

Developing a Pilot Data Trust for Open Access Ebook Usage (2020-2022) - Data Inventory

Revision History

Version	Date	Author	Comment
0.1	02/03/2022	Niamh Quigley	Draft compiled from documentation created by the Developing a Pilot Data Trust for Open Access Ebook Usage technical team (COKI)
0.2	12/03/2022	Niamh Quigley	Updated following review by Kathryn Napier (Lead Data Scientist, COKI) and Rebecca Handcock (Senior Data Scientist, COKI).
0.3	17/03/2022	Niamh Quigley	Updated following review by Lucy Montgomery and Cameron Neylon (COKI). Added data source properties provided by Aniek Roelofs (Developer, COKI).
0.4	22/03/2022	Niamh Quigley	Updated following review by Rebecca Handcock.
0.5	27/03/2022	Kathryn Napier, Aniek Roelofs, Rebecca Handcock	Updated following review from Christina Drummond, Kathryn Napier, Rebecca Handcock and Aniek Roelofs. Added details and tables for usage data analytic workflows following workflow updates. Updated sources and schemas.
0.6	12/04/2022	Kathryn Napier	Updated following review by Lucy Montgomery.



This project deliverable by the Curtin Open Knowledge Initiative is licensed under a Creative Commons Attribution 4.0 International License.

Suggested citation:

Quigley, N., Hosking, R., Handcock, R.N., Ozaygen, A., Diprose, J.P., Roelofs, A., Napier, K.R., Chien, T.-Y., Lange, R., Mata, S., Neylon, C. and Montgomery, L. (2022). Developing a Pilot Data Trust for Open Access Ebook Usage (2020-2022) - Data Inventory. Perth: Curtin University, pp.87. 10.5281/zenodo.6474480

¹Centre for Culture and Technology, Curtin University, ²Curtin Institute for Computation, Curtin University

Contributor statements

Conceptualization: Richard Hosking, Lucy Montgomery, and Cameron Neylon.

Data curation: Richard Hosking, Rebecca N. Handcock, and Alkim Ozaygen.

Formal analysis: Richard Hosking, Rebecca N. Handcock, Rebecca Lange, Alkim Ozaygen, and Aniek Roelofs.

Funding acquisition: Lucy Montgomery and Cameron Neylon.

Investigation: Richard Hosking, Rebecca N. Handcock, Cameron Neylon, Alkim Ozaygen, Aniek Roelofs, and James P. Diprose.

Methodology: Richard Hosking, Rebecca N. Handcock, Alkim Ozaygen, Aniek Roelofs, and James P. Diprose.

Project administration: Richard Hosking, Lucy Montgomery, Kathryn R. Napier, and Alkim Ozaygen.

Resources: Richard Hosking.

Software: Richard Hosking, Rebecca N. Handcock, Tuan-Yow Chien, Aniek Roelofs, and James P. Diprose.

Supervision: Richard Hosking, Rebecca N. Handcock, Lucy Montgomery, Cameron Neylon, Kathryn R. Napier, and James P. Diprose.

Validation: Richard Hosking, Rebecca N. Handcock, Cameron Neylon, Kathryn R. Napier, and Rebecca Lange.

Visualization: Richard Hosking, Rebecca N. Handcock, Kathryn R. Napier, and James P. Diprose.

Writing - original draft: Richard Hosking, Rebecca N. Handcock, Lucy Montgomery, Cameron Neylon, Tuan-Yow Chien, Sandra Mata, Alkim Ozaygen, Aniek Roelofs, and James P. Diprose.

Writing - review & editing: Rebecca N. Handcock, Lucy Montgomery, Kathryn R. Napier, and Niamh Quigley.

This project was funded by the [Andrew W. Mellon Foundation](#)

Email: coki@curtin.edu.au

Contents

1.	Glossary	7
1	Introduction	10
2	Project Description	11
2.1	Data workflow	12
2.2	Data storage and data quality	14
2.3	Data sharing agreements	14
2.4	Data access controls	15
3	Public Data Sources	16
3.1	Crossref – Metadata	18
3.1.1	Data source properties	18
3.1.2	Data imported from this source	18
3.2	Crossref – Events	19
3.2.1	Data source properties	19
3.2.2	Data imported from this source into the BigQuery Table <code>crossref.crossref_events</code>	20
3.3	Crossref – Fundref	24
3.3.1	Data source properties	24
3.3.2	Data imported from this source	24
3.4	Unpaywall	25
3.4.1	Data source properties	25
3.4.2	Data imported from this source	25
3.5	ORCID	26
3.5.1	Data source properties	26
3.5.2	Data imported from this source	26
3.6	Directory of Open Access Books	27
3.6.1	Data source properties	27
3.6.2	Data imported from this source	27
3.7	OAPEN Metadata	28
3.7.1	Data source properties	28
3.7.2	Data imported from this source into the BigQuery Table <code>opaen.metadata</code>	29
4	Pilot Project Dashboard Partner Data Sources	33
4.1	ONIX-FTP feed from Publishers	35
4.1.1	Data source properties	35
4.1.2	Data imported from this source into the BigQuery table <code>onix</code>	36
4.2	OAPEN IRUS-UK	53

4.2.1	Data source properties	53
4.2.2	Data imported from this source to the BiqQuery table oapen_irus_uk	54
4.3	JSTOR	56
4.3.1	Data source properties	56
4.3.2	Data imported from this source to the BiqQuery tables	57
4.3.2.1	jstor_country	57
4.3.2.2	jstor_institution	57
4.4	Google Books	58
4.4.1	Data source properties	58
4.4.2	Data imported from this source to the BigQuery tables	59
4.4.2.1	google_books_sales	59
4.4.2.2	google_books_traffic	60
4.5	Google Analytics	61
4.5.1	Data source properties	61
4.5.2	Data imported from this source to the BiqQuery table google_analytics	62
4.6	UCL Discovery	64
4.6.1	Data source properties	64
4.6.2	Data imported from this source to the BiqQuery table ucl_discovery	65
4.7	Fulcrum	67
4.7.1	Data imported from this source into the BigQuery table fulcrum	67
4.8	MUSE	69
4.8.1	Data imported from this source into the BigQuery table muse	69
4.9	EBSCO	70
4.9.1	Data imported from this source into the BigQuery table ebsco	70
4.10	SpringerLink	71
4.10.1	Data imported from this source into the BigQuery table springerlink	71
5	Book Usage Data Analytic Workflows	72
5.1	Book Usage Data Analytic Workflow Step 1	72
5.1.1	Onix workflow tables	72
5.1.1.1	BigQuery table onix_workflow.onix_workid_isbn	72
5.1.1.2	BigQuery table onix_workflow.onix_workfamilyid_isbn	72
5.1.1.3	BigQuery table onix_workflow.onix_workid_isbnerrors	72
5.2	Intermediate BigQuery Tables	73
5.3	Data Quality BigQuery Tables	74
5.3.1	BigQuery table oaebu_data_qa.onix_aggregate_metrics	74
5.3.2	BigQuery table oaebu_data_qa.onix_invalid_isbn	74

5.3.3	BigQuery table oaebu_data_qa.<platform>_invalid_isbn	74
5.3.4	BigQuery table oaebu_data_qa.<platform>_unmatched_isbn	74
5.4	Book Usage Data Analytic Workflow Step 2	76
5.4.1	BigQuery table Book table	76
5.5	Book Usage Data Analytic Workflow Step 3	83
5.5.1	Data Export BigQuery Tables	83
5.5.1.1	BigQuery Table oaebu-<publisher>-book-product-list	83
5.5.1.2	BigQuery Table oaebu_<publisher>_book_product_metrics	83
5.5.1.3	BigQuery Table oaebu_<publisher>_product_author_metrics	86
5.5.1.4	BigQuery Table oaebu_<publisher>_book_product_metrics_country	87
5.5.1.5	BigQuery Table oaebu-public-data-country-list	89
5.5.1.6	BigQuery Table oaebu_<publisher>_book_product_metrics_city	89
5.5.1.7	BigQuery Table oaebu_<publisher>_book_product_metrics_institution	90
5.5.1.8	BigQuery Table oaebu_<publisher>_institution_list	90
5.5.1.9	BigQuery Table oaebu_<publisher>_book_product_metrics_events	90
5.5.1.10	BigQuery Table oaebu_<publisher>_book_product_metrics_publisher	91
5.5.1.11	BigQuery Table oaebu_<publisher>_book_product_subject_year_metrics	92
5.5.1.12	BigQuery Tables oaebu_<publisher>_book_product_subject_bic_metrics, oaebu_<publisher>_book_product_subject_bisac_metrics, oaebu_<publisher>_book_product_subject_thema_metrics	94
5.5.1.13	BigQuery Table oaebu_<publisher>_book_product_year_metrics	95
5.5.1.14	BigQuery Table oaebu_<publisher>_unmatched_book_metrics	97
5.6	Elasticsearch/Kibana indexes	98

1. Glossary

Term	Meaning
API	Application Programming Interface – ‘a set of definitions and protocols for building and integrating application software’ ¹
COKI	Curtin Open Knowledge Initiative – a team of data scientists, software developers and researchers at Curtin University, Perth, Australia
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 ²
DAG	DAG is “a collection of organized tasks that you want to schedule and run” ³
dashboard	A dashboard is an interactive, up-to-date page of visualisations that aggregate and summarise data from different sources
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
DOAB	Directory of Open Access Books
DOI	Digital Object Identifier ⁴
edition	Is a new work, but is derived as a revision from an existing work as opposed to being entirely new ¹⁴
eISBN	An identifier for eBooks used by some publishers and platforms, specifically the manually-imported EBSCO data source in the 2020 - 2022 data dashboard pilot
Elasticsearch	Elasticsearch is a search and analytics engine that enables fast searches for large sets of data ⁵
Google Books	Google Books provides paid and free (open access) eBooks ⁶
Google Cloud Platform	Google Cloud Platform is a suite of public cloud computing services, including a range of hosted services for cloud compute, storage, and various applications such as BigQuery, a data warehouse.
ISBN-13	The International Standard Book Number (ISBN) is a 13-digit number that uniquely identifies books and book-like products published internationally ⁷
JSTOR	JSTOR is a digital library, which offers over 7000 open access eBooks ⁸
Kibana	Kibana is a free and open user interface to Elastic Search. ⁹ Kibana is used in the 2020-2022 data dashboard pilot to analyse, search, interact with and visualize the Elasticsearch data
OAeBU	Open Access eBook Usage (2020 - 2022) - a term used to refer to the Mellon Foundation funded pilot project Developing a Pilot Data Trust for Open Access Ebook Usage (2020-2022) ¹⁰

¹ <https://www.redhat.com/en/topics/api/what-are-application-programming-interfaces>

² <https://www.crossref.org/community/>

³ <https://cloud.google.com/composer/docs/run-apache-airflow-dag>

⁴ <https://www.doi.org/index.html>

⁵ <https://www.elastic.co/elasticsearch/>

⁶ <https://play.google.com/books/publish/>

⁷ https://www.isbn.org/faqs_general_questions#isbn_faql

⁸ <https://about.jstor.org/librarians/books/open-access-books-jstor/>

⁹ <https://www.elastic.co/kibana/>

¹⁰ https://educopia.org/data_trust/

Term	Meaning
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 Directory of Open Access Books ¹¹
OAPEN IRUS-UK	OAPEN IRUS-UK – a service for capturing and processing institutional repository usage data, making it possible for institutional repositories to generate COUNTER compliant usage data ¹²
open access	Open access (OA) is free access to information, and unrestricted use of electronic resources for all ¹³
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 ¹⁴
product or book product	A product is a manifestation of a work, and will have its own ISBN-13. There may be several DOIs linked to a single product though (or sometimes none at all) ¹⁵
publisher	A scholarly eBook publisher, who participated in the pilot project to provide metadata for their titles, and metrics of their usage
shard	A database shard is a way of storing data, so that the load for accessing the data can be spread for large amounts of data ¹⁶
SFTP	SSH File Transfer Protocol
technology stack	The collection of frameworks, services, tools and programming languages used to create a software solution
telescope	A telescope is a data workflow that fetches and ingests data from a data source. Some telescopes run workflows that process and output data on remote data locations. ¹⁷ Workflows are built on top of Apache Airflow's Directed Acyclic Graph (DAGs)
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 ¹⁴
work family	A collection of works which are different editions of each other ¹⁴

