Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **522,** 2584–2600 (2023) https://doi.org/10.1093/mnras/stad1021
Advance Access publication 2023 April 11

Radio galaxy zoo EMU: towards a semantic radio galaxy morphology taxonomy

Micah Bowles [1]★ Hongming Tang [2] Eleni Vardoulaki,[3] Emma L. Alexander [1] Yan Luo,[4] Lawrence Rudnick [5] Mike Walmsley [1] Fiona Porter,[1] Anna M. M. Scaife [1,6] Inigo Val Slijepcevic [1] Elizabeth A. K. Adams,[7,8] Alexander Drabent,[3] Thomas Dugdale,[1] Gülay Gürkan,[9,3,10] Andrew M. Hopkins,[11] Eric F. Jimenez-Andrade,[12] Denis A. Leahy [13] Ray P. Norris,[14,15] Syed Faisal ur Rahman [16] Xichang Ouyang,[4] Gary Segal,[15,17] Stanislav S. Shabala [18] and O. Ivy Wong [9,19,20]

[1]*Jodrell Bank Centre for Astrophysics, Department of Physics and Astronomy, University of Manchester, Oxford Road, Manchester M13 9PL, UK*
[2]*Department of Astronomy, Tsinghua University, Beijing 100084, China*
[3]*Thüringer Landessternwarte, Sternwarte 5, 07778 Tautenburg, Germany*
[4]*School of Physics and Astronomy, Sun Yat-sen University, 2 Daxue Road, Zhuhai 519082, China*
[5]*Minnesota Institute for Astrophysics, University of Minnesota, 116 Church St, SE, Minneapolis, MN 55455, USA*
[6]*The Alan Turing Institute, Euston Road, London NW1 2DB, UK*
[7]*ASTRON, the Netherlands Institute for Radio Astronomy, Oude Hoogeveesedijk 4, 7991 PD Dwingeloo, the Netherlands*
[8]*Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700 AV Groningen, the Netherlands*
[9]*CSIRO Space and Astronomy, ATNF, PO Box 1130, Bentley WA 6102, Australia*
[10]*Centre for Astrophysics Research, University of Hertfordshire, College Lane, Hatfield AL10 9AB, UK*
[11]*Australian Astronomical Optics, Macquarie University, 105 Delhi Rd, North Ryde, NSW 2113, Australia*
[12]*Instituto de Radioastronomía y Astrofísica, Universidad Nacional Autónoma de México, Antigua Carretera a Pátzcuaro # 8701, Ex-Hda. San José de la Huerta, Morelia, Michoacán, C.P. 58089, México*
[13]*Department of Physics and Astronomy, University of Calgary, Calgary, Canada*
[14]*School of Science, Western Sydney University, Locked Bag 1797, Penrith, NSW 2751, Australia*
[15]*CSIRO Space and Astronomy, PO Box 76, Epping, NSW 1710, Australia*
[16]*Institute of Space Science and Technology, University of Karachi, Pakistan*
[17]*School of Mathematics and Physics, University of Queensland, St Lucia, Brisbane, QLD 4072, Australia*
[18]*School of Natural Sciences, University of Tasmania, Private Bag 37, Hobart, TAS 7001, Australia*
[19]*ICRAR-M468, The University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia*
[20]*ARC Centre of Excellence for All-Sky Astrophysics in 3 Dimensions (ASTRO 3D), Australia*

Accepted 2023 March 29. Received 2023 February 27; in original form 2022 October 14

ABSTRACT

We present a novel natural language processing (NLP) approach to deriving plain English descriptors for science cases otherwise restricted by obfuscating technical terminology. We address the limitations of common radio galaxy morphology classifications by applying this approach. We experimentally derive a set of semantic tags for the Radio Galaxy Zoo EMU (Evolutionary Map of the Universe) project and the wider astronomical community. We collect 8486 plain English annotations of radio galaxy morphology, from which we derive a taxonomy of tags. The tags are plain English. The result is an extensible framework, which is more flexible, more easily communicated, and more sensitive to rare feature combinations, which are indescribable using the current framework of radio astronomy classifications.

Key words: standards – methods: statistical – catalogues – galaxies: statistics – radio continuum: galaxies.

