Book Analytics Service documentation May 2024
https://zenodo.org/doi/10.5281/zenodo.8388648

Book Analytics Service

Dashboard overview

Book Analytics Service

How can the Book Analytics Dashboard help me?

Whether you're a publisher, librarian, funder, administrator, or other stakeholder in the scholarly communications community, the Dashboard can help you gain a fuller view of book usage data.

Books are made available through a multitude of different platforms, and each has its own way of providing usage statistics. Using our Dashboard gives you a single point of synthesis of usage data from a wide range of sources, allowing you a consolidated view of usage.

We take care to bring together a synthesis of usage data of known provenance to allow you to compare data across your chosen time span. Learn more about how we collect and process book usage data.

Who's behind the Dashboard and how is it operated?

The Dashboard was initially developed as a pilot project 2020-2022, which developed a prototype for gathering book usage information from multiple data sources, and combining and presenting it in interactive visualisation dashboards for publisher partners.

Funded again by the Mellon Foundation, the pilot project was scaled up to become the Book Analytics Dashboard project (2022-2025), focused on creating a sustainable OA book focused analytics service.

The fully-functional Dashboard is now operated by OAPEN, a trusted infrastructure for OA books. OAPEN is a not-for-profit organisation which cannot be sold, and has committed to the Principles of Open Scholarly Infrastructure (POSI) with a public

self-audit of their practices across the themes of governance, sustainability, and insurance. OAPEN was chosen as the host organisation by the Book Analytics Dashboard (BAD) project Advisory Board, following a series of focus groups with publishers which endorsed OAPEN as an organisation which could be trusted by the community to run the Dashboard in an open and consultative way.

What's the difference between BAD and BAS?

BAD refers to the Book Analytics Dashboard project (2022-2025), whereas BAS means the Book Analytics Service. BAD is the project to launch BAS from a functional pilot Dashboard to a self-sustaining service with appropriate governance and cost-recovery financing. As well as being key elements of the Principles of Open Scholarly Infrastructure (POSI), these are also essential in ensuring that the future of BAS does not depend on grant funding.

How the Dashboard works

How does the Dashboard collect and process book usage data?

At the heart of the Dashboard's technology stack is a valid ONIX feed which includes the metadata of the works a partner wishes to represent in the dashboard. We use ONIX because this is the book industry's standard metadata interchange format that publishers use to share information about the books that they have published.

Our workflows collect book usage data from multiple sources (learn more about our data sources) and public bibliographic metadata and Event Data from Crossref. Data from these sources is integrated with the ONIX feed, using the ISBN-13 identifier to identify works and combine usage data from multiple sources. The partner's ONIX feed serves as the source of truth for a work's metadata, such as book title, authors, and related works. Crossref bibliographic metadata is used to match Event Data DOIs with book ISBNs, which are then matched with the ISBNs in the partner's ONIX feed.

Our workflows are the code which controls the data integration; all of which is built on an open-source workflow system. The workflows fetch, process, disambiguate, and analyse data about books from multiple sources, and this data is saved to Google Cloud's BigQuery data warehouse. The next steps of data processing include:

1. Ingesting data via telescope workflows from Crossref metadata, Crossref Event Data, Google Analytics, Google Books, JSTOR, IRUS Fulcrum, IRUS OAPEN, a publisher's ONIX feed (obtained via SFTP, or from the OAPEN Library, or from Thoth), UCL Discovery, and
2. A series of analytic workflows to process and combine the data ingested by the telescope workflows. The processed data in the Google Cloud BigQuery data warehouse is then visualised in dashboards provided by Looker Studio, a dashboarding solution offered by Google.

The information from our data sources is refreshed on a regular basis, keeping the Dashboard up-to-date. Updated usage data for all sources is available on the dashboard typically on the first Monday after the fourth of the month. Crossref Event Data is typically updated weekly.

Is the Dashboard data COUNTER-conformant?

Please see our Dashboard data sources overview which gives you a detailed overview of each source.

How do we deal with bot activity?

Bot identification is the responsibility of the platforms themselves, as they have access to the individual usage data, which we do not. Platforms that are using COUNTER-conformant standards (such as IRUS OAPEN usage statistics) should only include genuine, user-driven usage, as activity generated by internet robots and crawlers must be excluded from all COUNTER usage reports.

How is the Dashboard data protected?

We receive usage data from platform providers in aggregated and anonymised format: individual usage data is stripped out so that the data we receive is an aggregation and can't be traced back to individuals. In the event that platform usage reports contain location information such as individual IP address, this information is anonymised before it is provided to the Dashboard. For example, IRUS OAPEN usage reports do contain IP addresses, therefore this data is downloaded and anonymised within an OAPEN Google Cloud project located in Europe. The transformed data, with IP addresses removed and replaced with city or country information, is then sent to the Dashboard.

Since we do not collect any personally identifiable information, GDPR does not apply to our data.

Each partner's data is kept in a separate Google Cloud project (located in the USA).

Access to each is controlled with user access permissions (username and password credentials), providing strong security and privacy. Only Dashboard staff have access to this partner data.

Data provided is used only for the purposes of the Dashboards: it is not sold on or made available to any other parties for any reason.

Dashboard data sources

What are the Dashboard's data sources?

To see the data sources for a specific Dashboard, click on About & FAQ on the Dashboard, and consult the list at Data Sources. The only obligatory data source is title metadata in ONIX format; each publisher then chooses the other data sources they wish to include.

The data sources currently available to be visualised in the Dashboard are detailed in the tables below. The standard data sources and variables used are included, other data sources and variables may be supported as an extra add-on service.

Where the Dashboard gets title metadata from

Data source	Status	Access
Crossref metadata	Current	Public
OAPEN metadata	Current	Public
ONIX-FTP feed from publishers	Current	Private
Thoth	Current	Public

Where the Dashboard gets usage and mentions data from

Data source	Status	Access	COUNTER-conformant?	Time aggregation
Crossref Event Data	Current	Public	n/a	Monthly
Google Analytics Universal	Not current	Private	No	Monthly
Google Books	Current	Private	No	Monthly
IRUS Fulcrum	Current	Private	Yes	Monthly
IRUS OAPEN	Current	Private	Yes	Monthly
JSTOR	Current	Private	Yes	Monthly
UCL Discovery	Current	Public	No	Monthly

Mentions and page views

Data source	Events	Page Views
Crossref Event Data	Y (count of event [id])	
Google Analytics Universal		Y [page_views]

Book views and downloads

Data source	Book Views	Book Downloads
Google Analytics Universal		Y (with custom dimensions)
Google Books	Y [BV_with_Pages_Viewed]	Y [qty]
IRUS Fulcrum		[total_item_requests]
IRUS OAPEN		Y [title_requests] and [total_item_requests]
UCL Discovery		Y [total_downloads]

Chapter downloads

Data source	Chapter Downloads
JSTOR	Y [total_item_requests]

Public access data sources

The public access data sources are those where data is made publicly available by the data source. No additional access permission is required from Dashboard partners for the Dashboard to access the following data sources if partners want them to be included on their dashboard/s.

Crossref Event Data

Crossref Event Data captures online discussion about research outputs, such as 'a citation in a dataset or patent, a mention in a news article, Wikipedia page or on a blog, or discussion and comment on social media'. Event Data is retrieved using the Crossref Event Data API. Crossref Event Data must be queried using a DOI, which BAS obtains from Crossref metadata.

Crossref metadata

Crossref is a not-for-profit membership organisation, and an official Digital Object Identifier (DOI) Registration Agency of the International DOI Foundation. They make metadata available for all DOIs registered with Crossref. BAS uses Crossref metadata to match ISBNs obtained from a publisher's ONIX feed to DOIs to query Crossref Event Data.

OAPEN metadata

OAPEN enables libraries and aggregators to use the metadata of all available titles in the OAPEN Library. The metadata is available in different formats and BAS harvests the data in XML format and converts it into ONIX format for the OAPEN platform.

Thoth

Thoth is a free, open metadata service that publishers can use as a metadata storage solution. Thoth can provide metadata in a number of formats. BAS uses the Thoth export API to download metadata for publishers in ONIX format.

UCL Discovery

University College London (UCL) is an eBook publisher, and partner in the BAD project. UCL Discovery is UCL's open access repository, showcasing and providing access to the full texts of UCL research publications.

Private access data sources - access permission required

Google Analytics Universal

Google Analytics Universal monitors and records web traffic for specific websites. If a Dashboard partner had configured Google Analytics on their publisher website, the Google Analytics data can be used to find out which countries and territories website visitors are from.

Google Books

The Google Books Partner program hosts eBooks, including some free open access eBooks. eBook publishers can then download usage reports from Google Books. BAS uses data from the Google Play sales transaction report and the Google Books Traffic Report.

JSTOR

JSTOR is a digital library offering over 7,000 open access eBooks. Publisher usage reports offer details about the use (views and downloads) of eBooks by institution, and country.

ONIX-FTP feed from publishers

ONIX is a standard that book publishers use to share information about the books that they have published. BAS dashboard partners that have ONIX feeds are given credentials and access to their own upload folder on the Mellon SFTP server. Each publisher uploads their ONIX feed to their upload folder on a weekly, fortnightly, or monthly basis. The BAS data workflow downloads the ONIX data, transforms it (with the ONIX parser Java command line tool) and then loads it into BigQuery for further processing.

Private access data sources - no additional access permission required

IRUS Fulcrum

IRUS provides COUNTER standard access reports for eBooks hosted on the Fulcrum platform. Fulcrum is a “community-developed, open source platform for digital scholarship” which provides “users the ability to read books with associated digital enhancements, such as: 3-D models, embedded audio, video, and databases; zoomable online images, and interactive media”.

IRUS OAPEN

IRUS provides COUNTER standard access reports for eBooks hosted on the OAPEN library and platform. OAPEN "promotes and supports the transition to open access for academic books by providing open infrastructure services to stakeholders in scholarly communication". Almost all eBooks on OAPEN are provided as a PDF file for the whole book. The reports show access figures for each month, and the location (IP address) of the access. Within the OAPEN Google Cloud project (located in Europe), IP addresses are replaced with geographical information (city and country). This means that IP addresses are not stored within BAS data, and only de-identified geographical information is transferred to BAS.

How to use your Dashboard

What can I see in the Dashboard?

Publishers, in your Dashboard, you can see usage data for your own published works. You can see the usage of the books you have published in terms of views, downloads, and online mentions and events. You can also view which countries and institutions are using your books, and which subjects are represented in your collections.

Features to explore:

- **Reset** - click “Reset” at the top of the page, or “Reset filters” to the right of the filters selection
- Search by **author name** (currently displayed as *last name*, *first name*) - learn more in ORCID’s documentation about first and last names)
- **Track usage over time** - from the “Overview” page, see which month has the highest usage numbers when all book titles are selected
- See **usage for a specific title** - from the “Overview” page, select a book title from the filter and see the total access number
- See **usage across geographic regions** - from the “Global Reach” page, select a country from the country filter and see on the number of book downloads and the number of chapter downloads. Countries and territories are based on ISO standard 3166.

Which data sources have country, institution, and city information?

Google Books, JSTOR, and OAPEN Counter 5 usage data have information about usage by **country**.

JSTOR usage data has information about usage by **institution**.

OAPEN Counter 4 and Counter 5 usage data has information about usage by **city**.

How do I use the 'Authors' dashboard?

Select an author of interest to see the usage metrics aggregated for each unique author.

How do I use the 'Subjects' dashboard?

Select a subject of interest to see the usage metrics aggregated for each unique subject.

Does the OAPEN data include both Counter 4 and Counter 5 data?

Unless otherwise noted, OAPEN data includes both Counter 4 data (C4) and Counter 5 data (C5).

How do I use the filters?

By default, filters are set to show all entries.

> Click on a filter and select the checkbox at the top to select or unselect all items in that list.

Select Country / Territory filter

- Click “only” to the right of any item to see results for that item only. You may select multiple items from the list.
- Use the text field to enter text to search for a particular item.
- To reset the filter, click "Reset" on the top menu

Reset Filters button

I’m not yet participating, but can I see a demo?

You may view the BAD template Dashboard, powered by usage data from the University of Michigan Press which they have very kindly made publicly available, with no login details required.

How should I interpret the data from the Dashboard?

Presenting a holistic view of usage information can be difficult, because eBooks can be hosted in multiple repositories and platforms, in different file formats (PDF, EPUB, MOBI, HTML), and in different levels (whole book or by chapter). Each repository provides book content to different audiences in different ways.

Before studying data from the Dashboard, we suggest that you take some time to understand the data sources, including the limitations of comparing different data sources. Learn more about how usage is influenced by the language of the work, its subject, its platform, and seasonal differences in Ronald’s blog post and research paper, “Measured in a context: making sense of open access book data”.

Caveats:

- For publishers with a smaller number of titles, it's harder to see a pattern and understand why some get more downloads than others
- Incomplete data can hamper making accurate comparisons

But all that said, the Dashboard can help you explore some interesting questions!

Publishers, here are some questions the Dashboard data can help you consider:

- In which countries and territories are my publications most and least downloaded?
- Does this correspond to the languages in which I'm publishing?
- Which subjects are most or least popular in different areas?
- How does this change over time?

How can I share information from the Dashboard?

It's easy to share information from the Dashboard - you can export data for further exploration in a spreadsheet, and visualisations to share with colleagues.

If sharing is enabled for your Dashboard, click *more* (three dots icon) > *Export* (to export in CSV and other formats); and click *Download report* (at the bottom of the page) to export a PDF report.

When sharing Dashboard information, we recommend including some explanation to help your recipient interpret it in context. For this, you are welcome to reuse information from the section How should I interpret the data from the Dashboard? above, and include links to this documentation.

More information and contact us

For all your enquiries and questions, please contact us at info@book-analytics.org.

Book Analytics Service key information page, including how to get started.

Information about the Mellon Foundation-funded Book Analytics Dashboard (BAD) project (2022-2025):

- Visit the BAD project website
- Follow the BAD project on X (formerly Twitter) @BookAnalytics
- Join the BAD project mailing list
- Visit the BAD project Zenodo community
- See the BAD template dashboard, powered by the University of Michigan Press's data
- Our book-focused GitHub repository

Running in parallel with the Book Analytics Dashboard project is the OA Book Usage Data Trust, a project to formalise community governance mechanisms, quantify data trust participation benefits, and understand the full operational costs related to an international data space for OA book usage.

Glossary

BAD

The Book Analytics Dashboard project (2022-2025) - the Mellon Foundation-funded project that is focused on creating a sustainable open access book focused analytics service

BAS

The Book Analytics Service, the product whose development from pilot to full service is supported by the BAD project

COKI

Curtin Open Knowledge Initiative - a team of data scientists, software developers and researchers at Curtin University, Perth, Australia

COUNTER

COUNTER provides the standard that enables the knowledge community to count the use of electronic resources. To have their usage statistics and reports designated COUNTER compliant, report providers MUST provide usage statistics that conform to the current Code of Practice

Crossref

Crossref is a Digital Object Identifier (DOI) Registration Agency of the International DOI Foundation, that makes metadata available for all DOIs registered with them

Dashboard

A dashboard is an interactive, up-to-date page of visualisations that aggregate and summarise data from different sources. In the context of this documentation and written with a capital letter, “Dashboard” refers specifically to BAS

Data source

A public or pilot project dashboard partner source of data about open access eBooks and their usage, such as views, downloads and online mentions

eBook

A book publication made available in electronic or digital form

Google Books

Google Books provides paid and free (open access) eBooks

IRUS

A service for capturing and processing institutional repository usage data, making it possible for institutional repositories and platforms to generate COUNTER compliant usage data

IRUS Fulcrum

Fulcrum is a “community-developed, open source platform for digital scholarship”. IRUS provides COUNTER standard usage reports for eBooks hosted on the Fulcrum platform

IRUS OAPEN

IRUS provides COUNTER standard usage reports for eBooks hosted on the OAPEN library and platform

JSTOR

A digital library which offers over 7,000 open access eBooks

Looker Studio

A dashboarding solution provided by Google

OAeBU

Open Access eBook Usage (2020-2022) - the Mellon Foundation-funded pilot project "Developing a Pilot Data Trust for Open Access Ebook Usage", the precursor project to BAD

OAPEN

OAPEN is a not-for-profit organisation dedicated to open access, peer-reviewed books, operating three platforms: OAPEN Library; OAPEN Open Access Books Toolkit; and the Directory of Open Access Books (DOAB)

Open Access (OA)

From the Budapest Open Access Initiative

> By “open access” to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search,

> or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself.

ONIX

ONIX for Books (ONline Information eXchange) is a standard format that book publishers use to share information about the books that they have published

SFTP

SSH File Transfer Protocol

Telescope

A data workflow that fetches and ingests data from a data source. Some telescopes run workflows that process and output data to other places. Workflows are built on top of Apache Airflow's Directed Acyclic Graph (DAG), where a DAG is “a collection of organised tasks that you want to schedule and run”

License

the work and Derivative Works thereof.

"Contribution" shall mean any work of authorship, including the original version of the Work and any modifications or additions to that Work or Derivative Works thereof, that is intentionally submitted to Licensor for inclusion in the Work by the copyright ow or by an individual or Legal Entity authorized to submit on behalf the copyright owner. For the purposes of this definition, "submitte means any form of electronic, verbal, or written communication sent to the Licensor or its representatives, including but not limited t communication on electronic mailing lists, source code control syst and issue tracking systems that are managed by, or on behalf of, th Licensor for the purpose of discussing and improving the Work, but excluding communication that is conspicuously marked or otherwise designated in writing by the copyright owner as "Not a Contribution

"Contributor" shall mean Licensor and any individual or Legal Entit on behalf of whom a Contribution has been received by Licensor and subsequently incorporated within the Work.

2. Grant of Copyright License. Subject to the terms and conditions of this License, each Contributor hereby grants to You a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable copyright license to reproduce, prepare Derivative Works of, publicly display, publicly perform, sublicense, and distribute the Work and such Derivative Works in Source or Object form.

3. Grant of Patent License. Subject to the terms and conditions of this License, each Contributor hereby grants to You a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable (except as stated in this section) patent license to make, have mad use, offer to sell, sell, import, and otherwise transfer the Work, where such license applies only to those patent claims licensable by such Contributor that are necessarily infringed by their Contribution(s) alone or by combination of their Contribution(s) with the Work to which such Contribution(s) was submitted. If You institute patent litigation against any entity (including a cross-claim or counterclaim in a lawsuit) alleging that the Work or a Contribution incorporated within the Work constitutes direct or contributory patent infringement, then any patent licenses granted to You under this License for that Work shall terminate as of the date such litigation is filed.

4. Redistribution. You may reproduce and distribute copies of the Work or Derivative Works thereof in any medium, with or without modifications, and in Source or Object form, provided that You meet the following conditions:

(a) You must give any other recipients of the Work or

```
        Derivative Works a copy of this License; and

    (b) You must cause any modified files to carry prominent notices
        stating that You changed the files; and

    (c) You must retain, in the Source form of any Derivative Works
        that You distribute, all copyright, patent, trademark, and
        attribution notices from the Source form of the Work,
        excluding those notices that do not pertain to any part of
        the Derivative Works; and

    (d) If the Work includes a "NOTICE" text file as part of its
        distribution, then any Derivative Works that You distribute mus
        include a readable copy of the attribution notices contained
        within such NOTICE file, excluding those notices that do not
        pertain to any part of the Derivative Works, in at least one
        of the following places: within a NOTICE text file distributed
        as part of the Derivative Works; within the Source form or
        documentation, if provided along with the Derivative Works; or,
        within a display generated by the Derivative Works, if and
        wherever such third-party notices normally appear. The contents
        of the NOTICE file are for informational purposes only and
        do not modify the License. You may add Your own attribution
        notices within Derivative Works that You distribute, alongside
        or as an addendum to the NOTICE text from the Work, provided
        that such additional attribution notices cannot be construed
        as modifying the License.

    You may add Your own copyright statement to Your modifications and
    may provide additional or different license terms and conditions
    for use, reproduction, or distribution of Your modifications, or
    for any such Derivative Works as a whole, provided Your use,
    reproduction, and distribution of the Work otherwise complies with
    the conditions stated in this License.

5. Submission of Contributions. Unless You explicitly state otherwise,
   any Contribution intentionally submitted for inclusion in the Work
   by You to the Licensor shall be under the terms and conditions of
   this License, without any additional terms or conditions.
   Notwithstanding the above, nothing herein shall supersede or modify
   the terms of any separate license agreement you may have executed
   with Licensor regarding such Contributions.

6. Trademarks. This License does not grant permission to use the trade
   names, trademarks, service marks, or product names of the Licensor,
   except as required for reasonable and customary use in describing t
   origin of the Work and reproducing the content of the NOTICE file.
```

7. Disclaimer of Warranty. Unless required by applicable law or
   agreed to in writing, Licensor provides the Work (and each
   Contributor provides its Contributions) on an "AS IS" BASIS,
   WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or
   implied, including, without limitation, any warranties or condition
   of TITLE, NON-INFRINGEMENT, MERCHANTABILITY, or FITNESS FOR A
   PARTICULAR PURPOSE. You are solely responsible for determining the
   appropriateness of using or redistributing the Work and assume any
   risks associated with Your exercise of permissions under this Licen

8. Limitation of Liability. In no event and under no legal theory,
   whether in tort (including negligence), contract, or otherwise,
   unless required by applicable law (such as deliberate and grossly
   negligent acts) or agreed to in writing, shall any Contributor be
   liable to You for damages, including any direct, indirect, special,
   incidental, or consequential damages of any character arising as a
   result of this License or out of the use or inability to use the
   Work (including but not limited to damages for loss of goodwill,
   work stoppage, computer failure or malfunction, or any and all
   other commercial damages or losses), even if such Contributor
   has been advised of the possibility of such damages.

9. Accepting Warranty or Additional Liability. While redistributing
   the Work or Derivative Works thereof, You may choose to offer,
   and charge a fee for, acceptance of support, warranty, indemnity,
   or other liability obligations and/or rights consistent with this
   License. However, in accepting such obligations, You may act only
   on Your own behalf and on Your sole responsibility, not on behalf
   of any other Contributor, and only if You agree to indemnify,
   defend, and hold each Contributor harmless for any liability
   incurred by, or claims asserted against, such Contributor by reason
   of your accepting any such warranty or additional liability.

END OF TERMS AND CONDITIONS

APPENDIX: How to apply the Apache License to your work.