¹¹ <https://www.oapen.org/oapen/1891940-organisation>

¹² <https://www.jisc.ac.uk/irus>

¹³ <https://en.unesco.org/open-access/what-open-access>

¹⁴ <https://bisg.org/general/custom.asp?page=ONIXforBooks>

¹⁵ https://oaeu-workflows.readthedocs.io/en/latest/oaeu_workflows/workflows/onix_workflow_step_1.html

¹⁶ [https://en.wikipedia.org/wiki/Shard_\(database_architecture\)](https://en.wikipedia.org/wiki/Shard_(database_architecture))

¹⁷ <https://github.com/The-Academic-Observatory/oaeu-workflows>

1 Introduction

This data inventory is accurate at the time of compilation in March 2022. The related code versions are:

- The-Academic-Observatory/observatory-platform: 0.3.0¹⁸
- The-Academic-Observatory/academic-observatory-workflows: 2022.03.0¹⁹
- The-Academic-Observatory/oaebu-workflows: 2022.03.0²⁰
- ONIX parser v1.2²¹

The sources used to compile this data inventory are:

- Book Usage Data Workflows user documentation
<https://oaebu-workflows.readthedocs.io/en/latest/>
- Academic Observatory Workflows user documentation
<https://academic-observatory-workflows.readthedocs.io/en/latest/>
- Book Usage Data Workflows code
<https://github.com/The-Academic-Observatory/oaebu-workflows>
- Academic Observatory workflows code
<https://github.com/The-Academic-Observatory/academic-observatory-workflows>
- Onix Parser code <https://github.com/The-Academic-Observatory/onix-parser>
- 2020 - 2022 pilot project webpage https://educopia.org/data_trust/
- Zenodo community
<https://zenodo.org/communities/2020to2022-developing-pilot-data-trust-for-oa-book-usage/>
- Internal COKI technical documentation
- Google Cloud Platform BigQuery schemas

The structure of this data inventory is:

- Project description
- Public data sources
 - Data source properties
 - Data imported from this source
- Pilot project dashboard partner data sources
 - Data source properties
 - Data imported from this source
- Detailed descriptions of tables and indexes produced by the book usage data workflows

¹⁸ <https://doi.org/10.5281/zenodo.6366701>

¹⁹ <https://doi.org/10.5281/zenodo.6366695>

²⁰ <https://doi.org/10.5281/zenodo.6366691>

²¹ <https://doi.org/10.5281/zenodo.6388112>

2 Project Description

Developing a Pilot Data Trust for Open Access Ebook Usage (January 2020 to March 2022) was [a two-year project funded by the Andrew W. Mellon Foundation](#).²² The pilot project sought to address a growing analytics capability gap in scholarly book publishing. The vast majority of scholarly book publishers lack the technical and staffing capacity to engage with a growing tide of potentially valuable information relating to how OA books are used. This limits publishers' ability to improve the provision of OA books to the communities most likely to benefit from them; as well as to advocate for the value of OA scholarly books to authors and funders.

The pilot project developed systems for gathering information about open access eBooks and their usage from multiple data sources, combining and presenting it in a series of online interactive visualization dashboards for eBook publisher partners.

Presenting usage information can be complex, because open access eBooks can be hosted in multiple repositories, 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.

The data sources in the dashboard pilot included general bibliographic data from public datasets (Crossref and OAPEN) and usage data from multiple platforms (OAPEN, JSTOR, Google Books, Google Analytics, UCL Discovery). The data from these sources are integrated with publisher data via the book-specific metadata standard, ONIX. These data sources are continuously refreshed, so that the online dashboards are up to date. See Figure 1 for an overview of the data architecture of the pilot project. Note that the pilot included preliminary benchmarking only.

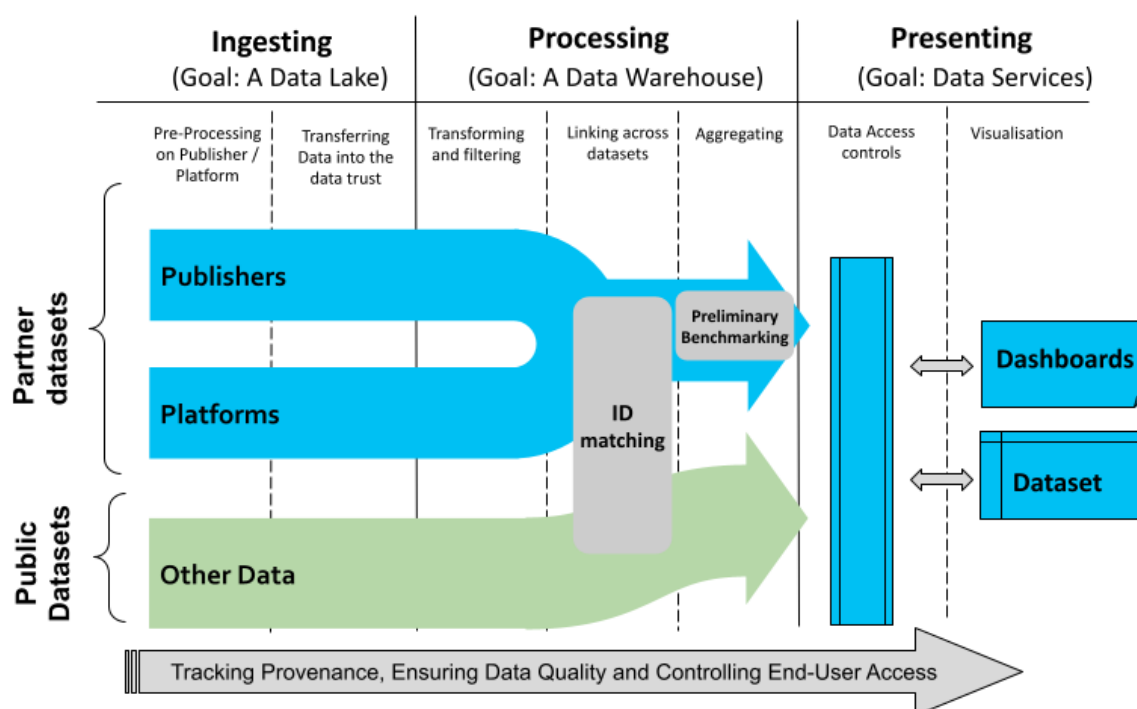


Figure 1 – Data architecture of the pilot project

The dashboard partners that took part in the pilot project were: the University of Michigan Press, Wits University Press, UCL Press, ANU Press, SpringerNature and OAPEN.

²² <https://digital.library.unt.edu/ark:/67531/metadc1596980/>

The online dashboard enables pilot project dashboard partners to see the usage of the eBooks they have published in terms of views, downloads and online mentions and events. Pilot project dashboard partners can also view which countries and institutions are using their eBooks, and which subjects are represented in their open access eBook collections. The data shown in the dashboards is visible only to each pilot project dashboard partners about their own collection. However, the University of Michigan Press have shared their online dashboards publicly, with no login details required:

Link to public dashboards embedded in the University of Michigan Press website:

<https://ebc.press.umich.edu/impact/#oa-book-usage>

Direct link to the University of Michigan Press public dashboards:

<https://tinyurl.com/umpress-public>

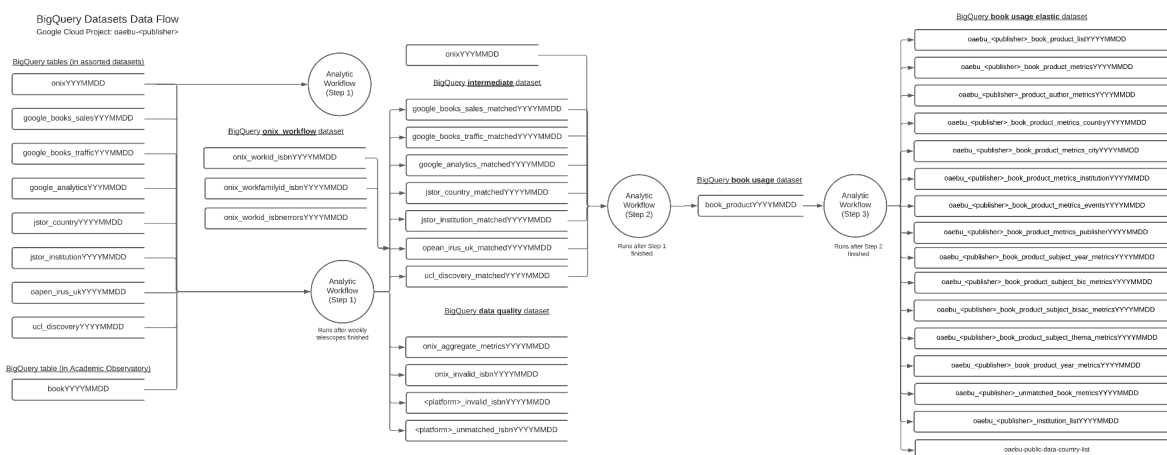
More information about this project is available on the pilot project webpage,²³ and in the Zenodo repository community ‘Developing a Data Trust for Open Access Ebook Usage’.²⁴

2.1 Data workflow

The pilot project’s technology stack uses the book industry metadata interchange standard ONIX, combined with open bibliographic metadata (Crossref and OAPEN) to integrate usage data from multiple sources. Data integration through the pilot project workflows code base is built on an open-source workflow system written in Python. Data workflows (or telescopes) fetch, process, disambiguate and analyse data about open access eBooks from multiple sources. This data is saved to Google Cloud’s BigQuery data warehouse. The multiple workflows include:

2. Ingesting data via telescope workflows from DOAB, Crossref-Metadata, Crossref-Fundref, Crossref-Events, Unpaywall, ORCID, Google Analytics, Google Books, JSTOR, OAPEN IRUS UK, OAPEN Metadata, ONIX, UCL Discovery, and
3. A series of analytic workflows to process and combine the data ingested by the telescope workflows.²⁵

During the pilot project, data sources were automatically ingested via telescopes, or manually imported via data uploads. The manual data uploads were completed for a select number of data sources for two of the pilot project dashboard partners (Project Muse, Fulcrum and EBSCO for the University of Michigan Press, and SpringerLink for Springer Nature). Additionally, several public data sources (DOAB, Crossref-Fundref, Unpaywall, ORCID) have been ingested via telescopes, but have not been aggregated into the book usage data analytic workflows. While this ingested data is stored as part of the pilot project, it is not used in the Kibana visualisations, but may become part of future phases of the book usage data dashboard project.



²³ <https://educopia.org/data-trust/>

²⁴ <https://zenodo.org/communities/2020to2022-developing-pilot-data-trust-for-oa-book-usage/>

²⁵ <https://oaebu-workflows.readthedocs.io/en/latest/workflows/index.html>

Figure 2 – Google Cloud Platform BigQuery datasets data flow

The processed data in the Google Cloud BigQuery data warehouse is then pushed into Elasticsearch where it is accessible to Kibana, an open-source data analytics and dashboarding system. See Figure 2 for the BigQuery datasets dataflow, including table names. Dashboards in Kibana were developed through a visual interface in collaboration with pilot project dashboard partners, and specified in JSON format so they can be maintained and versioned in a code repository. Data access permissions flow through from the underlying sources into the cloud database and Elastic/Kibana. Stakeholder data is sandboxed into separate areas with user access permissions controlling access for each area, providing strong security and privacy. See Figure 3 for an overview of the data workflow for the pilot.

