Mon. Not. R. Astron. Soc. **000**, 1–3 (2016) Printed 9 May 2016 (MN LATEX style file v2.2)

Radio Galaxy Zoo: Data Release 1 of 82,071 radio sources

Radio Galaxy Zoo science team et al.

Accepted year month day. Received year month day; in original form year month day

ABSTRACT
Data Release 1 for the Radio Galaxy Zoo (RGZ) project.

Key words: methods: data analysis — radio continuum: galaxies — infrared: galaxies.

1 INTRODUCTION

A complete introduction and description of the project is provided in Banfield et al. (2015, hereafter B15).

2 RADIO GALAXY ZOO

2.1 FIRST + WISE

FIRST images are 3×3 arcmin.

2.2 ATLAS + SWIRE

ATLAS images are 2×2 arcmin.

3 DATA REDUCTION

The time period for the classifications from RGZ Data Release 1 (DR1) runs from 17 Dec 2013 (project launch) to 30 Mar 2016. In that time period, 11,214 registered users provided at least 1 classification of an image. A total of 97,807 images have been fully classified and retired from 1,692,415 classifications. Anonymous users (not registered with the system) provided 25.4% of the total classifications. Among registered users, the distribution of effort was highly unequal, with a Gini coefficient of $G = 0.887$. While indicating that the bulk of classifications are done by the most prolific classifiers, this is consistent with values measured for a wide range of citizen science projects (Cox et al. 2015) and signals the presence of a dedicated user base.

The limit for retiring an individual subject was initially set at 20 classifications for every image. After analyzing the early results of the project, it became clear that 20 thresholds oversampled the number of classifications needed to accurately characterize images with only a single radio component in the frame. In these cases, the only useful input by the users is identifying the location of the infrared counterpart (if present), for which we determine that typically only a few independent classifications suffice. To increase efficiency, the retirement threshold for images with only 1 radio component was lowered to 5 classifications on 20 Jun 2014, after $\sim 750,000$ classifications were complete.

A *source* is a single astronomical object identified by users, consisting of one or more discrete radio components along with a possible infrared counterpart. A *component* refers to a discrete area of radio emission predefined by a cut above the noise level and represented as a set of enclosed contours in the RGZ interface. The *counterpart* refers to the identification of the probable source of the radio emission as seen in infrared.

Images in RGZ are presented for classification by independent users. The users are treated as having equivalent levels of skill, with consensus accomplished by a simple majority vote. To find the consensus, the algorithm first separates classifications by the number of *sources* (N_s) identified in the image. For each classifier who identified some number of sources $N_{s,i}$, the most common combination of possible radio *components* is selected and the number of votes recorded. The algorithm then compares the total number of votes for each $N_{s,i}$, with the overall highest value selected as the consensus identification. Ties between vote counts are broken (possibly incorrectly) by randomly selecting among combinations with the same number of votes.

Once the consensus radio source(s) have been identified, the IR data is separately considered. For each classifier selecting the same combination of radio components as the overall consensus, the location of their corresponding *counterpart* is marked. If the most common response for those radio components was to select "No Infrared", then the source is labeled as having no counterpart. If not, then the positions are then used in a 2-D Gaussian kernel-density estimator (KDE) to estimate the probability density function of the *counterpart* in pixel coordinates. If there is enough data to calculate the KDE (requiring at least 3 non-colinear points), we evaluate the KDE on the same grid size as the original infrared image and apply a 10×10 pixel maximum filter to locate peaks. The location of the highest peak (corresponding to the maximum of the probability distribution function) is used for the position of the IR *counterpart*.

Users who marked "No Infrared": potentially very problematic, since (for example) 80% of users could say there was no IR counterpart, and 20% selected some image. That means we'd be going with the IR position of a strong minority. Should be changed ASAP.

3.1 Duplicates in overlapping fields

Out of 40,270 entries in the consensus catalog (177,461 total) with overlapping areas in the $3' \times 3'$ images, 10,778 have identical consensus answers. Explore whether these are compact sources or have multiple radio components.

4 CATALOG

There are three basic types of data products for Radio Galaxy Zoo: raw classifications, consensus catalogs, and static versions.

Raw classifications are the individual clicks that each user performs; they contain the raw pixel information corresponding to the selection of radio components and the IR counterpart, if available. These are unlikely to be used by most science team members, since they don't have consensus or weighting, require linking to the subject, and are stored only in MongoDB format. Raw classifications are updated daily on the Zooniverse servers.

The consensus catalog is the aggregated classifications over all users, sorted for each subject. This is run throug a Python pipeline, combining the 20 total votes (or 5, in the case of single-component radio sources) and finding the most common answer. Only retired subjects with the full number of classifications are analyzed. We then add physical parameters to each match by measuring the properties in the radio image and positionally cross-matching to the AllWISE (Cutri 2013) and SDSS DR12 (Alam et al. 2015) catalogs. The consensus is updated whenever the latest raw classifications are re-run against the consensus algorithm (every couple weeks, usually). The data is stored in MongoDB format.

Static versions of the catalogs can be generated from the MongoDB versions. These are "flat" versions that are more like the data products typically used in astronomy; a data table in CSV or FITS format where each row corresponds to a unique source and each column is a measured parameter. It's different from the consensus catalog in two ways: firstly, it's not updated as often and so represents a "static" version of the total classifications. Secondly, there are parameters that will have different numbers of elements for each source — for example, the number of distinct radio sources or peaks in a given source. Since that can't be included in a flat table, these data are not included — use the MongoDB version of the consensus if you want data on that.

Neither the consensus nor static catalogs have the ATLAS subjects incorporated yet; they only contain FIRST images.