   To apply the Apache License to your work, attach the following
   boilerplate notice, with the fields enclosed by brackets "[]"
   replaced with your own identifying information. (Don't include
   the brackets!) The text should be enclosed in the appropriate
   comment syntax for the file format. We also recommend that a
   file or class name and description of purpose be included on the
   same "printed page" as the copyright notice for easier
   identification within third-party archives.

```
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implie
See the License for the specific language governing permissions and
limitations under the License.
```

Contributing Guide

Contributing

The oaebu-workflows is a dependent repository of the observatory-platform.

We welcome contributions to the project, please see the CONTRIBUTING.md file for the Observatory Platform for details about how to contribute to this repository.

Installing BAD Workflows

Dependencies:

- Python 3.10
- pip
- Python virtualenv (recommended)
- Docker Engine or Docker Desktop
- The Observatory Platform

See below for more a detailed walk-through of the installation process.

Installing System Dependencies

Ubuntu 20.04+

Update packages list and install software-properties-common:

```
sudo apt update
sudo apt install software-properties-common
```

Add deadsnakes PPA which contains Python 3.10 for Ubuntu 20.04; press `Enter` when prompted:

```
sudo add-apt-repository ppa:deadsnakes/ppa
```

Install Python 3.10:

```
sudo apt install python3.10 python3.10-dev
```

Install pip:

```
curl https://bootstrap.pypa.io/get-pip.py -o get-pip.py
python3.10 get-pip.py
```

Install virtualenv:

```
pip install --upgrade virtualenv
```

Install Docker Engine:

- Following the Install Docker Engine on Ubuntu tutorial.
- Make sure that Docker can be run without sudo, e.g.
  `sudo usermod -aG docker your-username`

MacOS

Install Homebrew with the following command:

```
/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/
```

Install Python 3.10 with brew:

```
brew install python@3.10
```

Add Python 3.10 to path:

```
echo 'export PATH="/usr/local/opt/python@3.8/bin:$PATH"' >> ~/.bash_p
```

Install virtualenv 20 or greater:

```
pip3.10 install --upgrade virtualenv
```

Install Docker Desktop:

> Follow the Install Docker Desktop on Mac tutorial.

Creating a Python Virtual Environment

It is highly recommended to create a Python virtual environment before installing the Observatory Platform and BAD workflows. To create a new virtual environment, execute the following command:

```
python3.8 -m venv MY_VENV_NAME # Change MY_VENV_NAME to whatever you plea
```

To activate/deactivate the environment:

```
source MY_VENV_NAME/bin/activate # Activate
deactivate # Deactivate
```

While active, all changes to Python packages will be contained to the environment. Similarly, any Python run commands will be done in the context of the active environment. **It is recommended that the following packages are installed to an active virtual environment**, although there is no reason that the BAD Workflows could not be run without one.

The Observatory Platform

Make sure that you have followed the above instructions before installing the observatory-platform.

Clone the Observatory Platform repository:

```
git clone https://github.com/The-Academic-Observatory/observatory-platfor
```

Install the `observatory-api` package:

```
pip install observatory-platform/observatory-api --constraint https://raw
```

Install the `observatory-platform` package:

```
pip install observatory-platform/observatory-platform --constraint https:
```

BAD Workflows

Clone the BAD workflows repository (oaebu-workflows)

```
git clone https://github.com/The-Academic-Observatory/oaebu-workflows.git
```

Install the workflows to the active python environment

```
pip install oaebu-workflows
```

Workflows & Telescopes

Workflow Schedule

Schedule

All workflows/telescopes are schedules to run at 00:00 UTC unless specified otherwise

Workflow/Telescope	Schedule	Notes
Google Analytics	N/A	Decommissioned
Google Books	Weekly - Sunday 12:00 UTC	Data collection monthly
IRUS Fulcrum	Monthly on the 4th	
IRUS OAPEN	Monthly on the 4th	
JSTOR	Monthly on the 4th	Reports sent monthly
UCL Discovery	Monthly on the 4th	
OAPEN Metadata	Weekly - Sunday 12:00 UTC	
ONIX Telescope	Weekly - Sunday 12:00 UTC	
Thoth	Weekly - Sunday 12:00 UTC	
ONIX Workflow	Weekly (Monday) 00:00 UTC	Waits for dependent telescopes before running
Crossref Metadata	Weekly (Monday) 00:00 UTC	Runs during the ONIX WF
Crossref Events	Weekly (Monday) 00:00 UTC	Runs during the ONIX WF

Data Telescopes

Telescopes that pull book usage and access data

Google Analytics Universal

Google Analytics was a web analytics service offered by Google that tracks and reports website traffic (now replaced with Google Analytics 4).
This telescope obtained data from Google Analytics for 1 view id per publisher and for several combinations of metrics and dimensions.
It is possible to add a regex expression to filter on pagepaths, so only data on relevant pagepaths is collected.
Note that Google Analytics data is only available for the last 26 months, see Data retention - Analytics Help for more info.

To get access to the analytics data a publisher needs to add the relevant google service account as a user.

Dataset Name	google
Table Names	google_analytics
Table Type	Partitioned
Average Runtime	10 min
Average Download Size	10-20 MB
Harvest Type	API
Run Schedule	Monthly
Catch-up Missed Runs	✅
Each Run Includes All Data	❌

Custom dimensions for ANU Press

ANU Press was using custom dimensions in their google analytics data. To ensure that the telescope processes these custom dimensions, the organisation name needs to be set to exactly 'ANU Press'.

The organisation name is used directly inside the telescope and if it matches 'ANU Press' additional dimensions will be added and a different BigQuery schema is used.

A note on the API metrics

We use the python client for the The Google Analytics API in order to retrieve the data on several metrics (such as page views) per country. It appears as though the API does not return a result for every country. We would have expected any data without a country field to be labelled with a country name of **not set**, however this does not appear to be the case.
At this time, we have no other way of retrieving country-level data on the desired metrics, so we must acknowledge that the numbers returned by the API are slightly different to those found on the Google Analytics web page.
A ticket has been created with google in the hope of resolving this issue.

Telescope kwargs

Organisation Name (organisation_name)

The name of the organisation as displayed on Google Analytics

View ID (view_id)

The View ID points to the specific view on which Google Analytics data is collected. See the google support page for more information on the hierarchy of the Analytics account.

Pagepath Regex (pagepath_regex)

This is a regular expression that is used to filter on pagepaths for which analytics data is collected.

The regular expression can be set to an empty string if no filtering is required. Note that the Google Analytics API uses 're2', so it is not possible to use e.g. negative lookaheads. See the google support page and github wiki for more information.

Setting up service account

- Create a service account from IAM & Admin - Service Accounts
- Create a JSON key and download the file with key
- For each organisation/publisher of interest, ask them to add this service account as a user for the correct view id

Getting the view ID (after given access)

```
from googleapiclient.discovery import build
from oauth2client.service_account import ServiceAccountCredentia

scopes = ['https://www.googleapis.com/auth/analytics.readonly']
credentials_path = '/path/to/service_account_credentials.json'

creds = ServiceAccountCredentials.from_json_keyfile_name(credent

# Build the service object.
service = build('analytics', 'v3', credentials=creds)

account_summaries = service.management().accountSummaries().list
view_ids = []
for account in account_summaries['items']:
    account_name = account['name']
    profiles = account['webProperties'][0]['profiles']
    website_url = account['webProperties'][0]['websiteUrl']
    for profile in profiles:
        view_id_info = {'account': account_name, 'websiteUrl': w
                        'view_name': profile['name']}
        view_ids.append(view_id_info)
```

Airflow connections

Name	Description
oaebu_service_account	The credentials for the service account that has been given access to the google analytics view.

Telescope Tasks

Data Download & Transform

Downloads a single month of reporting data using the GA3 analytics reporting API. Transforms the data from the report structure into the schema structure and saves it to a .jsonl file.

Big Query Load

The transformed data is loaded from the Google Cloud bucket into a partitioned BigQuery table under the *google* dataset (which will be created should it not exist yet). Since the data is partitioned on the release month, there will only be a single table named g*oogle_analytics3*.

Table Schema

name	type	mode	description
url	STRING	REQUIRED	Base URL of the book pages.
title	STRING	REQUIRED	Title o f the book.
start_date	DATE	REQUIRED	Start date for period of analytics info.
end_date	DATE	REQUIRED	End date for period of analytics info.
average_time	FLOAT	REQUIRED	Average time (in seconds) spent on each page.
unique_views	RECORD	NULLABLE	Unique views for several different dimensions. Unique views is the number of sessions during which the specified page was viewed at least once. A unique pageview is counted for each page URL + page title combination.
unique_views.country	RECORD	REPEATED	Unique views per users' country, derived from their IP addresses or Geographical IDs.
unique_views.country.name	STRING	NULLABLE	Country name.
unique_views.country.value	INTEGER	NULLABLE	Number of unique views.
unique_views.referrer	RECORD	REPEATED	Unique views per referrer, the full referring URL including the hostname and path.
unique_views.referrer.name	STRING	NULLABLE	Referrer name.
unique_views.referrer.value	INTEGER	NULLABLE	Number of unique views.

unique_views.social_network	RECORD	REPEATED	Unique views per social network. This is related to the referring social network for traffic sources; e.g., Google+, Blogger.
unique_views.social_network.name	STRING	NULLABLE	Social network name.
unique_views.social_network.value	INTEGER	NULLABLE	Number of unique views.
page_views	RECORD	NULLABLE	The total number of pageviews for the property
page_views.country	RECORD	REPEATED	Page views per users' country, derived from their IP addresses or Geographical IDs.
page_views.country.name	STRING	NULLABLE	Country name.
page_views.country.value	INTEGER	NULLABLE	Number of page views.
page_views.referrer	RECORD	REPEATED	Page views per referrer, the full referring URL including the hostname and path.
page_views.referrer.name	STRING	NULLABLE	Referrer name.
page_views.referrer.value	INTEGER	NULLABLE	Number of page views.
page_views.social_network	RECORD	REPEATED	Page views per social network. This is related to the referring social network for traffic sources; e.g., Google+, Blogger.
page_views.social_network.name	STRING	NULLABLE	Social network name.

page_views.social_network.value	INTEGER	NULLABLE	Number of page views.
sessions	RECORD	NULLABLE	Total number of sessions for several different dimensions.
sessions.country	RECORD	REPEATED	Unique views per users' country, derived from their IP addresses or Geographical IDs.
sessions.country.name	STRING	NULLABLE	Country name.
sessions.country.value	INTEGER	NULLABLE	Number of sessions.
sessions.source	RECORD	REPEATED	Sessions per source of referrals. For manual campaign tracking, it is the value of the utm_source campaign tracking parameter. For AdWords autotagging, it is google. If you use neither, it is the domain of the source (e.g., document.referrer) referring the users. It may also contain a port address. If users arrived without a referrer, its value is (direct)..
sessions.source.name	STRING	NULLABLE	Source name.
sessions.source.value	INTEGER	NULLABLE	Number of sessions.
release_date	DATE	REQUIRED	Last day of the release month. Table is partitioned on this column.

Google Books

Documentation for the Google Books telescope

The Google Books Partner program enables selling books through the Google Play store and offering a preview on Google books.
The program makes books discoverable to Google users around the world on Google books. When readers find a book on Google Books, they can preview a limited number of pages to decide if they're interested in it. Readers can also follow links to buy the book or borrow or download it when applicable.

As a publisher Google Books download reports are available at https://play.google.com/books/publish/

Currently there are 3 report types available:

- Google Play sales summary report
- Google Play sales transaction report
- Google Books Traffic Report

The telescope collects data from the last 2 reports.

Dataset Name	google
Table Names	google_books_sales, google_books_traffic
Table Type	Partitioned
Average Runtime	10 min
Average Download Size	1-100 MB
Harvest Type	SFTP
Run Schedule	Weekly
Catch-up Missed Runs	✅
Each Run Includes All Data	❌

Airflow connections

Name	Description
sftp_service	The username, password and host name used to connect to the SFTP server

Authentication

The reports are downloaded from `https://play.google.com/books/publish/`. To get access to the reports the publisher needs to give access to a google service account. This service account can then be used to login on this web page and download each report manually.

Downloading Reports Manually

There is no API available to download the Google Books report and it is quite

challenging to automate the Google login process through tools such as Selenium, because of Google's bot detection triggering a **reCAPTCHA**. Until this step can be automated, the reports need to be downloaded manually each month. For each publisher and for both the sales transaction report and the traffic report:

- A report should be created for exactly 1 month (e.g. starting 2021-01-01 and ending 2021-01-31).
- All titles should be selected.
- All countries should be selected.
- The traffic report is organised by 'Book'.
- It is important to save the file with the right name, this should be in the following format (<file_suffix> is optional):
  - `GoogleSalesTransactionReport_<file_suffix>YYYY_MM.csv` or
  - `GoogleBooksTrafficReport_<file_suffix>YYYY_MM.csv`
- Upload each report to the SFTP server.
  - Add it to the folder `/google_books_<publisher>/upload`
  - Files are automatically moved between folders; do not move files between folders manually

Telescope Tasks

Data Download & Transform

The download step connects to the SFTP server. The telescope looks in the relevant publisher's **upload** folder for the file format specified above. Any telescope DAG run will harvest **all** instances of the matching files (regardless of the date associated). Before downloading, the files on the SFTP server are moved to the **in_progress** folder.

Once downloaded, each report is transformed. The transform process re-formats headings and dates such that they are consistent. It also performs an integrity check on the reported dates. None of the raw data is modified in any way. The partition date (the report's associated month) is appended to each row at the end

of the transform step.

Big Query Load

The transformed data is loaded from the Google Cloud bucket. There are two resulting datasets from each telescope run, both of which will be loaded into their own partitioned BigQuery table under the *google* dataset (which will be created should it not exist yet). Then, the *google_books_sales* and *google_books_traffic* table partitions are loaded. Since the data is partitioned on the release month, there will only be a single table for each of these report types.

Table Schema - Google Books Sales

name	type	mode	description
Transaction_Date	DATE	REQUIRED	The date of the transaction.
Id	STRING	REQUIRED	A unique identifier for this transaction.
Product	STRING	NULLABLE	In UCL Press case "Single Purchase" (a normal sale). Can also be "Rental".
Type	STRING	NULLABLE	Type of transaction (can be 'sale' or 'refund').
Preorder	STRING	NULLABLE	Whether this transaction applied to a preorder. In UCL Press case 'None': The transaction didn't involve a preorder.
Qty	INTEGER	NULLABLE	The number of units in the transaction. Negative for refunds.
Primary_ISBN	STRING	NULLABLE	The primary ISBN or other identifier the book, prefixed by a single quotation mark so spreadsheet programs will display the entire ISBN.
Imprint_Name	STRING	REQUIRED	The template used for the book.
Title	STRING	REQUIRED	The title of the book.
Author	STRING	NULLABLE	The author of the book.
Original_List_Price_Currency	STRING	NULLABLE	The original currency of the book's list price.
Original_List_Price	FLOAT	NULLABLE	The original list price of the book.
			The currency of the book's

List_Price_Currency	STRING	NULLABLE	list price. If currency conversion was enabled, this is the currency of purchase as seen by the buyer.
List_Price_tax_inclusive_	FLOAT	NULLABLE	The book's list prince including tax.
List_Price_tax_exclusive_	FLOAT	NULLABLE	The book's list price excluding tax.
Country_of_Sale	STRING	NULLABLE	The country where the buyer bought the book.
Publisher_Revenue_Perc	FLOAT	NULLABLE	The publisher's percentage of the list price.
Publisher_Revenue	FLOAT	NULLABLE	The amount of revenue earned by the publisher. This will be negative if the transaction was a refund. Negative for refunds. The currency is the same as the payment currency.
Payment_Currency	STRING	NULLABLE	The currency of the publisher's earnings.
Payment_Amount	FLOAT	NULLABLE	The amount earned by the publisher for this transaction. Negative for refunds.
Currency_Conversion_Rate	FLOAT	NULLABLE	If the list price and payment amount are in different currencies, the rate of exchange between the two currencies.
Line_of_Business	STRING	NULLABLE	This field is not present for some publishers (UCL Press). For ANU Press the field value is “E-Book”.
release_date	DATE	REQUIRED	Last day of the release month. Table is

			partitioned on this column.

Table Schema - Google Books Traffic

name	type	mode	description
Primary_ISBN	STRING	NULLABLE	The primary identifier (e.g., ISBN) of the book. This column appears in the report if data is organized by book.
Title	STRING	REQUIRED	The title of the book.
Book_Visits_BV_	INTEGER	NULLABLE	A Book Visit is registered each time a unique user views one of your books on Google Books. This number includes informational page views (such as the “About this book” page) as well as preview content page views.
BV_with_Pages_Viewed	INTEGER	NULLABLE	The number of Book Visits in which users accessed preview pages of your book. This doesn’t include visits where a user accessed only informational pages for your books.
Non_Unique_Buy_Clicks	INTEGER	NULLABLE	The number of clicks on links for purchasing the book on retailer websites (including your website, if you provided a buy link).
BV_with_Buy_Clicks	INTEGER	NULLABLE	The number of visits which included a click on a purchase link.
Buy_Link_CTR	FLOAT	NULLABLE	The clickthrough rate for purchase links. The values are percentages.
Pages_Viewed	INTEGER	NULLABLE	The total number of unique preview content pages that a user viewed in a given session (counted as a 24-hour period). If a user views

			the same page of your book twice during a session, only a single page view is registered.
release_date	DATE	REQUIRED	Last day of the release month. Table is partitioned on this column.

IRUS Fulcrum

Documentation for the IRUS Fulcrum telescope

The IRUS Fulcrum telescope collects usage statistics for titles accessed via the Fulcrum Platform. Usage data is accessible through IRUS in much the same way as the IRUS OAPEN telescope. Unlike IRUS OAPEN, IRUS Fulcrum does not record sensitive IP address information. This makes dealing with the data much simpler.

The earliest available data for the Fulcrum platform is April 2022. It follows that all data is of COUNTER 5 standard.

Dataset Name	irus
Table Name	irus_fulcrum
Table Type	Partitioned
Average Runtime	10 min
Average Download Size	1-10 MB
Harvest Type	API
Run Schedule	Monthly on the 4th
Catch-up Missed Runs	✅
Each Run Includes All Data	❌

Airflow connections

The following airflow connections are required:

Name	Description
irus_api	The IRUS requestor_id/api_key - required to access the IRUS platform

Telescope kwargs

Fields passed as keyword arguments to the telescope upon instantiation.

Publishers (publishers)

A list of publisher names. Usage stats from Fulcrum will be filtered on these publisher names. Many institutions have many publisher names associated with them, so it is important that all related names are provided.

Telescope Tasks

Data Download

The download is done via an API call to IRUS:

```
https://irus.jisc.ac.uk/api/v3/irus/reports/irus_ir/?platform=235&request
```

Where the requestor ID is the API key for the IRUS API. The telescope will use the same begin and end dates (YYYY-MM) in order to retrieve data on a per-month basis. The requestor ID is the *irus_api* airflow connection.

A second call to the API is made with the following appended to the above URL:

```
&attributes_to_show=Country
```

Which splits the data by country, leaving us with two datasets. These datasets will be referred to as the *total* and *country* datasets.

Before making any changes to the data, these datasets are uploaded to a Google storage bucket

Data Transform

The transform step has a few things to achieve:

- Collate the *total* and *country* datasets into a single object
- Remove columns that are not of interest to us
- Add the release month to each row as a partitioning column
- Remove rows from the data that do not relate to the publisher of interest

The result of points 1 → 3 are evident in the schema. The final point requires some communication with the publisher. This is because a single publisher may have published titles under more than one name. For example, University of Michigan has 10 associated publishing names. These names are listed as part of a dictionary in the telescope.

The resulting transformed file is uploaded to a Google Cloud bucket.

BigQuery Load

The transformed data is loaded from the Google Cloud bucket into a partitioned BigQuery table in the *irus* dataset, which will be created if it does not yet exist. Since the data is partitioned on the release month, there will only be a single table named *irus_fulcrum*.

Table Schema

name	type	mode	description
proprietary_id	STRING	NULLABLE	Proprietary identifier of the book.
ISBN	STRING	NULLABLE	ISBN of the book.
book_title	STRING	NULLABLE	Title of the book
publisher	STRING	NULLABLE	The publisher
authors	STRING	NULLABLE	The names of the authors
event_month	STRING	NULLABLE	The investigated month.
total_item_investigations	INTEGER	NULLABLE	The total number of item investigations.
total_item_requests	INTEGER	NULLABLE	The total number of item requests.
unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations.
unique_item_requests	INTEGER	NULLABLE	The number of unique item requests.
country	RECORD	REPEATED	Record to store statistics on the country level.
country.name	STRING	NULLABLE	The country name of the client registered by IRUS.
country.code	STRING	NULLABLE	The country code of the client registered by IRUS.
country.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations.
country.total_item_requests	INTEGER	NULLABLE	The total number of item requests.
country.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations.

investigations			item investigations.
country.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests.
release_date	DATE	REQUIRED	Last day of the release month. Table is partitioned on this column.

IRUS OAPEN

Documentation for the IRUS OAPEN telescope

IRUS provides OAPEN COUNTER standard access reports. Almost all books on OAPEN are provided as a whole book PDF file. The reports show access figures for each month as well as the location of the access.

Since the location info includes an IP-address, the original data is handled only from within the OAPEN Google Cloud project.

Using a Cloud Function, the original data is downloaded and IP-addresses are replaced with geographical information, such as city and country.
After this transformation, the data without IP-addresses is uploaded to a Google Cloud Storage Bucket.

This is all done from within the OAPEN Google Cloud project. The Cloud Function is created and called from the telescope, when the Cloud Function has finished the data is copied from the Storage Bucket inside the OAPEN project, to a Bucket inside the main airflow project.

Dataset Name	irus
Table Name	oapen_irus
Table Type	Partitioned
Average Runtime	10 min
Average Download Size	1-10 MB
Harvest Type	API
Run Schedule	Monthly on the 4th
Catch-up Missed Runs	✅
Each Run Includes All Data	❌

Airflow connections

Name	Description
irus_login	Login credentials for legacy C4 data
irus_api	The IRUS requestor_id/api_key - required to access the IRUS platform
geoip_license_key	Key for GeoIP services

Telescope kwargs

Fields passed as keyword arguments to the telescope upon instantiation.

Counter 4 Publisher Name (publisher_name_v4)

The publisher_name_v4 can be found by going to the OAPEN page to manually create reports. On this page there is a drop down list with publisher names, to get the publisher name simply url encode the publisher name from this list.

