Onboarding Handbook for Data Providers

AWS Open Data Sponsorship Program

(Revised: 2022-10-14)

Welcome

This document is intended for participants or prospective participants of the Amazon Web Services (AWS) Open Data Sponsorship Program (the program). Here you will find an overview, step-by-step instructions, guidance & recommendations, as well as answers to frequently asked questions.

If you are new to the program, welcome! If you are considering applying to the program, we hope this guide will answer your questions about how to apply, and what will be expected from you if accepted to the program. As a team, we are excited to help you share your data and to expand its reach and impact.

—The AWS Open Data Team

Contents

Overview & timeline for onboarding

At a high level, joining the AWS Open Data Sponsorship Program consists of the following steps:

1. **APPLY** - Submit applications via online web form. We review applications on a quarterly schedule, and communicate results via email within two weeks of our review date. A schedule of review dates can be found on the program website.[1]
2. **PREPARE DOCUMENTATION & DRAFT DATASET ENTRY FOR THE REGISTRY OF OPEN DATA** - Program participants must provide documentation and an example tutorial of working with their dataset on AWS. Additionally, a draft Registry of Open Data entry for your dataset must be summitted via Github.
3. **SIGN TERMS & CONDITIONS** - Agree to the program's legal agreement via the online web form.
4. **CREATE & LINK NEW AWS ACCOUNT** - Create a pristine AWS account that will be used exclusively for hosting your dataset.

   This account must then be linked to the program's billing organization[2] prior to the account's use. You will provide your new AWS account number to opendata@amazon, and we will issue an invitation to link your account. **Important:** You are responsible for any charges incurred prior to confirmed account linking.
5. **PREPARE AND UPLOAD DATA** - Upload and organize your data into a newly created Amazon S3 bucket created using a supplied Cloud Formation template.
6. **LAUNCH!** - The AWS Open Data team will publish your finalized entry on the Registry of Open Data, and will include your dataset in our quarterly PR announcement.

Once we have communicated your successful application and eligibility for the AWS Open Data Sponsorship Program, we require that your documentation, draft dataset entry for the Registry of Open Data, and signed terms & conditions are completed within two weeks (steps 2 & 3).

[1] Program website with application review schedule: https://aws.amazon.com/opendata/open-data-sponsorship-program/

[2] https://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/consolidated-billing.html

Applicants that do not complete these items within two weeks of acceptance are elligible for release in the next quarter. Eligibility for unlaunched datasets expires after two quarterly release cycles, and applicants wishing to proceed are invited to reapply.

1 Applying to the AWS Open Data Sponsorship Program

Applications to host a new dataset are submitted via our online form.[3] We use this form to collect basic information about your dataset to assess its eligibility to the program.

The program covers the cost of storage and data transfer for publicly available high-value cloud-optimized datasets. We work with data providers who seek to:

- Democratize access to data by making it available for analysis on AWS,
- Develop new cloud-native techniques, formats, and tools that lower the cost of working with data, and
- Encourage the development of communities that benefit from access to shared datasets.

AWS evaluates applications to the program on a quarterly basis.[4] Application decisions are typically communicated within two weeks following the quarterly review period. New datasets are announced publicly on the AWS What's New page[5] on a quarterly basis.

If we bring your dataset into the program, we will cover the costs of storage and data transfer for a period of two years, in accordance with the program's Terms & Conditions.[6] Program enrollment is renewable after the two-year period under agreement from all parties, in accordance with the Open Data Sponsorship Program Terms and Conditions.

[3] https://application.opendata.aws/

[4] See program webpage for dates of upcoming review cycles: https://aws.amazon.com/opendata/open-data-sponsorship-program/

[5] https://aws.amazon.com/new/

[6] https://aws.amazon.com/opendata/open-data-sponsorship-program/terms/

2 Documentation, tutorial(s), & Registry of Open Data entry

A main goal of the program is to help expand the reach and usage of your dataset. One component of this is ensuring that data consumers understand the dataset well enough to make effective use of it. You, the data provider, are in a unique position to orient new users your data. Therefore before launch, we will look for three forms of documentation from you:

- A document describing how the data is organized and how users can find the data they need. Examples of this kind of documentation include Amazon Berkeley Objects,[7] Speedtest by Ookla,[8] Southern California Earthquake Data Center,[9] and Global Biodiversity Information Facility.[10]
- A tutorial on how to use the data on AWS. Tutorials differ from dataset documentation, in that tutorials illustrate some end-to-end usage of the data on AWS in a "how to" format. Beyond a factual description of the dataset, tutorial content should guide the reader through some interesting use of the dataset.
- An entry on the Registry of Open Data [11]. The Registry is a mechanism for users of your dataset to learn about the data, learn how to access the data, and find documentation, tutorials, usage examples, and publications that demonstrate using your dataset.