The fundamental entry in the DR1 catalog is a radio source, which contains one or more radio components and a possible IR counterpart (Tables 2 and 4).

Column 1 contains the unique ID for the RGZ source. Columns 2 and 3 contain the J2000.0 coordinates for the infrared counterpart of the radio source. Column 4 gives the kernel width (in arcsec) of the aggregate clicks used to pinpoint the IR source, providing a measure of positional uncertainty for the host identification. Columns 2-4 are only populated if at least 50% of the users positively identified an infrared counterpart from the WISE data. Columns 5-6 give the total integrated flux and error (in mJy) for all radio components associated with this source. Columns 7-8 give the peak integrated flux density and error (in mJy/beam) for the brightest radio peak in the source. Columns 9-10 give the integrated luminosity and error for all radio components associated with the source. Column 11 gives the maximum angular extent (in arcsec) of the bounding boxes for all radio components, as measured corner-to-corner. Column 12 is the transverse physical size (in kpc) corresponding to the maximum angular extent. Column 13 is the total solid angle for the radio source, calculated by summing the individual solid angles subtended by the outermost contours for each radio component. Column 14 gives the cross-sectional area (in kpc^2) corresponding to the total solid angle. Column 15 gives the total number of radio peaks in the source, defined as the sum of the number of individual components plus any additional local maxima within a single component.

All components relating to the radio luminosity, transverse size, or cross-sectional area are only calculated if a redshift has been detected for the radio source's optical counterpart, since all such values require a distance.

LR: List and give examples of the failure points in the catalog (contrast with success points, too!)

5 RESULTS

Emphasis of the analysis should be on the *statistics* of the sample.

5.1 FIRST

5.2 ATLAS

Suggestion from LR: Plot number of agreements between RGZ, Norris et al. (2006) for multi-component sources as a function of minimum flux level in components. There's an explicit floor in RGZ depending on the noise level we set, and likely an *implicit* one based on visual classification in Norris et al. (2006).

6 SUMMARY

REFERENCES

Alam S. et al., 2015, *Ap. J. Suppl.*, 219, 12
Banfield J. K. et al., 2015, *MNRAS*, 453, 2326
Cox J., Oh E., Simmons B., Lintott C., Masters K., Greenhill A., Graham G., Holmes K., 2015, Computing in Science Engineering, 17, 28
Cutri R. M., 2013, VizieR Online Data Catalog, 2328
Norris R. P. et al., 2006, *A. J.*, 132, 2409

Table 1. RGZ consensus classifications of FIRST radio morphologies

RGZ ID	FIRST ID	Zooniverse ID	N_{class}	C_l	N_{comp}	IR counterpart
1	FIRSTJ145834.5+140942	ARG0002qe4	18	0.833	1	Y
2	FIRSTJ130905.4+433849	ARG0000yc4	5	1.000	1	Y
3	FIRSTJ102805.7+542412	ARG0000dcs	4	0.800	1	Y

Note. — The full, machine-readable version of this table is available at on the journal website and at http://data.galaxyzoo.org/radio. A portion is shown here for guidance on form and content.

Table 2. Matched catalog for RGZ-FIRST consensus sources

RGZ ID	IR counterpart			Radio										
	RA	dec	e_{IR}	S_ν	σ_{S_ν}	$S_{\nu,\mathrm{peak}}$	$\sigma_{S_{\nu,\mathrm{peak}}}$	L_ν	σ_{L_ν}	θ_{max}	$D_{\mathrm{A,max}}$	Ω_{tot}	A_{tot}	N_{peaks}
	J2000	J2000	[arcsec]	[mJy]	[mJy]	[mJy beam^{-1}]	[mJy beam^{-1}]	[W/Hz]	[W/Hz]	[arcmin]	[kpc]	[arcsec^2]	[kpc^2]	
1	23.38219	251.6794	err	10.37	0.20	7.21	0.02	2.06e+24	3.92e+22	0.28	65.80	105.9	1666.7	1

Note. — The full, machine-readable version of this table is available at on the journal website and at http://data.galaxyzoo.org/radio. A portion is shown here for guidance on form and content.

Table 3. RGZ consensus classifications of ATLAS radio morphologies

RGZ ID	ATLAS ID	Zooniverse ID	N_{class}	C_l	N_{comp}	IR counterpart
1	CI002	ARG0002qe4	18	0.833	1	Y
2	CI003	ARG0000yc4	5	1.000	1	Y
3	CI004	ARG0000dcs	4	0.800	1	Y

Note. — The full, machine-readable version of this table is available at on the journal website and at http://data.galaxyzoo.org/radio. A portion is shown here for guidance on form and content.

Table 4. Matched catalog for RGZ-ATLAS consensus sources

RGZ ID	IR counterpart			Radio									
	RA	dec	e_{IR}	S_ν	σ_{S_ν}	$S_{\nu,\mathrm{peak}}$	$\sigma_{S_{\nu,\mathrm{peak}}}$	L_ν	σ_{L_ν}	θ_{max}	$D_{\mathrm{A,max}}$	Ω_{tot}	A_{tot}	N_{peaks}
	J2000	J2000	[arcsec]	[mJy]	[mJy]	[mJy beam^{-1}]	[mJy beam^{-1}]	[W/Hz]	[W/Hz]	[arcmin]	[kpc]	[arcsec^2]	[kpc^2]	
1	23.38219	251.6794	err	10.37	0.20	7.21	0.02	2.06e+24	3.92e+22	0.28	65.80	105.9	1666.7	1

Note. — The full, machine-readable version of this table is available at on the journal website and at http://data.galaxyzoo.org/radio. A portion is shown here for guidance on form and content.