Note that occasionally there are multiple publisher names for one publisher.
For example to get all data from Edinburgh University Press, you need data from both publishers `Edinburgh University Press` and `Edinburgh University Press,`.
Multiple publisher names can be passed on by delimiting them with a '|' character.

Counter 5 Publisher ID (publisher_uuid_v5)

The publisher_uuid_v5 can be found by querying the OAPEN API and creating a list of unique Publisher names and UUIDs.

This API request will return all items including their Publisher name and UUID:
https://irus.jisc.ac.uk/api/oapen/reports/oapen_ir/?platform=215&requestor_id=<requestor_id>&api_key=<api_key>&granularity=totals

&begin_date=2020-04&end_date=2021-11

To get a file with mappings between Publisher Name and UUID, use the following Python snippet:

```
import requests
import pandas as pd

# Set up your credentials, the start & end date and path to output file
requestor_id = "YOUR_REQUESTOR_ID"
api_key = "YOUR_API_KEY"
start_date = "2020-04"
end_date = "2021-11"
out_file = "/path/to/output_mapping.csv"

# Query the OAPEN API
url = f"https://irus.jisc.ac.uk/api/oapen/reports/oapen_ir/?platform=215&
response = requests.get(url)
response_json = response.json()

# Store result in dataframe, get unique publisher values and sort
df = pd.DataFrame(response_json['Report_Items'])
result = df.drop_duplicates(["Publisher", "Publisher_ID"])[["Publisher",

# Save result to csv file
result.to_csv(out_file, index=False)
```

From this file look up the publisher UUIDs of interest. Similar to the publisher names described above, multiple publisher UUIDs can be passed on by delimiting them with a '|' character.

Telescope Tasks

The IRUS OAPEN telescope makes use of a Google Cloud Function that resides in the OAPEN Google Cloud Platform project.

There is a specific airflow task that will create the Cloud Function if it does not exist yet, or update it if the source code has changed.

The source code for the Cloud Function can be found inside a separate repository

that is part of the same organization (https://github.com/The-Academic-Observatory/oapen-irus-uk-cloud-function).

Download access stats data

The Cloud Function downloads IRUS OAPEN access stats data for 1 month and for a single publisher. Usage data after April 2020 is hosted on a new platform.

The newer data is obtained by using their API, this requires a `requestor_id` and an `api_key`.
Data before April 2020 is obtained from an URL, this requires an `email` and a `password`.

The required values for either the newer or older way of downloading data are passed on as a `username` and `password` to the Cloud Function. The `username` and `password` are obtained from an airflow connection, which should be set in the config file (see below).

Replace IP addresses

Once the data is downloaded, the IP addresses are replaced with geographical information (corresponding city and country).
This is done using the GeoIp database, which is downloaded from inside the Cloud Function. The license key for this database is passed on as a parameter as well, `geoip_license_key`.
The `geoip_license_key` is also obtained from an airflow connection, which should be set in the config file (see below).

Upload data to storage bucket

Next, the data without the IP addresses is upload to a bucket inside the OAPEN project. All files in this bucket are deleted after 1 day. In the next airflow task, the data can then be copied from this bucket to the appropriate bucket in the project where airflow is hosted.

BigQuery Load

The transformed data is loaded from the Google Cloud bucket into a partitioned BigQuery table in the *irus* dataset, which will be created if it does not exist. Since the data is partitioned on the release month, there will only be a single table named *irus_oapen*.

Set-up OAPEN Google Cloud project

To make use of the Cloud Function described above it is required to enable two APIs and set up permissions for the Google service account that airflow is using.

See the Google support answer for info on how to enable an API. The API's that need to be enabled are:

- Cloud Functions API
- Cloud build API
- Cloud Run Admin API
- Artifact Registry API

Inside the OAPEN Google project, add the airflow Google service account (<airflow_project_id>@<airflow_project_id>.iam.gserviceaccount.com, where airflow_project_id is the project where airflow is hosted). This can be done from the 'IAM & Admin' menu and 'IAM' tab. Then, assign the following permissions to this account:

- Cloud Functions Developer (to create or update the Cloud Function)
- Cloud Functions Invoker (to call/invoke the Cloud Function)
- Storage Admin (to create a bucket)
- Storage Object Admin (to list and get a blob from the storage bucket)

Additionally, it is required to assign the role of service account user to the service account of the Cloud Function, with the airflow service account as a member. The

Cloud SDK command for this is:

```
gcloud iam service-accounts add-iam-policy-binding <OAPEN_project_id>-
compute@developer.gserviceaccount.com --
member=<airflow_project_id@airflow_project_id.iam.gserviceaccount.com> --
role=roles/iam.serviceAccountUser
```

Alternatively, it can be done with the Google Cloud console, from the 'IAM & Admin' menu and 'Service Accounts' tab.
Click on the service account of the Cloud Function: `<OAPEN_project_id>-compute@developer.gserviceaccount.com`.
In the 'permissions' tab, click 'Grant Access', add the airflow service account as a member `<airflow_project_id@airflow_project_id.iam.gserviceaccount.com>` and assign the role 'Service Account User'.

geoip_license_key

To get the user*id/license_key, first sign up for geolite2. From your account, in the 'Services' section, click on 'Manage License Keys'. The user_id is displayed on this page.*
Then, click on 'Generate new license key', this can be used for the 'license_key'.
Answer _No for the question: "Old versions of our GeoIP Update program use a different license key format. Will this key be used for GeoIP Update?"

Table Schema

name	type	mode	description
proprietary_id	STRING	NULLABLE	Proprietary identifier of the book.
URI	STRING	NULLABLE	URI of the book. Only available for data since 2020-04-01.
DOI	STRING	NULLABLE	DOI of the book.
ISBN	STRING	NULLABLE	ISBN of the book.
book_title	STRING	NULLABLE	Title of the book
grant	STRING	NULLABLE	Grant. Only available for data before 2020-04-01.
grant_number	STRING	NULLABLE	Grant number. Only available for data before 2020-04-01.
publisher	STRING	NULLABLE	The publisher
begin_date	DATE	NULLABLE	The begin date of the investigated period.
end_date	DATE	NULLABLE	The end date of the investigated period.
title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01.
total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01.
total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01.
			The number of unique

unique_item_investigations	INTEGER	NULLABLE	item investigations. Only available for data since 2020-04-01.
unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01.
country	RECORD	REPEATED	Record to store statistics on the country level.
country.name	STRING	NULLABLE	The country name of the client registered by oapen irus uk.
country.code	STRING	NULLABLE	The country code of the client registered by oapen irus uk.
country.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01.
country.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01.
country.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01.
country.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01.
country.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01.
locations	RECORD	REPEATED	Record to store statistics on the location level.

locations.latitude	FLOAT	NULLABLE	The latitude geolocated from the client's ip address.
locations.longitude	FLOAT	NULLABLE	The longitude geolocated from the client's ip address.
locations.city	STRING	NULLABLE	The city geolocated from the client's ip address.
locations.country_name	STRING	NULLABLE	The country name geolocated from the client's ip address.
locations.country_code	STRING	NULLABLE	The country code geolocated from the client's ip address.
locations.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01.
locations.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01.
locations.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01.
locations.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01.
locations.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01.
version	STRING	REQUIRED	Version of the OAPEN IRUS UK API, corresponds to the COUNTER report version.
			Last day of the release

release_date	DATE	REQUIRED	month. Table is partitioned on this column.

JSTOR

Documentation for the JSTOR telescope

JSTOR provides publisher usage reports, the reports offer details about the use of journal or book content by title, institution, and country. Journal reports also include usage by issue and article. Usage is aligned with the COUNTER 5 standard of Item Requests (views + downloads).
Reports can be run or scheduled weekly, monthly, or quarterly with custom date ranges.

To directly get access to the analytics data a publisher needs to grant access to e.g. a Gmail account. This account can then be used to login to the JSTOR portal and set-up the scheduled reports (see below) that are mailed to a G-suite account. Alternatively, the publisher can set-up a schedule to create reports that are sent to the G-suite account.
In the telescope the Gmail of the G-suite account is parsed for messages with a download link to the JSTOR report.

The server running this telescope needs to be white listed by JSTOR to avoid bot detection.

Dataset Name	jstor
Table Name	jstor_country, jstor_institution
Table Type	Partitioned
Average Runtime	10 min
Average Download Size	5-10 MB
Harvest Type	Gmail
Run Schedule	Monthly on the 4th
Catch-up Missed Runs	✅
Each Run Includes All Data	❌

Airflow connections

The following airflow connections are required

Name	Description
gmail_api	The api credentials for access to the delegated gmail account.

Telescope kwargs

This telescope is created using the Observatory API. There is one 'extra' field that is required for the corresponding Telescope object, namely the 'publisher_id'.

Publisher ID (publisher_id)

A mapping is required between the JSTOR publisher ID and the organisation name obtained from the observatory API. The JSTOR publisher_id can be found in the

original filename of a JSTOR report, for example:
`PUB_<publisher_id>_PUBBIU_20210501.tsv`

It is possible to get the original filename by directly downloading a (previous) report from the JSTOR portal.

Entity Type (entity_type)

This is a special input that must be set to either *publisher* or *collection*. By default, this is set to *publisher.* This option determines what the telescope looks for in Gmail as well as how it transforms the data. For traditional publishers, this should remain untouched (or explicitly set to *publisher*). For a collection of titles, this should be set to *collection*. If *collection* is set, the *country_partner* and *institution_partner* kwargs should be changed as described below. This is because the information supplied by collections reports differ to traditional publisher reports and therefore require a slightly different schema.

Country Partner (country_partner)

The **data_partner** to use for the country report. Should be set to *jstor_country* when running with *publisher* entity, or *jstor_country_collection* if using the *collection* entity.

Institution Partner (institution_partner)

The **data_partner** to use for the institution report. Should be set to *jstor_institution* when running with *publisher* entity, or *jstor_institution_collection* if using the *collection* entity.

Retrieving Reports

JSTOR has no automated method for directly accessing reports from their website or any API. Reports must be sent via email in either TSV or CSV format. This can be done as either a one-off or as part of a schedule. In any case, log in to the JSTOR

website and set up a report schedule at their portal to get started.

Scheduling Reports

It will be easiest to set the report frequency the same as the schedule interval of the telescope (monthly). For this telescope only the 'Book Usage by Country' (PUB_BCU) and 'Book Usage by Institution' (PUB_BIU) are used.

The format needs to be set to 'TSV' and the recipient to the Gmail account that will be used with the Gmail API. The title of the report is not used in the telescope, so set this to anything you'd like (it does not show up in the email).

Downloading previous reports

Above is described how to set up a report schedule. Unfortunately this schedule can only be set up starting from the current date.
To get previous reports (from before the start date of the schedule) it is possible to create a 'one-time' report and mail this to the relevant gmail account. It will then still be processed by this Telescope.
The settings are the same as for the scheduled report.

Using the Gmail API

See the google support answer for info on how to enable an API. Search for the Gmail API and enable this.

Creating the Gmail API connection and credentials

Currently, the telescope works only with a Gmail account that is an internal user (a G-suite account).
It is possible to create credentials for an external user with a project status of 'Testing' in the OAuth screen, however refresh tokens created in such a project expire after 7 days and the telescope does not handle expired refresh tokens. See

the documentation for more info on OAuth refresh token expiration.

Create OAuth credentials

- In the IAM section add the G-suite account you would like to use as a user.
- From the 'APIs & Services' section, click the 'Credentials' menu item.
- Click 'Create Credentials' and choose OAuth client ID.
- In the form, enter the following information:
  - Application type: Web application
  - Name: Can be anything, e.g. 'Gmail API'
  - Authorized redirect URIs: add the URI: http://localhost:8080/
  - Click 'Create'
- Download the client secrets file for the newly created OAuth 2.0 Client ID, by clicking the download icon for the client ID that you created. The file will be named something like `client_secret_token.apps.googleusercontent.com.json`
- Get the credentials info using the JSON file with client secret info by executing the following python code.

Note that there is currently a limit of 50 refresh tokens per client ID. If the limit is reached, creating a new refresh token automatically invalidates the oldest refresh token without warning. Additionally, tokens are invalidated whenever an account's password is reset.

```
import urllib.parse
from google_auth_oauthlib.flow import InstalledAppFlow

# When modifying these scopes, recreate the file token.json
SCOPES = ['https://www.googleapis.com/auth/gmail.readonly', 'https://www.
flow = InstalledAppFlow.from_client_secrets_file('/path/to/client_secret_

# This will open a pop-up, authorize the Gmail account you want to use
creds = flow.run_local_server(access_type='offline', approval_prompt='for

# Get the necessary credentials info
token = urllib.parse.quote(creds.token, safe='')
refresh_token = urllib.parse.quote(creds.refresh_token, safe='')
client_id = urllib.parse.quote(creds.client_id, safe='')
client_secret = urllib.parse.quote(creds.client_secret, safe='')

# This connection can be used in the config file
gmail_api_conn = f'google-cloud-platform://?token={token}&refresh_token={
```

Telescope Tasks

The telescope tasks revolve around the retrieval and processing of the correct JSTOR report for each telescope run. It should be noted that the JSTOR report format has been slightly altered on more than one occasion. So depending on the date of the telescope run (and the corresponding date of the report), there may be a variation in the report structure. The telescope is smart enough to account for the different structures.

Data Download

All reports sent by JSTOR over email appear identical until they are opened. This presents an unfortunate issue, as it is therefore necessary to download and open all of the reports to see which of them pertain to the relevant partner and have the expected date intervals. To avoid an increasing amount of unnecessary downloads, the telescope will filter the emails (from which it extracts the report) based on their label. Any email that contains the *processed report* label will be ignored. This label is added to the emails processed at the final (*cleanup*) step of the telescope.

Once all of the possible reports have been examined, the matching reports (there should be only two for each telescope run - country and institution) are moved to a permanent location in the local file system ready for transforming.

Data Transform

The transformation of the report data is quite simple. The data is read from its .tsv format into a python dictionary, keys are converted into BigQuery-compatible column names and the result is saved to a gzipped .jsonl.

BigQuery Load

The transformed data is loaded from the Google Cloud bucket. There are two resulting datasets from each telescope run, both of which will be loaded into their own partitioned BigQuery table under the *jstor* dataset (which will be created should it not exist yet). Then, the *jstor_country* and *jstor_institution* table partitions are loaded. Since the data is partitioned on the release month, there will only be a single table for each of these report types.

Table Schemas

Publisher

Country

name	type	mode	description
Country_Name	STRING	NULLABLE	Country Name.
Book_Title	STRING	REQUIRED	Title of the book.
Book_ID	STRING	NULLABLE	DOI of the book on JSTOR.
Authors	STRING	NULLABLE	Author of the book.
ISBN	STRING	NULLABLE	ISBN of the book (13 digits).
eISBN	STRING	NULLABLE	ISBN of the digital version of the book (13 digits).
Copyright_Year	INTEGER	NULLABLE	Publication year.
Disciplines	STRING	REQUIRED	Subject category of the book.
Usage_Type	STRING	NULLABLE	For our case it is Open Access.
Usage_Month	STRING	REQUIRED	Date (as month and year) of the request.
Total_Item_Requests	INTEGER	NULLABLE	Total number of request made from that specific country.
release_date	DATE	REQUIRED	Last day of the release month. Table is partitioned on this column.

Institution

name	type	mode	description
Institution	STRING	NULLABLE	Institution name.
Book_Title	STRING	REQUIRED	Title of the book.
Book_ID	STRING	NULLABLE	DOI of the book on JSTOR.
Authors	STRING	NULLABLE	Author of the book.
ISBN	STRING	NULLABLE	ISBN of the book (13 digits).
eISBN	STRING	NULLABLE	ISBN of the digital version of the book (13 digits).
Copyright_Year	INTEGER	NULLABLE	Publication year.
Disciplines	STRING	REQUIRED	Subject category of the book.
Usage_Type	STRING	NULLABLE	For our case it is Open Access.
Usage_Month	STRING	REQUIRED	Date (as month and year) of the request.
Total_Item_Requests	INTEGER	NULLABLE	Total number of request made from that specific country.
release_date	DATE	REQUIRED	Last day of the release month. Table is partitioned on this column.

Collection

Country

name	type	mode	description
Country_Name	STRING	NULLABLE	Country Name.
Book_Title	STRING	REQUIRED	Title of the book.
Book_ID	STRING	NULLABLE	DOI of the book on JSTOR.
Publisher	STRING	NULLABLE	The publisher of the book.
Authors	STRING	NULLABLE	Author of the book.
ISBN	STRING	NULLABLE	ISBN of the book (13 digits).
eISBN	STRING	NULLABLE	ISBN of the digital version of the book (13 digits).
Copyright_Year	INTEGER	NULLABLE	Publication year.
Disciplines	STRING	NULLABLE	Subject category of the book.
Usage_Type	STRING	NULLABLE	For our case it is Open Access.
Usage_Month	STRING	NULLABLE	Date (as month and year) of the request.
Total_Item_Requests	INTEGER	NULLABLE	Total number of request made from that specific country.
release_date	DATE	REQUIRED	Last day of the release month. Table is partitioned on this column.

Insitution

name	type	mode	description
Institution	STRING	NULLABLE	Institution name.
Book_Title	STRING	REQUIRED	Title of the book.
Book_ID	STRING	NULLABLE	DOI of the book on JSTOR.
Publisher	STRING	NULLABLE	The publisher of the book.
Authors	STRING	NULLABLE	Author of the book.
ISBN	STRING	NULLABLE	ISBN of the book (13 digits).
eISBN	STRING	NULLABLE	ISBN of the digital version of the book (13 digits).
Copyright_Year	INTEGER	NULLABLE	Publication year.
Disciplines	STRING	NULLABLE	Subject category of the book.
Usage_Type	STRING	NULLABLE	For our case it is Open Access.
Usage_Month	STRING	NULLABLE	Date (as month and year) of the request.
Total_Item_Requests	INTEGER	NULLABLE	Total number of request made from that specific country.
release_date	DATE	REQUIRED	Last day of the release month. Table is partitioned on this column.

UCL Discovery

UCL Discovery is UCL's open access repository, showcasing and providing access to the full texts of UCL research publications.

Dataset Name	ucl
Table Name	ucl_discovery
Table Type	Partitioned
Average Runtime	5-10 min
Average Download Size	~1 MB
Harvest Type	API
Run Schedule	Monthly on the 4th
Catch-up Missed Runs	✅
Each Run Includes All Data	❌

The Google Sheet

UCL's titles are referenced via their identifier - the eprint ID. Their metadata maps the eprint ID to an ISBN13, but not consistently. For this reason, we forgo the use of their metadata and instead employ a semi-manual process to reliably map the two identifiers. The telescope references a Google sheet that contains all of the titles available in the UCL Discovery repository under the following headings:

ISBN13	The title's ISBN13
date	The date of publication
title_list_title	The title of the publication
discovery_eprint_id	The eprint ID of the publication

Some notes:

- These headings are hardcoded into the telescope. Any change in the sheet will break the telescope without prior intervention.
- Entries without a publication date or with a publication date in the future (where the current time is determined by the airflow scheduler) will be ignored.
- Entries missing either an ISBN13 or eprint ID will be ignored.

For the aforementioned reasons, it is important that **the google sheet remains up to date**. Otherwise, the usage for a title may be missed and require a rerun.

Access

Access to the sheet can be granted using the sheet UI (*Share* at the top right of the page). The telescope will access the sheet via a service account, which will need to be given read access (*Viewer*) by supplying the account's email address.

Telescope kwargs

Sheet ID (sheet_id)

The ID of the google sheet. The ID can be found in its URI, which will have the form of `https://docs.google.com/spreadsheets/d/[SHEET_ID]`.

UCL Discovery Usage API

UCL Discovery provides free and open access to their usage REST API. Unfortunately, I can't find any documentation on its use and design. We utilise two endpoints:

> **Countries URI** =
>
> ```
> https://discovery.ucl.ac.uk/cgi/stats/get?
> from=[YYYYMMDD]&to=[YYYYMMDD]&irs2report=eprint&set_name=eprint&set_val
> ue=[EPRINT_ID]&datatype=countries&top=countries&view=Table&limit=all&ex
> port=JSON
> ```

> **Totals URI** =
>
> ```
> https://discovery.ucl.ac.uk/cgi/stats/get?
> from=[YYYYMMDD]&to=[YYYYMMDD]&irs2report=eprint&set_name=eprint&set_val
> ue=[EPRINT_ID]&datatype=downloads&graph_type=column&view=Google%3A%3AGr
> aph&date_resolution=month&title=Download+activity+-
> +last+12+months&export=JSON
> ```

Where *from*, *to* and *set_value* are appropriately set. The countries URI returns statistics pertaining to the number of downloads of the provided eprint ID broken down by country. The totals URI returns statistics pertaining to the number of downloads of the provided eprint ID aggregated over all regions. It should be noted that the *totals* data is not necessarily a simply aggregation of the *countries* data. This is because country data is omitted for downloads that are not attributed to a region. It is therefore not uncommon to have a total download count (derived from the totals URI) that is greater than the sum of all downloads from all listed countries (from the countries URI).

Telescope Tasks

Data Download

Acquires the eprint IDs and publication dates from the Google Sheet. For each ID that has a publication date that is before the current scheduled run date, download

the country and totals data using the API. Then upload to the GCS *download* bucket.

Data Transform

Acquires the eprint IDs, ISBN13s and titles from the Google Sheet. For each ID, load the downloaded data (both countries and totals) into a single data structure and include the title (whether it is empty or not does not matter - the title exists for completeness only). Add an additional field to each row - the *release_date* which is determined by the scheduled runtime. Upload this transformed structure to GCS *transform* bucket.

BigQuery Load

The transformed data is loaded from the Google Cloud bucket into a partitioned BigQuery table. The table is in the *ucl* dataset (which will be created should it not exist yet). Since the data is partitioned on the release month, there will only be a single table named *ucl_discovery*.

Table Schema

name	type	mode	description
ISBN	STRING	REQUIRED	ISBN13 of the book.
eprint_id	STRING	REQUIRED	eprint ID of the book.
title	STRING	NULLABLE	Title of the book.
timescale	RECORD	NULLABLE	Timescale of the statistics as reported by the origin.
timescale.format	STRING	NULLABLE	Format of the 'to' and 'from' fields
timescale.from	STRING	NULLABLE	Beginning of date range for the statistics
timescale.to	STRING	NULLABLE	End of date range for the statistics
origin	RECORD	NULLABLE	Origin of the statistics
origin.url	STRING	NULLABLE	The URL of the origin
origin.name	STRING	NULLABLE	The name of the origin
total_downloads	INTEGER	NULLABLE	The aggregated statistics for the reported period
country	RECORD	REPEATED	The aggregated statistics for each reported country
country.value	STRING	NULLABLE	The two letter country code.
country.count	INTEGER	NULLABLE	The total number of item downloads for the reported period from this country.
release_date	DATE	REQUIRED	Last day of the release month. Table is partitioned on this column.