1 INTRODUCTION

Language is often difficult to define or use. When new concepts arise and demand their own terminology, terms can be adopted from similar ideas (e.g. 'entropy' in information theory and physics; Natal et al. 2021), invented (e.g. 'utopia'; Romm 1991), or named after the discoverers (e.g. 'Newtonian physics'). Individual terms often have multiple accepted definitions within a given field, and especially across fields (e.g. 'modern' in philosophy and art; von Schelling, von Schelling & Schelling 1994; Adajian 2022). The construction of language in science is especially important, as language is believed to affect how we think (Wolff & Holmes 2010).

The terminology used in astronomy struggles with these same issues. Some terms are obfuscated, poorly defined, or so specific that experts can often only engage in the subject matter if a definition

★ E-mail: micah.r.bowles@gmail.com

is provided on each use. In radio astronomy, a field which began in the 1930s (Southworth 1956), the language used to describe celestial objects has been developed almost entirely in tandem with the instruments and corresponding scientific understanding. Consequently, some terms are limited by our physical understanding (e.g. Little Green Man 1; Hewish et al. 1968) or the sample inspected at the time (e.g. FRI/FRII; Fanaroff & Riley 1974).

In the case of radio galaxy morphologies, language is becoming increasingly difficult to use, especially as technological and scientific advancements provide deeper insight into the vast range of radio morphologies. The gap between the diverse range of observed radio galaxy morphologies and the classification schemes used is widening. The current morphological classifications carry information, which cannot be quantified under current frameworks, meaning the use of non-numeric features, i. e. language based schemes, is unavoidable. Accordingly, the current classification schemes fall victim to obfuscated language. Additionally, the terms used tend to describe abstract classes, which lack the ability to capture the increasingly complex features of radio galaxies observed with the newest generation of instruments. Rudnick (2021) urges the radio astronomy community to develop a tagging system rather than forcibly attempting to create classes, which neatly separate objects. Such a tagging system would allow an object to be assigned plain English descriptors capturing the semantics of the object's features through tags rather than be assigned a distinct class to which it belongs. Additionally, this tagging system would be able to consolidate instrument specific morphologies within the same framework without producing conflicts. As an example, a source could be tagged as 'compact' in a low-resolution survey while having specific morphological features captured by tags referring to observations made by higher resolution instruments. This work aims to build the framework for such a tagging system for the first time.

The newest radio instruments in operation are producing maps of sources which are so deep, resolved, and with such high dynamic range that our existing classification schemes are failing. An updated, and extensible, radio morphology taxonomy of tags would be a tremendous benefit moving forward because deeper and wider surveys are expected to be a massive driver of scientific development in the coming decades. If the scientific community had a framework and terminology, which were not intrinsically limited by sensitivity or resolution, it would mean that we could work with the same framework regardless of the technological improvements to the instruments in the field. We therefore expect this work could have major implications in various scientific contexts, including population studies and rare object searches in observations made by current and future radio instruments including the Australian Square Kilometre Array Pathfinder (ASKAP; Johnston et al. 2008), the Low Frequency Array (LOFAR; van Haarlem et al. 2013), the Deep Synopic Array 2000 (DSA-2000; Hallinan, Ravi & Deep Synoptic Array Team 2021), the Murchison Widefield Array (MWA; Tingay et al. 2013), MeerKAT (Jonas & MeerKAT Team 2016), the next generation Very Large Array (ngVLA; Murphy & ngVLA Science Advisory Council 2020), and the Square Kilometre Array (SKA; Dewdney et al. 2009).

This work uses data from the Evolutionary Map of the Universe (EMU; Norris et al. 2011), a radio survey being conducted with the ASKAP telescope. ASKAP's large field of view means that it can map a large portion of the sky at once. Because of this, EMU is currently planned to map three quarters of the sky, the first two-thirds of which are planned to be completed in the first five years. EMU is estimated to catalogue 40 million sources (estimate made using the Tiered Radio Extragalactic Continuum Simulation method, T-RECS; Bonaldi et al. 2019). In an effort to classify these sources at scale, we are launching 'Radio Galaxy Zoo EMU' (RGZ EMU). RGZ EMU is a citizen science project designed to allow the public to provide valuable and essential insights into these sources, including host identification, source assembly, and source classification (full details on RGZ EMU in Tang, Vardoulaki et al. in preparation). While designing this project, we discussed what classifications we would ask the citizen scientists to use. It became clear that there was no consensus on the terms to use. In part as a response to this dilemma, we collect plain English annotations on radio galaxies and implement a novel framework to derive semantic plain English tags.