We ask that you host your descriptive documentation and tutorial at a publicly available location such as on your institution's website or a public Github project.

Creating an entry on the Registry of Open Data for your dataset requires creating a structured (YAML-formatted) text file that contains information about your dataset. You can then submit that YAML via a draft pull request[12] on our Github project.

[7] https://amazon-berkeley-objects.s3.us-east-1.amazonaws.com/index.html
[8] https://github.com/teamookla/ookla-open-data
[9] https://scedc.caltech.edu/data/cloud.html
[10] https://github.com/gbif/occurrence/blob/master/aws-public-data.md
[11] https://registry.opendata.aws
[12] https://docs.github.com/en/pull-requests/collaborating-with-pull-requests/proposing-changes-to-your-work-with-pull-requests/creating-a-pull-request

The Registry's project and documentation for the format of the YAML file can be found at the Registry of Open Data's Github site.[13]

If you are new to working with Github or would like to see a video tutorial on submitting your YAML via a pull request, we have provided one on YouTube.[14]

Please also send a copy of your logo to be used on the dataset landing page to opendata@amazon.com. This logo is displayed at a width of 160px, so something larger than that will be required. An example of how this logo is used can be seen on this example dataset landing page.[15]

As you prepare your dataset, we will review your entry via Github and work with with you on any needed revisions. After your dataset is ready to launch, you will take your pull request out of draft state, and AWS Open Data will do a final review before merging your changes and making the entry live at the Registry of Open Data.

3 Terms & conditions

Someone in your organization must signify their agreement to the AWS Open Data Sponsorship Program Terms and Conditions[16] via the AWS Open Data Sponsorship Program agreement form.[17]

Please contact the Open Data team via email at opendata@amazon.com if you have questions,

[13] https://github.com/awslabs/open-data-registry/.
[14] https://www.youtube.com/watch?v=5nocWdjN1DA
[15] https://registry.opendata.aws/noaa-goes/
[16] https://aws.amazon.com/opendata/open-data-sponsorship-program/terms/
[17] https://pages.awscloud.com/open-data-sponsorship-program.html

or if you are unable to agree to the terms and conditions for some reason.

4 Creating & linking a new AWS account

In order for the program to pay for dataset hosting, program participants must create a new AWS account. That AWS account will be used solely to host the dataset. We will then connect the new account to an internal AWS account that will pay the bills for the new account.

You will be the legal owner of the account that hosts the data and, as such, will have complete ownership over the data shared through the account. We will not have the ability to modify any of the data.

Please note that we do monitor account charges for adherence to our Terms & Conditions. Unless specifically authorized by the Open Data team, usage of any service other than Amazon S3 is not permitted.

Even though account charges will be covered by our program, you will need to enter billing information when you create a new account.[18]

Once you have created your account, please confirm to opendata@amazon.com that you have done so. When doing so, please include the AWS account number and email address used to create the account.

Once we have received your account details, we will send an invitation for you to join our AWS Organization for consolidated billing. This is the step that will allow the Open Data Program to cover the costs of this AWS account. Please accept this invitation as soon as possible, as the invitation will expire after a few days. Note that if you generated your AWS account under a different AWS Organization, you will need to leave that Organization in order to view and accept the invitation to join the Open Data Program's Organization. Please be careful not to remove yourself from this Organization in the future or you will be responsible for the account billing.

After accepting the invitation, please email opendata@amazon.com for confirmation that your account was successfully linked. This will ensure that subsequent charges can be covered by the program.

[18] If you would prefer to set up this account for invoicing rather than credit card, please request this conversion via email to the Open Data Team at opendata@amazon.com. Please use the **root email** of your new account when you make this request.