Metadata Telescopes

Telescopes that pull book metadata for a publisher or collection

OAPEN Metadata

The OAPEN Metadata telescope collects data from the OAPEN Metadata feed. OAPEN enables libraries and aggregators to use the metadata of all available titles in the OAPEN Library.
OAPEN metadata is available in different formats and this telescope harvests the data in the XML format.
See the OAPEN Metadata webpage for more information.

Dataset Name	onix
Table Name	onix
Table Type	Sharded
Average Runtime	10 min
Average Download Size	1-200 MB
Harvest Type	URI
Run Schedule	Weekly
Catch-up Missed Runs	❌
Each Run Includes All Data	✅

Telescope Configuration

Telescope kwargs

Fields passed as keyword arguments to the telescope upon instantiation.

Metadata URI (metadata_uri)

This field holds the URI for the publisher/collection. For example, the OAPEN's ONIX feed URI is as follows:

```
kwargs:
    metadata_uri: "https://library.oapen.org/download-export?format=onix"
```

The URI must either be known internally, or in some cases can be retrieved from the OAPEN Metadata page.

Elevate Related Products (related_products_elevation)

```
kwargs:
      related_product_elevation: True
```

A boolean value ("True" | "False") that determines whether the transform step should elevate the feed's related products.

Related Product Manipulation

Some steps of the transform process manipulate the Related Products of the metadata feed in a way that alters the state of the input data. These processes deserve an explanation as it's not obvious what they're doing and why.

Related Product Elevation

The OAPEN Metadata telescope has the option to manipulate the feed's related product entries. **Each Related Product in each entry is turned into its own entry**. For each of these fabricated entries, the reference to itself as a related product is removed and replaced with the product identifier of the original entry. This can lead to many more entries than the original.

There elevation process will only apply to a Related Product if:

- The related product has an ISBN (product identifier code 15 as described in the

codelist)

> The relation code is "06" (ie. alternative format as described in the codelist)

For example, if we have a Product that looks like this:

```
<Product>
    <ProductIdentifier>
            <IDValue>100</IDValue>
            <ProductIDType>15</ProductIDType>
    </ProductIdentifier>
    <RelatedProduct>
        <ProductRelationCode>06</ProductRelationCode>
        <ProductIdentifier>
            <IDValue>200</IDValue>
            <ProductIDType>15</ProductIDType>
        </ProductIdentifier>
    </RelatedProduct>
</Product>
```

The related product is elevated, creating a new product and keeping the original:

```
<!-- Original -->
<Product>
    <ProductIdentifier>
            <IDValue>100</IDValue>
            <ProductIDType>15</ProductIDType>
    </ProductIdentifier>
    <RelatedProduct>
        <ProductRelationCode>06</ProductRelationCode>
        <ProductIdentifier>
            <IDValue>200</IDValue>
            <ProductIDType>15</ProductIDType>
        </ProductIdentifier>
    </RelatedProduct>
</Product>

<!-- Elevated Related Product -->
<Product>
    <ProductIdentifier>
            <IDValue>200</IDValue>
            <ProductIDType>15</ProductIDType>
    </ProductIdentifier>
    <RelatedProduct>
        <ProductRelationCode>06</ProductRelationCode>
        <ProductIdentifier>
            <IDValue>100</IDValue>
            <ProductIDType>15</ProductIDType>
        </ProductIdentifier>
    </RelatedProduct>
</Product>
```

Normalise Related Products

Related products retrieved from the OAPEN feed tend to have a format that is not consistent with ONIX 3.0 standards. For example, the following is an invalid implementation:

```
<RelatedProduct>
    <ProductRelationCode></ProductRelationCode>
    <ProductIdentifier>
        <IDValue></IDValue>
        <ProductIDType></ProductIDType>
    </ProductIdentifier>
    <ProductIdentifier>
        <IDValue></IDValue>
        <ProductIDType></ProductIDType>
    </ProductIdentifier>
</RelatedProduct>
```

The only exception to this format is if the `IDValue` elements are identical and the `ProductIDType` elements are different.

The correct way is the following:

```
<RelatedProduct>
    <ProductRelationCode></ProductRelationCode>
    <ProductIdentifier>
        <IDValue></IDValue>
        <ProductIDType></ProductIDType>
    </ProductIdentifier>
</RelatedProduct>
<RelatedProduct>
    <ProductRelationCode></ProductRelationCode>
    <ProductIdentifier>
        <IDValue></IDValue>
        <ProductIDType></ProductIDType>
    </ProductIdentifier>
</RelatedProduct>
```

This is an important distinction as the ONIX parser will ignore any consecutive `ProductIdentifier` tags that have the same `ProductIDType`.

Related Product Deduplication

Occasionally, a product can have multiple of the same Related Products, or it may have a Related Product that has the same identifier as its parent - effectively referencing itself as a Related Product. Under either of these circumstances, these duplicated Related Products are removed.

For example:

```
<Product>
    <ProductIdentifier>
            <IDValue>100</IDValue>
            <ProductIDType>15</ProductIDType>
    </ProductIdentifier>
    <RelatedProduct>
        <ProductRelationCode>06</ProductRelationCode>
        <ProductIdentifier>
            <IDValue>100</IDValue>
            <ProductIDType>15</ProductIDType>
        </ProductIdentifier>
    </RelatedProduct>
    <RelatedProduct>
        <ProductRelationCode>06</ProductRelationCode>
        <ProductIdentifier>
            <IDValue>200</IDValue>
            <ProductIDType>15</ProductIDType>
        </ProductIdentifier>
    </RelatedProduct>
    <RelatedProduct>
        <ProductRelationCode>06</ProductRelationCode>
        <ProductIdentifier>
            <IDValue>200</IDValue>
            <ProductIDType>15</ProductIDType>
        </ProductIdentifier>
    </RelatedProduct>
</Product>
```

The above product has a copy of the Product as a Related Product (ID 100). It also has two identical Related Products (with ID 200). After the deduplication process, the product will look like the following:

```
<Product>
    <ProductIdentifier>
            <IDValue>100</IDValue>
            <ProductIDType>15</ProductIDType>
    </ProductIdentifier>
    <RelatedProduct>
        <ProductRelationCode>06</ProductRelationCode>
        <ProductIdentifier>
            <IDValue>200</IDValue>
            <ProductIDType>15</ProductIDType>
        </ProductIdentifier>
    </RelatedProduct>
</Product>
```

Telescope Tasks

Data Download

This is where the metadata is downloaded. The XML file containing metadata is downloaded using the Metadat URI.

Note that if the metadata file is part-way through an update (occurring daily at +0000GMT and taking upwards of one hour), the XML file will be incomplete and invalid. The telescope has a failesafe to attempt to resolve this during runtime, which can lead to much longer than normal 'download' times.

Data Transform

The transform step modifies the downloaded metadata into a valid ONIX format. This is done in a few steps:

1. The XML is loaded and all unnecessary fields are removed. The necessary fields for the BAD workflows are described by a corresponding schema (.json) file.
2. The resulting XML is parsed through the Python onixcheck. This reveals any remaining invalid products. These products are **removed** from the file. The removed products are saved to a separate file and uploaded to the transform

bucket for storage/archiving.

3. Any Related Products that are incorrectly formatted will be fixed through the normalisation process.
4. Duplicated Related Products are removed via the deduplication process.
5. Optionally, the Related Products can be elevated through the Related Product elevation process.
6. The XML is then parsed throuh the Java ONIX Parser, which results in a .jsonl file.
7. The *PersonName* and *InvertedPersonName* fields are created (where possible) from the *KeyNames* and *NamesBeforeKey* fields.
8. Subject fields are collapsed (converted to a single semicolon-separated string) to match our expected input

BigQuery Load

The valid ONIX feed can now be loaded from the transform bucket into a BigQuery date-sharded table in the *onix* dataset (which will be created if it does not yet exist). There will be multiple *onix_YYYYMMDD* tables.

Table Schema

The OAPEN Metadata table uses the same schema as the ONIX Telescope's table. See the ONIX Telescope schema.

ONIX

Documentation for the ONIX telescope

The ONIX telescope downloads, transforms and loads publisher ONIX feeds into BigQuery. ONIX is a standard format that book publishers use to share information about the books that they have published.

Book publishers with ONIX feeds are given credentials and access to their own upload folder on the OAeBU SFTP server. They then configure ONIX Suite to upload their ONIX feeds to the SFTP server on a weekly basis. The ONIX feeds need to be **full dumps every time,** not incremental updates.

Dataset Name	onix
Table Name	onix
Table Type	Sharded
Average Runtime	10 min
Average Download Size	10-100 MB
Harvest Type	SFTP Server
Run Schedule	Weekly
Catch-up Missed Runs	❌
Each Run Includes All Data	✅

Telescope Configuration

Airflow connections

Name	Description
sftp_service	The username, password and host name used to connect to the SFTP server

Telescope kwargs

Fields passed as keyword arguments to the telescope upon instantiation.

Date Regular Expression (date_regex)

This field is used to extract the date from the ONIX feed file name. For example, the regex `\\d{8}` will extract the date from the file name `20220301_CURTINPRESS_ONIX.xml`.

```
kwargs:
    date_regex: "\\d{8}"
```

Telescope Tasks

Data Download

Discovers all files in the partner's SFTP server folder that match the supplied date regex pattern. These files are downloaded the the local file system for transforming.

Data Transform

In order to convert from the .xml format into one suitable for loading into BigQuery, the ONIX telescope utilises the the Java ONIX parser. The parser is Java based in order to leverage the Jonix-onix3 library. The output of the parser is a .jsonl file, which makes for simple Pythonic interpretation.

An additional step in the transform task collapses the subjects (Subjects.SubjectHeadingText) into a semicolon-separated string.

BigQuery Load

The transformed data is loaded from the Google Cloud bucket into a date-sharded BigQuery table in the *onix* dataset, which will be created if it does not yet exist. There will be multiple *onix_YYYYMMDD* tables.

Table Schema

name	type	mode	description
CountryOfManufacture	STRING	NULLABLE	An ISO code identifying the country of manufacture of a single-item product, or of a multiple-item product when all items are manufactured in the same country. This information is needed in some countries to meet regulatory requirements. Optional and non-repeating.
RecordSourceName	STRING	NULLABLE	The name of the party which issued the record, as free text. Optional and non-repeating, independently of the occurrence of any other field.
RecordSourceType	STRING	NULLABLE	An ONIX description which indicates the type of source which has issued the ONIX record. Optional and non-repeating, independently of the occurrence of any other field.
LCCN	STRING	NULLABLE	Library of Congress Control Number
Collections	RECORD	REPEATED	A bibliographic collection in ONIX 3.0 means a fixed or indefinite number of products, published over a fixed or indefinite time period, which share collective attributes (including a collective title) that are required as part of the bibliographic record of each individual product. In this respect, such a collection is most often thought of as a series. A bibliographic collection may, however, also be traded as a single product (often thought of as a set), but this does not alter the way in which its collective attributes

			which its collective attributes are described in the ONIX records for the individual products.
Collections.TitleDetails	RECORD	REPEATED	A group of data elements which together give the text of a title and specify its type. At least one title detail element is mandatory in each occurrence of the composite, to give the primary form of the product title. The composite is repeatable with different title types.
Collections.TitleDetails.TitleElements	RECORD	REPEATED	A group of data elements which together represent an element of a title. At least one title element is mandatory in each occurrence of the composite. The composite is repeatable with different sequence numbers and/or title element levels, each repeat carrying a different part of the title. An instance of the composite must include at least one of: ; ; , together with , or together with . In other words it must carry either the text of a title or a part or year designation; and it may carry both.
Collections.TitleDetails.TitleElements.PartNumber	RECORD	NULLABLE	When a title element includes a part designation within a larger whole (eg Part I, or Volume 3), this field should be used to carry the number and its ‘caption’ as text. Optional and non-repeating.
Collections.TitleDetails.TitleElements.PartNumber.Value	STRING	NULLABLE	PartNumber value.

Collections.TitleDetails.TitleElements.TitleElementLevel	STRING	NULLABLE	An ONIX description indicating the level of a title element: collection level, subcollection level, or product level. Mandatory in each occurrence of the composite, and non-repeating.
Collections.TitleDetails.TitleElements.TitlePrefix	RECORD	NULLABLE	Text at the beginning of a title element which is to be ignored for alphabetical sorting. Optional and non- repeating; can only be used when is omitted, and if the element is also present. These two elements may be used in combination in applications where it is necessary to distinguish an initial word or character string which is to be ignored for filing purposes, eg in library systems and in some bookshop databases.
Collections.TitleDetails.TitleElements.TitlePrefix.Value	STRING	NULLABLE	TitlePrefix value.
Collections.TitleDetails.TitleElements.TitleWithoutPrefix	STRING	NULLABLE	The text of a title element without the title prefix; and excluding any subtitle. Optional and non-repeating; can only be used if one of the or elements is also present.
Collections.TitleDetails.TitleElements.SequenceNumber	INTEGER	NULLABLE	A number which specifies a single overall sequence of title elements, which is the preferred order for display of the various title elements when constructing a complete title. Optional and non-repeating. It is strongly recommended that

be.			is strongly recommended that where there are multiple title elements within a composite, each occurrence of the composite should carry a .
Collections.TitleDetails.TitleElements.TitleText	STRING	NULLABLE	The text of a title element, excluding any subtitle. Optional and non-repeating, may only be used where , and are not used.
Collections.TitleDetails.TitleElements.Subtitle	STRING	NULLABLE	The text of a subtitle, if any. ‘Subtitle‘ means any added words which appear with the title element given in an occurrence of the composite, and which amplify and explain the title element, but which are not considered to be part of the title element itself. Optional and non-repeating.
Collections.TitleDetails.TitleType	STRING	NULLABLE	An ONIX description indicating the type of a title. Mandatory in each occurrence of the composite, and non-repeating.
Collections.CollectionIdentifers	RECORD	REPEATED	A repeatable group of data elements which together specify an identifier of a bibliographic collection. The composite is optional, and may only repeat if two or more identifiers of different types are sent for the same collection. It is not permissible to have two identifiers of the same type.
Collections.CollectionIdentifers.CollectionIdType	STRING	NULLABLE	An ONIX description identifying a scheme from which an identifier in the element is taken. Mandatory in each occurrence of the composite, and non-repeating.

Collections.CollectionIdentifers.IDValue	STRING	NULLABLE	An identifier of the type specified in the field. Mandatory in each occurrence of the composite, and non-repeating.
Collections.CollectionIdentifers.IDTypeName	STRING	NULLABLE	A name which identifies a proprietary identifier scheme (ie a scheme which is not a standard and for which there is no individual ID type code). Must be used when, and only when, the code in indicates a proprietary scheme, eg a publisher's own code. Optional and non-repeating.
Collections.CollectionType	STRING	NULLABLE	An ONIX description indicating the type of a collection: publisher collection, ascribed collection, or unspecified. Mandatory in each occurrence of the composite, and non-repeating.
EditionNumber	INTEGER	NULLABLE	The number of a numbered edition. Optional and non-repeating. Normally sent only for the second and subsequent editions of a work, but by agreement between parties to an ONIX exchange a first edition may be explicitly numbered.
RecordRef	STRING	NULLABLE	Two mandatory data elements must be included at the beginning of every product record or update. The first, , is a string of text which uniquely identifies the record. The second, , is a code which specifies the type of notification or update.

RelatedWorks	RECORD	REPEATED	A group of data elements which together describe a work which has a specified relationship to a content item. Optional and repeatable.
RelatedWorks.WorkRelationCode	STRING	NULLABLE	An ONIX description which identifies the nature of the relationship between a product and a work. Mandatory in each occurrence of the composite, and non-repeating.
RelatedWorks.WorkIdentifiers	RECORD	REPEATED	A group of data elements which together define an identifier of a work in accordance with a specified scheme. Mandatory in each occurrence of the composite, and repeatable only if two or more identifiers for the same work are sent using different identifier schemes (eg ISTC and DOI).
RelatedWorks.WorkIdentifiers.WorkIDType	STRING	NULLABLE	An ONIX description identifying the scheme from which the identifier in the element is taken. Mandatory in each occurrence of the composite, and non-repeating.
RelatedWorks.WorkIdentifiers.IDValue	STRING	NULLABLE	An identifier of the type specified in the element. Mandatory in each occurrence of the composite, and non-repeating.
RelatedWorks.WorkIdentifiers.IDTypeName	STRING	NULLABLE	A name which identifies a proprietary identifier scheme (ie a scheme which is not a standard and for which there is no individual ID type code). Must be used when, and only when, the code in the element indicates a proprietary scheme

			indicates a proprietary scheme. Optional and non-repeating.
TextContent	RECORD	REPEATED	An optional group of data elements which together carry text related to the product, repeatable in order to deliver multiple texts (often of different types, though for some text types, there may be multiple instances of that type).
TextContent.TextType	STRING	NULLABLE	An ONIX description which identifies the type of text which is sent in the element. Mandatory in each occurrence of the composite, and non-repeating.
TextContent.Text	STRING	REPEATED	The text specified in the element. Mandatory in each occurrence of the composite, and repeatable when essentially identical text is supplied in multiple languages. The language attribute is optional for a single instance of , but must be included in each instance if is repeated.
CityOfPublications	STRING	REPEATED	The name of a city or town associated with the imprint or publisher. Optional, and repeatable if parallel names for a single location appear on the title page in multiple languages, or if the imprint carries two or more cities of publication.
DOI	STRING	NULLABLE	The product's Digital object identifier.
EditionType	STRING	REPEATED	An ONIX description, indicating the type of a version or edition. Optional, and repeatable if the

EditionType	STRING	REPEATED	product has characteristics of two or more types (eg 'revised' and 'annotated').
Imprints	RECORD	REPEATED	An optional group of data elements which together identify an imprint or brand under which the product is marketed. The composite must carry either a name identifier or a name or both, and is repeatable to specify multiple imprints or brands.
Imprints.ImprintIdentifiers	RECORD	REPEATED	A group of data elements which together define the identifier of an imprint name. Optional, but mandatory if the composite does not carry an . The composite is repeatable in order to specify multiple identifiers for the same imprint or brand.
Imprints.ImprintIdentifiers.ImprintIDType	STRING	NULLABLE	An ONIX description which identifies the scheme from which the value in the element is taken. Mandatory in each occurrence of the composite.
Imprints.ImprintIdentifiers.IDValue	STRING	NULLABLE	A code value taken from the scheme specified in the element. Mandatory in each occurrence of the composite, and non-repeating..
Imprints.ImprintIdentifiers.IDTypeName	STRING	NULLABLE	A name which identifies a proprietary identifier scheme (ie a scheme which is not a standard and for which there is no individual ID type code). Must be used when, and only when, the code in the element indicates a proprietary scheme. Optional and non-repeating.

Imprints.ImprintName	STRING	NULLABLE	The name of an imprint or brand under which the product is issued, as it appears on the product. Mandatory if there is no imprint identifier in an occurrence of the composite, and optional if an imprint identifier is included. Non-repeating.
Publishers	RECORD	REPEATED	An optional group of data elements which together identify an entity which is associated with the publishing of a product. The composite allows additional publishing roles to be introduced without adding new fields. Each occurrence of the composite must carry a publishing role code and either a name identifier or a name or both, and the composite is repeatable in order to identify multiple entities.
Publishers.PublisherName	STRING	NULLABLE	The name of an entity associated with the publishing of a product. Mandatory if there is no publisher identifier in an occurrence of the composite, and optional if a publisher identifier is included. Non-repeating.
Publishers.Websites	RECORD	REPEATED	An optional group of data elements which together identify and provide a pointer to a website which is related to the publisher identified in an occurrence of the composite. Repeatable in order to provide links to multiple websites.

Publishers.Websites.WebsiteDescriptions	STRING	REPEATED	Free text describing the nature of the website which is linked through the element. Optional, and repeatable to provide parallel descriptive text in multiple languages. The language attribute is optional for a single instance of , but must be included in each instance if is repeated.
Publishers.Websites.WebsiteRole	STRING	NULLABLE	An ONIX description which identifies the role or purpose of the website which is linked through the element. Optional and non-repeating.
Publishers.Websites.WebsiteLinks	STRING	REPEATED	The URL for the website. Mandatory in each occurrence of the composite, and repeatable to provide multiple URLs where the website content is available in multiple languages. The language attribute is optional for a single instance of , but must be included in each instance if is repeated.
Publishers.PublishingRole	STRING	NULLABLE	An ONIX description which identifies a role played by an entity in the publishing of a product. Mandatory in each occurrence of the composite, and non-repeating.
RelatedProducts	RECORD	REPEATED	A group of data elements which together describe a product which has a specified relationship to a content item. Optional and repeatable.
RelatedProducts.ISBN13	STRING	NULLABLE	The related product's 13-digit International Standard Book Number.

RelatedProducts.ProductForm	STRING	NULLABLE	An ONIX description which indicates the primary form of a related product. Optional in an occurrence of , and non-repeating. If supplied, should be identical to the element supplied in the block of the full ONIX record describing the related product itself.
RelatedProducts.DOI	STRING	NULLABLE	The related product's digital object identifier.
RelatedProducts.GTIN_13	STRING	NULLABLE	The related product's 13-digit global trade item number.
RelatedProducts.ProductRelationCodes	STRING	REPEATED	An ONIX description which identifies the nature of the relationship between two products, eg ‘replaced-by’. Mandatory in each occurrence of the composite, and repeatable where the related product has multiple types of relationship to the product described.
RelatedProducts.PID_Proprietary	STRING	NULLABLE	The related product's proprietary product ID.
PID_Proprietary	STRING	NULLABLE	The product's proprietary product identifier.
ISBN10	STRING	NULLABLE	The product's 10-digit International Standard Book Number.
ISBN13	STRING	NULLABLE	The product's 13-digit International Standard Book Number.
			A group of data elements which together give the text

TitleDetails	RECORD	REPEATED	of a title and specify its type. At least one title detail element is mandatory in each occurrence of the composite, to give the primary form of the product title. The composite is repeatable with different title types.
TitleDetails.TitleElements	RECORD	REPEATED	A group of data elements which together represent an element of a title. At least one title element is mandatory in each occurrence of the composite. The composite is repeatable with different sequence numbers and/or title element levels, each repeat carrying a different part of the title. An instance of the composite must include at least one of: ; ; , together with , or together with . In other words it must carry either the text of a title or a part or year designation; and it may carry both.
TitleDetails.TitleElements.TitleWithoutPrefix_TextCaseFlags	STRING	NULLABLE	TitleWithoutPrefix textcase attribute.
TitleDetails.TitleElements.TitleText_TextCaseFlags	STRING	NULLABLE	TitleText textcase attribute.
TitleDetails.TitleElements.Subtitle_TextCaseFlags	STRING	NULLABLE	Subtitle textcase attribute.
TitleDetails.TitleElements.TitleText	STRING	NULLABLE	The text of a title element, excluding any subtitle. Optional and non-repeating, may only be used where , and are not used.

