OAeBU Data Trust Pilot Response to June 2021 COUNTER Consultation on OA Usage

***About this document**: This project response to the June 2021 Reporting Global Usage and Usage of Open Content Not Attributed to Institutions COUNTER Consultation notes the survey prompts and the OA eBook Usage Data Trust Pilot Project response submitted by project Co-PI Cameron Neylon on 14 June 2021. Responses were prepared in consultation with the project's technical development team and the project's open Technical Standards and Norms Working Group.*

COUNTER Consultation SECTION 1

Q1 Content Type: what kind of content do you provide?

- Journal
- Book
- Journal and Book
- Other – please specify

A1: Book - the OAeBU Data Trust effort is focussed on supporting the reporting and analytics of book usage data and related linked data sources

Q2 What is your business model? [multiselect]

- Fully Open Access
- Hybrid
- Other

A2: Other - While our project is focussed on usage of open access books, members of our community represent both open access only and mixed model publishers.

Q3 To report global usage, aggregating all institutional and non-institutional usage, a publisher or vendor can create a report with "The World" as Institution_Name in the report header. This would be useful:

- Strongly agree | Agree | Neither agree nor disagree | Disagree | Strongly disagree

A3: Strongly agree

Q4 COUNTER should define a customer ID for "The World" for requesting the report via SUSHI:

- Yes | No

A4: Yes

Q5 Global usage can be broken down by geolocation using ISO 3166-1 (country names and codes) and ISO 3166-2 (country subdivision names and codes). A usage report by country and country subdivision would be useful:

- Strongly agree | Agree | Neither agree nor disagree | Disagree | Strongly disagree

A5: Strongly agree

Notes: It will be important to provide some information on how the geolocation was determined. Geolocation is not straightforward and can be very variable. Recommending best practice and supporting the community development of consistent approaches for IP address to ISO3166-1/2 mapping will be valuable. There are a range of systems and services that offer IP-to-geolocation conversion using a range of datasets that map IP ranges to locations, but public and open datasets are often outdated, leading to systematic errors in geolocation. Therefore, transparent provenance of the geolocation data processing and of the origin and ideally version of mapping datasets is critical. Geolocation to country level is more reliable than for subdivisions of countries. Care should be taken to articulate the reliability and stability of these mappings over time and to note limitations. Common data-privacy measures such as adding noise and flooring the counts (omitting or merging segments with small numbers of counts) should be explicitly encouraged. It should also be noted that using geolocation services to convert IP to location has its own privacy implications and will require its own assessment by report providers. The 2016 Joint Research Centre Guidelines for Location privacy (https://joinup.ec.europa.eu/sites/default/files/news/attachment/jrc103110_1-dc246-d3.2_eulf_guideline_on_location_privacy_v1.00_final_-_pubsy.pdf) and Future of Privacy Forum Policy Brief (https://fpf.org/wp-content/uploads/2020/12/FPF_Guide_Location_Data_v2.2.pdf) are useful resources.

Q6 COUNTER should define a value "Unknown" for usage that cannot be attributed to a country or country subdivision:

- Yes | No

A6: Yes
Notes: Given that IP to country mapping will be limited it is important to provide uncategorised usage to allow proportions to be accurate determined.

Q7 Institution_Name and Customer_ID extensions can be used to break usage down by institution, with "All Other Usage" for usage not attributed to institutions. COUNTER reports broken down by institution would be valuable:

- Strongly agree | Agree | Neither agree nor disagree | Disagree | Strongly disagree

A7: Agree

Q8 As a publisher or provider, I would be able to break down COUNTER reports by institution without breaching confidentiality agreements / contracts:

- Yes | No

A8: N/A - No response

Notes: The visibility and use of institution-specific usage data It should be noted that this will raise legal and ethical concerns around privacy in data that remain unresolved in the sector in the context of large-scale data linking, aggregation, and repurposing.

Our project is not a publisher or service provider at this stage. Based on our research into the value propositions for an international data space focused on OA book usage data, we note that data controlling, processing, stewardship, and downstream terms of use are implied by this question. In our project, we have identified a need to develop standard contractual clauses to facilitate the data transfer and use of usage data across parties and hope to foster common language development via a Research Data Alliance Working Group in the coming year.

Q9 COUNTER should define a customer ID for "All Other Usage":

- Yes | No

A9: N/A - No response

Notes: We did not reach a full consensus on this question. We see value in allowing for an "other" category for some use cases, but note that for an identifier based column this can lead to problems for downstream data aggregators. We note that the utilising "fake values" within existing identifier schema has a history of contaminating downstream metadata (e.g. the use of ISSN 0000-0000 or strings that mimix DOIs). We suggest a more detailed examination of use cases and potential risks. For any implementation such an element should explicitly not be anything that would appear to mimic (or worse validate) against a formal identifier schema.

COUNTER Consultation SECTION 2
Your feedback to these questions will inform our future planning. If adopted these reports would also NOT be a mandatory requirement

Q10 Not all usage can be attributed to an institution or customer (e.g. Open Access content). We propose including an 'Attributed' element to help distinguish usage which may be attributed to an institution from all other usage. This report would be useful:

- Strongly agree | Agree | Neither agree nor disagree | Disagree | Strongly disagree

A10: No strong view. Is this redundant with other proposed data elements? If not this is a useful distinction.

