GetFTR High Performance Ingestion Specification

This document details how to "get started" by providing your entitlements to GetFTR for the High Performance GetFTR Service. The High Performance service provides entitlement decisions, determining if the user's Institution entitles them to one or more DOI, in less than 40 ms.

GetFTR achieves this by ingesting "Google Subscriber Links Archive" files through either SFTP or Remote Ingestion, and using the central store of entitlements to take quick decisions.

GetFTR has both staging and production SFTP servers that should align with your own target environments. You will need to connect to the SFTP server to configure your integration, see “Configuration”.

- Production: `sftp.prod.central.getft.io`
- Staging: `sftp.staging.central.getft.io`

Connecting

In order to connect to the GetFTR SFTP server, we must create a user for you. To do this, you will need to provide an SSH Public Key, beginning with `ssh-rsa <string>`. For instructions on how to generate an SSH key pair, see Generate SSH Keys.

As you are sharing a public key with us, feel free to send this as an attachment over email, or through your GetFTR Slack channel (if we have set you up).

Once we have created a user for you, we will let you know the username. You will use both your username and the corresponding private key to connect to the SFTP servers for **both staging and production**.

We support both manual and programmatic/automated connections to our SFTP servers.

AWS Transfer

The SFTP servers are provided by AWS Transfer and are backed by an S3 storage bucket.

Directory Structure

Your home directory has three root directories:

`upload`	All files are uploaded into folders inside this directory. Files are deleted from this directory as they are ingested.
`processed`	This directory contains all of the files our system has successfully ingested from the upload folder.
`errors`	Contains all files that produced an error during ingestion. A text file containing the service error with the same filename is also generated.

Configuration

A `config.toml` template file located in your home directory allows you to configure your integration with the centralised entitlements service.

The configuration file follows TOML Format. If additional keys or invalid syntax is uploaded, a `config.toml.err` file will be created containing information about the error.

Once a configuration file has been uploaded, if it has been successfully processed a `config.toml.curr` file will be created, indicating the active configuration for your identity.

The schema you can provide for the configuration file is documented below:

```
# Optional. Use the [api] configuration if you intend
to
# integrate with the gRPC API to send entitlement
# requests
[api]
# Required. The secret you will use to integrate
# with the gRPC API
key = "your-secret-value"


# Optional. Provide the [archive] key if you wish to
configure how subscribers links archives are ingested.
[archive]
# Optional. Notify one or more email addresses when the
ingestion process succeeds or fails.
notify = ["<email address>"]


# Optional. Provide the [remote] key if you want to
configure one remote ingestion (see "Remote
Ingestion"). Note that you cannot provide [remote] and
[remote.<tenant>] keys.
[remote]
# Required. The HTTP URL where our ingestion will look
for  a subscribers.xml file.
url = "https://example.com/subscribers.xml"
# Required. Provide a schedule string, this is in the
form of a day (SUN, MON, TUE, etc.) and a 24-hour time.
schedule = "MON 02:00"


# Optional. Provide the [remote.<tenant>] key if you
want to configure one or more remote ingestions.
[remote.<tenant>]
```

```
url = "https://first-archive.com/subscribers.xml"
schedule = "FRI 18:00"
```

Option 1: SFTP Upload

All uploads should be -placed into **`/upload`**. The type of file(s) you are uploading should be placed into the folder paths detailed below.

Subscribers Links Files

Note: *Currently, we do not support any variant of Subscribers Links Files, such as "SL+".*

Once a valid Subscribers Links file has been uploaded to the SFTP server, GetFTR takes responsibility for ingesting the content, producing the subsequent Institution Entitlements file and providing access to those entitlements to integrators via our API.

Uploaded entitlements expire after two weeks. This provides a mechanism for off-boarding Publishers and for sunsetting entitlements, if necessary. Therefore, we require that the entire archive is uploaded at least every 14 days.

Uploaded Subscribers Links files become effective on the following **Monday**. This includes any uploaded files in the days before, up to and including the previous Monday. Entitlements will not be queryable by integrators until that day.

A Publishers responsibility for those files ends once **all** uploaded files have been moved into the **`/processed`** folder. This indicates that GetFTR has successfully ingested all of the uploaded files. If you have configured notification emails (see “Configuration”), you will receive an email indicating the completion of the ingestion.

It is the Publishers responsibility to monitor the ingestion process for the presence of any files placed into the **`/errors`** folder; this indicates that one or more files have failed to be ingested. This is made easier by enabling notifications, as you will receive an email to one or more addresses when the process fails or succeeds.

All Subscribers Links files should be uploaded to the following folder path. We do not currently support a compressed archive.

`/upload/YYYY-MM-DD[/<subdirectory>]/<filename>.xml`

Here are some example file paths:

`/upload/2022-06-23/1234567890.xml`
`/upload/2022-06-23/subscribers.xml`
`/upload/2022-01-01/tenant-a/987612345.xml`
`/upload/2022-01-01/tenant-a/subscribers.xml`
`/upload/2022-01-01/tenant-b/987612345.xml`
`/upload/2022-01-01/tenant-b/subscribers.xml`
`/upload/2022-12-31/0987654321.xml`

When uploading files into a subdirectory, identical files can exist within each subdirectory, as described in the example above.

We do not begin to ingest files until up to five minutes *after the last file* has successfully been uploaded. This is done by counting the number of files in the directory and then waiting 5 minutes; after this time, we count again. If the number of files remains the same, we begin the ingestion process. Once files begin to be moved into the `/processed` folder, any additional files that are uploaded will be ignored.

In the case where duplicate filenames may exist within an archive, files can be placed into a subdirectory. **Multiple subdirectories are not supported.**

Troubleshooting

- An error message file is generated by our service and placed alongside any file that is created inside the `/errors` folder. If you believe this error to be something that you are able to resolve, you can begin a partial ingestion process, as documented below. *If not, please raise the issue with us over Slack or by email and we will look into it.*

- If files are placed into the `/errors` folder and all remaining files that were uploaded to the `/uploads` folder have been moved, you can upload those files again, along with the accompanying `subscribers.xml` file, to begin a partial ingestion process, for only the affected files.

Option 2: Remote Ingestion

To ease the onboarding of new and existing publishers to the Centralised Entitlements platform, we have implemented a remote ingestion process. This allows us to download and upload subscribers links archives for you at a specific day and time, from a URL you provide.

In-order to enable remote ingestion on the publisher side, you will need to whitelist the following IP addresses:

- Production: `34.194.224.65`
- Staging: `34.203.150.249`

The URL you provide must be a `subscribers.xml` file. The ingestion process will then read this file before uploading it into your SFTP upload directory, continuing to download each of the HTTP URLs provided for institution files in this file.

The provided schedule must match the format `"DDD HH:MM"`. At the moment, only one schedule can be provided per ingestion. Accepted values for `DDD` are: `MON, TUE, WED, THU, FRI, SAT, SUN`. Valid values for `HH:MM` is any 24-hour time.

Each of the files is uploaded into your SFTP directory as if you were uploading the files yourself, into the directory path detailed under the “Uploads” heading. Tenanted ingestion also works identically.

As the arrival of files into your **`upload/`** directory automatically starts the process of parsing entitlements, the process of remotely downloading and uploading files consequently causes the archive to be immediately ingested.