TitleDetails.TitleElements.TitleText_Language	STRING	NULLABLE	TitleText language attribute.
TitleDetails.TitleElements.TitleElementLevel	STRING	NULLABLE	An ONIX description indicating the level of a title element: collection level, subcollection level, or product level. Mandatory in each occurrence of the composite, and non-repeating.
TitleDetails.TitleElements.Subtitle	STRING	NULLABLE	The text of a subtitle, if any. ‘Subtitle’ means any added words which appear with the title element given in an occurrence of the composite, and which amplify and explain the title element, but which are not considered to be part of the title element itself. Optional and non-repeating.
TitleDetails.TitleElements.SequenceNumber	INTEGER	NULLABLE	A number which specifies a single overall sequence of title elements, which is the preferred order for display of the various title elements when constructing a complete title. Optional and non-repeating. It is strongly recommended that where there are multiple title elements within a composite, each occurrence of the composite should carry a .
TitleDetails.TitleElements.TitleWithoutPrefix	STRING	NULLABLE	The text of a title element without the title prefix; and excluding any subtitle. Optional and non-repeating; can only be used if one of the or element is also present.
TitleDetails.TitleE			

lements.Subtitle_Language	STRING	NULLABLE	Language attribute.
TitleDetails.TitleElements.TitlePrefix	RECORD	NULLABLE	Text at the beginning of a title element which is to be ignored for alphabetical sorting. Optional and non- repeating; can only be used when is omitted, and if the element is also present. These two elements may be used in combination in applications where it is necessary to distinguish an initial word or character string which is to be ignored for filing purposes, eg in library systems and in some bookshop databases.
TitleDetails.TitleElements.TitlePrefix.Value	STRING	NULLABLE	TitlePrefix value.
TitleDetails.TitleType	STRING	NULLABLE	An ONIX description indicating the type of a title. Mandatory in each occurrence of the composite, and non-repeating.
TitleDetails.TitleStatement	STRING	NULLABLE	Free text showing how the overall title (including any collection level title, if the collection title is treated as part of the product title and included in P.6) should be presented in any display, particularly when a standard concatenation of individual title elements from Group P.6 (in the order specified by the data elements) would not give a satisfactory result. Optional and non-repeating. When this field is sent, the recipient should use it to replace all title detail sent in Group P.6 for

			display purposes only. The individual title element detail must also be sent, for indexing and retrieval purposes.
PublishingDates	RECORD	REPEATED	A group of data elements which together specify a date associated with the publishing of the product. Optional, but where known, at least a date of publication must be specified either here (as a 'global' pub date) or in (P.25). Other dates related to the publishing of a product can be sent in further repeats of the composite
PublishingDates.PublishingDateRole	STRING	NULLABLE	An ONIX description indicating the significance of the date, eg publication date, announcement date, latest reprint date. Mandatory in each occurrence of the composite, and non-repeating.
PublishingDates.Date	INTEGER	NULLABLE	The date specified in the field. Mandatory in each occurrence of the composite, and non-repeating. may carry a dateformat attribute: if the attribute is missing, then indicates the format of the date; if both dateformat attribute and element are missing, the default format is YYYYMMDD.
GTIN_13	STRING	NULLABLE	The product's 13-digit global trade item number.
Languages	RECORD	REPEATED	A group of data elements which together represent a language, and specify its role and, where required, whether it is a country variant.

			Optional, and repeatable to specify multiple languages and their various roles.
Languages.CountryCode	STRING	NULLABLE	A code identifying the country when this specifies a variant of the language, eg US English. Optional and non-repeating.
Languages.LanguageRole	STRING	NULLABLE	An ONIX description indicating the ‘role’ of a language in the context of the ONIX record. Mandatory in each occurrence of the composite, and non-repeating.
Languages.LanguageCode	STRING	NULLABLE	An ISO code indicating a language. Mandatory in each occurrence of the composite, and non- repeating.
ProductForm	STRING	NULLABLE	An ONIX description which indicates the primary form of a related product. Optional in an occurrence of , and non-repeating. If supplied, should be identical to the element supplied in the block of the full ONIX record describing the related product itself.
Contributors	RECORD	REPEATED	A group of data elements which together describe a personal or corporate contributor to the product. Optional, and repeatable to describe multiple contributors.
Contributors.LettersAfterNames	STRING	NULLABLE	The seventh part of a structured name of a person who contributed to the creation of the product: qualifications and honors following a person’s names, eg ‘CBE FRS’. Optional and non-repeating.

Contributors.Gender	STRING	NULLABLE	An optional ONIX code specifying the gender of a personal contributor. Not repeatable. Note that this indicates the gender of the contributor's public identity (which may be pseudonymous) based on designations used in ISO 5218, rather than the gender identity, biological sex or sexuality of a natural person.
Contributors.Proprietary	INTEGER	NULLABLE	The contributor's proprietary identifier.
Contributors.NameType	STRING	NULLABLE	An ONIX description indicating the type of a primary name. Optional, and non-repeating. If omitted, the default is 'unspecified'.
Contributors.ProfessionalAffiliations	RECORD	REPEATED	An optional group of data elements which together identify a contributor's professional position and/or affiliation, repeatable to allow multiple positions and affiliations to be specified.
Contributors.ProfessionalAffiliations.Positions	STRING	REPEATED	A professional position held by a contributor to the product at the time of its creation. Optional, and repeatable to provide parallel text in multiple languages. The language attribute is optional for a single instance of , but must be included in each instance if is repeated.
Contributors.ProfessionalAffiliatio	STRING	NULLABLE	An organization to which a contributor to the product was affiliated at the time of its creation. and – if the

ns.Affiliations			element is also present – where s/he held that position. Optional and non-repeating.
Contributors.ORCID	STRING	NULLABLE	A 16-digit ORCID ID that uniquely identifies the author.
Contributors.BiographicalNotes	RECORD	REPEATED	A biographical note about a contributor to the product. (See the composite in Group P.14 for a biographical note covering all contributors to a product in a single text.) Optional, and repeatable to provide parallel biographical notes in multiple languages. The language attribute is optional for a single instance of , but must be included in each instance if is repeated. May occur with a person name or with a corporate name. A biographical note in ONIX should always contain the name of the person or body concerned, and it should always be presented as a piece of continuous text consisting of full sentences. Some recipients of ONIX data feeds will not accept text which has embedded URLs. A contributor website link can be sent using the composite below.
Contributors.BiographicalNotes.TextFormat	STRING	NULLABLE	The textformat attribute.
Contributors.BiographicalNotes.Note	STRING	NULLABLE	The biographical note.
			The first part of a structured name of a person who

Contributors.TitlesBeforeNames	STRING	NULLABLE	name of a person who contributed to the creation of the product: qualifications and/or titles preceding a person's names, eg 'Professor' or 'HRH Prince' or 'Saint'. Optional and non- repeating: see Group P.7 introductory text for valid options.
Contributors.Roles	STRING	REPEATED	An ONIX description indicating the role played by a person or corporate body in the creation of the product. Mandatory in each occurrence of a composite, and may be repeated if the same person or corporate body has more than one role in relation to the product.
Contributors.Websites	RECORD	REPEATED	An optional group of data elements which together identify and provide a pointer to a website which is related to the person or organization identified in an occurrence of the composite. Repeatable to provide links to multiple websites.
Contributors.Websites.WebsiteDescriptions	STRING	REPEATED	Free text describing the nature of the website which is linked through the element. Optional, and repeatable to provide parallel descriptive text in multiple languages. The language attribute is optional for a single instance of , but must be included in each instance if is repeated.
Contributors.Websites.WebsiteRole	STRING	NULLABLE	An ONIX description which identifies the role or purpose of the website which is linked through the element.

			Optional and non-repeating.
Contributors.Websites.WebsiteLinks	STRING	REPEATED	The URL for the website. Mandatory in each occurrence of the composite, and repeatable to provide multiple URLs where the website content is available in multiple languages. The language attribute is optional for a single instance of , but must be included in each instance if is repeated.
Contributors.PersonNameInverted	STRING	NULLABLE	The name of a person who contributed to the creation of the product, presented with the element used for alphabetical sorting placed first ('inverted order'). Optional and non-repeating: see Group P.7 introductory text for valid options.
Contributors.Dates	RECORD	REPEATED	A group of data elements which together specify a date associated with the person or organization identified in an occurrence of the composite, eg birth or death. Optional, and repeatable to allow multiple dates to be specified.
Contributors.Dates.Date	INTEGER	NULLABLE	The date specified in the field. Mandatory in each occurrence of the composite, and non-repeating. may carry a dateformat attribute: if the attribute is missing, then indicates the format of the date; if both dateformat attribute and element are missing, the default format is YYYYMMDD.

Contributors.Dates.Role	STRING	NULLABLE	An ONIX description indicating the significance of the date in relation to the contributor name. Mandatory in each occurrence of the composite, and non-repeating.
Contributors.SequenceNumber	INTEGER	NULLABLE	A number which specifies a single overall sequence of title elements, which is the preferred order for display of the various title elements when constructing a complete title. Optional and non-repeating. It is strongly recommended that where there are multiple title elements within a composite, each occurrence of the composite should carry a .
Contributors.PrefixToKey	STRING	NULLABLE	The third part of a structured name of a person who contributed to the creation of the product: a prefix which precedes the key name(s) but which is not to be treated as part of the key name, eg ‘van’ in Ludwig van Beethoven. This element may also be used for titles that appear after given names and before key names, eg ‘Lord’ in Alfred, Lord Tennyson. Optional and non-repeating.
Contributors.KeyNames	STRING	NULLABLE	The fourth part of a structured name of a person who contributed to the creation of the product: key name(s), ie the name elements normally used to open an entry in an alphabetical list, eg ‘Smith’ or ‘Garcia Marquez’ or ‘Madonna’ or ‘Francis de Sales’ (in Saint Francis de Sales). Non-

			repeating. Required if name part elements P.7.11 to P.7.18 are used.
Contributors.TitlesAfterNames	STRING	NULLABLE	The eighth part of a structured name of a person who contributed to the creation of the product: titles following a person's names, eg 'Duke of Edinburgh'. Optional and non-repeating.
Contributors.AlternativeNames	STRING	REPEATED	A group of data elements which together represent an alternative name of a contributor, and specify its type. The composite is optional, and is repeatable to provide multiple alternative names for the contributor.
Contributors.NamesBeforeKey	STRING	NULLABLE	The second part of a structured name of a person who contributed to the creation of the product: name(s) and/or initial(s) preceding a person's key name(s), eg James J. Optional and non-repeating.
Contributors.Places	RECORD	REPEATED	An optional group of data elements which together identify a geographical location with which a contributor is associated, used to support 'local interest' promotions. Repeatable to identify multiple geographical locations, each usually with a different relationship to the contributor.
Contributors.Places.CountryCode	STRING	NULLABLE	A code identifying a country with which a contributor is particularly associated. Optional and non-repeatable. There must be an occurrence

			of either the or the elements in each occurrence of .
Contributors.Places.Locations	STRING	REPEATED	The name of a city or town location within the specified country or region with which a contributor is particularly associated. Optional, and repeatable to provide parallel names for a single location in multiple languages (eg Baile Átha Cliath and Dublin, or Bruxelles and Brussel). The language attribute is optional for a single instance of , but must be included in each instance if is repeated.
Contributors.Places.Relation	STRING	NULLABLE	An ONIX description identifying the relationship between a contributor and a geographical location. Mandatory in each occurrence of and non-repeating.
Contributors.PersonName	STRING	NULLABLE	The name of a person who contributed to the creation of the product, unstructured, and presented in normal order. Optional and non-repeating: see Group P.7 introductory text for valid options.
Contributors.ISNI	STRING	NULLABLE	16-digit International Standard Name Identifier number.
Contributors.CorporateName	STRING	NULLABLE	The name of a corporate body which contributed to the creation of the product, unstructured. Optional and non-repeating: see Group P.7 introductory text for valid options.
[illegible]	[illegible]	[illegible]	The product's internal

COKI_ID	STRING	NULLABLE	COKI identifier.
Subjects	RECORD	REPEATED	An optional and repeatable group of data elements which together specify a subject classification or subject heading.
Subjects.SubjectHeadingText	STRING	REPEATED	The text of a subject heading taken from the scheme specified in the element, or of free language keywords if the scheme is specified as 'keywords'; or the text equivalent to the value, if both code and text are sent. Either or or both must be present in each occurrence of the composite.
Subjects.SubjectSchemeIdentifier	STRING	NULLABLE	A number which identifies a version or edition of the subject scheme specified in the associated element. Optional and non-repeating.
Subjects.SubjectSchemeVersion	FLOAT	NULLABLE	A number which identifies a version or edition of the subject scheme specified in the associated element. Optional and non-repeating.
Subjects.SubjectSchemeName	STRING	NULLABLE	A name identifying a proprietary subject scheme (ie a scheme which is not a standard and for which there is no individual identifier code) when is coded '24'. Optional and non-repeating.
Subjects.SubjectCode	STRING	NULLABLE	A subject class or category code from the scheme specified in the element. Either or or both must be present in each occurrence of

			present in each occurrence of the composite. Non-repeating.
Subjects.MainSubject	BOOLEAN	NULLABLE	An empty element that identifies an instance of the composite as representing the main subject category for the product. The main category may be expressed in more than one subject scheme, ie there may be two or more instances of the composite, using different schemes, each carrying the flag, so long as there is only one main category per scheme. Optional and non-repeating in each occurrence of the composite.
Extent	RECORD	REPEATED	A group of data elements which together describe an extent pertaining to the product. Optional, but in practice required for most products, eg to give the number of pages in a printed book or paginated e-book, or to give the running time of an audiobook. Repeatable to specify different extent types or units.
Extent.ExtentType	STRING	NULLABLE	An ONIX description which identifies the type of extent carried in the composite, eg running time for an audio or video product. Mandatory in each occurrence of the composite, and non-repeating. From Issue 9 of the code lists, an extended set of values for has been defined to allow more accurate description of pagination.
			The numeric value of the extent specified in . Optional,

Extent.ExtentValue	INTEGER	NULLABLE	and non-repeating. However, either or must be present in each occurrence of the composite; and it is very strongly recommended that should always be included, even when the original product uses Roman numerals.
Extent.ExtentUnit	STRING	NULLABLE	An ONIX description indicating the unit used for the and the format in which the value is presented. Mandatory in each occurrence of the composite, and non-repeating.
Extent.ExtentValueRoman	STRING	NULLABLE	The value of the extent expressed in Roman numerals. Optional, and non-repeating. Used only for page runs which are numbered in Roman.

Thoth

The Thoth Telescope downloads, transforms and loads publisher ONIX feeds from Thoth into BigQuery. ONIX is a standard format that book publishers use to share information about the books that they have published.

Thoth is a free, open metadata service that publishers can choose to utilise as a solution for metadata storage. Thoth can provide metadata upon request in a number of formats. The Thoth Telescope used the Thoth Export API to download metadata in an ONIX format. This API provides a snapshot of a specified publisher's metadata at the time of request.

The Thoth telescope downloads the ONIX metadata files and then transforms the data into a format suitable for loading into BigQuery with the ONIX parser Java command line tool. This is a near-identical process to how the ONIX telescope's data-transformation step is executed. The transformed data is loaded into BigQuery, where it can be picked up and used by the ONIX Workflow.

Dataset Name	onix
Table Name	onix
Table Type	Sharded
Average Runtime	5 min
Average Download Size	1-200 MB
Harvest Type	API
Run Schedule	Weekly
Catch-up Missed Runs	❌
Each Run Includes All Data	✅

Telescope Configuration

Telescope kwargs

Fields passed as keyword arguments to the telescope upon instantiation.

Publisher ID (publisher_id)

This field holds the Thoth internal ID for the publisher. For example, Open Book Publisher's ID would be presented as follows:

```
kwargs:
    publisher_id: "85fd969a-a16c-480b-b641-cb9adf979c3b"
```

Format Specification (format_specification)

Thoth can output the metadata feed in a number of different formats. Refer to the Thoth export API for more information. The format specification should be provided to the Telescope.

```
kwargs:
      format_specification: "onix_3.0::jstor"
```

Elevate Related Products (related_products_elevation)

```
kwargs:
      related_product_elevation: True
```

A boolean value ("True" | "False") that determines whether the transform step should elevate the feed's related products. See Related Products Manipulation for more information.

Retrieving the Publisher ID

To get a publisher's internal identifier, navigate to the Thoth's GraphiQL page and supply the following query:

```
{
  publishers{
    publisherName
    publisherId
  }
}
```

Which will output all available publishers with their name and internal ID. Inputs to the query are available to narrow down the search. For example, the 'filter' input can be used to filter by a string.

```
{
  publishers(filter: "Press"){
    publisherName
    publisherId
  }
}
```

This will only output publishers with *Press* in their name.

Telescope Tasks

Data Download

The download step is simple, thanks to Thoth's export API. The telescope uses the API to gain a publisher's metadata record. All that is required is to query the API with the proper URI:

```
https://export.thoth.pub/specifications/{format_specification}/publisher/
```

Where the `format_specification` and `publisher_id` are those supplied in the telescope kwargs.

Data Transform

The transform step consists of a few steps:

1. The downloaded XML file is parsed though the Java ONIX Parser, which results in a .jsonl file
2. The *SubjectHeadingText* field is collapsed into a single semicolon-separated string (for downstream)
3. The *PersonName* field is created (where possible) from the *KeyNames* and *NamesBeforeKey* fields.

BigQuery Load

The valid ONIX feed can now be loaded from the transform bucket into a BigQuery date-sharded table in the *onix* dataset, which will be created if it does not yet exist. There will be multiple *onix_YYYYMMDD* tables.

Table Schema

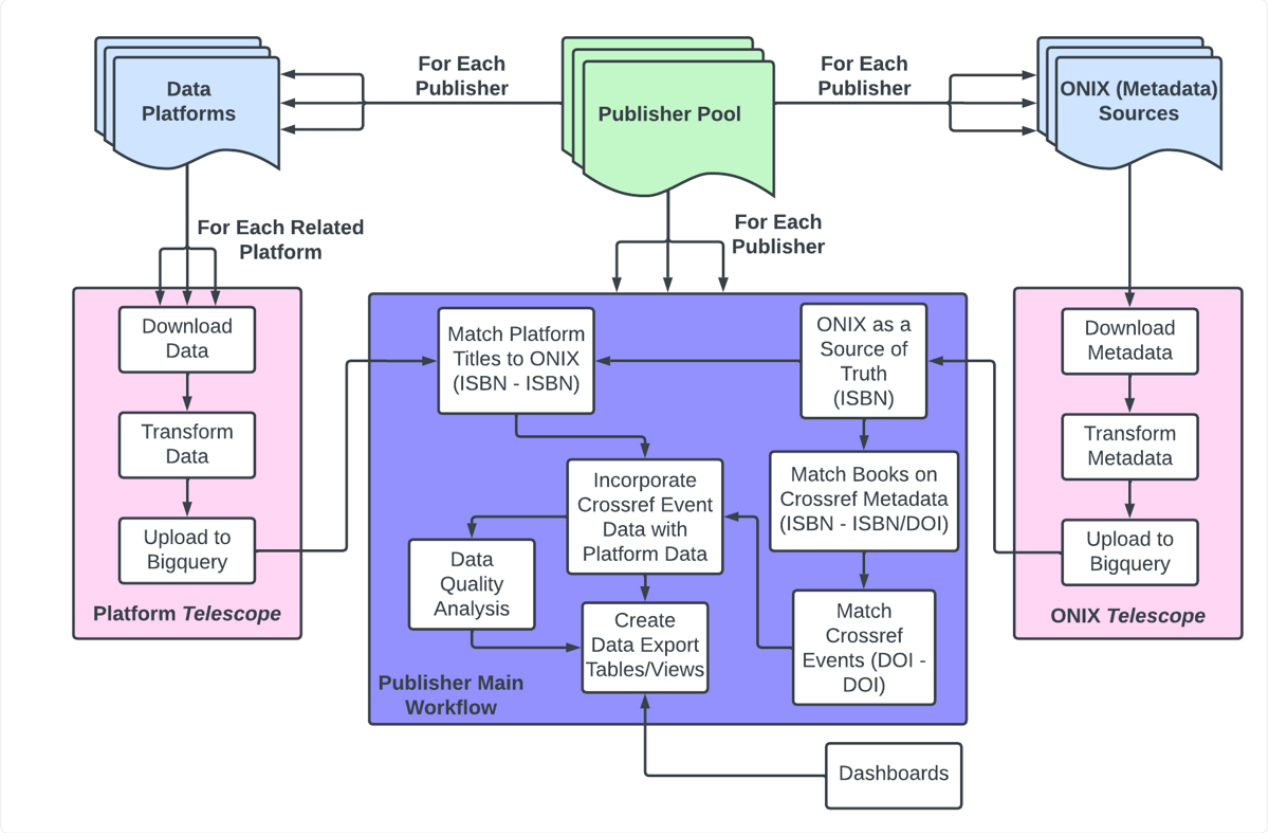
The Thoth table uses the same schema as the ONIX Telescope's table. See the ONIX Telescope schema.

ONIX Workflow

The ONIX Workflow filters and aggregates the data ingested by the telescopes.

BAD has a single core workflow - the ONIX Workflow which is responsible for filtering and aggregating all data for a publisher. The workflow prepares the data so that it can be easily accessed/imported by a dashboarding tool. The ONIX Workflow can broadly broken into three parts:

1. Aggregating and Mapping book products into works and work families
2. Linking data from metric providers to book products
3. Creating export tables for visualisation in dashboards



An overview of BAD workflows. The ONIX Workflow is represented in purple.