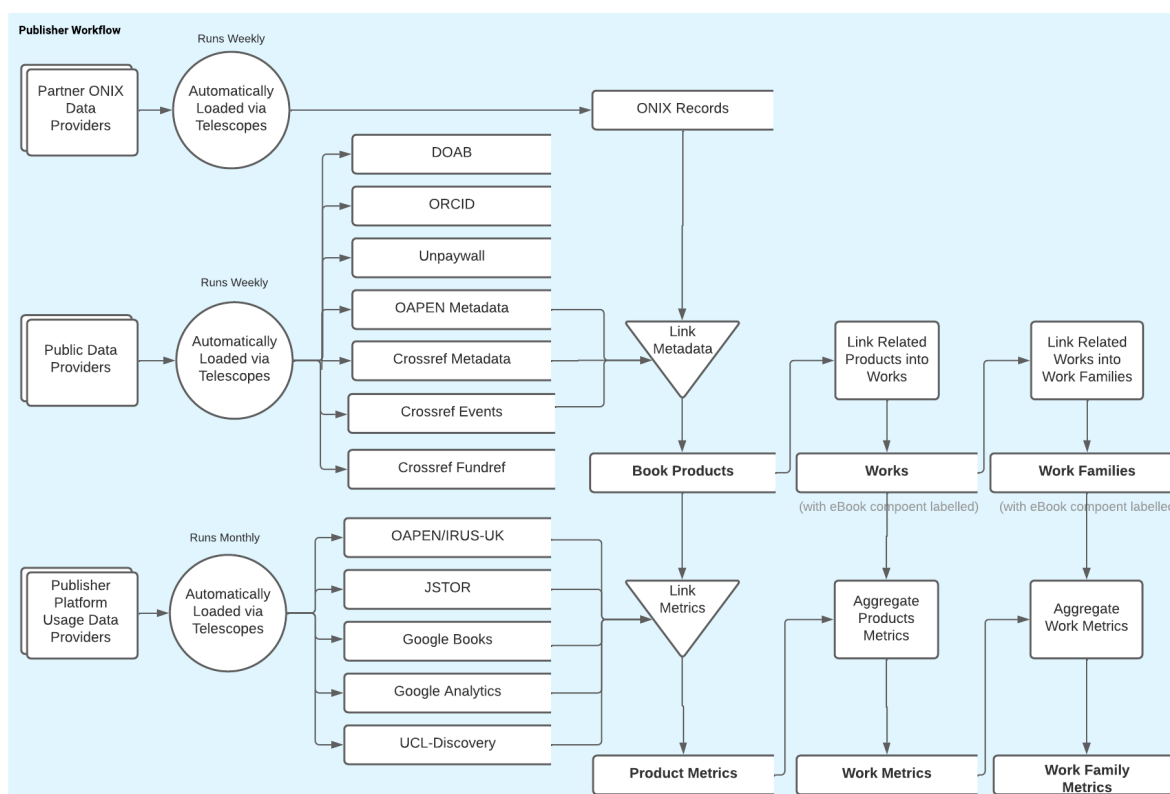


Figure 3 – Overall data workflow for the pilot project

2.2 Data storage and data quality

Data from public and pilot project dashboard partner data sources for the pilot project is stored in the Google Cloud Platform and Elasticsearch. Feedback and recommendations about data quality (particularly in relation to their ONIX feeds) are also provided to individual pilot project dashboard partners via emails and meetings. Pilot project dashboard partners can also view their own Data Quality dashboard to identify any ISBNs for their eBooks that are not listed in their ONIX feeds. Note, only deidentified geographical information is supplied to the pilot project.

Data storage in workflow processing/BigQuery

- Data is saved in Google Cloud storage buckets every time the workflows are run (scheduled weekly), ensuring BigQuery tables can be re-generated if needed
- Backups of the BigQuery SQL instance are taken every 24 hours and stored for 7 days
- BigQuery tables and cloud storage buckets are located in the USA in multiple locations

Data storage in Elasticsearch/Kibana

- Elasticsearch instances are located in the USA.

- Snapshots (backups of a running Elasticsearch cluster) are taken once every 24 hours and stored for 60 days
- Elasticsearch stores snapshots in an off-cluster cloud storage location called a snapshot repository

2.3 Data sharing agreements

Data sharing agreements were put in place between Curtin University (where the Curtin Open Knowledge Initiative is located), and each dashboard pilot project partner. The forms of data included in the data sharing agreements are:

- Individual events of usage
- Usage of individual books by month
- Patterns of usage in the collection
- Benchmarks calculated across categories of books in the collection
- Derived analysis conclusions on the basis of research data

2.4 Data access controls

In addition to legal security provided by data sharing agreements, access control measures were implemented to ensure pilot project dashboard partners could only access their own data.

Data protection

- Access controlled Google Cloud buckets are used for some pilot project dashboard partner data
- A specific pilot project G-suite address is used to transfer some pilot project dashboard partner data to the Google Cloud Storage Buckets (Google Books, JSTOR)

Google Cloud Platform access

- Access to each of the pilot project dashboard partner projects on the Google Cloud Platform is restricted to specific members of the COKI technical team

Dashboard access (Kibana)

- Access to view the Kibana dashboards is manually administered by the COKI team via defined user, role and space permissions
- The publicly available dashboard for the University of Michigan Press at <https://ebc.press.umich.edu/impact/#oa-book-usage> is on a separate Kibana/Elasticsearch instance
- Kibana space permissions are configured so that pilot project dashboard partners can only view data for their own eBook collections, and it is restricted to specific members of the COKI technical team

3 Public Data Sources

The public data sources are where data is publicly available, rather than data provided by a specific pilot project dashboard partner. Each of the data sources and their properties will be presented in this section. If descriptions are blank, this is due to a lack of detail in the original schema from the data source.

Public data source	What is used from this data source
Crossref – Metadata	Ingested, and aggregated into intermediate tables
Crossref – Events	Ingested and aggregated into final tables
Crossref – Fundref	Ingested, not currently aggregated into final tables
Unpaywall	Ingested, not currently aggregated into final tables
ORCID	Ingested, not currently aggregated into final tables
Directory of Open Access Books	Ingested, not currently aggregated into final tables
OAPEN Metadata	Ingested and aggregated into final tables

The following properties are presented for each data source:

Data source property	Property meaning
Provider webpage	Webpage of a data source
Link to online documentation	Link to user documentation created for the pilot project
Link to telescope code	Link to the relevant open-source code in the pilot project
Harvest type	The method by which the data source is harvested: <ul style="list-style-type: none">• API• URL• SFTP
Harvest frequency	How often the data source is harvested
Update frequency	How often the data source is harvested by pilot project workflows
Average runtime	How often the data source is updated at the source side
Average download size	The average time for a DAG run to complete from start to finish (in minutes or hours)
Runs on remote worker	The average size of the raw data that is downloaded (in MB or GB)
Catchup missed runs	If True, any missed DAG runs will be scheduled and executed ²⁶
Table write disposition	When writing query results to tables in BigQuery, the write disposition of the destination table can be one of the following: <ul style="list-style-type: none">• Write if empty• Append to table• Overwrite table/truncate²⁷
Credentials required	Is special access required to harvest the data source i.e., a password or API key

²⁶ <https://airflow.apache.org/docs/apache-airflow/stable/dag-run.html?highlight=runs#catchup>

²⁷ <https://cloud.google.com/bigquery/docs/writing-results>

Data source property	Property meaning
Uses telescope template	Which template is used for this telescope (if any), this can be one of Workflow/Snapshot/Stream/Organisation ²⁸
Each shard includes all data	When the data is stored in a BigQuery table shard ²⁹ , this describes whether the shard represents a snapshot of all the data at that point or the shard only has entries for a specific time period and all shards together form the complete data
Corresponding BigQuery table(s)	If relevant, the BigQuery tables that are created from a data source
Notes on data	Any other notes about this data source

²⁸

<https://github.com/The-Academic-Observatory/observatory-platform/tree/develop/observatory-platform/observatory/platform/workflows>

²⁹ https://cloud.google.com/bigquery/docs/partitioned-tables#dt_partition_shard

3.1 Crossref – Metadata

Crossref is a non-for-profit membership organization, and an official Digital Object Identifier (DOI) Registration Agency of the International DOI Foundation. They make metadata available for all DOIs registered with Crossref.³⁰

The Crossref metadata is available in snapshots that are released monthly and can be downloaded through their API, each snapshot contains all metadata up until that date. Each snapshot is stored in a separate BigQuery shard. Although this data source is ingested into the pilot project, it is not currently aggregated into the final tables.

3.1.1 Data source properties

Crossref Metadata – data source properties	
Provider webpage	https://www.crossref.org/services/metadata-retrieval/
Link to online documentation	https://academic-observatory-workflows.readthedocs.io/en/latest/telescopes/crossref_metadata.html
Link to telescope code	https://github.com/The-Academic-Observatory/academic-observatory-workflows/blob/develop/academic_observatory_workflows/workflows/crossref_metadata_telescope.py
Harvest type	API
Harvest frequency	Monthly
Update frequency	Monthly
Average runtime	12 hours
Average download size	150 GB
Runs on remote worker	True
Catchup missed runs	True
Table write disposition	Truncate
Credentials required	No
Uses telescope template	Snapshot
Each shard includes all data	Yes
Corresponding BigQuery table(s)	crossref.crossref_metadataYYYYMMDD
Notes on data	None

3.1.2 Data imported from this source

CrossRef Metadata data is ingested, but not aggregated into the final tables in the pilot project.

³⁰ <https://www.crossref.org/community/>

3.2 Crossref – Events

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’.³¹ The event data³² is retrieved using the Crossref Events API.³³ This data source is ingested into the pilot project, and aggregated into the final tables.

3.2.1 Data source properties

Crossref Events – data source properties	
Provider webpage	https://www.crossref.org/services/event-data/
Link to online documentation	https://academic-observatory-workflows.readthedocs.io/en/latest/telescopes/crossref_events.html
Link to telescope code	https://github.com/The-Academic-Observatory/academic-observatory-workflows/blob/develop/academic_observatory_workflows/workflows/crossref_events_telescope.py
Harvest type	API
Harvest frequency	Weekly
Update frequency	Daily
Average runtime	2 hours
Average download size	10 GB
Runs on remote worker	True
Catchup missed runs	False
Table write disposition	Append
Credentials required	No
Uses telescope template	Stream
Corresponding BigQuery tables	crossref.crossref_events crossref.crossref_events_partitions
Notes on data	For pilot project dashboard partner dashboards, event data is available from different time points in 2018 onwards for Crossref Events

³¹ <https://www.crossref.org/services/event-data/>

³² <https://www.eventdata.crossref.org/guide/data/events/>

³³ <https://www.eventdata.crossref.org/guide/service/query-api/>

3.2.2 Data imported from this source into the BigQuery Table crossref.crossref_events

Field 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	BOOLEAN	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	

3.3 Crossref – Fundref

Crossref maintains an open Funder Registry, containing persistent identifiers for grant-giving organizations.³⁴ Although this data source is ingested into the OAeBU data dashboard pilot, it is not currently aggregated into the final tables.

3.3.1 Data source properties

Crossref Fundref – data source properties	
Provider webpage	https://www.crossref.org/services/funder-registry/
Link to online documentation	https://academic-observatory-workflows.readthedocs.io/en/latest/telescopes/crossref_fundref.html
Link to telescope code	https://github.com/The-Academic-Observatory/academic-observatory-workflows/blob/develop/academic_observatory_workflows/workflows/crossref_fundref_telescope.py
Harvest type	API
Harvest frequency	Weekly
Update frequency	Random
Average runtime	5 min
Average download size	50 MB
Runs on remote worker	True
Catchup missed runs	True
Table write disposition	Truncate
Credentials required	No
Uses telescope template	Snapshot
Each shard includes all data	Yes
Corresponding BigQuery tables	crossref.crossref_fundrefYYYYMMDD
Notes on data	None

3.3.2 Data imported from this source

Crossref Fundref data is ingested, but not aggregated into the final tables in the pilot project..

³⁴ <https://www.crossref.org/services/funder-registry/>

3.4 Unpaywall

Unpaywall is an open database of metadata for open access research outputs, including journal articles, books and book chapters.³⁵ Although this data source is ingested into the pilot project, it is not currently aggregated into the final tables.

3.4.1 Data source properties

Unpaywall – data source properties	
Provider webpage	https://unpaywall.org/products
Link to online documentation	https://academic-observatory-workflows.readthedocs.io/en/latest/telescopes/unpaywall.html
Link to telescope code	https://github.com/The-Academic-Observatory/academic-observatory-workflows/blob/develop/academic_observatory_workflows/workflows/unpaywall_telescope.py
Harvest type	URL
Harvest frequency	Daily
Update frequency	Every release
Average runtime	15 min
Average download size	100 MB
Runs on remote worker	False
Catchup missed runs	True
Table write disposition	Append
Credentials required	Yes
Corresponding BigQuery tables	Stream
Notes on data	None

3.4.2 Data imported from this source

Unpaywall data is ingested, but not aggregated into the final tables in the pilot project..