The proposed process uniquely combines existing natural language processing (NLP) methods. NLP has garnered significant research interest over the last 20 yr (see for instance Mishra & Kumar 2020). Thomas et al. (2022) use NLP and a form of topic modelling called latent Dirichlet allocation (LDA; Vayansky & Kumar 2020) with the aim of guiding the planning process of research priorities by analysing trends in previous publications. Grezes et al. (2021) use deep learning-based NLP techniques with the aim of improving the SAO/NASA Astrophysics Data System (ADS.[1]).

The method we present is not bound to radio astronomy. It can be applied to any domain. The code used in this work is publicly available at https://github.com/mb010/Text2Tag and is written to be transferable to other fields.

Our work is structured as follows. In Section 2, we detail the data used in and collected through our experiments. We present the proposed method in full in Section 3 before presenting the details of its application and the resulting taxonomy in Section 4. Initial physical results using the semantic taxonomy are presented in Section 5. The results and impact are discussed in Section 6 and conclusions are made in Section 7.

2 DATA

Two experiments were designed and executed. Both made use of early versions of cutouts prepared for the RGZ EMU project as described in Section 2.1. The intent and design are detailed for the *Plain English Annotations* experiment as well as the *Expert Classification* experiment in Sections 2.2 and 2.3, respectively. An anonymized version of the data is publicly available.[2]

2.1 Image data

To produce the data analysed in this work, users were asked to consider individual images in turn. An example of one of these images is presented in Fig. 1. These images are early versions of the data to be used in the RGZ EMU project. This version consists of three panels containing a 6 arcmin by 6 arcmin cutout from the EMU pilot survey. The panels show EMU contours with the false colour EMU image, a Digitized Sky Survey (DSS; Lasker et al. 1990) cutout, and a Wide-Field Infrared Survey Explorer (WISE; Wright et al. 2010) cutout. Each image is centred on an EMU Selavy catalogue component (Norris et al. 2011, 2021b).

The cutouts are subject to a number of criteria designed to select a small number of sources for early testing. Components which had an angular extent of less than 27 arcsec (1.5 beamwidths) were removed. Components which were within 45 arcsec (2.5 beamwidths) of another catalogued component were also removed, as these are largely simple doubles with little to no morphological features and

[1] See: https://ui.adsabs.harvard.edu
[2] https://zenodo.org/record/7254123#.Y7VvGtLP3Lp

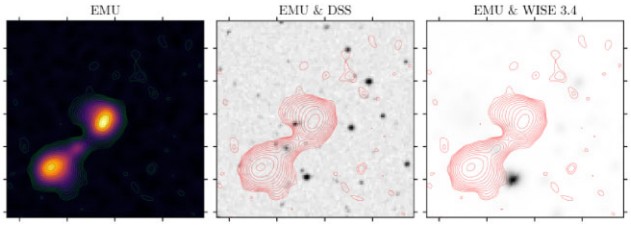

Figure 1. Example cutout as presented to the participants of the experiments detailed in Sections 2.2 and 2.3. Cutout centre: 21 h 02 min 16 s – 54°23′36″ (J2000). The lowest contour is fixed at 0.12 mJy. Subsequent contours are multiples of $\sqrt{2}$ higher.

can be classified algorithmically. This resulted in a list of 306 sources, for which cutout images were made. Our final sample consists of 299 of these cutouts because an undetected upload error caused seven cutouts to not be uploaded to the Zooniverse platform.

2.2 Plain English annotations

To derive the desired plain English taxonomy, we started with plain English descriptions (annotations) of the given object or phenomenon. Using the 299 sets of images outlined in Section 2.1, we built a private Zooniverse[3] project, where we presented users with an empty text box and a prompt reading:

Please describe the source:

(i) *in the***middle** *of the frame and any***associated** *emission.*
(ii) *use***simple** *English.*
(iii) *avoid***jargon**.

(a) *e.g. refrain from typing FRI, WAT, and so on.*

(iv) *descriptions should be***separated by ‘’,**.

This data was intentionally collected to be relatively unconstrained to encourage the annotations to cover diverse ideas of source features. Thus users were enabled to highlight and describe whatever caught their attention within the image. The trade-off in this unstructured description approach is that the resulting data are unwieldy and noisy (the consistency and formatting of phrases is not constrained). As such, the method outlined in Section 3 contains a significant overhead of data cleaning, which is common with any unstructured natural language data.

