Data Dictionary and Methodology Notes

PKP Beacon: Measuring the Size and Location of the Global PKP Publishing Community
Saurabh Khanna, Jonas Raoni Soares da Silva, Alec Smecher, Juan Alperin, Jon Ball

Descriptive Fields

Column	Example Values	Description
application	OJS	Type of application: Open Journal Systems (OJS) is an open source software application for managing and publishing scholarly journals.
version	2.4.2.0, 3.2.1.4	Software version for OJS
country	id, br, us	ISO 3166-1 alpha-2 code for estimated country of origin of journal. See “Location” for more details.
country_marc	Macedonia	The MARC country as extracted from the LOC ISSN API
country_issn	US	The MARC country converted to ISO3166
country_tld	FR	The country extracted from the host TLD (top level domain) as ISO3166
country_ip	BR	The country extracted through geolocation as ISO3166
country_consolidated	CA	The resolved/final country which will be assigned to the journal (the first non-empty value from the fields country_issn, country_tld and country_ip)

Data Fields

Column	Example Values	Description
issn	2168-9660	An eight-digit International Standard Serial Number (ISSN) used to uniquely identify a serial publication. If the journal has multiple ISSNs, they will be separated by a line break.

oai_url	https://journals.sfu.ca/iccps/index.php/index/oai	A base URL for either study metadata or related citation metadata
journal_url	https://journals.sfu.ca/iccps/index.php/index/	URL for the journal website
domain	journals.sfu.ca	Domain name extracted from the OAI URL
admin_email	email@example.com	Contact information of the installation administrator
earliest_datestamp	2012-02-29 03:06:14	The timestamp of the earliest published submission
repository_name	Open Journal Systems	Name of repository hosting the journal
set_spec	publicknowledge:ART; publicknowledge:REV	A set is an optional construct for grouping items for the purpose of selective harvesting. The set_spec is as defined in the OAI specification. In OJS's implementation, the set spec is used to permit access to each journal's contents (in the case of a multi-journal site), and within each journal, the contents of each section (e.g. ART for articles and REV for reviews, with details to be set by each journal individually).
context_name	International Critical Childhood Policy Studies Journal	Full name of the journal
stats_id	594dcecd31704	Unique identifier of the OJS installation (an installation might host N journals)
total_record_count	26	Total articles published by the journal across all years
record_count_XXXX	12	Total articles published by the journal in year XXXX
last_completed_update	2021-09-22 04:14:34	Last time the journal was synchronized and had its records counted
first_beacon	2021-02-12 14:55:11	The first time the OJS installation contacted the Beacon
last_beacon	2021-06-11 10:56:51	The last time the OJS installation contacted the Beacon
last_oai_respo	2021-09-22 04:14:33	Last time the installation received a successful

nse		response from the OAI endpoint for the “Identify” verb
unresponsive_endpoint	0	Defines if the OJS installation as a whole has been unresponsive during the synchronization process for at least 30 days
unresponsive_context	0	Defines if the journal has been unresponsive during the synchronization process for at least 30 days

Methodology Notes

This section describes the approach used to identify, inspect, deduplicate, and locate uses of PKP’s scholarly publishing software. The packages included in this dataset support a feature built into the software for this purpose; these include:

- Open Journal Systems (OJS), version 2.4.7 (October 2015) or newer
- Open Monograph Press (OMP), version 1.2.0 (April 2016) or newer
- Open Preprint Server (OPS), all releases.

Identification

Supported OJS, OMP, and OPS installations are distributed on the Internet with no registration process required. In order to collect information about these installations, PKP first needs to identify them. This is done using the “beacon” feature in the packages and versions listed above; this feature permits the installation to identify itself to the Public Knowledge Project web server during a pre-existing exchange of data between the two whereby the application checks with the PKP website to see if a new version is available. When this happens, the application provides PKP with the following data:

- OAI-PMH endpoint URL
- Disambiguating identifier (a uniqid)

These items are captured in the PKP web server’s access log as URL parameters.

(In older releases of OJS, OMP, and OPS, disabling the version check also disables the beacon data exchange. This has been corrected in newer releases so that each can be enabled/disabled without affecting the other. See #6457.)

Inspection

Fetching descriptive information from each installation via OAI-PMH

Deduplication

Each row in the data contains a unique combination of “oai_url, repository_name, and set_spec”. Duplicated combinations of “oai_url, repository_name, and set_spec” were handled by choosing the most recent observation based on “last_oai_response”.

The Beacon might receive URLs that end up pointing to the same OJS installation, just hosted under a different hostname, as well as installations which are replicas, staging environments, etc. In order to avoid counting the same journals N times, we’ve established a deduplication process during the synchronization, which basically just attempts to merge the duplicate entries.

For each journal found in an installation, we try to find a similar existing journal in our archive using the following heuristics:

- Same ISSN
- Same journal name or identifier (set spec) and at least two matching fields from the following list:
    - Stats ID
    - Hostname
    - Earliest date stamp
    - Administrator email
    - Repository name

If a duplicated entry is found, then we select one of them to be the leader. The preference is based on these checks:

- The journal which is already a “leader”
- The installation isn’t disabled
- The journal isn’t disabled or removed
- The journal isn’t unresponsive (failed to be synchronized for at least 30 consecutive days)
- The journal which was last successfully synchronized
- The journal which the installation failed to synchronize less often
- The journal that sent the latest beacon signal

The data can still contain duplicated values of single columns (like ISSN or journal name) because we were unable to detect a duplication automatically due to slightly non-conforming field values.

Location

Country of origin of a journal is determined by (in order of preference):

1. Looking at the ISSN of a journal. More specifically, we query the LOC ISSN lookup API for location, which provides us with a MARC code. This MARC code can then be mapped to a country (mapping details here). MARC codes that map to within-country regions or states are handled manually to reflect corresponding countries.

2. The top level domain mapped to ISO 3166-1 alpha-2 country code.
3. Geolocating the IP address of the journal host using a weekly updated copy of the GeoLite2 Free Geolocation Data.

This method is not perfect and is known to be skewed towards the location of hosting servers.

Filtering

This dataset *excludes* journals based on the following criteria:

- Journals for which we cannot resolve their hosts into IPv4 addresses, such as intranet hosts
- Journals which are apparently test instances (the journal's name or its URL contain the words "test", "dev", "demo", "theme")
- Journals hosted under IP addresses (e.g. http://255.255.255.255)

Only for the map:

- Journals which we were unable to reach during our synchronization process (e.g. unexpected errors, timeouts, network issues, ...) for at least 30 days
- Journals which haven't published at least 5 articles in a given year