web browsers support this, this violates RFC-1738.

3.9 What do the gatherer options "Search=Breadth" and "Search=Depth" do and which keywords are available for "Search=" option?

Search option selects an enumerator for http and gopher URLs. Harvest comes with breadth first (Search=Breadth) and depth first (Search=Depth) enumerator for http and gopher. They have different strategy when following the URLs to get a list of candidates for processing. The breadth first enumerator processes all links in a level before descending to next level. In case of limiting the number of URLs to gather from a site, it will give you a more representative overview of the site. The depth first enumerator will descend to next level as soon as possible. When there are no links left for the current branch, it will process the next branch. The depth first enumerator doesn't use as much memory as the breadth first enumerator. If you don't have compelling reasons to switch from an enumerator to the other, the default value should be a reasonable choice.

3.10 How can I index html pages generated by cgi scripts? How can I index URLs which has a "?" (question mark) in it?

Remove *HTTP-Query* from *$HARVEST_HOME/lib/gatherer/stoplist.cf* and *$HARVEST_HOME/gatherers/YOUR_GATHERER/lib/stoplist.cf*. For versions earlier than 1.7.5, you also have to create a (symbolic) link from `$HARVEST_HOME/lib/gatherer/HTML.sum` to `$HARVEST_HOME/lib/gatherer/HTTP-Query.sum`. To do this, type:

```
# cd $HARVEST_HOME/lib/gatherer
# ln -s HTML.sum HTTP-Query.sum
```

3.11 Why is the gatherer so slow? How can I make it faster?

The gatherer's default setting is to sleep one second after retrieving an URL. This is to avoid an overload of the webserver. If you gather from webservers under your control and know that they can handle the additional load caused by the gatherer add "Delay=0" in your root node specification to disable the sleep.

The lines should look like:

```
<RootNodes>
http://www.SOMESERVER.com/ Search=Breadth Delay=0
</RootNodes>
```

Alternatively, you can set the delay value for all root nodes by adding **Acces-Delay: 0** in your configuration file.

It should look like:

```
Gatherer-Name:  YOUR Gatherer
Gatherer-Port:  8500
Top-Directory:  /HARVEST_DIR/work1/gatherers/testgather
Access-Delay:   0

<RootNodes>
http://www.MYSITE.com/ Search=Breadth
</RootNodes>
```

3.12 Why is the gatherer still so slow?

Harvest's gatherer is designed to handle many types of documents and many types of protocols. To achieve this flexibility it uses external programs to handle the different types of documents and protocols. For example, when gathering HTML documents via HTTP, the document is parsed twice. First to get list of candidates to gather and then to get a summary of the document. The summarizer is started each time when a document arrives, quits after summarizing that document and has to be restarted for the next document. Compared to more HTTP/HTML oriented approaches this causes a significant overhead when gathering HTTP/HTML only.

Harvest retrieves one document at a time which causes slowdown if you encounter a slow site. Due to implementation, the Gathering process is quite heavyweight and uses up to 25 MB of RAM per Gatherer. For this reason, there were no attempts to spawn more gatherers to optimize the bandwidth usage.

3.13 How do I request "304 Not Modified" answers from HTTP servers?

To send "Last Modified: xx" headers and get "304 Not Modified" answers from HTTP servers, add following line to the gatherer's configuration file:

```
HTTP-If-Modified-Since: Yes
```

If the document hasn't changed since last gathering, the gatherer will use the data from its database, instead of retrieving it again. This will save bandwidth and speed up gathering significantly.

3.14 Why does Harvest gather different URLs between gatherings?

When **HTTP-If-Modified-Since** is enabled, the candidate selection scheme of the http enumerators will change for successful database lookups. For unchanged URLs, the enumerators will behave more like depth first gatherer. The result of the gatherings should be the same if you are gathering all URLs of a site, but if you gather only parts of a site by using **URL=n** with **n** < **number of URLs of a site** you will get different subset of the system you gather.

3.15 Why has the Gatherer's database vanished after gathering?

The Gatherer uses GDBM databases to store its data on disk. Database files for Gatherer can grow very large depending on how much data you gather. On some systems, (e.g. i386 based Linux) the maximum file size is 2GB. If the amount of data surpasses this limit, the GDBM database file will be wiped from the disk.