The data collection for this experiment ran from 2021 December 17 to 2022 January 27. We offered users who processed more than 100 sources co-authorship on this publication, which is a direct result of their efforts. In total, we had 19 users annotate an average of 154 sources each, resulting in a total of 2920 descriptions consisting of a 8486 comma separated annotations. Almost all of these users are astronomers, and more than three quarters of them have at least some academic experience of radio morphologies.

2.3 Expert classification

We conducted a second experiment to collect expert classifications on the same sets of images. This experiment was conducted with the aim of extracting ideas represented by annotations which are relevant to the expert’s science cases.

We established a separate private Zooniverse project and invited a number of experts to participate in classifying the radio morphologies of the objects in the images described in Section 2.1. To classify objects with predefined classes participants were prompted with:

Radio morphology*: Please describe the source:*

(i) *in the***middle** *of the frame and any***associated** *emission*
(ii) *select one or more tags that fit object radio morphology.*

We presented the participants with 22 classes, which they could use as they wished, including assigning none or all of them to the subject. The abstract classes listed were selected from a compiled list of radio morphology classes presented in Rudnick (2021) and were: Single, Double, Classical Double, Triple, Narrow-angle tail (NAT), Wide-angle tail (WAT), Bent tail, Fanaroff & Riley Class 1 (FR I), Fanaroff & Riley Class 2 (FR II), Fanaroff & Riley Class 0 (FR 0), Hybrid, X-shaped, S-shaped, C-shaped, Diffuse, Double–double (DDRG), Core-dominant, Core-jet, Compact Symmetric Object (CSO), 1-sided, Odd Radio Circle (ORC), and Star-Forming Galaxy (SFG). This experiment ran from 2022 January 27 to 2022 May 19. Five experts made a total of 1257 multilabel classifications of an average 251 objects each.

3 METHOD

To the best of our knowledge, there is no existing process or NLP approach, which produces a semantic taxonomy from a corpus of short annotations. The closest approaches are widely used topic modelling approaches. These approaches capture topics within a corpus through distributions of terms in documents. Vayansky & Kumar (2020) present a helpful review of topic modelling variants. These models are designed to return a distribution of terms, which belong to each discovered topic. They are not designed to return terms, which communicate what a given topic is. We explicitly want to build a taxonomy on a certain subject. Therefore, the terms which effectively capture the meaning of a topic are essential.

Although a panel of experts may be able to manually define a semantic set of terms for a given problem, the success of such an approach would depend on whether the panel agree, the backgrounds of the experts, and their ability to distil complex ideas into simple plain English effectively. This manual approach would likely also lack the reproducibility and tractability that is expected by the physical sciences.

We therefore propose a method through which short annotations are distilled into semantic tags in accordance with a specific science case and its respective features (including classes). The workflow of the method is presented in Fig. 2. The derived taxonomy should provide wide coverage of objects of interest, have the ability to distinguish features, be clear in what semantic feature it describes, and be appropriate to the science cases.

Conceptually, in the framework similar annotations are aggregated to produce a single term, which we call ‘tag’. We rank how important tags are based on the impact they have in classifying the existing abstracted science classes. A selection is made on the most important tags to form a taxonomy.

A technical outline of the method is presented in Section 3.1. The implementation details for our data and project are presented in Section 3.2.

3.1 Technical outline

Sequences of words are processed where w_i represents the ith word of N in a given annotation, $\mathbf{a}_j = (w_1, w_2, ..., w_N)$. Here $\mathbf{a}_j$ is the jth annotation of the M annotations in our corpus. Note that in the NLP literature, the equivalent of annotations would be ‘documents’. This