Q11 COUNTER reports should indicate the mechanism used to attribute usage to institutions:

- Yes | No

A11: Yes. It will be important to identify how data were processed. This is also true of geolocation processes. The details of how to achieve this is complex and some possibilities are identified below. These mechanisms will require further community discussion.

Q12 COUNTER reports should distinguish different types of institutions (e.g. academic and corporate):

- Yes | No

A12: No

Notes: With respect to categorisation of institutions our experience has been that it is far more useful to have stable organisational identifiers which a user can subsequently categorise based on their own needs. Categories of organisations are not stable and different use cases require very different categorisations. As a general principle using standard open identifiers (with a preference in this case for ROR/GRID as identifiers) is a better approach and avoids COUNTER taking on the role of an authority for organisation categories. If categories of organisations are provided then the basis and provenance of that categorisation by the report provider must be provided.

Q13 I would want to be able to filter or restrict the content of reports for "The World" (e.g. by country):

- Strongly agree | Agree | Neither agree nor disagree | Disagree | Strongly disagree

A13: Neither agree nor disagree

Notes: From the perspective of the OAeBU project as both a data exchange and as potential provider of dashboard services, we prefer to take all available data and therefore prefer not to filter data on ingest. The downstream users for both a data exchange and dashboarding services will be best supported by being able to manipulate and filter comprehensive and granular data.

As a result we are concerned with obtaining comprehensive and comparable data and therefore have a preference for processing and providing access and security controls on top of unfiltered granular data. While there are likely to be cases where the providers of COUNTER reports cannot provide comprehensive granular data for privacy, safety, or ethical reasons, there is a significant risk of confusion if restriction of data or filtering is not transparent. For this reason it may be preferable to recommend not providing reports which raise such issues.

In our research on the uses of OA eBook usage data, participants noted multiple use cases involving the need to be able to analyse usage data by country. Please see (Forthcoming Zenodo site URL)

Q14 Please provide a list of fields on which you wish to filter "The World" reports:

A14: Our data trust infrastructure functions best when receiving granular data from data providers instead of filtered reports. The OAeBU data use cases surfaced many related data domains of interest where stakeholders would benefit from being able to flexibly analyse, integrate and (within the analysis system) filter usage data. While not an exhaustive list, included among them are:

- Institution (Grid or ROR ID)
- Country code
- Author (ORCID ID)
- Subject (BISAC)
- Work Identifiers (ISBN, DOI, URI)
- Parent Identifiers (ISBN, DOI, URI)
- Platform
- Publication date (YOP)
- Access Type
- Monthly Usage Details
- Work Type

Q17 Please provide any other comments or suggestions.

A17:

1. Usage of both open access resources and usage of books have been areas that are not a traditional focus for COUNTER, and this has meant that the intersection of these two spaces is not as well covered as for other content and access types. Adoption of COUNTER in this space is therefore patchy and dominated by large players. Within the OA book ecosystem there are many providers that have no current engagement with COUNTER standards. Improving adoption of COUNTER amongst smaller book publishers has the potential to add significant value, and work in this space is therefore important.
2. Adapting COUNTER to more readily represent usage in a primarily open access world is crucial and represents an ongoing challenge given the standard's roots in subscription management and information. One area that may require future consideration is the degree to which the reporting on institutional usage reflects subscription or content purchasing arrangements. For instance, the usage of open access content from an institution that is also a subscriber to content from the same publisher will generally be captured to that institution and "attributed" to a customer ID. However, this is not a customer relationship. Future consideration of whether and how to distinguish between subscribed (or purchased) access and other access from an institution would be valuable. This distinction is potentially critical information for publishers and content suppliers as well as for institutions, particularly as the mechanisms for financial support of content diversify. This is particularly relevant to books, where an increasing range of

financing systems are being developed (e.g., Open the Future, Subscribe to Open, Knowledge Unlatched and other collective support arrangements).

3. We strongly encourage the adoption of community-developed standard identifiers and the adoption of best practice from relevant standards organisations. Specific to this consultation we strongly recommend the adoption of GRID/ROR as an open identifier standard with openly available metadata that will allow downstream users to provide their own categorisations of user identification type. The use of ISO3166 is supported for similar reasons.
4. Geo-location data provenance and data transformation algorithmic transparency information is crucial. Geo-location from IP is a challenging technical problem for a variety of reasons including IP ranges changing ownership and control, institutions having multiple locations, use of proxy services and the resultant curation challenges. There can be a substantial differential between geolocation accuracy and currency of high-level service offerings that are generally more expensive, and cheaper but often less accurate offerings. In addition, combined data from both institutionally known customers and “the world” will likely use two different data pipelines to infer location (the known location for known institutions and IP geolocation for others). These will have differing levels of accuracy. Adoption of improved COUNTER standards may be put at risk if low-quality geolocation processes create a pool of unreliable, lower-quality data. It will be crucial to provide a means for describing how both institutional attribution and geolocation from IP for usage not attributed to institutions was determined to provide confidence in the data. Provenance description is not straightforward and we do not have a simple recommendation for how to achieve this. At a minimum providing sufficient information to identify the service, date, and ideally a clearly versioned identification of a mapping dataset would provide a starting point. An ability to point to versioned code (alongside mapping dataset) would be an aspirational target.