³⁵ <https://unpaywall.org/>

3.5 ORCID

ORCID is a non-profit organization that provides researchers with a unique and persistent digital identifier. This enables researchers to keep a record of their research contributions.³⁶ Although this data source is ingested into the OAeBU data dashboard pilot, it is not currently aggregated into the final tables.

3.5.1 Data source properties

ORCID – data source properties	
Provider webpage	https://info.orcid.org/documentation/
Link to online documentation	https://academic-observatory-workflows.readthedocs.io/en/latest/telescopes/orcid.html
Link to telescope code	https://github.com/The-Academic-Observatory/academic-observatory-workflows/blob/develop/academic_observatory_workflows/workflows/orcid_telescope.py
Harvest type	API
Harvest frequency	Monthly
Update frequency	Daily
Average runtime	15 hours
Average download size	70 GB
Runs on remote worker	True
Catchup missed runs	False
Table write disposition	Append
Credentials required	Yes
Uses telescope template	Stream
Corresponding BigQuery tables	orcid.orcid orcid.orcid_partitions
Notes on data	None

3.5.2 Data imported from this source

ORCID data is ingested, but not aggregated into the final tables in the pilot project..

³⁶ <https://orcid.org/>

3.6 Directory of Open Access Books

The Directory of Open Access Books (DOAB) is a directory of open-access peer reviewed scholarly books, with the aim of increasing discoverability of books. Academic publishers provide metadata of their Open Access books to DOAB.³⁷ DOAB services are free of charge, and all data is freely available under a CC0 1.0 license.³⁸

The DOAB data is downloaded from a CSV file that is updated daily. All records are initially loaded into BigQuery. After the initial run, any new records are appended to the table while any edited records are updated in-place. Although this data source is ingested into the pilot project., it is not currently aggregated into the final tables.

3.6.1 Data source properties

Directory of Open Access Books – data source properties	
Provider webpage	https://www.doabooks.org/en/doab/metadata-harvesting-and-content-dissemination
Link to online documentation	https://oaeu-workflows.readthedocs.io/en/latest/oaeu_workflows/telescopes/doab.html
Link to telescope code	https://github.com/The-Academic-Observatory/oaeu-workflows/blob/develop/oaeu_workflows/workflows/doab_telescope.py
Harvest type	API
Harvest frequency	Weekly
Update frequency	Daily
Average runtime	15 min
Average download size	50 MB
Runs on remote worker	False
Catchup missed runs	False
Table write disposition	Append
Credentials required	No
Corresponding BigQuery tables	doab.doabYYYYMMDD
Notes on data	None

3.6.2 Data imported from this source

DOAB data is ingested, but not aggregated into the final tables in the pilot project..

³⁷ <https://www.doabooks.org/en>

³⁸ <https://www.doabooks.org/en/doab/metadata-harvesting-and-content-dissemination>

3.7 OAPEN Metadata

OAPEN operates OAPEN Library – a central repository that hosts and disseminates OA books; and the Directory of Open Access Books – a discovery service that indexes OA books, in partnership with OpenEdition.³⁹ OAPEN provides their metadata under a CC0 1.0 license.⁴⁰

The OAPEN Metadata telescope collects data from the OAPEN Metadata feeds. OAPEN enables libraries and aggregators to use the metadata of all available titles in the OAPEN Library. The metadata is available in different formats, with this telescope harvesting the data in the CSV format. This data source is ingested into the pilot project, and aggregated into the final tables.

3.7.1 Data source properties

OAPEN Metadata – data source properties	
Provider webpage	https://www.oapen.org/resources/15635975-metadata
Link to online documentation	https://oaeu-workflows.readthedocs.io/en/latest/oaeu_workflows/telescopes/oapen_metadata.html
Link to telescope code	https://github.com/The-Academic-Observatory/oaeu-workflows/blob/develop/oaeu_workflows/workflows/oapen_metadata_telemeter.py
Harvest type	URL
Harvest frequency	Weekly
Update frequency	Daily
Average runtime	5 min
Average download size	50 MB
Runs on remote worker	True
Catchup missed runs	False
Table write Disposition	Append
Credentials required	No
Uses telescope template	Stream
Corresponding BigQuery tables	oapen.metadataYYYYMMDD
Notes on data	<p>For pilot project dashboard partner dashboards, usage data is from April 2020 onwards for OAPEN (COUNTER 5).</p> <p>For pilot project dashboard partner dashboards, the field 'title_requests' is used from different points in 2018 to March 2020 (COUNTER 4), then the field 'total_item_requests' is used from April 2020 for OAPEN</p>

³⁹ <https://oapen.org/oapen/1891940-organisation>

⁴⁰ <https://oapen.org/librarians/15635975-metadata>

3.7.2 Data imported from this source into the BigQuery Table opaen.metadata

Field name	Type	Mode	Description
id	STRING	NULLABLE	ID for the book.
collection	STRING	REPEATED	Collections
BITSTREAM_Download_URL	STRING	REPEATED	URL of the book's PDF file on OAPEN repository.
BITSTREAM_ISBN	STRING	REPEATED	ISBN of the book. Can have multiple isbn separated with .
BITSTREAM_License	STRING	REPEATED	Text describing the licence.
BITSTREAM_Webshop_URL	STRING	REPEATED	URL of the book's web page on publisher site.
dc	RECORD	REQUIRED	dc
dc.contributor	RECORD	NULLABLE	contributor
dc.contributor.advisor	STRING	NULLABLE	Advisor name.
dc.contributor.author	STRING	REPEATED	Author name.
dc.contributor.editor	STRING	REPEATED	Editor name.
dc.contributor.other	STRING	REPEATED	Other contributor(s) name.
dc.date	RECORD	REQUIRED	date
dc.date.accessioned	STRING	REPEATED	Date made available in OAPEN. Starts from June 2010
dc.date.available	TIMESTAMP	REPEATED	Date of upload to the new DSPACE platform. Starts from April of 2020.
dc.date.issued	DATE	REPEATED	Year of publication of the book.
dc.description	RECORD	NULLABLE	description
dc.description.value	STRING	NULLABLE	description
dc.description.abstract	STRING	REPEATED	Abstract of the book.
dc.description.provenance	STRING	REPEATED	Upload information, on who created the record, when it is uploaded to DSPACE.
dc.description.version	STRING	NULLABLE	?
dc.identifier	RECORD	NULLABLE	identifier
dc.identifier.value	STRING	REPEATED	'Old' OAPEN ID or ONIX import ID. It can be a number or can be in the form of a string ("ONIX.20200714.9789811555732.9")
dc.identifier.isbn	STRING	REPEATED	Additional ISBN(s) number of the book.
dc.identifier.issn	STRING	NULLABLE	ISSN number.
dc.identifier.uri	STRING	NULLABLE	Web page of the book on OAPEN (HANDLE = unique ID).
dc.identifier.urlwebshop	STRING	NULLABLE	[Blank]
dc.language	STRING	REPEATED	Language of the book.
dc.relation	RECORD	NULLABLE	relation
dc.relation.isnodouble	STRING	REPEATED	?

dc.relation.ispartofseries	STRING	REPEATED	Name of the book series.
dc.relation.isreplacedbydouble	STRING	REPEATED	?
dc.rights	RECORD	NULLABLE	rights
dc.rights.uri	STRING	NULLABLE	[Blank]
dc.source	STRING	NULLABLE	[Blank]
dc.subject	RECORD	NULLABLE	subject
dc.subject.classification	STRING	REPEATED	BIC subject category.
dc.subject.classification_code	STRING	REPEATED	BIC subject category code.
dc.subject.other	STRING	REPEATED	Keywords.
dc.title	RECORD	NULLABLE	title
dc.title.value	STRING	REPEATED	Title of the book.
dc.title.alternative	STRING	NULLABLE	Alternative.
dc.type	STRING	REPEATED	Can be 'book', 'chapter', 'book book'.
dcterms	RECORD	NULLABLE	[Blank]
dcterms.abstract	STRING	NULLABLE	[Blank]
eperson	RECORD	NULLABLE	[Blank]
eperson.firstname	STRING	NULLABLE	[Blank]
eperson.language	STRING	NULLABLE	[Blank]
eperson.lastname	STRING	NULLABLE	[Blank]
eperson.phone	STRING	NULLABLE	[Blank]
grantor	RECORD	NULLABLE	[Blank]
grantor.acronym	STRING	NULLABLE	[Blank]
grantor.doi	STRING	NULLABLE	[Blank]
grantor.jurisdiction	STRING	NULLABLE	[Blank]
grantor.name	STRING	NULLABLE	[Blank]
grantor.number	STRING	REPEATED	[Blank]
oopen	RECORD	NULLABLE	oopen
oopen.abstract	RECORD	NULLABLE	abstract
oopen.abstract.otherlanguage	STRING	NULLABLE	Abstract in other language.
oopen.autodoi	STRING	NULLABLE	
oopen.chapternumber	INTEGER	NULLABLE	Number of chapter.
oopen.collection	STRING	REPEATED	OOPEN Collection.
oopen.description	RECORD	NULLABLE	Description in another language.
oopen.description.otherlanguage	STRING	NULLABLE	Description in another language.
oopen.embargo	STRING	NULLABLE	[Blank]
oopen.grant	RECORD	NULLABLE	Grant.
oopen.grant.acronym	STRING	REPEATED	Acronym of the grant.
oopen.grant.number	STRING	REPEATED	Grant number, can include strings.
oopen.grant.program	STRING	REPEATED	Grant program name.
oopen.grant.project	STRING	REPEATED	Grant project name.
oopen.identifier	RECORD	NULLABLE	Identifier

oopen.identifier.value	STRING	NULLABLE	Identifier value.
oopen.identifier.doi	STRING	NULLABLE	DOI of the book. Can also include full URL.
oopen.identifier.isbn	STRING	NULLABLE	ISBN of the book.
oopen.identifier.ocn	STRING	REPEATED	
oopen.imprint	STRING	REPEATED	Imprint of the book.
oopen.notes	STRING	REPEATED	
oopen.pages	STRING	NULLABLE	Number of pages. Can include arabic page notation, as well as strings.
oopen.place	RECORD	NULLABLE	Place
oopen.place.publication	STRING	NULLABLE	City of publisher.
oopen.redirect	STRING	REPEATED	
oopen.relation	RECORD	NULLABLE	Relation
oopen.relation.funds	STRING	NULLABLE	ID of book - part of funder record
oopen.relation.hasChapter	STRING	NULLABLE	[ID of chapter - part of book record
oopen.relation.hasChapter_dc	RECORD	NULLABLE	
oopen.relation.hasChapter_dc.title	STRING	REPEATED	Chapter name of chapter records.
oopen.relation.isFundedBy	STRING	NULLABLE	ID of funder - part of book or chapter record
oopen.relation.isFundedBy_grantor	RECORD	NULLABLE	
oopen.relation.isFundedBy_grantor.name	STRING	REPEATED	Funder name.
oopen.relation.isPartOfBook	STRING	NULLABLE	ID of book - part of chapter record
oopen.relation.isPartOfBook_dc	RECORD	NULLABLE	
oopen.relation.isPartOfBook_dc.title	STRING	NULLABLE	Chapter's book title.
oopen.relation.isPublishedBy	STRING	NULLABLE	Publisher name
oopen.relation.isPublishedBy_publisher	RECORD	NULLABLE	
oopen.relation.isPublishedBy_publisher.name	STRING	NULLABLE	
oopen.relation.isPublisherOf	STRING	REPEATED	ID of book - part of publisher record
oopen.relation.isbn	STRING	REPEATED	
oopen.remark	RECORD	NULLABLE	Remark
oopen.remark.private	STRING	NULLABLE	
oopen.remark.public	STRING	REPEATED	
oopen.series	RECORD	NULLABLE	
oopen.series.number	STRING	NULLABLE	
publisher	RECORD	NULLABLE	Publisher
publisher.country	STRING	NULLABLE	[Blank]
publisher.name	STRING	NULLABLE	[Blank]

publisher.peerreviewpolicy	STRING	NULLABLE	Text describing policy
publisher.website	STRING	NULLABLE	

4 Pilot Project Dashboard Partner Data Sources

The pilot project dashboard partner data sources are where permission has been granted for the pilot project to access usage data about publishers' open access eBook collections. If descriptions are blank, this is due to a lack of detail in the original schema from the data source.