Average Runtime	10-20 min
Run Schedule	Weekly
Catch-up Missed Runs	❌
Each Run Includes All Data	✅

Workflow kwargs

Sensor DAG IDs (sensor_dag_ids)

A list of DAG IDs. Upon instantiation, the workflow will create a sensor task for each of the supplied DAGs. Each of the sensors will look back 7 days for a completed DAG. If there are no runs in the last 7 days, the sensor will be marked as a success. Otherwise, the sensor will wait for the DAG to complete before marking itself as a success.

Data Partners (data_partners)

A list of data partners that the workflow will use to aggregate and filter the data. The corresponding sensors for each data partner should be present. This should contain one *onix* type partner and at least one *data* type partner.

Workflow Tasks

Aggregate Works

Dataset Name	onix_workflow
Table Names	`onix_workfamilyid_isbn`, `onix_workid_isbn`, `onix_workid_isbn_errors`
Average Runtime	1-5 min
Average Download Size	0-1 MB
Table Type	Sharded

The ONIX workflow uses the ONIX table created by an ONIX telescope (ONIX Telescope, Thoth, OAPEN Metadata) to do the following:

1. Aggregate book product records into works records. Works are equivalence classes of products, where each product in the class is a manifestation of each other. For example, a PDF and a paperback of the same work.
2. Aggregate work records into work family records. A work family is an equivalence class of works where each work in the class is just a different edition.
3. Produce intermediate lookup tables mapping ISBN13 → WorkID and ISBN13 → WorkFamilyID.
4. Produce intermediate tables that append work_id and work_family_id columns to different data tables with ISBN keys.

<details>
<summary>Definitions - Product, Work and Work Families</summary>

> **Product**: A product is a manifestation of a work, and will have its own ISBN. There may be several DOIs linked to a single product though (or sometimes none at all).

> **Work**: Can be a collection of products, which are each different manifestation of the same work. Some datasets have unique IDs assigned to the concept of a work, but these are not as clear as the usage of ISBN for a product.

> **Edition**: Is a new Work, but is derived as a revision from an existing work as opposed to being entirely new.
</details>

> **Work Family** is a collection of works which are different editions of each other.

Work ID

The Work ID will be an arbitrary ISBN representative from a product in the equivalence class.

Work Family ID

The Work Family ID will be an arbitrary Work ID (ISBN) representative from a work in the equivalence class.

Work ID ISBN Errors

Create Crossref Metadata table

Crossref Metadata is required to proceed. The ISBNs for each work is obtained from the publisher's Onix table. For each of these ISBNs, the Crossref Metadata table produced by the Academic Observatory workflows is queried. Refer to the Crossref Metadata task

Create Crossref Events table

Similarly to Crossref Metadata, Crossref Event Data is retrieved through Crossref's dedicated event REST API through the Crossref Event Data telescope. The API accepts queries based on DOI only, which we retrieve by matching the appropriate ISBN13 from the metadata. Refer to the Crossref Events task.

Create Book table

The book table (schema) is a collection of works and their relevant details for the relative publisher. The table accommodates a title's Crossref metadata, events and

separate chapters.

Dataset Name	oaebu
Table Name	`book`
Average Runtime	~1 min
Average Download Size	5-50 MB
Table Type	Sharded

Create intermediate tables

For each data partner's tables containing ISBN, create new *matched* tables which extend the original data with new *work_id* and *work_family_id* columns.

The **schemas** for these tables are identical to the raw Telescope's schemas, with the addition of work_ids and work_family_ids.

Dataset Name	oaebu_intermediate
Table Names	`{data_source}_matched`
Average Runtime	1-5 min
Average Download Size	0-1 MB
Table Type	Sharded

Create Book Product table

The ONIX workflow takes the metrics fetched through various telescopes, then aggregates and joins them to the book records in the publisher's ONIX feed.

The output is the book product table (schema), containing one row per unique

book, with a nested month field, which groups all the metrics relating to that book for each calendar month. This table is the main output of the workflow and contains all of the aggregated and filtered data from all of a a publisher's data partners/ sources.

Dataset Name	oaebu
Table Names	`book_product`
Average Runtime	1 min
Average Download Size	100-2000 MB
Table Type	Sharded

Create QA ISBN tables

For each data source, including the intermediate tables, we perform basic quality assurance checks on the data, and output the results to tables that are easy to export for analysis by the publisher (e.g. to CSV). For example we verify if the provided ISBNs are valid, or if there are unmatched ISBNs indicating that there are missing ONIX product records.

Dataset Name	oaebu_data_qa
Table Names	`{data_source}_unmatched_{isbn}`, `{data_source}_invalid_{isbn}`
Average Runtime	~1 min
Average Download Size	1 MB
Table Type	Sharded

ONIX Invalid ISBN

Details ISBN13s in the ONIX feed that are not valid.

Data Platform Invalid ISBN

Details ISBN13s in the data source that are not valid. An example schema is below, as data platforms may use different name fields (e.g, 'ISBN', 'publication_id', 'Primary_ISBN').

Data Platform Unmatched ISBN

Details ISBN\13s in the data source that were not matched to ISBN-13s in the ONIX feed.

Create QA Aggregate tables

ONIX Aggregate Metrics

Dataset Name	oaebu_data_qa
Table Names	`onix_aggregate_metrics`
Average Runtime	~1 min
Average Download Size	1 MB
Table Type	Sharded

Create Export tables

Step three of the ONIX workflow is to export the book_product table to a sequence of flattened data export tables. The data in these tables is not materially different to the book product table, just organised in a way better suited for dashboards in Looker Studio.

Since these are date-sharded tables, their names will be updated each time the

workflow is run. When using Google's Looker (previously Data Studio), it is preferable for us to use a static naming scheme. For this reason, after creating the (sharded) *export* and *quality analysis* tables, we also create/update a *view* for table. These views have a static name. By referencing the view, we can keep the Looker dashboards up-to-date without manual intervention.

Dataset Name	data_export
Table Names*	`author_metrics`, `list`, `metrics`, `metrics_city`, `metrics_country`, `metrics_events`, `metrics_institution`, `metrics_referrer`, `publisher_metrics`, `subject_bic_metrics`, `subject_bisac_metrics`, `subject_thema_metrics`, `subject_year_metrics`, `year_metrics`
Table Names	`institution_list`, `unmatched_book_metrics`
Average Runtime	~1 min
Average Download Size	1 MB
Table Type	Sharded

***Table names prefixed with** `oaebu_{publisher}_book_product`

Book Product List

This table is a list of each Book Product. It is primarily used for drop-down fields, or where a list of all the books independent of metrics is desired.

Book Product Metrics

This table contains metrics, organised by month, that are linked to each book. The country, city, institution, events and referrals expand on this to provided further useful breakdowns of metrics.

Book Product Author Metrics

This table contains metrics, organised by month and author, that are linked to each author.

Book Product Year Metrics

This table contains metrics, organised by published year and month, that are linked to each book.

Book Product Metrics Events

This table contains metrics, organised by month and crossref event type, that are linked to each book.

Book Product Metrics City

This table contains metrics, organised by month and city of measured usage, that are linked to each book.

Book Product Metrics Country

This table contains metrics, organised by month and country of measured usage, that are linked to each book.

Book Product Metrics Institutions

This table contains metrics, organised by month and institution for which there is measured activity linked to each book.

Book Product Publisher Metrics

This index contains a summary of metrics, organised by month that are linked to each publisher.

Book Product Subjects BIC

This table contains metrics, organised by month and BIC subject type, that are

linked to each book.

Book Product Subjects BISAC

This table contains metrics, organised by month and BISAC subject type, that are linked to each book.

Book Product Subjects THEMA

This table contains metrics, organised by month and THEMA subject type, that are linked to each book.

Book Product Subject Year

This table contains metrics, organised by published year and month and currently just the BIC subject type, that are linked to each book.

Institution List

This table is a list of each unique Institution where metrics are linked too. It is primarily used for drop-down fields, or where a list of all the institutions independent of metrics is desired.

Unmatched Book Metrics

This dataset is helpful for understanding where metrics and books defined in the onix feed are not matched. Helping target data quality tasks upstream of this workflow.

Create Latest Views

Because the export tables are all sharded by date, once the workflow has run the export table names will be updated. This is an issue for Looker Studio, which looks for a specific table name to pull data from. For this reason, the final step of the workflow is to create/update a set of views for both the export tables and their QA

counterparts. The first run will create the views, subsequent runs will update each view to point to the appropriate (latest) table.

Dataset Name	`data_export_latest`, `oaebu_data_qa_latest`
Table Names*	
Average Runtime	~1 min
Average Download Size	0 MB
Table Type	View

*The **table names** are a copy of the tables created in the Create Export Tables (data_export_latest dataset) and the Create QA ISBN and Create QA Aggregate (oaebu_data_qa_latest dataset) tasks with the date shard removed.

Data Partners

Metadata (ONIX) Type

ONIX (onix)

Name	Value
type_id	onix
bq_dataset_id	onix
bq_table_name	onix
isbn_field_name	ISBN13
title_field_name	TitleDetails.TitleElements.TitleText
sharded	✅

Thoth (thoth)

Name	Value
type_id	onix
bq_dataset_id	onix
bq_table_name	onix
isbn_field_name	ISBN13
title_field_name	TitleDetails.TitleElements.TitleText
sharded	✅

OAPEN Metadata (oapen_metadata)

Name	Value
type_id	oapen_metadata
bq_dataset_id	onix
bq_table_name	onix
isbn_field_name	ISBN13
title_field_name	TitleDetails.TitleElements.TitleText
sharded	✅

Data Type

Google Analytics (google_analytics)

Name	Value
type_id	google_analytics
bq_dataset_id	google
bq_table_name	google_analytics
isbn_field_name	publication_id
title_field_name	title
sharded	❌

Google Books Sales (google_books_sales)

Name	Value
type_id	google_book_sales
bq_dataset_id	google
bq_table_name	google_books_sales
isbn_field_name	Primary_ISBN
title_field_name	Title
sharded	❌

Google Books Traffic (google_books_traffic)

Name	Value
type_id	google_book_traffic
bq_dataset_id	google
bq_table_name	google_books_traffic
isbn_field_name	Primary_ISBN
title_field_name	Title
sharded	❌

JSTOR Country (jstor_country)

Name	Value
type_id	jstor_country
bq_dataset_id	jstor
bq_table_name	jstor_country
isbn_field_name	ISBN
title_field_name	Book_Title
sharded	❌

JSTOR Institution (jstor_institution)

Name	Value
type_id	jstor_institution
bq_dataset_id	jstor
bq_table_name	jstor_institution
isbn_field_name	ISBN
title_field_name	Book_Title
sharded	❌

JSTOR Country Collection (jstor_country_collection)

Used only when running the JSTOR telescope in collection mode

Name	Value
type_id	jstor_country_collection
bq_dataset_id	jstor
bq_table_name	jstor_country_collection
isbn_field_name	ISBN
title_field_name	Book_Title
sharded	❌

JSTOR Institution Collection (jstor_institution_collection)

Used only when running the JSTOR telescope in collection mode

Name	Value
type_id	jstor_institution_collection
bq_dataset_id	jstor
bq_table_name	jstor_institution_collection
isbn_field_name	ISBN
title_field_name	Book_Title
sharded	❌

IRUS OAPEN (irus_oapen)

Name	Value
type_id	irus_oapen
bq_dataset_id	irus
bq_table_name	irus_oapen
isbn_field_name	ISBN
title_field_name	book_title
sharded	❌

IRUS Fulcrum (irus_fulcrum)

Name	Value
type_id	irus_fulcrum
bq_dataset_id	irus
bq_table_name	irus_fulcrum
isbn_field_name	ISBN
title_field_name	book_title
sharded	❌

UCL Discovery (ucl_discovery)

Name	Value
type_id	ucl_discovery
bq_dataset_id	ucl
bq_table_name	ucl_discovery
isbn_field_name	ISBN
title_field_name	title
sharded	❌

Schemas

Aggregate Works

Work ID ISBN (onix_workid_isbn)

name	type	mode	description
isbn13	STRING	NULLABLE	ISBN13
work_id	STRING	NULLABLE	The WorkID. Likely to be an ISBN.

Work ID ISBN Errors (onix_workid_isbn_errors)

name	type	mode	description
Error	STRING	NULLABLE	Error string

Work Family ID ISBN (onix_workfamilyid_isbn)

name	type	mode	description
isbn13	STRING	NULLABLE	ISBN13
work_family_id	STRING	NULLABLE	The Work Family ID. Likely to be an ISBN.

Book (book)

name	type	mode	description
isbn	STRING	NULLABLE	ISBN of the Book
crossref_objects	RECORD	REPEATED	Crossref Objects (that are not chapter types) associated with the primary ISBN
crossref_objects.doi	STRING	NULLABLE	DOI
crossref_objects.title	STRING	REPEATED	Title of the Book
crossref_objects.type	STRING	NULLABLE	Crossref Type
crossref_objects.publisher	STRING	NULLABLE	Publisher Name
crossref_objects.published_year	INTEGER	NULLABLE	Year of Publication
crossref_objects.published_year_month	STRING	NULLABLE	Month and Year of Publication
crossref_objects.work_isbns	STRING	REPEATED	Full list of Assoicated ISBNs
chapters	RECORD	REPEATED	Crossref Objects (that are of type chapter) associated with the primary ISBN
chapters.doi	STRING	NULLABLE	DOI of chapter
chapters.title	STRING	REPEATED	Title of Chapter
chapters.type	STRING	NULLABLE	Crossref Type
events	RECORD	NULLABLE	Crossref events assoicated with ISBN
events.overall	RECORD	REPEATED	Overall Event Count
events.overall.source	STRING	NULLABLE	Event Source

events.overall.count	INTEGER	NULLABLE	Count of Events
events.months	RECORD	REPEATED	Event counts broken down by month
events.months.month	STRING	NULLABLE	Month of the count
events.months.source	STRING	NULLABLE	Event Source
events.months.count	INTEGER	NULLABLE	Event Count in time period
events.years	RECORD	REPEATED	Event counts broken down by month
events.years.year	INTEGER	NULLABLE	Year of count
events.years.source	STRING	NULLABLE	Event Source
events.years.count	INTEGER	NULLABLE	Event count in time period

Book Product (book_product)

name	type	mode	description
ISBN13	STRING	NULLABLE	ISBN13
onix	RECORD	NULLABLE	Fields Pulled from the ONIX Record for this Book Product
onix.Doi	STRING	NULLABLE	DOI
onix.ProductForm	STRING	NULLABLE	The product form, such as digital, print etc
onix.EditionNumber	INTEGER	NULLABLE	The edition number of this book product
onix.title	STRING	NULLABLE	The Book's Title
onix.published_year	STRING	NULLABLE	The year the book was published
onix.bic_subjects	STRING	REPEATED	A list of BIC subjects
onix.bisac_subjects	STRING	REPEATED	A list of BISAC subjects
onix.thema_subjects	STRING	REPEATED	A list of THEMA subjects
onix.keywords	STRING	REPEATED	A list of Keywords
onix.authors	RECORD	REPEATED	Book Authors
onix.authors.PersonName	STRING	NULLABLE	The Authors Full Name
onix.authors.ORCID	STRING	NULLABLE	Authors ORCID ID, if present
onix.publisher.publisher_name	STRING	NULLABLE	The name of an entity associated with the publication process. See publisher_role for type of entity the name refers to
onix.publisher.publishing_role	STRING	NULLABLE	The role of the publishing entity

string_role			publishing entity
work_id	STRING	NULLABLE	The dervied Work_ID that we calculate
work_family_id	STRING	NULLABLE	The Dervied Work_Family_ID that we calculate
metadata	RECORD	NULLABLE	Metadata on this book, derived and organised by source
metadata.crossref_objects	RECORD	REPEATED	Linked Objects from Crossref and their values
metadata.crossref_objects.doi	STRING	NULLABLE	The DOI from crossref
metadata.crossref_objects.title	STRING	REPEATED	The title from crossref
metadata.crossref_objects.type	STRING	NULLABLE	The type from crossref
metadata.crossref_objects.publisher	STRING	NULLABLE	The publisher from crossref
metadata.crossref_objects.published_year	INTEGER	NULLABLE	The published year from crossref
metadata.crossref_objects.published_year_month	STRING	NULLABLE	The published year-month from crossref
metadata.crossref_objects.work_isbns	STRING	REPEATED	ISBNs
metadata.chapters	RECORD	REPEATED	Linked Objects from Crossref where they are of type book-chapter only
metadata.chapters.doi	STRING	NULLABLE	The Book Chapter DOI

metadata.chapters.title	STRING	REPEATED	The Book Chapter title
metadata.chapters.type	STRING	NULLABLE	The Book Chapter type
metadata.events	RECORD	REPEATED	Count of events from Crossref Events
metadata.events.source	STRING	NULLABLE	Event Source Type
metadata.events.count	INTEGER	NULLABLE	Count of events
metadata.google_books_sales	RECORD	NULLABLE	Metadata derived from Google Books Sales
metadata.google_books_sales.ISBN13	STRING	NULLABLE	ISBN
metadata.google_books_sales.Imprint_Name	STRING	NULLABLE	The template used for the book.
metadata.google_books_sales.Title	STRING	NULLABLE	The title of the book.
metadata.google_books_sales.Author	STRING	NULLABLE	The author of the book.
metadata.google_books_traffic	RECORD	NULLABLE	Metadata derived from Google Books Sales
metadata.google_books_traffic.ISBN13	STRING	NULLABLE	ISBN
metadata.google_books_traffic.Title	STRING	NULLABLE	The title of the book
metadata.jstor_metadata	RECORD	NULLABLE	Metadata derived from JSTOR

metadata.jstor_metadata.ISBN13	STRING	NULLABLE	ISBN of the book (13 digits)
metadata.jstor_metadata.Book_Title	STRING	NULLABLE	Title of the book
metadata.jstor_metadata.Book_ID	STRING	NULLABLE	DOI of the book on JSTOR
metadata.jstor_metadata.Authors	STRING	NULLABLE	Author of the book
metadata.jstor_metadata.ISBN	STRING	NULLABLE	ISBN of the book
metadata.jstor_metadata.eISBN	STRING	NULLABLE	ISBN of the digital version of the book (13 digits)
metadata.jstor_metadata.Copyright_Year	INTEGER	NULLABLE	Publication year
metadata.jstor_metadata.Disciplines	STRING	NULLABLE	Subject category of the book
metadata.jstor_metadata.Usage_Type	STRING	NULLABLE	For our case it is Open Access
metadata.jstor_institution_metadata	RECORD	NULLABLE	Metadata derived from JSTOR Institutions
metadata.jstor_institution_metadata.ISBN13	STRING	NULLABLE	ISBN of the book (13 digits)
metadata.jstor_institution_metadata.Book_Title	STRING	NULLABLE	Title of the book
metadata.jstor_institution_metadata.Book_ID	STRING	NULLABLE	DOI of the book on JSTOR

metadata.jstor_institution_metadata.Authors	STRING	NULLABLE	
metadata.jstor_institution_metadata.ISBN	STRING	NULLABLE	ISBN of the book (13 digits)
metadata.jstor_institution_metadata.eISBN	STRING	NULLABLE	ISBN of the digital version of the book (13 digits)
metadata.jstor_institution_metadata.Copyright_Year	INTEGER	NULLABLE	Publication year
metadata.jstor_institution_metadata.Disciplines	STRING	NULLABLE	Subject category of the book
metadata.jstor_institution_metadata.Usage_Type	STRING	NULLABLE	For our case it is Open Access
metadata.irus_oapen_metadata	RECORD	NULLABLE	Metadata derived from IRUS OAPEN
metadata.irus_oapen_metadata.ISBN13	STRING	NULLABLE	ISBN of the book
metadata.irus_oapen_metadata.book_title	STRING	NULLABLE	Title of the book
metadata.irus_oapen_metadata.publisher	STRING	NULLABLE	The publisher
metadata.irus_fulcrum_metadata	RECORD	NULLABLE	Metadata derived from IRUS Fulcrum
metadata.irus_fulcrum_metadata.ISBN13	STRING	NULLABLE	ISBN of the book

metadata.irus_fulcrum_metadata.book_title	STRING	NULLABLE	Title of the book
metadata.irus_fulcrum_metadata.publisher	STRING	NULLABLE	The publisher
metadata.ucl_discovery_metadata	RECORD	NULLABLE	Metadata derived from UCL Discovery
metadata.ucl_discovery_metadata.ISBN13	STRING	NULLABLE	ISBN of the book
metadata.ucl_discovery_metadata.eprint_id	STRING	NULLABLE	The UCL Discovery eprint ID
months	RECORD	REPEATED	Linked Metrics from all sources, organised by month of occurance
months.month	DATE	NULLABLE	Month of Recorded Metrics
months.crossref_events	RECORD	REPEATED	Metrics Derived From Crossref Events
months.crossref_events.source	STRING	NULLABLE	The event source
months.crossref_events.count	INTEGER	NULLABLE	The count of events
months.google_analytics	RECORD	NULLABLE	Metrics derived from Google Analytics
months.google_analytics.views_total_country	RECORD	REPEATED	The total number of views per country
months.google_analytics.views_total_co	STRING	NULLABLE	The country name

ytics.views_total_country.name	STRING	NULLABLE	The country name
months.google_analytics.views_total_country.value	INTEGER	NULLABLE	The total number of views
months.google_analytics.downloads_total_country	RECORD	REPEATED	The total number of downloads per country
months.google_analytics.downloads_total_country.name	STRING	NULLABLE	The country name
months.google_analytics.downloads_total_country.value	INTEGER	NULLABLE	The total number of downloads
months.google_analytics.downloads_pdf_book_country	RECORD	REPEATED	PDF book downloads per country
months.google_analytics.downloads_pdf_book_country.name	STRING	NULLABLE	The country name
months.google_analytics.downloads_pdf_book_country.value	INTEGER	NULLABLE	The total number of PDF book downloads
months.google_analytics.downloads_pdf_chapter_country	RECORD	REPEATED	PDF chapter downloads per country
months.google_analytics.downloads_pdf_chapter_country.name	STRING	NULLABLE	The country name
months.google_analytics.downloads_pdf_chapter_country.v	INTEGER	NULLABLE	The total number of PDF chapter downloads

