World Data System – International Technology Office

The World Data System at the Research Data Alliance

Compiled Dataset

Dr. Alicia Urquidi Díaz

Research Associate for the International Technology Office

aurquidi@oceannetworks.ca

1 October, 2021

Executive Summary

This document describes a dataset of RDA users and their links to WDS member organizations. The data was compiled from a number of sources and summarized into this package. The sources used were: User profile pages from the RDA website, rd-alliance.org; WDS member institutions' information from worlddatasystem.org; and country income levels were obtained from the World Bank Country and Lending Groups[1] dataset. This document describes how the data was crawled and processed to be visualized as Tableau dashboards by different parameters (region, type of institution, disciplines, etc.).

[1] World Bank. (2021). *World Bank Country and Lending Groups – World Bank Data Help Desk* [Dataset]. World Bank. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

Acknowledgements

I would like to thank Seiya Terada for his work in helping obtain and in curating the first two versions of this dataset.

Creating the Dataset

This section describes the scraping, parsing and curation steps undertaken to create this dataset. The section that follows, Contents of the Dataset, describes contents of the final dataset.

Dataset Versions

Three separate crawls were carried out to obtain RDA user data. The initial crawl was conducted in October 2019 and it retrieved 8191 user profile records. Of those, were identified as belonging to or representing current WDS member organizations. A second crawl in August 2020 retrieved 10935 users, of which 432 were identified as belonging to or representing current WDS member organizations.

The current version of this dataset is based on the data crawl that was carried out in September 2021. It contains 12092 de-identified RDA user records, of which 472 were deemed WDS members. It also contains 28099 records that link RDA users (also de-identified, from WDS and otherwise) to RDA groups in which they participate.

Scraping user data from the RDA website

As said above, the 2021 (present) version of the dataset is the third and most current one. It was obtained in September 2021. The python code used for this crawl has been packaged and documented in a JupyterNotebook, and it is available here. The script uses `requests` to grab HTML from URLs, which are generated by the script following the `https://www.rd-alliance.org/user/[counter]` schema. As of August 25, 2021, the RDA boasted 12009 registered members. This does not mean that the highest user number is 12009: As of the same date, the highest user number was in the 29550s. To reuse this script, I recommend starting with a pessimistic (= very high) counter and running the script a few times until the currently highest user number is found.

For each record found, the script employs `BeautifulSoup` to parse the HTML and find the information fields (see table 1). Then, the regex package `re` is used to clean the data into tab-separated rows. The output of this process are two tab-separated text files: A table of users and user details, and a long table linking users to the RDA groups in which they are members.

Table 1: Fields in output tables from RDA user data scrape.

Table: Users	Table: Groups
user_ID	**user_ID**
Professional title	**Group**
Primary Domain/Field of Expertise (Other)	
Organization name	
Organization type	
Country	

Merging with the World Bank's Country Income data

The **Country** variable holds country names as text. We converted each name to its equivalent ISO 3166-1 alpha-3 code, in a separate column labeled **CountryISO**. This column was used to match each user's record with a corresponding country income group (*High income, Upper middle income, Lower middle income, Low income*) according to the World Bank's classification. This data was saved as the **RDACountryIncomeGroup** variable.

Merging with WDS membership data

The step of identifying RDA users from WDS organizations required intensive, manual curation with a two-pronged approach. A separate RDA crawl was used to obtain all user names, which we (manually) compared against a contact list of WDS people. If the person's name was a match, we looked at whether **Organization Name** contained a WDS member *or* the host institution associated with that person and WDS member. After this name-matching steps, all RDA member names were removed from the dataset.

The second 'prong' looked exclusively at organization names. Because of the varied ways in which organization names appeared on RDA user profiles vs how they appear in official WDS records, this process also required manual curation. We reviewed each of the 12092 records, looking up names/acronyms when we suspected they could be referring to WDS member organizations (e.g. with alternative, new or obsolete names, or acronyms we were not familiar with).