Partner data source	What is used from this data source
ONIX-FTP feed from publishers	Information about eBooks from each publisher including Title, DOI, Edition, Publishing Dates, Contributors and Subjects
OAPEN IRUS-UK	eBook usage by title for OAPEN eBooks
JSTOR	eBook usage by title, institution, and country for specific publishers, where eBooks were accessed via JSTOR
Google Books	eBook views where eBooks were accessed via Google Books
Google Analytics	eBook views by country and territory for visitors to the ANU Press website
UCL Discovery	eBook downloads by title, country and Thema subject in UCL Discovery data
Fulcrum	eBook usage for top authors, top titles, top institutions and top publishers, for University of Michigan Press titles in Fulcrum data
MUSE	eBook usage by country, institution and format for both gated and open access titles, for University of Michigan Press titles in MUSE data
EBSCO	eBook usage by imprint, subjects, customers and markets for University of Michigan Press titles in EBSCO data
SpringerLink	eBook usage by title for Springer Nature titles in SpringerLink data

The following properties are presented for each data source:

Data source property	Property meaning
Provider webpage	Webpage of a data source
Link to online documentation	Link to user documentation created for the pilot project
Link to telescope code	Link to the relevant open-source code in the pilot project
Harvest type	The method by which the data source is harvested: <ul style="list-style-type: none">• API• URL• SFTP
Harvest frequency	How often the data source is harvested
Update frequency	How often the data source is harvested by pilot project workflows
Average runtime	How often the data source is updated at the source side
Average download size	The average time for a DAG run to complete from start to finish (in minutes or hours)
Runs on remote worker	The average size of the raw data that is downloaded (in MB or GB)
Catchup missed runs	If True, any missed DAG runs will be scheduled and executed ⁴¹

⁴¹ <https://airflow.apache.org/docs/apache-airflow/stable/dag-run.html?highlight=runs#catchup>

Data source property	Property meaning
Table write disposition	When writing query results to tables in BigQuery, the write disposition of the destination table can be one of the following: <ul style="list-style-type: none"> • Write if empty • Append to table • Overwrite table/truncate⁴²
Credentials required	Is special access required to harvest the data source i.e., a password or API key
Uses telescope template	Which template is used for this telescope (if any), this can be one of Workflow/Snapshot/Stream/Organisation ⁴³
Each shard includes all data	When the data is stored in a BigQuery table shard ⁴⁴ , this describes whether the shard represents a snapshot of all the data at that point or the shard only has entries for a specific time period and all shards together form the complete data
Corresponding BigQuery table(s)	If relevant, the BigQuery tables that are created from a data source
Notes on data	Any other notes about this data source

⁴² <https://cloud.google.com/bigquery/docs/writing-results>

⁴³

<https://github.com/The-Academic-Observatory/observatory-platform/tree/develop/observatory-platform/observatory/platform/workflows>

⁴⁴ https://cloud.google.com/bigquery/docs/partitioned-tables#dt_partition_shard

4.1 ONIX-FTP feed from Publishers

ONIX is a standard that book publishers use to share information about the books that they have published.⁴⁵ Publishers that have ONIX feeds are given credentials and access to their own upload folder on the Mellon SFTP server. The publisher uploads their ONIX feed to their upload folder on a weekly, fortnightly or monthly basis. The pilot project ONIX telescope downloads, transforms (with the ONIX parser Java command line tool) and then loads the ONIX data into BigQuery for further processing.

4.1.1 Data source properties

ONIX-FTP – data source properties	
Provider webpage	Direct from publisher
Link to online documentation	https://oaebu-workflows.readthedocs.io/en/latest/oaebu_workflows/telescopes/onix.html
Link to telescope code	https://github.com/The-Academic-Observatory/oaebu-workflows/blob/develop/oaebu_workflows/workflows/onix_telescope.py
Harvest type	SFTP server
Harvest frequency	Weekly
Average runtime	10-20 mins
Average download size	10-100 MB
Runs on remote worker	False
Catchup missed runs	False
Credentials required	Yes
Each shard includes all data	Yes
Corresponding BigQuery tables	onix.onixYYYYMMDD
Notes on data	None

⁴⁵ <https://www.editeur.org/83/Overview/>

4.1.2 Data imported from this source into the BigQuery table onix

Field 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.
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 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 <DescriptiveDetail> composite, to give the primary form of the product title. The composite is repeatable with different title types.
Collections.TitleDetails.Title Elements	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 <TitleDetail> 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 <TitleElement> composite must include at least one of: <PartNumber>; <YearOfAnnual>; <TitleText>, <NoPrefix/> together with <TitleWithoutPrefix>, or <TitlePrefix> together with

			<TitleWithoutPrefix>. 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 <TitleElement> 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 <TitleText> is omitted, and if the <TitleWithoutPrefix> 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 <NoPrefix/> or <TitlePrefix> 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 where there are multiple title elements within a <TitleDetail> composite, each occurrence of the <TitleElement> composite should carry a <SequenceNumber>.
Collections.TitleDetails.TitleElements.TitleText	STRING	NULLABLE	The text of a title element, excluding any subtitle. Optional and non-repeating, may only be used where <TitlePrefix>, <NoPrefix/> and <TitleWithoutPrefix> are not used.
Collections.TitleDetails.TitleType	STRING	NULLABLE	An ONIX description indicating the type of a title. Mandatory in each occurrence of the <TitleDetail> composite, and non-repeating.

Collections.CollectionIdentifiers	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.CollectionIdentifiers.CollectionIDType	STRING	NULLABLE	An ONIX description identifying a scheme from which an identifier in the <IDValue> element is taken. Mandatory in each occurrence of the <CollectionIdentifier> composite, and non-repeating.
Collections.CollectionIdentifiers.IDValue	STRING	NULLABLE	An identifier of the type specified in the <CollectionIDType> field. Mandatory in each occurrence of the <CollectionIdentifier> composite, and non-repeating.
Collections.CollectionIdentifiers.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 <CollectionIDType> 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 <Collection> 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, <RecordReference>, is a string of text which uniquely identifies the record. The second, <NotificationType>, 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 <RelatedWork> 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 <RelatedWork> 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 <IDValue> element is taken. Mandatory in each occurrence of the <WorkIdentifier> composite, and non-repeating.
RelatedWorks.WorkIdentifiers.IDValue	STRING	NULLABLE	An identifier of the type specified in the <WorkIDType> element. Mandatory in each occurrence of the <WorkIdentifier> 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 <WorkIDType> element 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 <Text> element. Mandatory in each occurrence of the <TextContent> composite, and non-repeating.
TextContent.Text	STRING	REPEATED	The text specified in the <TextType> element. Mandatory in each occurrence of the <TextContent> composite, and repeatable when essentially identical text is supplied in multiple languages. The language attribute is optional for a single instance of <Text>, but must be included in each instance if <Text> 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 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 <Imprint> composite does not carry an <ImprintName>. 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 <IDValue> element is taken. Mandatory in each occurrence of the <ImprintIdentifier> composite.
Imprints.ImprintIdentifiers.IDValue	STRING	NULLABLE	A code value taken from the scheme specified in the <ImprintIDType> element. Mandatory in each occurrence of the <ImprintIdentifier> 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 <ImprintIDType> 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 <Imprint> 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

			<Publisher> 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 <Publisher> 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 <WebsiteLink> element. Optional, and repeatable to provide parallel descriptive text in multiple languages. The language attribute is optional for a single instance of <WebsiteDescription>, but must be included in each instance if <WebsiteDescription> is repeated.
Publishers.Websites.WebsiteRole	STRING	NULLABLE	An ONIX description which identifies the role or purpose of the website which is linked through the <WebsiteLink> element. Optional and non-repeating.
Publishers.Websites.WebsiteLinks	STRING	REPEATED	The URL for the website. Mandatory in each occurrence of the <Website> 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 <WebsiteLink>, but must be included in each instance if <WebsiteLink> 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 <Publisher> 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 <RelatedProduct>, and non-repeating. If supplied, should be identical to the <ProductForm> element supplied in the <DescriptiveDetail> 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 <RelatedProduct> 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.
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 <DescriptiveDetail> 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 <TitleDetail> 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 <TitleElement> composite must include at least one of: <PartNumber>; <YearOfAnnual>; <TitleText>, <NoPrefix/> together with <TitleWithoutPrefix>, or <TitlePrefix> together with <TitleWithoutPrefix>. 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 <TitlePrefix>, <NoPrefix/> and <TitleWithoutPrefix> 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 <TitleElement> 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 <TitleElement> 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 <TitleDetail> composite, each occurrence of the <TitleElement> composite should carry a <SequenceNumber>.
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 <NoPrefix/> or <TitlePrefix> element is also present.
TitleDetails.TitleElements.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 <TitleText> is omitted, and if the <TitleWithoutPrefix> 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 <TitleDetail> 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 <SequenceNumber> 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 details 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 <MarketPublishingDetail> (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 <PublishingDate> composite, and non-repeating.
PublishingDates.Date	INTEGER	NULLABLE	The date specified in the <PublishingDateRole> field. Mandatory in each occurrence of the <PublishingDate> composite, and non-repeating. <Date> may carry a dateformat attribute: if the attribute is missing, then <DateFormat> indicates the format of the date; if both dateformat attribute and <DateFormat> 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 <Language> composite, and non-repeating.
Languages.LanguageCode	STRING	NULLABLE	An ISO code indicating a language. Mandatory in each occurrence of the <Language> composite, and non-repeating.
ProductForm	STRING	NULLABLE	An ONIX description which indicates the primary form of a related product. Optional in an occurrence of <RelatedProduct>, and

			non-repeating. If supplied, should be identical to the <ProductForm> element supplied in the <DescriptiveDetail> 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 <ProfessionalPosition>, but must be included in each instance if <ProfessionalPosition> is repeated.
Contributors.ProfessionalAffiliations.Affiliations	STRING	NULLABLE	An organization to which a contributor to the product was affiliated at the time of its creation, and – if the <ProfessionalPosition> 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 <TextContent> 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

			<p>biographical notes in multiple languages. The language attribute is optional for a single instance of <BiographicalNote>, but must be included in each instance if <BiographicalNote> 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 <Website> composite below.</p>
Contributors.BiographicalNotes.TextFormat	STRING	NULLABLE	The textformat attribute.
Contributors.BiographicalNotes.Note	STRING	NULLABLE	The biographical note.
Contributors.TitlesBeforeNames	STRING	NULLABLE	The first part of a structured name of a person who contributed to the creation of the product: qualifications and/or titles preceding a person's names, e.g. '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 <Contributor> 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 <Contributor> 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 <WebsiteLink> element. Optional, and repeatable to provide parallel descriptive text in multiple languages. The language attribute is optional for a single instance of <WebsiteDescription>, but must be included in each instance if <WebsiteDescription> is repeated.
Contributors.Websites.WebsiteRole	STRING	NULLABLE	An ONIX description which identifies the role or purpose of the website which is linked through the <WebsiteLink> element. Optional and non-repeating.