l_chapter_country.value			PDF chapter downloads
months.google_analytics.downloads_html_book_country	RECORD	REPEATED	HTML book downloads per country
months.google_analytics.downloads_html_book_country.name	STRING	NULLABLE	The country name
months.google_analytics.downloads_html_book_country.value	INTEGER	NULLABLE	The total number of HTML book downloads
months.google_analytics.downloads_html_chapter_country	RECORD	REPEATED	HTML chapter downloads per country
months.google_analytics.downloads_html_chapter_country.name	STRING	NULLABLE	The country name
months.google_analytics.downloads_html_chapter_country.value	INTEGER	NULLABLE	The total number of HTML chapter downloads
months.google_analytics.downloads_epub_book_country	RECORD	REPEATED	EPUB book downloads per country
months.google_analytics.downloads_epub_book_country.name	STRING	NULLABLE	The country name
months.google_analytics.downloads_epub_book_country.value	INTEGER	NULLABLE	The total number of EPUB book downloads

months.google_analytics.downloads_epub_chapter_country	RECORD	REPEATED	EPUB chapter downloads per country
months.google_analytics.downloads_epub_chapter_country.name	STRING	NULLABLE	The country name
months.google_analytics.downloads_epub_chapter_country.value	INTEGER	NULLABLE	The total number of EPUB chapter downloads
months.google_analytics.downloads_mobi_book_country	RECORD	REPEATED	MOBI book downloads per country
months.google_analytics.downloads_mobi_book_country.name	STRING	NULLABLE	The country name
months.google_analytics.downloads_mobi_book_country.value	INTEGER	NULLABLE	The total number of MOBI book downloads
months.google_analytics.downloads_mobi_chapter_country	RECORD	REPEATED	MOBI chapter downloads per country
months.google_analytics.downloads_mobi_chapter_country.name	STRING	NULLABLE	The country name
months.google_analytics.downloads_mobi_chapter_country.value	INTEGER	NULLABLE	The total number of MOBI chapter downloads
months.google_boo	RECORD	NULLABLE	Metrics derived from

ks_sales	RECORD	NULLABLE	Google Books Sales
months.google_books_sales.qty	INTEGER	NULLABLE	The number of units in the transaction. Negative for refunds
months.google_books_sales.countries	RECORD	REPEATED	The list of countries where buyers brought the book
months.google_books_sales.countries.Country_of_Sale	STRING	NULLABLE	The country where the buyer bought the book
months.google_books_sales.countries.qty	INTEGER	NULLABLE	The number of units in the transaction. Negative for refunds
months.google_books_traffic	RECORD	NULLABLE	Metrics derived from Google Books Traffic
months.google_books_traffic.Book_Visits_BV_	INTEGER	NULLABLE	A Book Visit is registered each time a unique user views one of your books on Google Books. This number includes informational page views (such as the “About this book” page) as well as preview content page views
months.google_books_traffic.BV_with_Pages_Viewed	INTEGER	NULLABLE	The number of Book Visits in which users accessed preview pages of your book. This doesn’t include visits where a user accessed only informational pages for your books
months.google_books_traffic.Non_Unique_Buy_Clicks	INTEGER	NULLABLE	The number of clicks on links for purchasing the book on retailer websites (including your website, if you provided a buy link)
months.google_boo			The number of visits

ks_traffic.BV_with_Buy_Clicks	INTEGER	NULLABLE	which included a click on a purchase link
months.google_books_traffic.Buy_Link_CTR	FLOAT	NULLABLE	The clickthrough rate for purchase links. The values are percentages
months.google_books_traffic.Pages_Viewed	INTEGER	NULLABLE	The total number of unique preview content pages that a user viewed in a given session (counted as a 24-hour period). If a user views the same page of your book twice during a session, only a single page view is registered
months.jstor_country	RECORD	REPEATED	Metrics derived from JSTOR Country
months.jstor_country.Country_name	STRING	NULLABLE	Country Name
months.jstor_country.Total_Item_Requests	INTEGER	NULLABLE	Total number of request made from that specific country
months.jstor_institution	RECORD	REPEATED	Metrics derived from JSTOR Institutions
months.jstor_institution.Institution	STRING	NULLABLE	Institution name
months.jstor_institution.Total_Item_Requests	INTEGER	NULLABLE	Total number of request made from that specific institution
months.irus_oapen	RECORD	NULLABLE	Metrics from OAPEN. Recorded with the IRUS-UK platform
months.irus_oapen.version	STRING	NULLABLE	Version of the OAPEN IRUS UK API, corresponds to the COUNTER report version

months.irus_oapen.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
months.irus_oapen.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01
months.irus_oapen.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01
months.irus_oapen.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01
months.irus_oapen.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01
months.irus_oapen.country	RECORD	REPEATED	Record to store statistics on the country level
months.irus_oapen.country.name	STRING	NULLABLE	The country name of the client registered by oapen irus uk
months.irus_oapen.country.code	STRING	NULLABLE	The country code of the client registered by oapen irus uk
months.irus_oapen.country.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
months.irus_oapen.country.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01
months.irus_oapen.country.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01

months.irus_oapen.country.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01
months.irus_oapen.country.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01
months.irus_oapen.locations	RECORD	REPEATED	Record to store statistics on the location level
months.irus_oapen.locations.latitude	FLOAT	NULLABLE	The latitude geolocated from the client's ip address
months.irus_oapen.locations.longitude	FLOAT	NULLABLE	The longitude geolocated from the client's ip address
months.irus_oapen.locations.city	STRING	NULLABLE	The city geolocated from the client's ip address
months.irus_oapen.locations.country_name	STRING	NULLABLE	The country name geolocated from the client's ip address
months.irus_oapen.locations.country_code	STRING	NULLABLE	The country code geolocated from the client's ip address
months.irus_oapen.locations.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
months.irus_oapen.locations.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01
months.irus_oapen.locations.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01
months.irus_oapen.locations.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01

months.irus_oapen.locations.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01
months.irus_fulcrum	RECORD	NULLABLE	Metrics from Fulcrum. Recorded with the IRUS platform
months.irus_fulcrum.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations
months.irus_fulcrum.total_item_requests	INTEGER	NULLABLE	The total number of item requests
months.irus_fulcrum.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations
months.irus_fulcrum.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests
months.irus_fulcrum.country	RECORD	REPEATED	Record to store statistics on the country level
months.irus_fulcrum.country.name	STRING	NULLABLE	The country name of the client
months.irus_fulcrum.country.code	STRING	NULLABLE	The country code of the client
months.irus_fulcrum.country.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations
months.irus_fulcrum.country.total_item_requests	INTEGER	NULLABLE	The total number of item requests
months.irus_fulcrum.country.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations

~~m_investigations~~			
months.irus_fulcrum.country.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests
months.ucl_discovery	RECORD	NULLABLE	Metrics derived from UCL Discovery
months.ucl_discovery.total_downloads	INTEGER	NULLABLE	Number of downloads
months.ucl_discovery.country	RECORD	REPEATED	Number of downloads per country
months.ucl_discovery.country.country_code	STRING	NULLABLE	Country code
months.ucl_discovery.country.country_name	STRING	NULLABLE	Country name
months.ucl_discovery.country.country_downloads	INTEGER	NULLABLE	Number of downloads for the given country

Data Export Tables

Book List (book_list)

name	type	mode	description
product_id	STRING	NULLABLE	Book Product ID
work_id	STRING	NULLABLE	Book Work ID
work_family_id	STRING	NULLABLE	Book Work Family ID
ProductForm	STRING	NULLABLE	The product form of the book
usage_flag	BOOLEAN	NULLABLE	Was there any usage detected, from any source, for this book
EditionNumber	INTEGER	NULLABLE	The edition number of the book
published_year	INTEGER	NULLABLE	The published year of the book
published_date	DATE	NULLABLE	The date the book was published
publisher_name	STRING	NULLABLE	The name of the publisher
title	STRING	NULLABLE	The Books Title
bic_subjects	STRING	REPEATED	A list of BIC subjects
bisac_subjects	STRING	REPEATED	A list of BISAC subjects
thema_subjects	STRING	REPEATED	A list of thema subjects
keywords	STRING	REPEATED	A list of keywords
authors.PeronName	STRING	NULLABLE	The author's full name in the format '[first name] [last name]'
authors.PeronName Inverted	STRING	NULLABLE	The author's full name in the format '[last name], [first name]'

ORCID	STRING	NULLABLE	The Authors ORCID ID

Book Metrics (book_metrics)

name	type	mode	description
product_id	STRING	NULLABLE	Book Product ID
work_id	STRING	NULLABLE	Book Work ID
work_family_id	STRING	NULLABLE	Book Work Family ID
title	STRING	NULLABLE	The title of the book
authors	RECORD	REPEATED	A list of Book Authors
authors.PersonName	STRING	NULLABLE	The author's full name in the format '[first name] [last name]'
authors.PersonNameInverted	STRING	NULLABLE	The author's full name in the format '[last name], [first name]'
authors.ORCID	STRING	NULLABLE	The Authors ORCID ID
published_year	INTEGER	NULLABLE	The Books published year
published_name	STRING	NULLABLE	The name of the publisher
month	DATE	NULLABLE	The month in which the metrics took place
google_analytics	RECORD	NULLABLE	Metrics from Google Analytics
google_analytics.page_views	INTEGER	NULLABLE	Number of page views
google_analytics.downloads	INTEGER	NULLABLE	Number of total downloads
google_analytics.downloads_pdf_book	INTEGER	NULLABLE	Number of PDF book downloads
google_analytics.downloads_pdf_chapter	INTEGER	NULLABLE	Number of PDF chapter downloads

pts.			
google_analytics.downloads_html_chapter	INTEGER	NULLABLE	Number of HTML chapter downloads
google_analytics.downloads_epub_book	INTEGER	NULLABLE	Number of ePUB book downloads
google_analytics.downloads_epub_chapter	INTEGER	NULLABLE	Number of ePUB chapter downloads
google_analytics.downloads_mobi_book	INTEGER	NULLABLE	Number of MOBI book downloads
google_analytics.downloads_mobi_chapter	INTEGER	NULLABLE	Number of MOBI chapter downloads
crossref_events	RECORD	NULLABLE	Metrics from Crossref Events
crossref_events.count	INTEGER	NULLABLE	Count of events
google_books_traffic	RECORD	NULLABLE	Metrics from Google Books Traffic
google_books_traffic.Book_Visits_BV_	INTEGER	NULLABLE	A Book Visit is registered each time a unique user views one of your books on Google Books. This number includes informational page views (such as the “About this book” page) as well as preview content page views
google_books_traffic.BV_with_Pages_Viewed	INTEGER	NULLABLE	The number of Book Visits in which users accessed preview pages of your book. This doesn’t include visits where a user accessed only informational

			pages for your books
google_books_traffic.Non_Unique_Buy_Clicks	INTEGER	NULLABLE	The number of clicks on links for purchasing the book on retailer websites (including your website, if you provided a buy link)
google_books_traffic.BV_with_Buy_Clicks	INTEGER	NULLABLE	The number of visits which included a click on a purchase link
google_books_traffic.Buy_Link_CTR	FLOAT	NULLABLE	The clickthrough rate for purchase links. The values are percentages
google_books_traffic.Pages_Viewed	INTEGER	NULLABLE	The total number of unique preview content pages that a user viewed in a given session (counted as a 24-hour period)
google_books_sales	RECORD	NULLABLE	Metrics from Google Books Sales
google_books_sales.qty	INTEGER	NULLABLE	Quantity of sales
google_books_sales.countries	RECORD	REPEATED	A list of Countries
google_books_sales.countries.Country_of_Sale	STRING	NULLABLE	Country in which sale occured
google_books_sales.countries.qty	INTEGER	NULLABLE	Quantity of sales
jstor	RECORD	NULLABLE	Metrics from JSTOR
jstor.Total_Item_Requests	INTEGER	NULLABLE	Total number of item requests
irus_oapen	RECORD	NULLABLE	Metrics from OAPEN. Recorded with the

irus_oapen	RECORD	NULLABLE	Recorded with the IRUS-UK platform
irus_oapen.version	STRING	NULLABLE	Version of the OAPEN IRUS UK API, corresponds to the COUNTER report version
irus_oapen.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
irus_oapen.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01
irus_oapen.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01
irus_oapen.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01
irus_oapen.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01
irus_fulcrum	RECORD	NULLABLE	Metrics from Fulcrum. Recorded with the IRUS platform
irus_fulcrum.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations.
irus_fulcrum.total_item_requests	INTEGER	NULLABLE	The total number of item requests.
irus_fulcrum.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations.
irus_fulcrum.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests.

ucl_discovery	RECORD	NULLABLE	Metrics from UCL Discovery
ucl_discovery.total_downloads	INTEGER	NULLABLE	

Book Metrics Author (book_metrics_author)

name	type	mode	description
PersonName	STRING	NULLABLE	The author's full name in the format '[first name] [last name]'
PersonNameInverted	STRING	NULLABLE	The author's full name in the format '[last name] [first name]'
orcid	STRING	NULLABLE	Author's ORCID ID
unique_books	INTEGER	NULLABLE	Number of unique Books matched to the author
month	DATE	NULLABLE	Month in which metrics took place
google_analytics	RECORD	NULLABLE	Metrics from Google Analytics
google_analytics.views_total_country	INTEGER	NULLABLE	Number of page views aggregated over all countries
google_analytics.downloads	INTEGER	NULLABLE	Number of total downloads
google_analytics.downloads_pdf_book	INTEGER	NULLABLE	Number of PDF book downloads
google_analytics.downloads_pdf_chapter	INTEGER	NULLABLE	Number of PDF chapter downloads
google_analytics.downloads_html_chapter	INTEGER	NULLABLE	Number of HTML chapter downloads
google_analytics.downloads_epub_book	INTEGER	NULLABLE	Number of ePUB book downloads
google_analytics.downloads_epub_c	INTEGER	NULLABLE	Number of ePUB chapter downloads

hapter			chapter downloads
google_analytics.downloads_mobi_book	INTEGER	NULLABLE	Number of MOBI book downloads
google_analytics.downloads_mobi_chapter	INTEGER	NULLABLE	Number of MOBI chapter downloads
crossref_events	RECORD	NULLABLE	Metrics from Crossref events
crossref_events.count	INTEGER	NULLABLE	Count of events
google_books_traffic	RECORD	NULLABLE	Metrics from Google Books Traffic
google_books_traffic.Book_Visits_BV_	INTEGER	NULLABLE	A Book Visit is registered each time a unique user views one of your books on Google Books. This number includes informational page views (such as the “About this book” page) as well as preview content page views
google_books_traffic.BV_with_Pages_Viewed	INTEGER	NULLABLE	The number of Book Visits in which users accessed preview pages of your book. This doesn’t include visits where a user accessed only informational pages for your books
google_books_traffic.Non_Unique_Buy_Clicks	INTEGER	NULLABLE	The number of clicks on links for purchasing the book on retailer websites (including your website, if you provided a buy link)
google_books_traffic.BV_with_Buy_Clicks	INTEGER	NULLABLE	The number of visits which included a click on a purchase link

google_books_traffic.Buy_Link_CTR	FLOAT	NULLABLE	The clickthrough rate for purchase links. The values are percentages
google_books_traffic.Pages_Viewed	INTEGER	NULLABLE	The total number of unique preview content pages that a user viewed in a given session (counted as a 24-hour period)
google_books_sales	RECORD	NULLABLE	Metrics from Google Books Sales
google_books_sales.qty	INTEGER	NULLABLE	Number of sales
jstor	RECORD	NULLABLE	Metrics from JSTOR
jstor.Total_Item_Requests	INTEGER	NULLABLE	The total number of item requests
irus_oapen	RECORD	NULLABLE	Metrics from OAPEN. Recorded with the IRUS-UK platform
irus_oapen.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
irus_oapen.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01
irus_oapen.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01
irus_oapen.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01
irus_oapen.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01

irus_fulcrum	RECORD	NULLABLE	Metrics from Fulcrum. Recorded with the IRUS platform
irus_fulcrum.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations.
irus_fulcrum.total_item_requests	INTEGER	NULLABLE	The total number of item requests.
irus_fulcrum.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations.
irus_fulcrum.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests.

Book Metrics Events (book_metrics_events)

name	type	mode	description
product_id	STRING	NULLABLE	Book Product ID
work_id	STRING	NULLABLE	Book Work ID
work_family_id	STRING	NULLABLE	Book Work Family ID
title	STRING	NULLABLE	The title of the book
published_year	INTEGER	NULLABLE	The publisher year of the book
publsiher_name	STRING	NULLABLE	The name of the publisher
month	DATE	NULLABLE	The month for which the metrics apply to
event_source	STRING	NULLABLE	Event Source
crossref_events	RECORD	NULLABLE	Metrics from Crossref Events
crossref_events.count	INTEGER	NULLABLE	Count of Events

Book Metrics City (book_metrics_city)

name	type	mode	description
product_id	STRING	NULLABLE	Book Product ID
work_id	STRING	NULLABLE	Book Work ID
work_family_id	STRING	NULLABLE	Book Work Family ID
title	STRING	NULLABLE	The title of the book
published_year	INTEGER	NULLABLE	The publisher year of the book
published_name	STTING	NULLABLE	The name of the publisher
month	DATE	NULLABLE	The month for which the metrics apply to
city	STRING	NULLABLE	The name of the city
coordinates	STRING	NULLABLE	Geographical coordinates of the city
irus_oapen	RECORD	NULLABLE	Metrics from OAPEN. Recorded with the IRUS-UK platform
irus_oapen.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
irus_oapen.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01
irus_oapen.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01
irus_oapen.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01

irus_oapen.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01

Book Product Metrics Country (book_product_metrics_country)

name	type	mode	description
product_id	STRING	NULLABLE	Book Product ID
work_id	STRING	NULLABLE	Book Work ID
work_family_id	STRING	NULLABLE	Book Work Family ID
title	STRING	NULLABLE	The title of the book
published_year	INTEGER	NULLABLE	The publisher year of the book
publisher_name	STRING	NULLABLE	The name of the publisher
month	DATE	NULLABLE	The month for which the metrics apply to
country_code	STRING	NULLABLE	The Country Code
country_name	STRING	NULLABLE	The Country Name
country_iso_name	STRING	NULLABLE	The ISO3166 (alpha 2) Country Name
country_wikipedia_name	STRING	NULLABLE	The Country Wikipedia Name
jstor	RECORD	NULLABLE	Metrics from JSTOR
jstor.Total_Item_Requests	INTEGER	NULLABLE	Total number of request made
irus_oapen	RECORD	NULLABLE	Metrics from OAPEN. Recorded with the IRUS-UK platform
irus_oapen.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
irus_oapen.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data

_investigations			Only available for data since 2020-04-01
irus_oapen.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01
irus_oapen.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01
irus_oapen.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01
google_analytics	RECORD	NULLABLE	Metrics from Google Analytics
google_analytics.page_views	INTEGER	NULLABLE	Number of page views
google_analytics.downloads	INTEGER	NULLABLE	Number of total downloads
google_analytics.downloads_pdf_book	INTEGER	NULLABLE	Number of PDF book downloads
google_analytics.downloads_pdf_chapter	INTEGER	NULLABLE	Number of PDF chapter downloads
google_analytics.downloads_html_chapter	INTEGER	NULLABLE	Number of HTML chapter downloads
google_analytics.downloads_epub_book	INTEGER	NULLABLE	Number of ePUB book downloads
google_analytics.downloads_epub_chapter	INTEGER	NULLABLE	Number of ePUB chapter downloads
google_analytics.downloads_mobi_book	INTEGER	NULLABLE	Number of MOBI book downloads
google_analytics.dow	INTEGER	NULLABLE	Number of MOBI

nloads_mobi_chapter	[illegible]	[illegible]	chapter downloads
irus_fulcrum	RECORD	NULLABLE	Metrics from Fulcrum. Recorded with the IRUS platform
irus_fulcrum.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations.
irus_fulcrum.total_item_requests	INTEGER	NULLABLE	The total number of item requests.
irus_fulcrum.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations.
irus_fulcrum.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests.
google_books_sales	RECORD	NULLABLE	Metrics from Google Books Sales
google_books_sales.qty	INTEGER	NULLABLE	The number of units in the transaction. Negative for refunds
ucl_discovery	RECORD	NULLABLE	Metrics from UCL Discovery
ucl_discovery.download_count	INTEGER	NULLABLE	Number of downloads

Book Metrics Institution (book_metrics_institution)

name	type	mode	description
product_id	STRING	NULLABLE	Book Product ID
title	STRING	NULLABLE	The title of the book
published_year	INTEGER	NULLABLE	The publisher year of the book
month	DATE	NULLABLE	The month for which the metrics apply to
institution	STRING	NULLABLE	Institution Name
jstor	RECORD	NULLABLE	Metrics from JSTOR
jstor.Total_Item_Requests	INTEGER	NULLABLE	Total number of request made from that specific institution

Book Product Subjects BIC (book_product_subjects_bic_metrics)

name	type	mode	description
subject	STRING	NULLABLE	BIC Subject
subject_code	STRING	NULLABLE	BIC Subject Code
unique_books	INTEGER	NULLABLE	The number of unique books
month	DATE	NULLABLE	The month in which the metrics occurred
google_analytics	RECORD	NULLABLE	Metrics from Google Analytics
google_analytics.page_views	INTEGER	NULLABLE	Number of page views
google_analytics.downloads	INTEGER	NULLABLE	Number of total downloads
google_analytics.downloads_pdf_book	INTEGER	NULLABLE	Number of PDF book downloads
google_analytics.downloads_pdf_chapter	INTEGER	NULLABLE	Number of PDF chapter downloads
google_analytics.downloads_html_chapter	INTEGER	NULLABLE	Number of HTML chapter downloads
google_analytics.downloads_epub_book	INTEGER	NULLABLE	Number of ePUB book downloads
google_analytics.downloads_epub_chapter	INTEGER	NULLABLE	Number of ePUB chapter downloads
google_analytics.downloads_mobi_book	INTEGER	NULLABLE	Number of MOBI book downloads
google_analytics.d			Number of MOBI

ownloads_mobi_chapter	INTEGER	NULLABLE	Number of MOBI chapter downloads
crossref_events	RECORD	NULLABLE	Metrics from Crossref events
crossref_events.count	INTEGER	NULLABLE	Count of events
google_books_traffic	RECORD	NULLABLE	Metrics from Google Books Traffic
google_books_traffic.Book_Visits_BV_	INTEGER	NULLABLE	A Book Visit is registered each time a unique user views one of your books on Google Books. This number includes informational page views (such as the “About this book” page) as well as preview content page views
google_books_traffic.BV_with_Pages_Viewed	INTEGER	NULLABLE	The number of Book Visits in which users accessed preview pages of your book. This doesn’t include visits where a user accessed only informational pages for your books
google_books_traffic.Non_Unique_Buy_Clicks	INTEGER	NULLABLE	The number of clicks on links for purchasing the book on retailer websites (including your website, if you provided a buy link)
google_books_traffic.BV_with_Buy_Clicks	INTEGER	NULLABLE	The number of visits which included a click on a purchase link
google_books_traffic.Buy_Link_CTR	FLOAT	NULLABLE	The clickthrough rate for purchase links. The values are percentages
google_books_traf	INTEGER	NULLABLE	The total number of unique preview content pages that a

fic.Pages_Viewed			user viewed in a given session (counted as a 24-hour period)
google_books_sales	RECORD	NULLABLE	Metrics from Google Books Sales
google_books_sales.qty	INTEGER	NULLABLE	The number of units in the transaction. Negative for refunds
jstor	RECORD	NULLABLE	Metrics from JSTOR
jstor.Total_Item_Requests	INTEGER	NULLABLE	Total number of request made
irus_oapen	RECORD	NULLABLE	Metrics from OAPEN. Recorded with the IRUS-UK platform
irus_oapen.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
irus_oapen.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01
irus_oapen.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01
irus_oapen.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01
irus_oapen.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01
irus_fulcrum	RECORD	NULLABLE	Metrics from Fulcrum. Recorded with the IRUS platform
irus_fulcrum.total_i	INTEGER	NULLABLE	The total number of

tem_investigations	INTEGER	NULLABLE	item investigations.
irus_fulcrum.total_item_requests	INTEGER	NULLABLE	The total number of item requests.
irus_fulcrum.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations.
irus_fulcrum.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests.