As a result of this curation step, RDA user records were enhanced with four additional columns that contained WDS information:

- **WDSOrg**: A uniform name for each unique WDS member or candidate organization
- **WDSMembership**: Type of WDS member (or candidate)
- **WDSOrgLat:** Latitude of the WDS organization's headquarters
- **WDSOrgLong**: Longitude of the WDS organization's headquarters

The **Organization Name** variable from RDA user records was retained and renamed OrgName.

Once the WDS users were identified, their user IDs (**user_ID**) were used to retrieve all RDA groups in which WDS members participate, and generate a new groupmembers_nn file, where each record matches a (non-unique, de-identified user) with an RDA group and a WDS organization. For RDA Users not in any groups, a single record for that user will have *None* in the **RDAGroup**, **RDAGroupType** and **RDAGroupTypeName** columns. For RDA users with no WDS affiliation, columns **WDSOrg**, **WDSMembership**, **WDSOrgLat, WDSOrgLong** are empty.

Contents of the dataset

The current (2021) version of the RDA-and-WDS-visualization dataset contains two tab-separated tables: users_nn.tsv and groupmembers_nn.tsv. Each record in users_nn represents a unique, de-identified RDA member. If the RDA member has a WDS affiliation, the record contains information about the WDS institution of affiliation. Each record in groupmembers_nn represents a non-unique, de-identified RDA member who participates in an RDA group or, for RDA Users not in any groups, a single record for that user will have *None* in the **RDAGroup**, **RDAGroupType** and **RDAGroupTypeName** columns.

Table 2 describes the data variables in the dataset.

Table 2: Data variables in 2021 version of RDA-and-WDS-visualization dataset.

Variable	Description	Possible values	Source
CountryISO	Location on RDA User's profile	ISO alpha-3 (3-letter) code	RDA
CountryName	Location on RDA User's profile	Country name	RDA
InstitutionalRole	RDA user's professional title/role at their institution	*Student, Programme Manager/Project Manager, Other, CEO/Managing Director/Chief Executive, Librarian, Researcher, IT Specialist/IT Architect ,CTO/IT Director, Advisor/Consultant, Professor, Policy development manager/Policy Consultant, Journalist/Editor/Copywriter*	RDA
OrgName	RDA user's institutional affiliation	Free text: An institution’s name	RDA
OrgType	Type of institution (refers to **OrgName**)	*Academia/Research, Government/Public Services, Policy/Funding Agency, Other, IT Consultancy/Development, Small and Medium Enterprise, Press and Media, Large Enterprise*	RDA

PrimaryDomain	RDA user's primary domain of expertise	Free text: An area of knowledge or professional activity	RDA
RDAGroup	One of the groups of which the RDA User is a member. For RDA Users not in any groups, a single record for that user will have *None* in this column.	The name of one of the active or historical groups at RDA or *None*.	RDA
RDAGroupType	Short alias for **RDAGroupTypeName**	*None, WG, TF, Regional, List, Planning, Other, National, IG, FG, Event, COVID, CG, BoF*	RDA
RDAGroupTypeName	Type of RDA group for (refers to **RDAGroup**)	*None, Working Group, Task Force, Regional Group, RDA List, Planning Group, Other Group, National Group, Interest Group, Focus Group, Event, COVID 19 Group, Coordination Group, Birds-of-a-Feather*	RDA
RDAMemberCountryIncomeGroup	The income group of the RDA user's country of location	*High income, Upper middle income, Lower middle income, Low income*	World Bank
WDSOrg	If RDA User is from a WDS member, name of WDS Organization linked to the user	Name or acronym of a current or candidate WDS organization	WDS
WDSMembership	WDS Organization's membership type (refers to **WDSOrg**) (empty if not a WDS member)	*Associate, Candidate, Network, Partner, Regular*	WDS
WDSOrgLat	If RDA User is from a WDS member, latitude of WDS organization's headquarters (empty if not a WDS member)	Latitude (decimal)	WDS
WDSOrgLong	If RDA User is from a WDS member, longitude of WDS organization's headquarters (empty if not a WDS member)	Longitude (decimal)	WDS