Contributors.Websites.WebsiteLinks	STRING	REPEATED	The URL for the website. Mandatory in each occurrence of the <Website> 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 <WebsiteLink>, but must be included in each instance if <WebsiteLink> 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 <Contributor> 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 <ContributorDateRole> field. Mandatory in each occurrence of the <ContributorDate> composite, and non-repeating. <Date> may carry a dateFormat attribute: if the attribute is missing, then <DateFormat> indicates the format of the date; if both dateFormat attribute and <DateFormat> 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 <ContributorDate> 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 <TitleDetail> composite, each occurrence of the <TitleElement> composite should carry a <SequenceNumber>.
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, e.g. '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, e.g. '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 <AlternativeName> composite is optional, and is repeatable to provide multiple alternative names for the contributor.
Contributors.NamesBefore Key	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), e.g. 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 <CountryCode> or the <RegionCode> elements in each occurrence of <ContributorPlace>.
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 (e.g. Baile Átha Cliath and Dublin, or Bruxelles and Brussel). The language attribute is optional for a single instance of <LocationName>, but must be included in each instance if <LocationName> 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 <ContributorPlace> 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.
COKI_ID	STRING	NULLABLE	The product's internal 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 <SubjectSchemeIdentifier> element, or of free language keywords if the scheme is specified as 'keywords'; or the text equivalent to the <SubjectCode> value, if both code and text are sent. Either <SubjectCode> or <SubjectHeadingText> or both must be present in each occurrence of the <Subject> composite.
Subjects.SubjectSchemeIdentifier	STRING	NULLABLE	A number which identifies a version or edition of the subject scheme specified in the associated <SubjectSchemeIdentifier> 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 <SubjectSchemeIdentifier> element. Optional and non-repeating.
Subjects.SubjectSchemeName	STRING	NULLABLE	A name identifying a proprietary subject scheme (i.e. a scheme which is not a standard and for which there is no individual identifier code) when <SubjectSchemeIdentifier> is coded '24'. Optional and non-repeating.
Subjects.SubjectCode	STRING	NULLABLE	A subject class or category code from the scheme specified in the <SubjectSchemeIdentifier> element. Either <SubjectCode> or <SubjectHeadingText> or both must be present in each occurrence of the <Subject> composite. Non-repeating.

Subjects.MainSubject	BOOLEAN	NULLABLE	An empty element that identifies an instance of the <Subject> composite as representing the main subject category for the product. The main category may be expressed in more than one subject scheme, i.e. there may be two or more instances of the <Subject> composite, using different schemes, each carrying the <MainSubject/> flag, so long as there is only one main category per scheme. Optional and non-repeating in each occurrence of the <Subject> 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, e.g. 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, e.g. running time for an audio or video product. Mandatory in each occurrence of the <Extent> composite, and non-repeating. From Issue 9 of the code lists, an extended set of values for <ExtentType> has been defined to allow more accurate description of pagination.
Extent.ExtentValue	INTEGER	NULLABLE	The numeric value of the extent specified in <ExtentType>. Optional, and non-repeating. However, either <ExtentValue> or <ExtentValueRoman> must be present in each occurrence of the <Extent> composite; and it is very strongly recommended that <ExtentValue> 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 <ExtentValue> and the format in which the value is presented. Mandatory in each occurrence of the <Extent> 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.

4.2 OAPEN IRUS-UK

IRUS-UK provides OAPEN COUNTER standard access reports. 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 the pilot project data, and only de-identified geographical information transferred to the pilot project.

4.2.1 Data source properties

OAPEN IRUS-UK – data source properties	
Provider webpage	https://irus.jisc.ac.uk/r5/uk/
Link to online documentation	https://oaebu-workflows.readthedocs.io/en/latest/oaebu_workflows/telescopos/oapen_irus_uk.html
Link to telescope code	https://github.com/The-Academic-Observatory/oaebu-workflows/blob/develop/oaebu_workflows/workflows/oapen_irus_uk_telescope.py
Harvest type	API
Harvest frequency	Monthly
Update frequency	Daily
Average runtime	5 min
Average download size	5 MB
Runs on remote worker	False
Catchup missed runs	True
Table write disposition	Truncate
Credentials required	Yes
Uses telescope template	Snapshot
Each shard includes all data	No
Corresponding BigQuery tables	oapen.oapen_irus_ukYYYYMMDD oaebu_intermediate.oapen_irus_uk_matchedYYYYMMDD
Notes on data	None

4.2.2 Data imported from this source to the BigQuery table oapen_irus_uk

Field 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.
unique_item_investigations	INTEGER	NULLABLE	The number of unique 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.
release_date	DATE	REQUIRED	Last day of the release month. Table is partitioned on this column.

4.3 JSTOR

JSTOR is a digital library, offering over 7000 open access eBooks.⁴⁶ This includes titles from the University of Michigan Press, UCL Press and ANU Press.⁴⁷ Publisher usage reports offer details about the use (views and downloads) of eBooks by institution, and country.

4.3.1 Data source properties

JSTOR – data source properties	
Provider webpage	https://support.publishers.jstor.org/hc/en-us/articles/360043921594-Books-at-JSTOR-Publisher-Reports
Link to online documentation	https://oaebu-workflows.readthedocs.io/en/latest/oaebu_workflows/telescopes/jstor.html
Link to telescope code	https://github.com/The-Academic-Observatory/oaebu-workflows/blob/develop/oaebu_workflows/workflows/jstor_telescope.py
Harvest type	API
Harvest frequency	Monthly
Update frequency	Daily
Average runtime	5 min
Average download size	5 MB
Runs on remote worker	False
Catchup missed runs	True
Table write disposition	Truncate
Credentials required	Yes
Uses telescope template	Snapshot
Each shard includes all data	No
Corresponding BigQuery tables	jstor.jstor_countryYYYYMMDD jstor.jstor_institutionYYYYMMDD oaebu_intermediate.jstor_country_matchedYYYYMMDD oaebu_intermediate.jstor_institution_matchedYYYYMMDD
Notes on data	For the pilot project dashboard partner dashboards, usage data is from January 2018 onwards for JSTOR

⁴⁶ <https://about.jstor.org/librarians/books/open-access-books-jstor/>

⁴⁷ <https://about.jstor.org/librarians/books/participating-book-publishers/>

4.3.2 Data imported from this source to the BiqQuery tables

4.3.2.1 jstor_country

Field 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 requests made from that specific country.
release_date	DATE	REQUIRED	Last day of the release month. Table is partitioned on this column.

4.3.2.2 jstor_institution

Field 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 requests made from that specific country.
release_date	DATE	REQUIRED	Last day of the release month. Table is partitioned on this column.

4.4 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.⁴⁸ The pilot project uses data from the Google Play sales transaction report and the Google Books Traffic Report.

4.4.1 Data source properties

Google Books – data source properties	
Provider webpage	https://play.google.com/books/publish
Link to online documentation	https://oaeu-workflows.readthedocs.io/en/latest/oaeu_workflows/telescopes/google_books.html
Link to telescope code	https://github.com/The-Academic-Observatory/oaeu-workflows/blob/develop/oaeu_workflows/workflows/google_books_telescope.py
Harvest type	SFTP
Harvest frequency	Monthly
Update frequency	Daily
Average runtime	5 min
Average download size	1-100MB
Runs on remote worker	False
Catchup missed runs	True
Table write disposition	Truncate
Credentials required	Yes
Uses telescope template	Snapshot
Each shard includes all data	No
Corresponding BigQuery tables	google.google_books_salesYYYYMMDD google.google_books_trafficYYYYMMDD oaeu_intermediate.google_books_sales_matchedYYYYMMDD oaeu_intermediate.google_books_traffic_matchedYYYYMMDD
Notes on data	For the pilot project dashboard partner dashboards, usage data is from January 2018 onwards for Google Books

⁴⁸ <https://play.google.com/books/publish/>

4.4.2 Data imported from this source to the BigQuery tables

4.4.2.1 google_books_sales

Field 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 of 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.
List_Price_Currency	STRING	NULLABLE	The currency of the book's 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 price 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.

4.4.2.2 google_books_traffic

Field 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.

4.5 Google Analytics

Google Analytics monitors and records web traffic for specific websites. If a pilot project dashboard partner has 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.

4.5.1 Data source properties

Google Analytics – data source properties	
Provider webpage	https://analytics.google.com/analytics/web/provision/#/provision
Link to online documentation	https://oaebu-workflows.readthedocs.io/en/latest/oaebu_workflows/telescopes/google_analytics.html
Link to telescope code	https://github.com/The-Academic-Observatory/oaebu-workflows/blob/develop/oaebu_workflows/workflows/google_analytics_telescope.py
Harvest type	API
Harvest frequency	Monthly
Update frequency	Daily
Average runtime	5 min
Average download size	1 MB
Runs on remote worker	False
Catchup missed runs	True
Table write Disposition	Truncate
Credentials required	Yes
Uses telescope template	Snapshot
Each shard includes all data	No
Corresponding BigQuery tables	google.google_analyticsYYYYMMDD oaebu_intermediate.google_analytics_matchedYYYYMMDD
Notes on data	The Google Analytics retention period is a maximum of 26 months. ⁴⁹

⁴⁹ <https://support.google.com/analytics/answer/7667196?hl=en>

4.5.2 Data imported from this source to the BigQuery table google_analytics

Field name	Type	Mode	Description
url	STRING	REQUIRED	Base URL of the book pages.
title	STRING	REQUIRED	Title of the book.
publication_id	STRING	REQUIRED	Custom dimension Publication ID.
publication_type	STRING	REQUIRED	Custom dimension Publication type.
publication_imprint	STRING	REQUIRED	Custom dimension Publication imprint.
publication_group	STRING	REQUIRED	Custom dimension Publication group.
publication_whole_or_part	STRING	REQUIRED	Custom dimension Publication whole/part.
publication_format	STRING	REQUIRED	Custom dimension Publication format.
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.
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.

4.6 UCL Discovery

University College London (UCL) is an eBook publisher, and dashboard partner in the pilot project. UCL Discovery is UCL's open access repository, showcasing and providing access to the full texts of UCL research publications. This data source is ingested and aggregated into the final tables for the pilot project. While this is a pilot project dashboard partner data source specific to UCL, this data source is also publicly available via the provider web pages listed in the table below.

4.6.1 Data source properties

UCL Discovery – data source properties	
Provider web pages	https://discovery.ucl.ac.uk/cgi/search/advanced https://discovery.ucl.ac.uk/cgi/stats/report
Link to online documentation	https://oaebu-workflows.readthedocs.io/en/latest/oaebu_workflows/telescopes/ucl_discovery.html
Link to telescope code	https://github.com/The-Academic-Observatory/oaebu-workflows/blob/develop/oaebu_workflows/workflows/ucl_discovery_telescope.py
Harvest type	CSV and API
Harvest frequency	Monthly
Update frequency	Daily
Average runtime	10 min
Average download size	1.5 MB
Runs on remote worker	False
Catchup missed runs	True
Table write disposition	Truncate
Credentials required	No
Uses telescope template	Snapshot
Each shard includes all data	No
Corresponding BigQuery tables	ucl.ucl_discoveryYYYYMMDD oaebu_intermediate.ucl_discovery_matchedYYYYMMDD
Notes on data	First a CSV file is downloaded with a list of ids, then for each id an API is used to get metadata on downloads per country for that id. For pilot project dashboard partner dashboards, usage data is from January 2018 onwards for UCL Discovery.

4.6.2 Data imported from this source to the BigQuery table ucl_discovery

Field name	Type	Mode	Description
eprintid	STRING	REQUIRED	Eprint id.
book_title	STRING	REQUIRED	Title of the book.
creators_name_family	STRING	REPEATED	Family name of the creators
creators_name_given	STRING	REPEATED	Given name of the creators
ispublished	STRING	NULLABLE	Info on whether the book is published
subjects	STRING	REPEATED	Subjects
divisions	STRING	REPEATED	Divisions
keywords	STRING	REPEATED	Keywords
abstract	STRING	NULLABLE	Abstract
date	STRING	NULLABLE	Date
publisher	STRING	NULLABLE	Publisher
official_url	STRING	NULLABLE	Official URL
oa_status	STRING	NULLABLE	OA status
language	STRING	NULLABLE	Language
doi	STRING	NULLABLE	DOI
isbn	STRING	NULLABLE	ISBN
language_elements	STRING	NULLABLE	Language elements
series	STRING	NULLABLE	Series
pagerange	STRING	NULLABLE	Page range
pages	INTEGER	NULLABLE	Pages
editors_name_family	STRING	REPEATED	Family name of the editors
editors_name_given	STRING	REPEATED	Given name of the editors
lyricists_name_family	STRING	REPEATED	Family name of the lyricists
lyricists_name_given	STRING	REPEATED	Given name of the lyricists
begin_date	DATE	REQUIRED	Begin date.
end_date	DATE	REQUIRED	End date.
total_downloads	INTEGER	REQUIRED	Number of downloads
downloads_per_country	RECORD	REPEATED	Number of downloads per country
downloads_per_country.download_count	INTEGER	NULLABLE	Number of downloads for the given country
downloads_per_country.country_name	STRING	NULLABLE	Country name
downloads_per_country.country_code	STRING	NULLABLE	Country code
release_date	DATE	REQUIRED	Last day of the release month. Table is partitioned on this column.