To recap:

1. Create a new AWS account
2. Provide account details to opendata@amazon.com
3. Accept email invitation to join program billing organization
4. Confirm successful account linking with opendata@amazon.com

5 Preparing & uploading data

Note: Please do not proceed with this section until confirmation that your AWS account has successfully linked to the program billing organization. In this step, you will create an Amazon S3 bucket that will host your dataset. To facilitate creating an S3 bucket with optimal settings, we have developed a Cloud Formation template. The template sets up a number of features for you:

- S3 bucket for your dataset
- Intelligent tiering lifecycle configuration
- Delete aborted multi-part uploads after 7 days lifecycle configuration
- S3 bucket logs to a separate S3 bucket for data usage analysis
- Enables S3 metrics for your bucket for data usage insights
- S3 bucket policy that enables public access
- SNS topic for S3 event notifications for object creation
- SNS policy enabling public access via SQS or Lambda to the SNS topic

The Cloud Formation template can be used as follows:

1. Log in to your account.
2. Click or paste the provided Cloud Formation URL[19]

   Note you can change the region from the default "us-west-2" to whichever region you would like to work, by replacing "us-west-2" in the Cloud Formation URL. For geospatial datasets, unless there is a compelling reason to use a different region, we ask that you use the default us-west-2.

[19] https://console.aws.amazon.com/cloudformation/home?region=us-west-2#/stacks/new?stackName=pds-buckets&templateURL=https://s3-us-west-2.amazonaws.com/opendata.aws/pds-bucket-cf.yml

3. Hit the "Next" button.
4. In the Parameters section, enter the name of the new S3 bucket that you are creating you'd like to use for the dataset. Note that it may only contain lowercase letters, numbers, and ., -, or _ characters. Also note that S3 bucket names are unique, so you will get an error if you choose a name already in use by another bucket.
5. Hit "Next" two more times, and then hit "Create."
6. In the top-right of the next window, you should see a "Refresh" button that uses a circular arrow icon. Hit that a few times over the next minute or so to track progress.
7. Once it says the stack has been created, go to the S3 console[20] and you should see the bucket there (along with some others). It will have the right permissions for public sharing and will have logging enabled.
8. Lastly, we recommend enabling "Termination Protection" which can help avoid accidental S3 bucket deletions.[21]

Once your S3 bucket has been created, you can begin uploading data. For uploading data we recommend using the AWS CLI.[22] If you are moving data from another existing S3 bucket, the FAQ has tips for speeding up transfers.

For guidance on organizing your dataset files, please see Appendix A.

6 Launching your dataset

Having completed the prior steps in this guide, your dataset is ready to launch. There are only two things that must happen for us to consider your dataset launched.

First, with your data fully uploaded into S3 and ready for use, and your documentation & Registry entry complete, AWS Open Data will merge your finalized entry via Github. This will result your dataset having a live landing page (e.g. https://registry.opendata.aws/sentinel-2/). Once this landing page is live on the Registry of Open Data, we invite you to announce availability of your dataset via any mechanisms that you wish.

Second, your dataset will be included with our quarterly release announcement and may be blogged or promoted on social media by AWS. No action is required on your part for this step.

[20] https://console.aws.amazon.com/s3/
[21] https://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/using-cfn-protect-stacks.html
[22] https://aws.amazon.com/cli/

Post launch, we invite you to submit updates to your dataset's entry on the Registry of Open Data. Particularly if you are aware of new tutorials, blog posts, or research papers published, updating your dataset's entry enriches your dataset and its community of users.

Congratulations!

Appendices

A Bucket layout guidance

When uploading data, we recommend putting some thought into the structure of object prefixes in your bucket. This will not only aid users in discovery of your data, but will also open up more possibilities for using data with a wider variety of tools.

The "right" bucket layout will vary from dataset to dataset, and of course every layout is a collection of trade-offs with its own unique advantages and disadvantages. However, there are some general patterns to keep in mind when structuring objects in your S3 bucket.

In general a few general ideas to keep in mind are:

Create some top-level prefixes for organization

At the top-level prefix of your bucket, consider some high-level structural elements as applicable, e.g.:

```
s3://my-bucket/
├── README.txt
├── data/
├── metadata/
└── docs/
```

Even if a `data/` prefix and a `README.txt` file are the only applicable prefixes today, consider how your bucket structure may change over time and how you might want to add elements without needing to move existing objects.