[3]See: https://www.zooniverse.org/

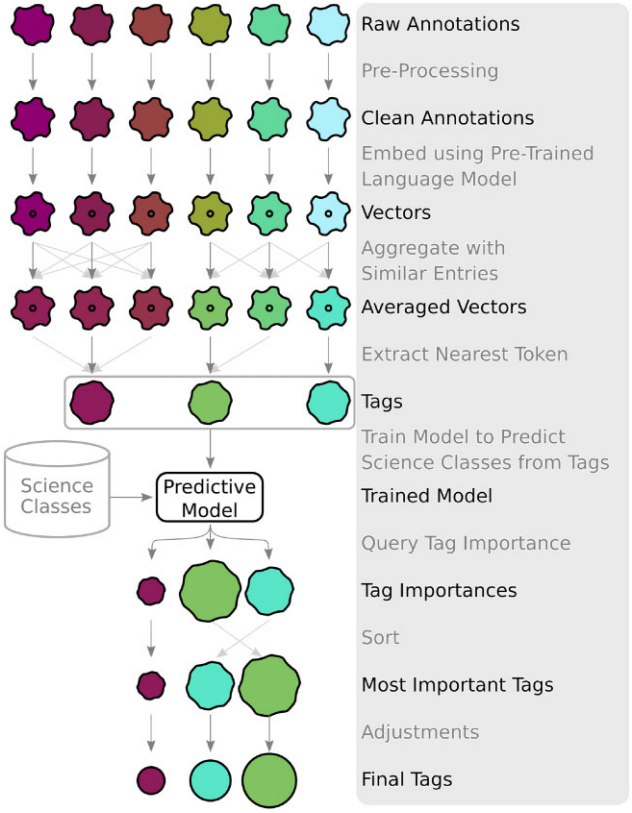

Figure 2. Proposed workflow to computationally derive a semantic plain English taxonomy from a set of annotations.

method is expected to work best on extremely short annotations (documents), where each annotation contains a single idea. The annotations are embedded into a k-dimensional vector through a pre-trained model, $f_{\rm emb}$:

$$f_{\rm emb}\left(\mathbf{a}_j\right) = \mathbf{v}_j \in \mathbb{R}^k. \tag{1}$$

This is currently implemented such that the order of the words does not affect the encoding (i.e. in a bag-of-words paradigm). We embed each word within an annotation through

$$f_{\rm emb}\left(\mathbf{a}_j\right) = \frac{1}{N}\sum_{i=1}^{N} f_{\rm emb}\left(w_i\right). \tag{2}$$

For pairs of annotations, $(i, j) \in [1, M]^2$, a similarity value is calculated using the cosine similarity, $g_{\rm cs\ sim} : \mathbb{R}^k \rightarrow [-1, 1]$, which takes the dot product of two vectors scaled by the inverse of the product of the Euclidean norms of those vectors:

$$g_{\rm cs\ sim}(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\|\|\mathbf{v}_j\|}. \tag{3}$$

According to a similarity threshold, σ, M averaged vectors are then calculated through:

$$v_i' = \frac{1}{M'}\sum_{j=1}^{M}\begin{cases}\mathbf{v}_j & \text{if} \quad g_{\rm sim}(\mathbf{v}_i, \mathbf{v}_j) \geq \sigma \\ 0 & \text{else}\end{cases}, \tag{4}$$

where M' is the number of non-zero elements being summed over. The model, $f_{\rm emb}$, used to embed the annotations is then used to produce the token, which is closest to $\mathbf{v}_i'$:

$$f_{\rm emb}^{-1}(\mathbf{v}_j') \approx s_q, \tag{5}$$

where s_q is the qth entry of all Q unique derived tags. As tags are derived from the M annotations $Q \leq M$.

For each annotated object, we define $\mathbf{t} = (t_1,..., t_Q)$ as a vector encoding of the tags, where t_q is 1 if that tag was present in an annotation associated with that object, or 0 if that tag is not associated with it. As each object has multiple individual annotations associated with it, it can be described through its derived tags $\mathbf{t} \in \{0, 1\}^Q$.

We consider each science class y in the set of science classes Y. Using the encoded tag vector, $\mathbf{t}$, for each object, we fit a model, $f_y : \{0, 1\}^Q \rightarrow \{0, 1\}$, to predict the presence of each science class $y \in Y$. For each model, f_y, and tag representation, t_q, we calculate an importance

$$f_{\rm Importance}(f_y, t_q) = I_{(q,y)} \in \mathbb{R}, \tag{6}$$

where a larger value of $I_{(q,y)}$ means that t_q, and subsequently s_q, are more important to the classification output.