4.7 Fulcrum

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.”⁵⁰ The Fulcrum data source is specific to the University of Michigan Press in the pilot project, and is a manual data upload (.csv file) so does not include ID-matching and linking to other data sources in the project.

The University of Michigan Press Ebook Collection can be accessed at: <https://www.fulcrum.org/michigan>

4.7.1 Data imported from this source into the BigQuery table fulcrum

Field name	Type	Mode	Description
Item	String	Nullable	Record ID (can be filename, title, etc).
Publisher	String	Nullable	Publisher name.
Platform	String	Nullable	Name of the platform (i.e. 'Fulcrum/University of Michigan Press')
Authors	String	Nullable	Authors' name.
Publication_Date	Integer	Nullable	Publication year.
DOI	String	Nullable	DOI of the work in URL format (https://hdl.handle.net/ or https://doi.org/).
Proprietary_ID	String	Nullable	ID of the work.
ISBN	String	Nullable	ISBN for each format of the work (hardcover, paper, ebook, open access, etc.) separated by ",".
URI	String	Nullable	URL of the work.
Parent_Title	String	Nullable	Title of the parent work.
Parent_Data_Type	String	Nullable	Type of the parent work.
Parent_DOI	String	Nullable	DOI of the parent work.
Parent_Proprietary_ID	String	Nullable	ID of the parent work.
Parent_ISBN	String	Nullable	Parent work's ISBN for each format (hardcover, paper, ebook, open access, etc.) separated by ",".
Parent_URL	String	Nullable	URL of the parent work.
Data_Type	String	Nullable	Can be 'Book' for book content, 'Multimedia' for enriching files such as video and images, and 'Other' for other types of files including interview, assignments.
Access_Type	String	Nullable	Is always 'OA_Gold' for Open Access. 'Controlled' is used when the time of access the content was not open, it was behind a paywall.
Access_Method	String	Nullable	A COUNTER attribute indicating whether the usage related to investigations and requests was generated by a human user browsing and searching a website (Regular) or by a computer (Machine).
Metric_Type	String	Nullable	Total Item Investigations.

⁵⁰ <https://www.press.umich.edu/librarians>

4.8 MUSE

Project MUSE provides a platform which hosts journals and books from multiple publishers including the University of Michigan Press, University College London and Wits University Press. Some of the MUSE offerings are open access eBooks⁵¹. The MUSE data source is specific to the University of Michigan Press in the pilot project, and is a manual data upload (.csv file) so does not include ID-matching and linking to other data sources in the project.

To look for open access book titles on Project MUSE, go to https://muse.jhu.edu/search?action=oa_browse, and select 'Content Type' = Books.

4.8.1 Data imported from this source into the BigQuery table muse

Field name	Type	Mode	Description
YEAR	Integer	Nullable	Access year.
MONTH	Integer	Nullable	Access month.
RESOURCE_TYPE	String	Nullable	'book'
RESOURCE_ID	Integer	Nullable	ID of the book given by the platform.
ISSN_ISBN	Float	Nullable	ISBN of the book.
RESOURCE	String	Nullable	Title of the book.
RESOURCE_URL	String	Nullable	URL of the book.
RESOURCE_LAUNCH	Date	Nullable	Resource upload date.
AUTHOR	String	Nullable	Author of the book.
FULLTEXT_TITLE	String	Nullable	Name of the book section.
FULLTEXT_URL	String	Nullable	URL of the book section.
FULLTEXT_LAUNCH	Date	Nullable	Full text upload date.
FORMAT	String	Nullable	Format of the book section (pdf, html)
ACCESS	String	Nullable	Access type (open_access/gated)
COUNTRY	String	Nullable	Accessing country.
INSTITUTION	String	Nullable	Accessing institution.
REQUESTS	Integer	Nullable	Number of requests.

⁵¹ <https://about.muse.jhu.edu/muse/open-access-overview/>

4.9 EBSCO

EBSCO hosts collections of different publications including eBooks, with some open access. They provide data about the usage of these ebooks to publishers, such as University of Michigan Press.⁵² This is a manual data upload (.csv file) so does not include ID-matching and linking to other data sources in the project, and is specific to the University of Michigan Press in the pilot project.

4.9.1 Data imported from this source into the BigQuery table ebsco

Field name	Type	Mode	Description
Month_of_Log_Month	String	Nullable	Date of retrieval (MMM-YY)
Contract_Publisher	String	Nullable	Contract publisher name.
Imprint_Publisher	String	Nullable	Imprint Publisher name (for University of Michigan Press it is the same as Contract_Publisher).
Title	String	Nullable	Title of the book.
Customer_Name	String	Nullable	Name of the customer (usually institution name).
CustId	String	Nullable	Customer ID.
Market	String	Nullable	Type of market according to customer (can be academic, corporate, schools, military, etc.).
Cust_Postal_Code	String	Nullable	Postal code of the customer.
Cust_State_Prov	String	Nullable	State of the customer (country name for customers outside of the US).
Country_Name	String	Nullable	Country of the customer.
ISBN	Integer	Nullable	ISBN of the book (no dashes).
EISBN	Integer	Nullable	EISBN of the book (no dashes).
Subjects	String	Nullable	Subject of the book.
Retrieval_Count	Integer	Nullable	Number of retrieval.

⁵² <https://more.ebsco.com/ebooks-open-access-2021.html>

4.10 SpringerLink

SpringerLink is an online collection of Springer Nature's electronic and printed journals and books.⁵³ This is a manual data upload (via .csv on a Google Cloud Platform bucket) so does not include ID-matching and linking to other data sources in the project, and is specific to Springer Nature.

4.10.1 Data imported from this source into the BigQuery table springerlink

Field name	Type	Mode	Description (supplied by the COKI technical team)
cal_year_month_no	String	Nullable	Year and month of usage, e.g. 202107
platform	String	Nullable	Source platform of data, i.e 'SPL'
bk_ebook_isbn	String	Nullable	Book ISBN
bk_title	String	Nullable	Book title
bk_ed_yr	Integer	Nullable	Publication year
imprint	String	Nullable	Imprint of book
bk_subj	String	Nullable	Book subject
item_doi	String	Nullable	Item DOI
item_doi_title	String	Nullable	Item title
sn_total_chapter_req	Integer	Nullable	Item Total chapter requests
unique_title_requests	Integer	Nullable	Book unique title requests

5 Book Usage Data Analytic Workflows

⁵³ <https://www.springer.com/gp/help/about-springerlink/18548>

The book usage data analytic workflows contains three steps that:⁵⁴

1. Aggregates and maps book products (i.e. unique ISBN-13) from the ONIX table into works and work families
2. Links data (ingested via telescopes) from metric providers to book products from the ONIX feed
3. Export of these linked metrics to Elasticsearch for viewing in Kibana Dashboards

5.1 Book Usage Data Analytic Workflow Step 1

The ONIX workflow uses the ONIX table created by the pilot project ONIX telescope to aggregate book products (designated by a unique International Standard Book Number (ISBN-13)) into works records and work family records. A book product is a manifestation of a work, and has its own ISBN-13. Works records are different manifestations of the same product, such as a PDF and a html of the same work. A work family designates different editions of the same work. The following intermediate BigQuery tables are produced that map a product identifier (ISBN-13) to a WorkID and WorkFamilyID.

5.1.1 Onix workflow tables

5.1.1.1 BigQuery table onix_workflow.onix_workid_isbn

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

Field name	Type	Mode	Description
isbn13	STRING	NULLABLE	ISBN13
work_id	STRING	NULLABLE	The WorkID. Likely to be an ISBN.

5.1.1.2 BigQuery table onix_workflow.onix_workfamilyid_isbn

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

Field name	Type	Mode	Description
isbn13	STRING	NULLABLE	ISBN13
work_family_id	STRING	NULLABLE	The Work Family ID. Likely to be an ISBN.

5.1.1.3 BigQuery table onix_workflow.onix_workid_isbnerrors

Details when product ISBN-13s have a related product that is not included as a separate product identifier in the ONIX feed.

Field name	Type	Mode	Description
Error	STRING	NULLABLE	Error string

5.2 Intermediate BigQuery Tables

For each publisher partner, the tables created for each data source (e.g. [OAPEN](#) IRUS-UK, [JSTOR](#), [Google Books](#), [Google Analytics](#), [UCL Discovery](#)) are 'matched' with the ISBN-13, and the new 'work_id' and 'work_family_id' fields are linked. The schemas for these tables are identical to the raw Telescope schemas (see the 'Data imported from this source' tables for each data source), with the addition of work_ids and work_family_ids, and are saved as BigQuery tables with '_matched' appended (e.g. 'oaeu_intermediate.google_books_sales_matched').

⁵⁴ https://oaeu-workflows.readthedocs.io/en/latest/oaeu_workflows/workflows/onix_workflow_intro.html

5.3 Data Quality BigQuery Tables

For each data source, including the intermediate tables, basic quality assurance checks are performed on the data. The data quality tables are designed to be easily exported to csv for further analysis by the pilot project dashboard partners if desired. For example, the provided ISBN-13's in the ONIX table are verified to ensure they are valid, and if there are unmatched ISBN-13's from public data sources (indicating that there are missing ONIX product records).

5.3.1 BigQuery table oaebu_data_qa.onix_aggregate_metrics

Field name	Type	Mode	Description
table_size	INTEGER	NULLABLE	Total Number of Book Products
no_isbns	INTEGER	NULLABLE	Count of how many rows are missing an ISBN
no_relatedworks	INTEGER	NULLABLE	Count of how many rows are a related work
no_relatedproducts	INTEGER	NULLABLE	Count of how many rows are a related product
no_doi	INTEGER	NULLABLE	Count of how many rows are missing a DOI
no_productform	INTEGER	NULLABLE	Count of how many rows are missing a product form
no_contributors	INTEGER	NULLABLE	Count of how many rows are missing contributors
no_titledetails	INTEGER	NULLABLE	Count of how many rows are missing title details
no_publisher_urls	INTEGER	NULLABLE	Count of how many rows are missing a publisher url

5.3.2 BigQuery table oaebu_data_qa.onix_invalid_isbn

Details ISBN-13s in the partners ONIX feed that are not valid.

Field name	Type	Mode	Description
Primary_ISBN	STRING	NULLABLE	ISBN-13

5.3.3 BigQuery table oaebu_data_qa.<platform>_invalid_isbn

Details ISBN-13s in the data source that are not valid.

Field name	Type	Mode	Description
ISBN	STRING	NULLABLE	ISBN-13

5.3.4 BigQuery table oaebu_data_qa.<platform>_unmatched_isbn

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

Field name	Type	Mode	Description
ISBN	STRING	NULLABLE	ISBN-13
title	STRING	NULLABLE	Book title from the data source
release_date	DATE	NULLABLE	The release date (month)

5.4 Book Usage Data Analytic Workflow Step 2

Step 2 of the ONIX workflow takes the metrics fetched through various telescopes, then aggregates and joins them to the book records in the pilot project dashboard partners ONIX feed. This produces the BigQuery book_product table, containing one row per unique book product, with a nested month field, which groups all the metrics relating to that book for each calendar month.