3.16 How can I avoid GDBM files growing very big during Gathering?

The Gatherer's temporary GDMB database file *WORKING.gdbm* will grow very rapidly when gathering nested objects like tar, tar.gz, zip etc. archives. GDBM databases keep growing when tuples are inserted and deleted from them, because GDBM reuses only fractions of the empty filespace. To get rid of unused space, the GDBM database has to be reorganized. The reorganization however is slow and will slow down the gathering, so the default is not to reorganize the gatherer's temporary database. This should work well for small to medium sized Gatherers, but for large Gatherers it may be necessary to reorganize the temporary database during gathering to keep the size of the database at manageable level. To reorganize the *WORKING.gdbm* every 100 deletions add following line to your gatherer configuration file:

```
Essence-Options: --max-deletions 100
```

Don't set this value too low, since it will consume significant share of CPU time and disk I/O. Reorganizing every 10 to 100 deletions seems to be a reasonable value.

3.17 Can I use Htdig as Gatherer? Can the Broker import data from Htdig?

The perl module *Metadata* from Dave Beckett can dump data from Htdig database into a SOIF stream. Metadata only supports GDBM databases, so this only works with versions earlier than Htdig 3.1, because newer versions of Htdig switched from GDBM to Sleepycat's Berkeley DB.

3.18 How can I control access to Gatherer's database?

Edit *$HARVEST_HOME/gatherers/YOUR_GATHERER/data/gatherd.cf* to allow or deny access. A line that begins with **Allow** is followed by any number of domain or host names that are allowed to connect to the Gatherer. If the word **all** is used, then all hosts are matched. **Deny** is the opposite of **Allow**. The following example will only allow hosts in the **cs.colorado.edu** or **usc.edu** domain access the Gatherer's database:

```
Allow  cs.colorado.edu usc.edu
Deny   all
```

3.19 Does Harvest's Gatherer support WAP/WML, Gnutella, Napster?

No. Harvest's Gatherer doesn't support WAP. Peer to peer services like Gnutella, Napster, etc. are also unsupported.

3.20 How do I gather ftp URLs from wu-ftp daemons?

Changes in wu-ftpd 2.6.x broke `ftpget`. There is a replacement for it in contrib directory which wraps any ftp client to behave like `ftpget`.

3.21 Why doesn't file URLs in LeafNodes work as expected?

File URLs pointing to directories like *file://misc/documents/* in LeafNodes are considered as nested object which will be unnested.

3.22 Why does gathering from a site fail completely or for parts of the site?

This may be caused by the site's *robots.txt*. You can check this by typing "http://www.SOME.SITE.com/robots.txt" into your favourite web browser.

Chapter 4

Summarizer

4.1 Why doesn't Post-Summarizing work?

The most common error is that the instructions are indented by spaces instead of a tab-stop. Check the **Post-Summarizing** rule file and make sure that instructions are indented by a tab-stop. The **Post-Summarizing** rule file uses a syntax like in *Makefile*. Conditions begin in the first column and instructions are indented by a tab-stop.

4.2 How can I summarize meta tags in HTML documents?

In Harvest 1.5.20.kj-0.3, the default summarizer for HTML data was switched to `HTML-lax.sum` which does not handle meta tags. Edit `$HARVEST_HOME/lib/gatherer/HTML.sum` and uncomment the SGML or Perl based summarizer.

4.3 Why are raw HTML tags in some query results?

If you see raw HTML tags in query results, the HTML summarizer was not able to parse the page correctly. Harvest comes with three different summarizers for HTML. If the default summarizer fails try the other two summarizers. To do this, edit `$HARVEST_HOME/lib/gatherer/HTML.sum` and uncomment one of the summarizers.

4.4 How can I summarize DVI files?

Use Harvest older than 1.5.20-kj-0.8 or newer than 1.7.2. The versions between these two versions have a bug which prevents DVI files being summarized.

4.5 How can I summarize Pdf files?

You need *xpdf* to summarize Pdf files. Harvest uses `pdftotext` from *xpdf* to summarize Pdf files.