To recover the importance of the qth tag, we take an average across all models, f_y, for a given tag, s_q. We take the support weighted average of the importance of each model

$$I_q = \frac{1}{Q}\sum_{y\in Y} n_y I_{(q,y)}. \tag{7}$$

Here, n_y is the number of (positive) class y entries. Note that other weightings may be preferable depending on the available data and purpose. For instance, if a multi-objective regression task were used instead to calculate $I_{(q,y)}$, then a uniform weighting across tasks may be more appropriate. We normalize the importance values, I_q, such that

$$\sum_{q=1}^{Q} I_q' = 1, \quad \text{and} \quad I_q' \in [0, 1]. \tag{8}$$

Finally, the tags, s_q, are sorted by their I_q'. Although tags are all expected to have some non-zero I_q', a majority of the information is contained within the top tags. Additionally, the tags ranked lowest in this scheme are expected to be noisy (e.g. annotations which contain incorrect spellings, reference otherwise irrelevant features, or are only impactful on a given prediction through the random association of its small sample size). We set an importance threshold to select the top $Q' < Q$ tags. These most important tags, consisting of Q' strings, $s \in S_{\rm Taxonomy}$, is the derived taxonomy.

Some of the tags, s, may require clarification to allow the tag to be clear upon first reading. To do so the raw annotations, which that tag was derived from are taken into consideration, in order to verify what it represents.

3.2 Implementation

The exact implementation of the method outlined in Section 3.1 will depend on the data being used. The details for our implementation are as follows.

3.2.1 Pre-processing

The goal of pre-processing the annotations is to format the data in a uniform manner without disrupting the content (i. e. standardizing grammar, spelling, and formatting). To do this, a number of common NLP data processes are applied. These are applied in the order they are presented in.

All annotations are set to lower case and all accents are removed from characters. Ampersands and new line commands are removed or replaced as appropriate. Forward slash and full stop

characters are replaced with commas as they are often observed to represent separate ideas, which our method assumes are comma separated. Double whitespaces are corrected and hyphens are removed.

Based on manual inspection, additional corrections are made to a number of annotations. These are spelling mistakes such as 'copact' being corrected to 'compact'. We then drop any annotations which mention 'DSS', 'WISE', or 'optical' as these annotations are not expected to be a comment on the radio morphology, which is our target of interest.

At this point, the sets of comma separated annotations are separated into individual annotations. Contractions are expanded. A list of stopwords[4] is extended to include 'like', as well as scientific terms, which the pipeline should not be affected by as they are both technical and not related to morphology. These additional stopwords include 'emu', 'galaxy', 'galactic', 'emission', and 'source'. Terms for cardinal directions ('north', 'south', 'east', and 'west') are also added to the stopwords since our focus is on the features themselves instead of their position relative to a specific source. This stopwords list is applied to the annotations.

We consider both lemmatized[5] and unlemmatized approaches to the data moving forward. Each annotation has now been cleaned, and the data are largely consistently formatted.

3.2.2 *Embeddings*

When embedding the cleaned annotations into vectors, we use SpaCy's large English language model,[6] which is the largest available model within the SpaCy package (v3.3.0) that has tokens (largely words) embedded. It contains 685k embeddings. For a given vector, the model can return the closest embedded word(s). This feature is essential to our process, and is a key factor in the decision to use this model. Other advanced, such as transformer based models, can take word order into account, but do not have these token embeddings. In the implemented SpaCy version, the vector embeddings are learned through the GloVe algorithm introduced by Pennington, Socher & Manning (2014). GloVe (Global Vectors for Word Representation) is an unsupervised representation learning algorithm, which aims to embed words in a space which presents various desirable features. These features include semantic and linguistic similarity, which is advantageous for our use case.

To decide what similarity threshold to use to aggregate over, we consider the histogram of cosine similarities of all annotations. This is presented in Fig. 3. This histogram is presented using annotation embeddings, which were lemmatized and does not include self-similarities. The peak at 1.0 is due to tags, which are identical (maximally similar). The rigorous cleaning process reducing short annotations to a few words before the annotations are embedded is likely a factor. Additionally, self-consistent vocabulary across multiple annotations may be embedded to an (essentially) identical vector.

To capture the excess tail of more similar vectors, we consider similarity thresholds above 0.5 for our models. Lower thresholds would contain the bulk of all annotation pairs, which would be counterproductive in trying to capture individual concepts. However, we would like to explore thresholds down to 0.5 to reduce the number of unique terms and maximize the number of entries per unique tag. The tags derived from the aggregated vector encodings, are returned by SpaCy as single tokens (words).