Group like data types by common prefix

For example if your data includes both JPEG and TIFF data, don't mix the two within the same prefix without good reason. Instead think of creating a prefix that differentiates the format, e.g.:

```
s3://my-bucket/
```

```
└─ data/
   ├─ jpeg/
   │  ├─ file1.jpg
   │  └─ (…).jpg
   └─ tiff/
      ├─ file1.tiff
      └─ (…).tiff
```

If, for example, you think you may add additional data formats in the future, this approach also allows you to do so without disrupting users of existing data, e.g.:

```
s3://my-bucket/
└─ data/
   ├─ jpeg/(…)
   ├─ tiff/(…)
   └─ png/(…)
```

Use Hive partition format where applicable

Often times datasets have natural groupings within a schema where users might only want a subset of the data. For example, think of a dataset that spans several years, but where users may often only want a day or two of data at a time.

There are lots of ways to group this kind of data by date, but using Hive conventions for the naming of objects makes the intent clear, and can make working with AWS analytics services such as Amazon Athena much easier. The Hive naming convention partitions the data using the key prefix format `variable=value`. In the above example of data partitioned by date, a bucket layout might look like this:

```
s3://my-bucket/
├─ README.txt
├─ metadata/
└─ data/
   └─ csv/
      ├─ date=2112-01-01/
      │  └─ (…)
      ├─ date=2112-01-02/
      │  └─ (…)
      └─ date=(…)/
```

Ask for individual guidance

The above suggestions are general suggestions that we think are applicable to many data providers. That said, every dataset is unique, and there may be additional considerations, or reasons to bend these suggestions.

The AWS Open Data team is happy to answer questions about the structure and layout of your S3 bucket. If you have questions, please email us at opendata@amazon.com.

B Frequently Asked Questions (FAQ)

B.1 The terms I signed said we aren't allowed to talk about our participation in the program without reaching out to you first, is this true?

Yes and no! If you want to issue a formal statement like a press release or similar large-scale announcement, yes, we would like you to reach out to us for review. There is more info on that process in the Press Kit section of this handbook. If you are looking to talk about this is a less formal setting (e.g., blog post, conference talk, publication), we definitely want you to talk about it! If you're looking for guidance on suggested language, you can find that in the Press Kit section.

B.2 How can I see metrics / usage statistics for my dataset?

You have a couple of options for understanding usage from your public Amazon S3 bucket:

1. First, Cloud Watch Metrics should be enabled on your bucket if the bucket was created with our Cloud Formation template. This is far and away the easiest option. From the AWS S3 console, you can browse to your bucket, and the metrics tab. On that tab only a couple of top line metrics are displayed, however there is a "View additional charts" link below the the default two. From there you can explore charts related to storage as well as by requests.

2. Next, for the purposes of self-service data usage analysis, we permit you to enable S3 bucket logs on your data bucket and to create a second S3 bucket for temporary storage of S3 logs within your Open Data-sponsored AWS account.

    a) To set up S3 bucket logs in your AWS Open Data sponsored account, follow official S3 user documentation.[23].

[23] Enable server access logging: https://docs.aws.amazon.com/AmazonS3/latest/userguide/enable-server-access-logging.html

b) We require setting up an object lifecycle that expires log objects after 1 month. To set up an object lifecycle, follow the Amazon S3 Object Lifecycle documentation.[24]

c) If you need to retain logs for longer than this period, you will need to copy the objects to an S3 bucket in a paid AWS account.[25]

Note, that our program terms only cover S3 storage / egress costs. If you elect to analyze S3 logs, our program's expectation is that you carry out that analysis outside of your AWS program-sponsored account.

B.3 Should I zip / tar / archive my files before uploading?

Generally speaking, no. Archived sets of files usually necessitate downloading the entire archive. If files are uploaded to S3 as individual objects, they can be used a la carte and potentially with a broader set of AWS tools.

B.4 How can I speed up large bulk transfers between S3 buckets with the AWS CLI?

