An effective synchronization method of GSI Tiles using a GSI Tile List

Hidenori FUJIMURA*

(Published online: 27 October 2015)

Abstract

"GSI Tiles" are tile-based geospatial information for web maps provided by the Geospatial Information Authority of Japan (GSI) under the standardized terms of use of the Japanese Government. GSI Tiles contain official topographic map data maintained under the Survey Act, and also more than 1,000 layers, covering topographic maps, aerial and satellite images, thematic maps, and disaster information. While these data are typically used for web maps such as the "GSI Maps" run by GSI, they are also available for download, and it has been decided that the same server which is used to run "GSI Maps" is to be used to provide these data for download. To keep the server highly available, an effective synchronization method for GSI Tiles has been developed.

To fulfill this purpose, the concept of "GSI Tile Lists" was proposed. A GSI Tile List contains metadata for the files for a layer of GSI tiles. File paths, modification times (mtime), size, and MD5 message-digests are recorded in a GSI Tile List. The download and synchronization of the GSI Tiles can be made more effective by using a GSI Tile List. GST Tile Lists are updated weekly to reflect the rapid update of the topographic data of GSI. A reference implementation of the GSI Tiles downloader using GSI Tile Lists (qdltc) has been developed and published via a social coding site. A layer of topographic map data covering the whole area of Japan, consisting of 50 million tiles, can be downloaded within around a week. A weekly update of these topographic data, consisting of tens of thousands of tiles, can be downloaded within around 7 hours.

1. Background

The Geospatial Information Authority of Japan (GSI) publishes more than a thousand layers of tiled geospatial data called "GSI Tiles". GSI Tiles contains the latest products of the topographic surveys conducted since the 1880s. GSI Tiles also contain orthophotos, thematic maps, and disaster information.

The Information Access Division also operates a web map of GSI called "GSI Maps" (Geospatial Information Authority of Japan, 2015a). GSI Maps functions as a showcase of the application of GSI Tiles. GSI Maps has been continuously in operation for more than 12 years, if we take the preceding service, Digital Japan Portal, started on July 15, 2003, into account. The emphasis is always on availability of the service, aiming at 24-hour 365-day operation from the beginning.

GSI provides a view of Japan to the nation of Japan and also to the globe, by providing GSI Tiles and GSI Maps.

As of July 2015, the Information Access Division considers the following three policies important for promoting the use of GSI Tiles: open data policy, open source policy, and open innovation policy.

The open data policy promotes wider use of the GSI Tiles by the adoption of terms of use which are easier to understand. GSI has enacted the GSI Contents Terms of Use following the Japanese Government Standard Terms of Use (version 1.0) on September 30, 2014. The terms of use for the GSI Tiles are updated by the GSI Contents Terms of Use on the very day of enacting the terms of use.

Because the terms of use cover only the copyright, naturally the procedure enacted in the Survey Act is

* Director, Information Access Division, Geospatial Information Department

required. The Information Access Division keeps contact with developers who use GSI Tiles to lower the difficulty of how they follow the procedure.

The open source policy means the use of open source software and the release of open source software in GSI Maps (Geospatial Information Authority of Japan, 2015b). The aim of the policy is to share functions and technologies fundamentally necessary for the use of GSI Tiles, and raise the minimum technological expertise of the industry.

Open innovation policy is to pursue innovative results by related parties by seeking collaboration with external bodies. We see the idea of open innovation is important especially because there are a variety of applications of geospatial information. A set of specialized skills is necessary for each field of application. Collaboration with parties for each field of application is necessary.

In order to realize an open innovation policy, the functions of GSI Maps have gradually been modularized, taking the various existing standards into account. With modularization, parties can take some modules for their application, or they can replace some modules for their application. One example of such modularization is the adoption of slippy map tilenames (XYZ) on 30 October 2013. By adopting a well-used tile-naming standard, more open competition for software can be initiated. In other words, the applicability of the GSI Tiles is widened. The number of desktop applications and mobile applications got significantly bigger. Another benefit of adopting slippy map tilenames is the successful migration from OpenLayers 2 to Leaflet in GSI Maps on 8 July, 2015.

The Information Access Division has also operated the “GSI Maps Partner Network” since July 11, 2014. The network consists of developers who use or plan to use GSI Tiles for their work. 121 parties have already joined as of 30 June, 2015. They contribute to the “GSI Maps Partner List”, a collection of information about the application of GSI Tiles. In order to promote sharing the latest information, a conference for the GSI Maps Partner Network has been held three times, in November 2014, February 2015 and June 2015.