Alternatively, you can use `acroread` to convert Pdf files to Postscript and pass it to Postscript summarizer. To do this, edit `$HARVEST_HOME/lib/gatherer/Pdf.sum` accordingly.

4.6 Where can I get pdftotext?

`pdftotext` is part of *xpdf*. It is available at *Xpdf homepage http://www.foolabs.com/xpdf/*.

4.7 How can I improve summarizer for Microsoft Word files?

Harvest uses *catdoc* to summarize Microsoft Word files. If you get bad summaries for Microsoft Word files, you might want to try `wvHtml`, which is part of *wvWare*, instead of *catdoc*.

4.8 Where can I get wvWare?

wvWare is available at *wvWare homepage http://www.wvware.com/*.

4.9 How can I add support for new file type?

Give the new file type a name and make Harvest know how to recognize the new file type by modifying *byname.cf* (to determine filetype by its name), *byurl.cf* (to determine filetype by the URL), or *magic* and *bycontent.cf* (to determine filetype by looking at the content of the file). You will find *bycontent.cf*, *byname.cf*, *byurl.cf* and *magic* in your *$HARVEST_HOME/lib/gatherer/* directory.

Create a summarizer (a programm or script) which takes the filename as first argument and prints a SOIF stream "Attributename{length of data}:<`tab`>your data" to stdout. For file type "Xyz", you have to create a summarizer called `Xyz.sum` in the *$HARVEST_HOME/lib/gatherer/* directory.

In most of the cases it might be easiest to convert filetype "Xyz" to a supported filetype like HTML, PostScript, etc. and use an existing summarizer on the converted file.

4.10 How can I use nsgmls instead of sgmls to summarize documents?

Edit `$HARVEST_HOME/lib/gatherer/SGML.sum` and set **$sgmls_cmd** = **"/usr/local/bin/nsgmls"** or where ever you have installed nsgmls.

Chapter 5

Broker

5.1 How can I start a Broker at boot time?

Some user contributed startup scripts are located in *contrib/etc/* directory of Harvest source distribution. Modify apropriate files and copy them to your startup script directory.

5.2 How can I start a Broker without starting a collection?

When a Broker starts, it starts collecting data, which can take some time. To avoid this, use the **-nocol** option when invoking `RunBroker`.

If you have installed Harvest in */usr/local/harvest/*, put following line into your startup file, e.g. `/etc/rc.local`:

```
/usr/local/harvest/brokers/YOUR_BROKER/RunBroker -nocol
```

Replace */usr/local/harvest/* with the directory where you have installed Harvest.

5.3 Why don't the documents which I have gathered right now show up in the Broker?

The Broker imports data from the Gatherer once in every 24 hours. If you want to import the data immediately after gathering, just restart the Broker or signal the Broker to import data.

You can signal the broker with the command line client `brkclient`, located in *$HARVEST_HOME/lib/broker/* by typing:

```
# brkclient localhost 8501 '#ADMIN #Password secret #collection'
```

Replace hostname, port and password if necessary.

Other easier method is to use the WWW based admin interface at: "http://www.YOUR_SERVER.com/Harvest/brokers/YOUR_BROKER/admin/admin.html".

5.4 Why do I get error messages when I try to access "http://some.host/Harvest/brokers/your-broker-path/" after running $HARVEST_HOME/RunHarvest?

Check the error log of your http daemon. The http daemon must be able to follow symbolic links. For apache httpd you can do this by adding:

```
<Location /Harvest/brokers/your-broker-path/>
        Options FollowSymLinks
</Location>
```

to your *httpd.conf*.

If you don't want symbolic links, delete the symbolic link and copy the file to the new name.

5.5 Why are NEWS URLs broken? Where are the hostnames in NEWS URLs? How can I follow NEWS URLs?

Harvest's Gatherer doesn't put hostnames into NEWS URLs. If your web browser complains about missing news server, configure your web browser to use the news server of your provider, company or organization as your default news server.

For more information why Harvest doesn't put hostnames into NEWS URLs, see RFC-1738 chapter 3.6 and 3.7.

5.6 Why don't I get any results if I use a long or complex query string?

The length of a query string is limited to 30 characters when using regluar expressions (wildcards), excluding the escape characters.