The AWS CLI must copy any object over 5GB in pieces as the maximum atomic file operation for S3 is 5GB.[26]

The good news is this all happens on the server side (no data is brought down the pipe to wherever your S3 client is running). But there is still a communications overhead. To speed up this kind of operation quite significantly, there's a few things we recommend:

- Adjust the settings of your AWS CLI client to increase the `multipart_chunksize` parameter.[27] The default is only 8MB, which means that for large files, it's making [size of objects in MB] / 8 requests for a mv operation. This is highly inefficient. Since the copy operation is happening on the server, you can typically change the default to 1024MB safely. Each chunk will take longer to copy, but S3 -> S3 operations are generally extremely fast. This will reduce the "chatter" dramatically.
- Run your mv operation on a computer in the same region as the S3 bucket. This will reduce the latency of the copy requests a little bit, but won't have nearly the impact of adjusting the chunk size.

[24] Object lifecycle management documentation: https://docs.aws.amazon.com/AmazonS3/latest/userguide/object-lifecycle-mgmt.html

[25] To learn more about how to copy objects from S3 bucket to another, see this re:Post answer: https://repost.aws/knowledge-center/move-objects-s3-bucket

[26] https://docs.aws.amazon.com/AmazonS3/latest/API/API_CopyObject.html

[27] https://docs.aws.amazon.com/cli/latest/topic/s3-config.html

- You can also increase the `max_concurrent_requests` parameter, but you'll get less milage here than reducing the number of requests.

Additionally, newer versions of the AWS CLI support using the "AWS Common Runtime" to accelerate transfers as compared to the default Python implementation of the copy functionality.[28]

B.5 How should we acknowledge the Registry of Open Data in our scientific publications?

Thanks for thinking of us! We would love it if you could include the URL to the Registry of Open Data page for the dataset in the "Data Availability" section of your manuscript. Sample text: "Data are available at registry.opendata.aws/cell-painting." We don't require any attribution beyond that, but you're always welcome to thank the AWS Open Data Sponsorship Program for supporting the storage and distribution of your dataset in your Acknowledgements section.

B.6 Does AWS offer a communication platform that we can leverage for questions, discussions, etc. about our dataset(s)?

AWS re:Post gives you access to a community that helps you be successful on AWS.[29] With AWS re:Post, you and your data users can communicate about a specific dataset, and can also contribute to topics relevant to AWS Open Data at large. All re:Post questions tagged with AWS Open Data can be found at:

https://repost.aws/tags/TApd0Wl5P8S9O6riTWmh-cGw/aws-open-data

Note that you do not need an AWS account to browse re:Post questions, but to post or respond to one, you do need to be logged in to your AWS account.[30] By adding AWS Open Data as a skill to your re:Post profile, you will receive a notification if a new question is tagged with AWS Open Data.

[28] https://awscli.amazonaws.com/v2/documentation/api/latest/topic/s3-config.html#preferred-transfer-client
[29] https://repost.aws/
[30] https://repost.aws/account/signin

C Press Kit

Welcome to the Amazon Web Services (AWS) Open Data Sponsorship Program! To help you communicate the news of your dataset being openly accessible on AWS, the AWS PR team has developed the following press kit.

The press kit includes:

1. General guidelines for working with AWS Communications
2. A template press release
3. Social media guidelines
4. Logos

If you have questions about any of the resources or guidelines in this kit, please reach out to your lead point of contact on the AWS Open Data team.

Note: Before beginning any external marketing or communications, you must have first agreed to the program Terms & Conditions and your data must be listed on the Registry of Open Data on AWS.

C.1 Guidelines for working with AWS Communications

Below are guidelines to keep in mind as you develop your press release and/or other announcement materials.

- Remember to share your draft press release for review/approval with your lead point of contact on the AWS Open Data team. We will deliver feedback and share any edits to your press release within three business days.
- AWS reserves the right to alter and/or decline any proposed language about AWS in your draft press release per our guidelines (outlined below).
- Once your announcement has gone live, you can coordinate press briefings discussing your announcement and work with AWS. We welcome the opportunity to discuss how AWS can help amplify your media efforts. As a courtesy, please inform your lead point of contact on the AWS Open Data team of the interviews ahead of time.