2. Challenges

The standard map of GSI Tiles was officially announced as an authoritative basic survey result on 1 July, 2014. GSI takes responsibility for keeping these survey results up to date, and takes responsibility for providing the data to the public. The server for GSI Maps remains the sole channel for providing GSI Tiles to the public. This has led to an urgent need to provide GSI Tiles more effectively, especially for usage where bulk downloading and/or frequent updates are required.

There were risks to the availability if we did not take any action to provide ways to download and synchronize GSI Tiles data.

2.1 Technical details

While the map-tiling method is a cost-effective way to access data at a specified geographic location, the way to download a tile-set of some volume is not very obvious.

A list of tile files with metadata for timeliness or uniqueness would be useful for downloading such data more effectively. We call such a list the “GSI Tile List”.

There are more than 50 million tile files for a single layer of a standard topographic map. The total number of GSI Tiles is about 200 million.

We use a content delivery network (CDN) to provide faster access to the tile data. It was agreed that to be cost-effective, the same CDN must be utilized for bulk download and rapid synchronization.

It was agreed that the frequency of the synchronization must be up to users at the downloading side, because timeliness requirements for applications will vary.

Because the announcement has already been made, it was agreed that the release of the method must be as soon as possible, especially for developers. For this reason, it was agreed that the data production process and data uploading process must not change for the development of the method.

3. GSI Tile List

3.1 The reason GSI Tile List is developed

The GSI Tile List is a CSV file which contains metadata for each tile file stored on one layer at the server.

Japan consists of a long archipelago. The HTTP access log for the tile server indicates that there have been a lot of requests for sea areas where no tile data exists. In the case of the Japan area, a "brute force" attempt to download the tile data is not effective. This ineffectiveness is considered harmful to both sides of the HTTP request. This is why the GSI Tile List was developed.

3.2 Specifications for GSI Tile List

When the template URL of the tile-set is http://server/t/{z}/{x}/{y}.ext , the URL of the GSI Tile List must be http://server/t/mokuroku.csv.gz as default. The format of the data is gzipped CSV.

As of July 2015, the format of the CSV data is:

[path],[modification date],[size],[MD5 message-digest]

The part for "{x}/{z}/{y}.ext" of the URL is defined as [path] field for the CSV data. The URL of the tile data would be http://server/{t}/[path].

The [modification date] is an integer number which represents UNIX epoch time. The unit for [size] is bytes.

The [MD5 message-digest] is the MD5 hash value of the specified tile data (Rivest, 1992). The reason for using the MD5 message-digest is to skip "phony updates," updates where only the modification time has changed, not the content. This occurs rather frequently because the rendering of the tiles runs for a specified area where only several parts are actually modified.

These specifications are available at a social coding site (Geospatial Information Authority of Japan, 2015c).

The size of mokuroku.csv.gz for the standard topographic map data (around 50 million tiles, around 350GB) is around 1.2GB when packed in a csv.gz file.

3.3 mokuroku generator

The tool to generate mokuroku.csv.gz at the server side is also released at a social coding site (Geospatial Information Authority of Japan, 2015d). This tool takes layers*.txt file (Geospatial Information Authority of Japan, 2014) on the server and scans the file system to generate mokuroku.csv.gz for each layer.

3.4 A viewer for GSI Tile List

A viewer for GSI Tile List, sl.rb, is available at the social coding site (Geospatial Information Authority of Japan, 2015e). The following is an example of viewing the current status of the GSI Tile List in the server.

```
$ ruby sl.rb
◯ ,std,2015-07-19 12:17:36,1.21 GB, 標準地図
◯ ,pale,2015-07-14 12:13:23,1.21 GB, 淡色地図
◯ ,blank,2015-07-14 13:15:27,9.12 MB, 白地図
◯ ,english,2015-07-14 13:25:46,602 KB,English
...
```

The following is an example to get a list of URLs of the GSI Tile List.

```
$ ruby sl.rb --mokuroku_urls
http://cyberjapandata.gsi.go.jp/xyz/std/mokuroku.csv.gz
http://cyberjapandata.gsi.go.jp/xyz/pale/mokuroku.csv.gz
http://cyberjapandata.gsi.go.jp/xyz/blank/mokuroku.csv.gz
http://cyberjapandata.gsi.go.jp/xyz/english/mokuroku.csv.gz
...
```

4. qdltc – a reference implementation of a downloader

An effective synchronization method of GSI Tiles using a GSI Tile List is proposed in this section. A reference implementation of this method, qdltc (Queued DownLoader with Timeline backup and MD5 message-digest Cache) is available as open source software (Geospatial Information Authority of Japan, 2015f).