5.7 Can I use wildcards in attribute value for structured queries?

No, regular expressions for attribute names and attribute values in structured queries aren't supported. So, queries like "Author: Smi.*" or "Auth.*: Smith" won't do what you might expect.

5.8 Are the attribute names case sensitive?

No, the attribute names are not case sensitiv. So, "Time-To-Live" is the same like "Time-to-Live", "Time-to-live", "time-to-live", etc.

5.9 Why doesn't collecting from broker work?

This is due to a bug introduced in Harvest 1.5.18. The bug was fixed in 1.7.8. To make it work again, update to 1.7.8 or higher.

5.10 How can I customize the Harvest user interface?

The query pages are located in *$HARVEST_HOME/brokers/YOUR_BROKER/query-**. Most likely, you don't want to make all the variables visible to users who want to query your broker. Edit *query-** and use the **hidden** type to set suitable defaults for variables you want to hide.

The result set presentation can be customized by choosing or modifying the configuration files located in *$HARVEST_HOME/cgi-bin/lib/* directory. The configuration files *Sample.cf, classic.cf, modern.cf* and some *LANGUAGE.cf* are already installed in *$HARVEST_HOME/cgi-bin/lib/* directory. You can either create a new configuration file or modify one of th configuration files to get the result set presentation you want. See the Harvest User's Manual for information about available options for the configuration file.

If you want to customize the result presentation even further, then edit `$HARVEST_HOME/cgi-bin/search.cgi`.

5.11 How do I localize/translate user interface?

To localize the user interface, do:

1. Create *src/broker/example/brokers/skeleton/query-glimpse-modern.html.xx.in*, where *xx* is a two letter abbreviation for your language/country, by translating either *query-glimpse-modern.html.in* or other *query-glimpse-modern.html.yy.in*. This is the localized query page.
2. Create *components/broker/standard/WWW/language.cf* by translating *modern.cf* or other translated configuration file like *spanish.cf*, *german.cf*, etc. This will localize the result pages and error messages.
3. Create *src/broker/example/brokers/skeleton/query-glimpse.html.xx.in* by translating *query-glimpse.html.in* or *query-glimpse.html.yy.in*. This is the advanced query page.
4. Translate *src/broker/example/brokers/*.html* to get localized additional help pages.

5.12 How can I replace the bundled Glimpse with an other version of Glimpse?

Edit *$HARVEST_HOME/brokers/YOUR_BROKER/admin/broker.conf* to let Harvest know the location of your `glimpse`, `glimpseindex`, and `glimpseserver`.

Chapter 6

Terms

6.1 What is a Gatherer?

A Gatherer is a system that retrieves documents from various sources (Web-, News-, FTP-server, local files) for processing. In HTML/HTTP context, it is also often called *crawler*, *robot*, or *spider*.

6.2 What is Local-Mapping?

To reduce the CPU load and speed up Gathering, Harvest can map local files to URLs. The gatherer can bypass the server and use local file, while pretending that the objects were gatherered as usual to the rest of the Harvest system.

6.3 What is a Summarizer?

A Summarizer transforms a document into a form which is more suitable for fulltext searching.

The HTML summarizer for example, extracts the title of a document, removes all HTML tags, generates a wordlist, etc.

6.4 What is a Broker?

A Broker processes search requests received from a user by a cgi-script and presents the search results.

Chapter 7

Miscellaneous

7.1 Who are the maintainers of Harvest?

Kang-Jin Lee `lee@arco.de` and Harald Weinreich `harald@weinreichs.de` are maintaining Harvest.

7.2 I have found a bug. What should I do?

Post a bug report to the newsgroup comp.infosystems.harvest or mail it to Kang-Jin Lee `lee@arco.de` and Harald Weinreich `harald@weinreichs.de`.

7.3 Is there a mailinglist for Harvest? What about a newsgroup?

There is a *Harvest developer's mailinglist http://lists.sourceforge.net/lists/listinfo/harvest-devel/* for Harvest users and developers. There also is a *Harvest newsgroup news:comp.infosystems.harvest* `<news:comp.infosystems.harvest>`.