To streamline the release review process, please keep the following guidance from the AWS PR team in mind. The release and other communications should **NOT**:

- Use the terms "partners" or "partnership" or "alliance" to describe the AWS relationship. We prefer the terms "collaborated," "teamed," or "in cooperation with."
- Use the name "Amazon" when referring to Amazon Web Services or AWS.
- Disclose proprietary information about Amazon or AWS.
- Include the Amazon stock ticker symbol.
- Use the "About AWS" boilerplate. These are reserved for AWS releases only.
- Use generic, catch all words to describe AWS as a "platform" or "ecosystem". Be specific in your description, such as "AWS services" or "AWS product offerings."
- Lastly, make sure it does not sound like your data is available "in" AWS. It is more appropriate to reference that it is available "on" AWS.

C.2 Template press release

Below is a press release template that you can customize and use to announce your participation in the AWS Open Data Sponsorship Program. Feel free to share the release with your media contacts and other stakeholders; publish the release on your website; send in email blasts, and more. Instructions for using the press release template are as follows:

1. Customize the below template for your organization. Please ensure that you're following AWS's general guidelines outlined in Section 1 of this document.
2. Share your draft press release for review/approval with your lead point of contact on the AWS Open Data team. Our Public Relations team will deliver feedback and share any edits to your press release within three business days.
3. Distribute your press release any time after AWS has approved the content. While we release quarterly announcements of new datasets - the data of which was shared with you when you were accepted into the program - you can announce your dataset at any point after approval of the press release.

[ORGANIZATION'S][NAME OF DATASET] Now Openly Accessible on the Amazon Web Services Cloud

[SHORT DESCRIPTION OF DATASET] now openly accessible via the Registry of Open Data on AWS

[CITY], [STATE](DATE) — [NAME] today announced that the [INSERT NAME OF DATA SET] is now openly accessible on the Amazon Web Services (AWS) cloud.

Better information sharing has the power to accelerate discoveries and improve the world around us. The research community can now access [DATASET NAME] on AWS without needing to pay to store their own copies of the dataset. Researchers will only pay for the computing services they use, and do not need to purchase storage to start a project using the dataset. AWS, through its Open Data Sponsorship Program, is covering the costs of the storage and transfer of the data, so that it can be accessed and analyzed in the cloud by researchers around the world.

[INSERT DETAILED DESCRIPTION OF THE DATASET – i.e. what is it, how large is it, what has it been/can it be used for, when was the dataset first established, what is new or novel about this dataset? What will having this data available on AWS unlock/enable?]

[INSERT ORGANIZATION QUOTE]

The AWS Open Data Sponsorship Program covers the cost of storage and egress for publicly available, high-value, cloud-optimized datasets. AWS works with data providers to democratize access to data by making it available for analysis on AWS; develop new cloud-native techniques, formats, and tools that lower the cost of working with data; and encourage the development of communities that benefit from access to shared datasets. Through the program, AWS has democratized access to petabytes of data, including satellite imagery, climate and weather data, genomic data, and data used for natural language processing. The full list of publicly available datasets is available on the Registry of Open Data on AWS.

To learn more and access [INSERT NAME OF DATASET], visit [INSERT REGISTRY LINK].

About [NAME] [INSERT YOUR BOILERPLATE/ORGANIZATIONAL OVERVIEW HERE]

C.3 Social media guidelines

You are welcome to share out the news of your participation in the AWS Open Data Sponsorship Program on your various social accounts. Please follow the general guidelines for AWS Communications outlined in Section 1, and please tag us at @awscloud on Twitter, Amazon Web Services (AWS) on LinkedIn, and/or Amazon Web Services on Facebook. AWS does not need to review all supporting social media posts as long as the messages contained in the social post follow our PR guidelines. Below are two suggested tweets for your consideration:

1. Check out the new [SHORT DESCRIPTION OF DATASET] now openly accessible via @AWSCloud [link to dataset].
2. [ORGANIZATION] and @AWSCloud are working together to make [SHORT DESCRIPTION OF DATASET] widely available through the Open Data program [link to dataset].

C.4 Logos

If you would like a logo to include in any marketing or communications materials, please feel free to use the "Powered By Amazon Web Services" logo. You can find this logo, along with other logos and additional logo guidance at the AWS Co-Marketing website.[31].

powered by aws

[31] https://aws.amazon.com/co-marketing/