4.1 Filter tiles by checking MD message-digest

When there is a local copy of the tile specified in a line of mokuroku.csv.gz, the MD5 message-digest of the local copy is calculated. If the MD5 message-digest of the local copy is the same as the MD5 message-digest written in the mokuroku.csv.gz, nothing needs to be done and thus

the line of the mokuroku.csv.gz is skipped (Fig-1).

By checking not the modification time (mtime) but the MD5 message-digest, we can also skip "phony updates" which constitute a significant portion of the update, at least under the current production process of GSI Tiles.

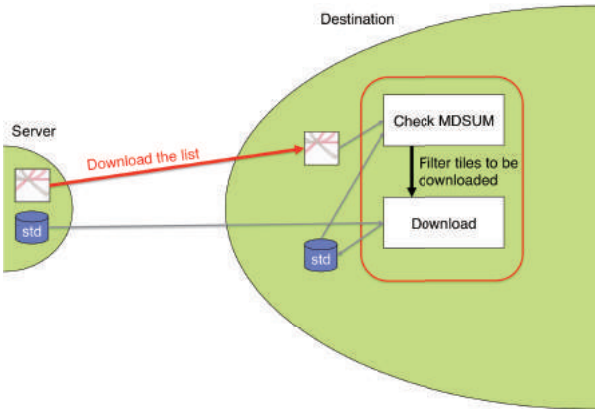

Fig. 1 Filter tiles by checking MD5 message-digest

4.2 Use of queue

Parallel and asynchronous HTTP requests are commonly used in web browsers. To take the same measure within this tool, the process to check MD5 message-digest and the process to download and write the tile data are separated as different threads. A queue is inserted to connect the MDS5SUM check thread and parallel download threads (Fig-2). The default number of download threads is set to 8, and the depth of the queue is set to 200. Where the Internet connection is not very good, the number of download threads can be set bigger, such as 32. The depth of the queue can be several tens of thousands, especially for less significant updates.

By introducing a queue, the performance of the tools got a few times faster.

4.3 Backup of old files

The process to backup old local copy is introduced. The old local copy {z}/{x}/{y}.ext is moved to bak/{z}/{x}/{y}.{yymmdd}.ext, where the modification date of the old local copy is {yyyymmdd}, so backups of old local copies will be accumulated.

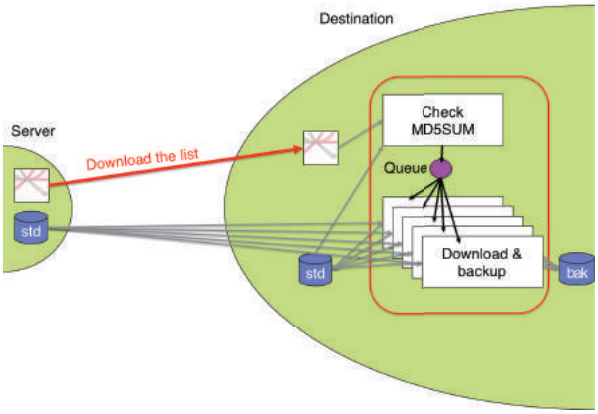

Fig. 3 Backup of old files

4.4 Caching local MD5 message-digest

The first step of the synchronization is to calculate the MD5 message-digest of the local copy. This is performed for tile files which have not been updated, which comprises the majority of the tile-set. To eliminate this, caching of the local MD5 message-digest