Book Product Subjects BISAC (book_product_subjects_bisac_metrics)

name	type	mode	description
subject	STRING	NULLABLE	BISAC Subject
subject_code	STRING	NULLABLE	BISAC Subject Code
unique_books	INTEGER	NULLABLE	The number of unique books
month	DATE	NULLABLE	The month in which the metrics occurred
google_analytics	RECORD	NULLABLE	Metrics from Google Analytics
google_analytics.page_views	INTEGER	NULLABLE	Number of page views
google_analytics.downloads	INTEGER	NULLABLE	Number of total downloads
google_analytics.downloads_pdf_book	INTEGER	NULLABLE	Number of PDF book downloads
google_analytics.downloads_pdf_chapter	INTEGER	NULLABLE	Number of PDF chapter downloads
google_analytics.downloads_html_chapter	INTEGER	NULLABLE	Number of HTML chapter downloads
google_analytics.downloads_epub_book	INTEGER	NULLABLE	Number of ePUB book downloads
google_analytics.downloads_epub_chapter	INTEGER	NULLABLE	Number of ePUB chapter downloads
google_analytics.downloads_mobi_book	INTEGER	NULLABLE	Number of MOBI book downloads
google_analytics.d			Number of MOBI

ownloads_mobi_chapter	INTEGER	NULLABLE	Number of MOBI chapter downloads
crossref_events	RECORD	NULLABLE	Metrics from Crossref events
crossref_events.count	INTEGER	NULLABLE	Count of events
google_books_traffic	RECORD	NULLABLE	Metrics from Google Books Traffic
google_books_traffic.Book_Visits_BV_	INTEGER	NULLABLE	A Book Visit is registered each time a unique user views one of your books on Google Books. This number includes informational page views (such as the “About this book” page) as well as preview content page views
google_books_traffic.BV_with_Pages_Viewed	INTEGER	NULLABLE	The number of Book Visits in which users accessed preview pages of your book. This doesn’t include visits where a user accessed only informational pages for your books
google_books_traffic.Non_Unique_Buy_Clicks	INTEGER	NULLABLE	The number of clicks on links for purchasing the book on retailer websites (including your website, if you provided a buy link)
google_books_traffic.BV_with_Buy_Clicks	INTEGER	NULLABLE	The number of visits which included a click on a purchase link
google_books_traffic.Buy_Link_CTR	FLOAT	NULLABLE	The clickthrough rate for purchase links. The values are percentages
google_books_traf	INTEGER	NULLABLE	The total number of unique preview content pages that a

fic.Pages_Viewed			user viewed in a given session (counted as a 24-hour period)
google_books_sales	RECORD	NULLABLE	Metrics from Google Books Sales
google_books_sales.qty	INTEGER	NULLABLE	The number of units in the transaction. Negative for refunds
jstor	RECORD	NULLABLE	Metrics from JSTOR
jstor.Total_Item_Requests	INTEGER	NULLABLE	Total number of request made
irus_oapen	RECORD	NULLABLE	Metrics from OAPEN. Recorded with the IRUS-UK platform
irus_oapen.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
irus_oapen.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01
irus_oapen.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01
irus_oapen.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01
irus_oapen.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01
irus_fulcrum	RECORD	NULLABLE	Metrics from Fulcrum. Recorded with the IRUS platform
irus_fulcrum.total_i	INTEGER	NULLABLE	The total number of

tem_investigations	INTEGER	NULLABLE	item investigations.
irus_fulcrum.total_item_requests	INTEGER	NULLABLE	The total number of item requests.
irus_fulcrum.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations.
irus_fulcrum.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests.

Book Product Subjects THEMA (book_product_subjects_thema_metrics)

name	type	mode	description
subject	STRING	NULLABLE	Thema Subject
subject_code	STRING	NULLABLE	Thema Subject Code
unique_books	INTEGER	NULLABLE	The number of unique books
month	DATE	NULLABLE	The month in which the metrics occurred
google_analytics	RECORD	NULLABLE	Metrics from Google Analytics
google_analytics.page_views	INTEGER	NULLABLE	Number of page views
google_analytics.downloads	INTEGER	NULLABLE	Number of total downloads
google_analytics.downloads_pdf_book	INTEGER	NULLABLE	Number of PDF book downloads
google_analytics.downloads_pdf_chapter	INTEGER	NULLABLE	Number of PDF chapter downloads
google_analytics.downloads_html_chapter	INTEGER	NULLABLE	Number of HTML chapter downloads
google_analytics.downloads_epub_book	INTEGER	NULLABLE	Number of ePUB book downloads
google_analytics.downloads_epub_chapter	INTEGER	NULLABLE	Number of ePUB chapter downloads
google_analytics.downloads_mobi_book	INTEGER	NULLABLE	Number of MOBI book downloads
google_analytics.d			Number of MOBI

ownloads_mobi_chapter	INTEGER	NULLABLE	Number of MOBI chapter downloads
crossref_events	RECORD	NULLABLE	Metrics from Crossref events
crossref_events.count	INTEGER	NULLABLE	Count of events
google_books_traffic	RECORD	NULLABLE	Metrics from Google Books Traffic
google_books_traffic.Book_Visits_BV_	INTEGER	NULLABLE	A Book Visit is registered each time a unique user views one of your books on Google Books. This number includes informational page views (such as the “About this book” page) as well as preview content page views
google_books_traffic.BV_with_Pages_Viewed	INTEGER	NULLABLE	The number of Book Visits in which users accessed preview pages of your book. This doesn’t include visits where a user accessed only informational pages for your books
google_books_traffic.Non_Unique_Buy_Clicks	INTEGER	NULLABLE	The number of clicks on links for purchasing the book on retailer websites (including your website, if you provided a buy link)
google_books_traffic.BV_with_Buy_Clicks	INTEGER	NULLABLE	The number of visits which included a click on a purchase link
google_books_traffic.Buy_Link_CTR	FLOAT	NULLABLE	The clickthrough rate for purchase links. The values are percentages
google_books_traf	INTEGER	NULLABLE	The total number of unique preview content pages that a

fic.Pages_Viewed			user viewed in a given session (counted as a 24-hour period)
google_books_sales	RECORD	NULLABLE	Metrics from Google Books Sales
google_books_sales.qty	INTEGER	NULLABLE	The number of units in the transaction. Negative for refunds
jstor	RECORD	NULLABLE	Metrics from JSTOR
jstor.Total_Item_Requests	INTEGER	NULLABLE	Total number of request made
irus_oapen	RECORD	NULLABLE	Metrics from OAPEN. Recorded with the IRUS-UK platform
irus_oapen.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
irus_oapen.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01
irus_oapen.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01
irus_oapen.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01
irus_oapen.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01
irus_fulcrum	RECORD	NULLABLE	Metrics from Fulcrum. Recorded with the IRUS platform
irus_fulcrum.total_i	INTEGER	NULLABLE	The total number of

tem_investigations	INTEGER	NULLABLE	item investigations.
irus_fulcrum.total_item_requests	INTEGER	NULLABLE	The total number of item requests.
irus_fulcrum.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations.
irus_fulcrum.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests.

Institution List (institution_list)

name	type	mode	description
institution	STRING	NULLABLE	Institution Name

Crossref Metadata

Crossref Metadata

Crossref is a non-for-profit membership organisation working on making scholarly communications better. It is an official Digital Object Identifier (DOI) Registration Agency of the International DOI Foundation. They provide metadata for every DOI that is registered with Crossref.

Crossref Members send Crossref scholarly metadata on research which is collated and standardised into the Crossref metadata dataset. This dataset is made available through services and tools for manuscript tracking, searching, bibliographic management, library systems, author profiling, specialist subject databases, scholarly sharing networks . - *source: Crossref Metadata* and schema details.

This table is created as part of the ONIX workflow. The Master Crossref Metadata table is created by Academic Observatory workflows and contains the entirety of the Crossref Metadata dataset. The ONIX workflow creates a subset of the Crossref Metadata by matching on the ISBNs of the workflow's publisher by querying the master table. This date-sharded table is placed in the *crossref* dataset in multiple *crossref_metadataYYYYMMDD* tables.

Dataset Name	crossref
Table Name	crossref_metadata
Table Type	Sharded
Average Runtime	2 min
Average Download Size	null
Harvest Type	Query
Each Run Includes All Data	✅

Table Schema

name	type	mode	description
DOI	STRING	NULLABLE	DOI of the work.
ISBN	STRING	REPEATED	
ISSN	STRING	REPEATED	
URL	STRING	NULLABLE	URL form of the work's DOI.
alternative_id	STRING	REPEATED	Other identifiers for the work provided by the depositing member
abstract	STRING	NULLABLE	Abstract as a JSON string or a JATS XML snippet encoded into a JSON string.
author	RECORD	REPEATED	
author.ORCID	STRING	NULLABLE	URL-form of an ORCID identifier
author.affiliation	RECORD	REPEATED	
author.affiliation.acronym	STRING	REPEATED	
author.affiliation.name	STRING	NULLABLE	
author.affiliation.id	RECORD	REPEATED	
author.affiliation.id.id	STRING	NULLABLE	
author.affiliation.id.id_type	STRING	NULLABLE	
author.affiliation.id.asserted_by	STRING	NULLABLE	
author.affiliation.place	STRING	REPEATED	

author.affiliation.department	STRING	REPEATED	
author.authenticated_orcid	BOOLEAN	NULLABLE	If true, record owner asserts that the ORCID user completed ORCID OAuth authentication.
author.family	STRING	NULLABLE	
author.given	STRING	NULLABLE	
author.name	STRING	NULLABLE	
author.sequence	STRING	NULLABLE	
author.suffix	STRING	NULLABLE	
clinical_trial_number	RECORD	REPEATED	
clinical_trial_number.clinical_trial_number	STRING	NULLABLE	Identifier of the clinical trial.
clinical_trial_number.registry	STRING	NULLABLE	DOI of the clinical trial regsitry that assigned the trial number.
clinical_trial_number.type	STRING	NULLABLE	One of preResults, results or postResults
container_title	STRING	REPEATED	Full titles of the containing work (usually a book or journal)
funder	RECORD	REPEATED	
funder.DOI	STRING	NULLABLE	Optional Open Funder Registry DOI uniquely identifing the funding body (http://www.crossref.org/

			fundingdata/registry.html)
funder.award	STRING	REPEATED	Award number(s) for awards given by the funding body.
funder.doi_asserted_by	STRING	NULLABLE	Either crossref or publisher
funder.name	STRING	NULLABLE	Funding body primary name
group_title	STRING	NULLABLE	Group title for posted content.
is_referenced_by_count	INTEGER	NULLABLE	Count of inbound references deposited with Crossref.
issn_type	RECORD	REPEATED	List of ISSNs with ISSN type information
issn_type.type	STRING	NULLABLE	ISSN type, can either be print ISSN or electronic ISSN.
issn_type.value	STRING	NULLABLE	ISSN value
issue	STRING	NULLABLE	Issue number of an article's journal.
published_print	RECORD	NULLABLE	
published_print.date_parts	INTEGER	REPEATED	
issued	RECORD	NULLABLE	Earliest of published-print and published-online
issued.date_parts	INTEGER	REPEATED	Contains an ordered array of year, month, day of month. Only year is required. Note that the field contains a nested array, e.g. [[2006, 5, 19]] to conform to citeproc JSON dates
license	RECORD	REPEATED	

license.URL	STRING	NULLABLE	Link to a web page describing this license
license.content_version	STRING	NULLABLE	Either vor (version of record,) am (accepted manuscript,) tdm (text and data mining) or unspecified.
license.delay_in_days	INTEGER	NULLABLE	Number of days between the publication date of the work and the start date of this license.
license.start	RECORD	NULLABLE	Date on which this license begins to take effect
license.start.date_parts	INTEGER	REPEATED	Contains an ordered array of year, month, day of month. Only year is required. Note that the field contains a nested array, e.g. [[2006, 5, 19]] to conform to citeproc JSON dates
license.start.date_time	TIMESTAMP	NULLABLE	ISO 8601 date time.
license.start.timestamp	INTEGER	NULLABLE	Seconds since UNIX epoch.
link	RECORD	REPEATED	URLs to full-text locations.
link.URL	STRING	NULLABLE	Direct link to a full-text download location.
link.content_type	STRING	NULLABLE	Content type (or MIME type) of the full-text object.
link.content_version	STRING	NULLABLE	Either vor (version of record,) am (accepted manuscript) or unspecified.
link.intended_appli	STRING	NULLABLE	Either text-mining, similarity-

cation	STRING	NULLABLE	checking or unspecified.
member	INTEGER	NULLABLE	Member identifier of the form http://id.crossref.org/member/MEMBER_ID
page	STRING	NULLABLE	Pages numbers of an article within its journal.
prefix	STRING	NULLABLE	DOI prefix identifier of the form http://id.crossref.org/prefix/DOI_PREFIX.
published	RECORD	NULLABLE	Date on which content was published.
published.date_parts	INTEGER	REPEATED	Contains an ordered array of year, month, day of month. Only year is required. Note that the field contains a nested array, e.g. [[2006, 5, 19]] to conform to citeproc JSON dates
publisher	STRING	NULLABLE	Name of work's publisher.
publisher_location	STRING	NULLABLE	Location of work's publisher
references_count	INTEGER	NULLABLE	Count of outbound references deposited with Crossref
short_container_title	STRING	REPEATED	Abbreviated titles of the containing work.
subject	STRING	REPEATED	Subject category names, a controlled vocabulary from Sci-Val. Available for most journal articles
title	STRING	REPEATED	Work titles, including translated titles.
type	STRING	NULLABLE	Enumeration, one of the type ids from https://

			api.crossref.org/v1/types.
volume	STRING	NULLABLE	Volume number of an article's journal.

Crossref Events

> When someone links their data online, or mentions research on a social media site, we capture that event and make it available for anyone to use in their own way. We provide the unprocessed data—you decide how to use it.

Before the expansion of the Internet, most discussion about scholarly content stayed within scholarly content, with articles citing each other. With the growth of online platforms for discussion, publication and social media, we have seen discussions extend into new, non-traditional venues. Crossref Event Data captures this activity and acts as a hub for the storage and distribution of this data. An event may be a citation in a dataset or patent, a mention in a news article, Wikipedia page or on a blog, or discussion and comment on social media.

When someone links their data online, or mentions research on, for example, Twitter, Wikipedia, or Reddit, Crossref's uses a set of APIs to captures and records those events in their 'Event dataset'. Events are tracked via their DOI and URLs, which enables Crossref to monitor where it's been shared, linked, bookmarked, referenced or commented on. Crossref Event Data currently contains events from a range of data sources, including Crossref Metadata, DataCite Metadata, F1000Prime (Recommendations of research publications, Hypothes.is, The Lens (Cambia), Newsfeed, Reddit, Reddit Links, Stack Exchange Network, Wikipedia, and Wordpress.com

See the crossref events page, and data details, for more information.

This date-sharded table is created during the Onix workflow, and placed in the *crossref* dataset in multiple *crossref_eventsYYYYMMDD* tables.

Dataset Name	crossref
Table Name	crossref_events
Table Type	Sharded
Average Runtime	10-120 min
Average Download Size	10-500MB
Harvest Type	API
Each Run Includes All Data	✅

Table Schema

name	type	mode	description
id	STRING	REQUIRED	Unique ID for the Event.
subj_id	STRING	NULLABLE	Subject persistent ID.
relation_type_id	STRING	NULLABLE	Type of the relationship between the subject and object.
obj_id	STRING	NULLABLE	Object persistent ID.
timestamp	TIMESTAMP	REQUIRED	Timestamp of when the Event was created.
occurred_at	TIMESTAMP	REQUIRED	Timestamp of when the Event is reported to have occurred.
experimental	BOOL	NULLABLE	
total	INTEGER	NULLABLE	
source_id	STRING	REQUIRED	A name for the source.
source_token	STRING	NULLABLE	Unique ID that identifies the Agent that generated the Event.
terms	STRING	NULLABLE	Terms of use for using the API at the point that you acquire the Event.
license	STRING	NULLABLE	A license under which the Event is made available.
evidence_record	STRING	NULLABLE	Link to an Evidence Record for this Event.
subj	RECORD	NULLABLE	Subject metadata.
subj.pid	STRING	NULLABLE	The persistent ID. Must correspond to 'subj_id' or 'obj_id'

subj.issued	TIMESTAMP	NULLABLE	Publication date.
subj.title	STRING	NULLABLE	The title of the webpage, comment, etc.
subj.author	RECORD	REPEATED	Author of the comment, blog etc.
subj.author.url	STRING	NULLABLE	
subj.author.name	STRING	NULLABLE	
subj.author.id	STRING	NULLABLE	
subj.url	STRING	NULLABLE	URL where this was found. May be different to 'pid'
subj.alternative_id	STRING	NULLABLE	
subj.original_tweet_author	STRING	NULLABLE	
subj.original_tweet_url	STRING	NULLABLE	
subj.type	STRING	NULLABLE	
subj.work_type_id	STRING	NULLABLE	
subj.work_subtype_id	STRING	NULLABLE	
subj.jurisdiction	STRING	NULLABLE	
subj.api_url	STRING	NULLABLE	
subj.publisher	RECORD	REPEATED	
subj.publisher.url	STRING	NULLABLE	
subj.publisher.name	STRING	NULLABLE	

subj.publisher.id	STRING	NULLABLE	
subj.publisher.type	STRING	NULLABLE	
subj.json_url	STRING	NULLABLE	
subj.name	STRING	NULLABLE	
subj.datePublished	STRING	NULLABLE	
subj.registrantId	STRING	NULLABLE	
subj.dateModified	TIMESTAMP	NULLABLE	
subj.id	STRING	NULLABLE	
subj.proxyIdentifiers	STRING	NULLABLE	
subj.funder	RECORD	NULLABLE	
subj.funder.id	STRING	NULLABLE	
subj.funder.type	STRING	NULLABLE	
subj.funder.name	STRING	NULLABLE	
subj.issueNumber	STRING	NULLABLE	
subj.periodical	RECORD	NULLABLE	
subj.periodical.id	STRING	NULLABLE	
subj.periodical.issn	STRING	NULLABLE	
subj.periodical.type	STRING	NULLABLE	
subj.periodical.name	STRING	NULLABLE	
subj.pagination	STRING	NULLABLE	
subj.version	STRING	NULLABLE	

subj.volumeNumber	STRING	NULLABLE	
subj.includedInDataCatalog	RECORD	NULLABLE	
subj.includedInDataCatalog.id	STRING	NULLABLE	
subj.includedInDataCatalog.type	STRING	NULLABLE	
subj.includedInDataCatalog.name	STRING	NULLABLE	
obj	RECORD	REPEATED	Object metadata.
obj.pid	STRING	NULLABLE	
obj.url	STRING	NULLABLE	
obj.method	STRING	NULLABLE	
obj.verification	STRING	NULLABLE	
obj.work_type_id	STRING	NULLABLE	
obj.publisher	RECORD	REPEATED	
obj.publisher.url	STRING	NULLABLE	
obj.publisher.name	STRING	NULLABLE	
obj.publisher.id	STRING	NULLABLE	
obj.publisher.type	STRING	NULLABLE	
obj.name	STRING	NULLABLE	
obj.datePublished	STRING	NULLABLE	
obj.registrantId	STRING	NULLABLE	

obj.dateModified	TIMESTAMP	NULLABLE	
obj.id	STRING	NULLABLE	
obj.proxyIdentifiers	STRING	NULLABLE	
obj.author	STRING	NULLABLE	
obj.type	STRING	NULLABLE	
obj.funder	RECORD	NULLABLE	
obj.funder.id	STRING	NULLABLE	
obj.funder.type	STRING	NULLABLE	
obj.funder.name	STRING	NULLABLE	
obj.issueNumber	STRING	NULLABLE	
obj.periodical	RECORD	NULLABLE	
obj.periodical.id	STRING	NULLABLE	
obj.periodical.issn	STRING	NULLABLE	
obj.periodical.type	STRING	NULLABLE	
obj.periodical.name	STRING	NULLABLE	
obj.pagination	STRING	NULLABLE	
obj.version	STRING	NULLABLE	
obj.volumeNumber	STRING	NULLABLE	
obj.includedInDataCatalog	RECORD	NULLABLE	
obj.includedInDataCatalog.id	STRING	NULLABLE	

obj.includedInDataCatalog.type	STRING	NULLABLE	
obj.includedInDataCatalog.name	STRING	NULLABLE	
updated	STRING	NULLABLE	will have a value of 'deleted' or 'edited'
updated_reason	STRING	NULLABLE	optional, may point to an announcement page explaining the edit
updated_date	TIMESTAMP	NULLABLE	ISO8601 date string for when the event was updated
message_action	STRING	NULLABLE	
action	STRING	NULLABLE	
jwt	STRING	NULLABLE	