5.4.1 BigQuery table Book table

Field 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
work_id	STRING	NULLABLE	The derived Work_ID that we calculate
work_family_id	STRING	NULLABLE	The derived 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.oapen_irus_uk_metadata	RECORD	NULLABLE	Metadata derived from IRUS_UK
metadata.oapen_irus_uk_metadata.ISBN13	STRING	NULLABLE	ISBN of the book
metadata.oapen_irus_uk_metadata.book_title	STRING	NULLABLE	Title of the book
metadata.oapen_irus_uk_metadata.publisher	STRING	NULLABLE	The publisher
months	RECORD	REPEATED	Linked Metrics from all sources, organised by month of occurrence
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.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.
months.google_analytics.unique_views.average_time	FLOAT	NULLABLE	Average time (in seconds) spent on each page
months.google_analytics.unique_views.country	RECORD	REPEATED	Unique views per users' country, derived from their IP addresses or Geographical IDs.
months.google_analytics.unique_views.country.name	STRING	NULLABLE	Country name
months.google_analytics.unique_views.country.value	INTEGER	NULLABLE	Number of sessions

months.google_analytics.unique_views.referrer	RECORD	REPEATED	Unique views per referrer, the full referring URL including the hostname and path
months.google_analytics.unique_views.referrer.name	STRING	NULLABLE	Referrer name
months.google_analytics.unique_views.referrer.value	INTEGER	NULLABLE	Number of unique views
months.google_analytics.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.
months.google_analytics.unique_views.social_network.name	STRING	NULLABLE	Social network name
months.google_analytics.unique_views.social_network.value	INTEGER	NULLABLE	Number of unique views
months.google_analytics.downloads	INTEGER	NULLABLE	Number of total downloads summed from downloads_pdf_book, downloads_pdf_chapter, downloads_html_chapter, downloads_epub_book, downloads_epub_chapter, downloads_mobi_chapter
months.google_analytics.downloads_pdf_book	INTEGER	NULLABLE	Number of PDF book downloads
months.google_analytics.downloads_pdf_chapter	INTEGER	NULLABLE	Number of PDF chapter downloads
months.google_analytics.downloads_html_chapter	INTEGER	NULLABLE	Number of HTML chapter downloads
months.google_analytics.downloads_epub_book	INTEGER	NULLABLE	Number of ePUB book downloads
months.google_analytics.downloads_epub_chapter	INTEGER	NULLABLE	Number of ePUB chapter downloads
months.google_analytics.downloads_mobi_chapter	INTEGER	NULLABLE	Number of MOBI chapter downloads
months.google_books_sales	RECORD	NULLABLE	Metrics derived from 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_books_traffic. BV_with_Buy_Clicks	INTEGER	NULLABLE	The number of visits 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_It em_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.Instituti on	STRING	NULLABLE	Institution name
months.jstor_institution.Total_It em_Requests	INTEGER	NULLABLE	Total number of request made from that specific institution
months.oapen_irus_uk	RECORD	NULLABLE	Metrics derived from IRUS-UK
months.oapen_irus_uk.version	STRING	NULLABLE	Version of the OAPEN IRUS UK API, corresponds to the COUNTER report version
months.oapen_irus_uk.title_re quests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
months.oapen_irus_uk.total_it em_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01
months.oapen_irus_uk.total_it em_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01
months.oapen_irus_uk.unique _item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01
months.oapen_irus_uk.unique _item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01
months.oapen_irus_uk.country	RECORD	REPEATED	Record to store statistics on the country level
months.oapen_irus_uk.country .name	STRING	NULLABLE	The country name of the client registered by oapen irus uk
months.oapen_irus_uk.country .code	STRING	NULLABLE	The country code of the client registered by oapen irus uk

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

5.5 Book Usage Data Analytic Workflow Step 3

The third step of the ONIX workflow is to export the book_product table to a sequence of flattened data export tables that can be exported to Elasticsearch. The data in these tables is not materially different to the book product table, just organised in a way that is better suited for creating dashboards in Kibana.

5.5.1 Data Export BigQuery Tables

5.5.1.1 BigQuery Table oaebu-<publisher>-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.

Field 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
time_field	DATE	NULLABLE	Required for Elasticsearch, generally not used though
published_year	INTEGER	NULLABLE	The published year of the book
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

5.5.1.2 BigQuery Table oaebu_<publisher>_book_product_metrics

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

Field 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 Name
authors.ORCID	STRING	NULLABLE	The Author's ORCID ID
published_year	INTEGER	NULLABLE	The Books published year
month	DATE	NULLABLE	The month in which the metrics took place
google_analytics	RECORD	NULLABLE	Metrics from Google Analytics
google_analytics.unique_views	INTEGER	NULLABLE	The number of unique 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_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 occurred
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
oapen_irus_uk	RECORD	NULLABLE	Metrics from IRUS-UK
oapen_irus_uk.version	STRING	NULLABLE	Version of the OAPEN IRUS UK API, corresponds to the COUNTER report version
oapen_irus_uk.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
oapen_irus_uk.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01
oapen_irus_uk.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01
oapen_irus_uk.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01
oapen_irus_uk.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01
ucl_discovery	RECORD	NULLABLE	Metrics from UCL Discovery
ucl_discovery.total_downloads	INTEGER	NULLABLE	

5.5.1.3 BigQuery Table oaebu_<publisher>_product_author_metrics

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

Field name	Type	Mode	Description
PersonName	STRING	NULLABLE	Author's 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.unique_views	INTEGER	NULLABLE	Number of unique 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_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
oapen_irus_uk	RECORD	NULLABLE	Metrics from IRUS-UK
oapen_irus_uk.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
oapen_irus_uk.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01
oapen_irus_uk.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01
oapen_irus_uk.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01
oapen_irus_uk.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01

5.5.1.4 BigQuery Table oaebu_<publisher>_book_product_metrics_country

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

Field 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
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
jstor	RECORD	NULLABLE	Metrics from JSTOR
jstor.Total_Item_Requests	INTEGER	NULLABLE	Total number of request made
oapen_irus_uk	RECORD	NULLABLE	Metrics from IRUS-UK
oapen_irus_uk.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
oapen_irus_uk.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01

oapen_irus_uk.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01
oapen_irus_uk.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01
oapen_irus_uk.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.unique_views	INTEGER	NULLABLE	Number of unique 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_chapter	INTEGER	NULLABLE	Number of MOBI chapter downloads
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

5.5.1.5 BigQuery Table oaebu-public-data-country-list

This table is a list of each unique Country or Territory. It is primarily used for drop-down fields.

Field name	Type	Mode	Description
country_code	STRING	NULLABLE	ISO 3166 country code (alpha 2)
country_name	STRING	NULLABLE	ISO 3166 country name

5.5.1.6 BigQuery Table oaebu_<publisher>_book_product_metrics_city

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

Field 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
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
oapen_irus_uk	RECORD	NULLABLE	Metrics from IRUS-UK
oapen_irus_uk.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
oapen_irus_uk.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01
oapen_irus_uk.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01
oapen_irus_uk.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01
oapen_irus_uk.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01

5.5.1.7 BigQuery Table oaebu_<publisher>_book_product_metrics_institution

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

Field 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

5.5.1.8 BigQuery Table oaebu_<publisher>_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.

Field name	Type	Mode	Description
institution	STRING	NULLABLE	Institution Name

5.5.1.9 BigQuery Table oaebu_<publisher>_book_product_metrics_events

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

Field 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
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

5.5.1.10 BigQuery Table oaebu_<publisher>_book_product_metrics_publisher

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

Field name	Type	Mode	Description
month	DATE	NULLABLE	Month for which these metrics apply
unique_books	INTEGER	NULLABLE	The number of unique books
google_analytics	RECORD	NULLABLE	Metrics from Google Analytics
google_analytics.unique_views	INTEGER	NULLABLE	The number of unique views
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.No_n_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 clickthrough rate for purchase links. The values are percentages
google_books_sales	RECORD	NULLABLE	Metrics from Google Books Sales
google_books_sales.qty	INTEGER	NULLABLE	Quantity of sales
jstor	RECORD	NULLABLE	Metrics from JSTOR
jstor.Total_Item_Requests	INTEGER	NULLABLE	Total number of item requests
oapen_irus_uk	RECORD	NULLABLE	Metrics from IRUS-UK
oapen_irus_uk.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
oapen_irus_uk.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01
oapen_irus_uk.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01
oapen_irus_uk.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01
oapen_irus_uk.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01

5.5.1.1.1 BigQuery Table

oaeu_<publisher>_book_product_subject_year_metrics

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

Field name	Type	Mode	Description
subject	STRING	NULLABLE	BIC subject (top level BIC subjects only)
published_year	INTEGER	NULLABLE	The published year of the book
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.unique_views	INTEGER	NULLABLE	The number of unique views
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	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
oapen_irus_uk	RECORD	NULLABLE	Metrics from OAPEN IRUS-UK
oapen_irus_uk.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
oapen_irus_uk.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01
oapen_irus_uk.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01
oapen_irus_uk.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01

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

5.5.1.12 BigQuery Tables oaebu_<publisher>_book_product_subject_bic_metrics, oaebu_<publisher>_book_product_subject_bisac_metrics, oaebu_<publisher>_book_product_subject_thema_metrics

These tables contain metrics, organised by month and BIC, BISAC or THEMA subject type, that are linked to each book.

Field name	Type	Mode	Description
subject	STRING	NULLABLE	BIC Subject / BISAC Subject / THEMA subject
subject_code	STRING	NULLABLE	BIC Subject Code / BISAC Subject Code / 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.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_chapter	INTEGER	NULLABLE	Number of MOBI chapter downloads
google_analytics.unique_views	INTEGER	NULLABLE	The number of unique views
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.No_n_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	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
oapen_irus_uk	RECORD	NULLABLE	Metrics from IRUS-UK
oapen_irus_uk.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
oapen_irus_uk.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01
oapen_irus_uk.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01
oapen_irus_uk.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01
oapen_irus_uk.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01

5.5.1.13 BigQuery Table oaebu_<publisher>_book_product_year_metrics

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

Field name	Type	Mode	Description
published_year	INTEGER	NULLABLE	Published Year
unique_books	INTEGER	NULLABLE	The number of unique books published that year in the dataset
month	DATE	NULLABLE	The month for which the metrics apply to
google_analytics	RECORD	NULLABLE	Metrics from Google Analytics
google_analytics.unique_views	INTEGER	NULLABLE	Unique Views
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	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
oapen_irus_uk	RECORD	NULLABLE	Metrics from IRUS-UK
oapen_irus_uk.title_requests	INTEGER	NULLABLE	The total number of title requests. Only available for data before 2020-04-01
oapen_irus_uk.total_item_investigations	INTEGER	NULLABLE	The total number of item investigations. Only available for data since 2020-04-01
oapen_irus_uk.total_item_requests	INTEGER	NULLABLE	The total number of item requests. Only available for data since 2020-04-01
oapen_irus_uk.unique_item_investigations	INTEGER	NULLABLE	The number of unique item investigations. Only available for data since 2020-04-01
oapen_irus_uk.unique_item_requests	INTEGER	NULLABLE	The number of unique item requests. Only available for data since 2020-04-01

5.5.1.14 BigQuery Table oaebu_<publisher>_unmatched_book_metrics

This table 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.

Field name	Type	Mode	Description
ISBN	STRING	NULLABLE	The ISBN13
release_date	DATE	NULLABLE	The release date (month)
google_analytics_title	STRING	NULLABLE	The title of book, as specified in Google Analytics
google_books_title	STRING	NULLABLE	The title of book, as specified in Google Books
jstor_title	STRING	NULLABLE	The title of book, as specified in JSTOR
oapen_title	STRING	NULLABLE	The title of book, as specified in OAPEN/IRUS-UK
ucl_discovery_title	STRING	NULLABLE	The title of book, as specified in UCL Discovery
in_google_analytics	BOOLEAN	NULLABLE	Was this ISBN contained in Google Analytics
in_google_books	BOOLEAN	NULLABLE	Was this ISBN contained in Google Books
in_jstor	BOOLEAN	NULLABLE	Was this ISBN contained in JSTOR
in_oapen	BOOLEAN	NULLABLE	Was this ISBN contained in OAPEN
in_ucl_discovery	BOOLEAN	NULLABLE	Was this ISBN contained in UCL Discovery

5.6 Elasticsearch/Kibana indexes

Once the data export tables have been created, they are exported to Elasticsearch for visualisation in Kibana dashboards.

The pilot project Elasticsearch/Kibana indexes contain additional fields specific to Elasticsearch, such as time fields and keywords. The pilot project Elasticsearch/Kibana indexes as of 2022-03-31 are detailed here: [2022-03-31 pilot project indexes](#)