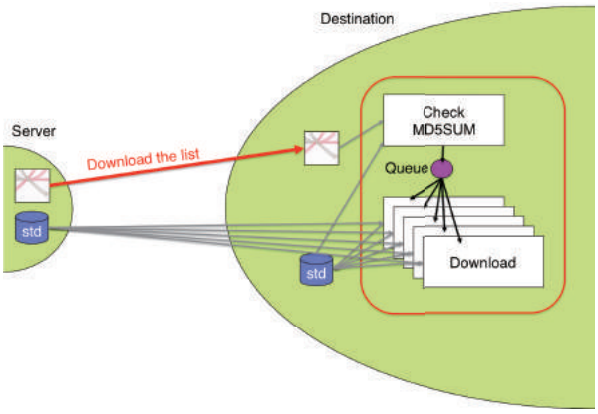

Fig. 2 Use of queue

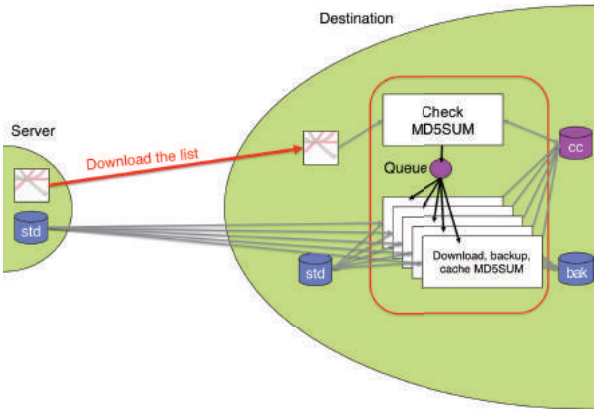

Fig. 4 Caching local MD5 message-digest

was introduced. The modified first step is to compare the cached local MD5 message-digest and MD5 message-digest in mokuroku.csv.gz. The calculation of MD5 message-digest for the local copy is performed only when the first comparison fails. By introducing this, the synchronization process becomes a few times faster.

4.5 Performance

The performance of qdltc depends mainly on the quality of the Internet connection and the performance of the hard disk drive where the tiles are stored.

It takes about a week to download a whole tile-set of the standard topographic maps, and it takes about 3 to 9 hours to reflect weekly updates with a typical consumer-level PC. The performance is known to be significantly better under a server-level environment. The performance of the anti-virus software sometimes has effects on the performance of the software.

5. Future Directions

The idea of differential updates by using the GSI Tile List can be applied for not only downloading but also uploading of the tiles. Use of the idea of the GSI Tile List in the uploading process may lead to more rapid updates of the served data.

The backup data stored by qdltc can be used as an archive of the GSI Tiles. Ways to visualize and access the backup need to be developed. Simultaneously, the specification of the modification time (mtime) of the tile-file needs refinement. Possibly the modification time of the tile-file should be the official publication time of the geospatial information.

By taking these measures, it is preferable to consider tile data as a key currency of geospatial information, working online and offline across web maps, desktop GIS, mobile applications, and tablet applications.

It is worth noting that the proposed synchronization method is independent of the contents of the tile-set. This means the method is applicable for vector tiles or tiles from other sources. The Information Access Division would like to modularize the implementation of the proposed method as much as possible, and contribute to open innovation in the application of geospatial information.

References

Geospatial Information Authority of Japan(2014): Convention of layers.txt – definition file for layers for web maps - , https://github.com/gsi-cyberjapan/layers-dot-txt-spec, (accessed 31 Jul. 2015) (in Japanese)

Geospatial Information Authority of Japan(2015a): GSI Maps, http://maps.gsi.go.jp/, (accessed 31 Jul. 2015) (in Japanese)

Geospatial Information Authority of Japan(2015b): gsimaps (GSI Maps), https://github.com/gsi-cyberjapan/gsimaps, (accessed 31 Jul. 2015) (in Japanese, partially in English)

Geospatial Information Authority of Japan(2015c): Specifications for GSI Tile List, https://github.com/gsi-cyberjapan/mokuroku-spec, (accessed 31 Jul. 2015) (in Japanese)

Geospatial Information Authority of Japan(2015d): GSI Tile List Generator, https://github.com/gsi-cyberjapan/mokuroku-generator, (accessed 31 Jul. 2015) (in Japanese)

Geospatial Information Authority of Japan(2015e): Supplied layers from layers*.txt, for Ruby, https://github.com/gsi-cyberjapan/sl, (accessed 31 Jul. 2015) (in Japanese)

Geospatial Information Authority of Japan(2015f): A reference implementation of a downloader using GSI Tile List, https://github.com/gsi-cyberjapan/qdltc, (accessed 31 Jul. 2015) (in Japanese)

Rivest, R. (1992): The MD5 Message-Digest Algorithm, https://www.ietf.org/rfc/rfc1321.txt, (accessed 31 Jul. 2015)