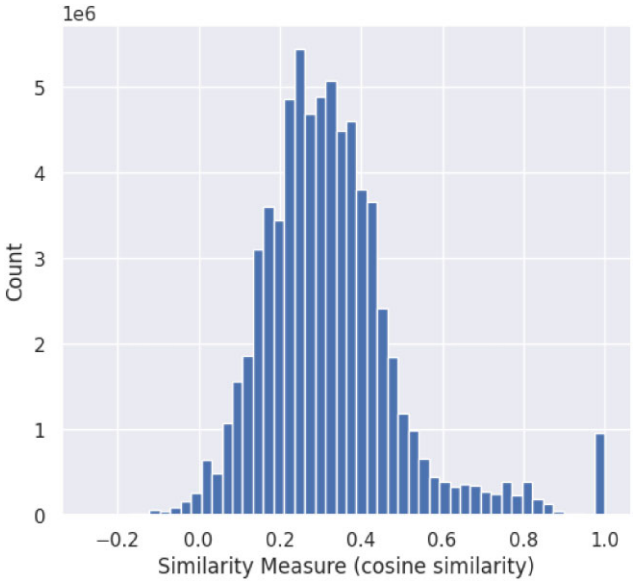

Figure 3. The histogram of the cosine similarities of the embedded annotations. Data is lemmatized and does not include self-similarity values.

3.2.3 *Model definition*

We predict science classes on a source by source basis. We use 22 science classes to train the models to classify science classes from our derived tags. Three science classes presented to our expert classifiers are functionally removed as they are either not usable without spectra (*CSO*), extremely contested and largely disused (*Fanaroff-Riley Class 0*; *FR0*, Hardcastle & Croston 2020; Rudnick 2021), or largely uninformative when considering extended radio morphology (*Single*). Furthermore, the following three science classes have insufficient positive cases for training: Double–Double Radio Galaxies (DDRG; Schoenmakers et al. 2000), Hybrid Morphology Radio Sources (HyMoRS; Banfield et al. 2015; Kapińska et al. 2017), and ORC (Norris et al. 2021a). The experts often did not agree with their usage of the terms. We explore what degree of agreement (expert threshold) beyond which we will consider a source as being positively classified with a certain science class in Section 4.1.

To train models that predict science classes from our derived tags, we have 299 sources. This is a small data set. We chose a relatively simple model in response. We train random forests in a one-vs-rest scheme, i. e. one random forest model to classify one science class. We treat a set of these models as a single model, which predicts the multilabel target of an input. The random forests use the Gini impurity criterion, with the aim of improving the explainability of the selected features (Menze et al. 2009). The random forests are configured with 500 estimators, no maximum depth, and a seeded random state to allow for reproducible results. Unspecified features of the models are inherited from default values as implemented by Scikit-Learn v1.1.0.

[4]Stopwords are a list of superfluous words, which when applied to a text removes instances of those words from the text. Examples include 'it' and 'the'. A list of default stopwords can be found in the SpaCy source code under spacylangenstop_words.py.

[5]Lemmatization is replacing a word with its root word, e.g. 'apples' becomes 'apple'.

[6]https://spacy.io/models/en#en_core_web_lg

3.2.4 Evaluation

Another challenge of the relatively small data set is the evaluation of the trained models. We use cross-validation to maximize our use of the data. The model is evaluated through 10-fold cross-validation, where we train 10 models on 10 different sets of nine-tenths of the available data and evaluate each on the respectively withheld final tenth. With predictions for each tenth from one of the 10 trained models, we recover predictions for each data point. These predictions result in approximate generalized performance metrics for the models.

We choose to track the performance of our models with macro and weighted F1 scores. An F1 score is the harmonic mean of precision and recall, and can be written as

$$F1 = 2 * \frac{(precision \cdot recall)}{(precision + recall)} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}, \quad (9)$$

where *TP* is the number of true positives, *FP* is the number of false positives, and *FN* is the number of false negatives. The macro and weighted F1 scores are extensions to enable evaluations of multilabel problems. The macro F1 score is calculated by averaging F1 scores calculated in a one-vs-rest scheme for each target class. The weighted F1 score is calculated identically to the macro F1 score, except that we take the weighted mean where each score is first scaled by the number of samples of a given class, which may be more telling in an imbalanced data classification problem.

3.2.5 Importance

To estimate the importance of each of the tags, we use Shapley values. Shapley values are a common explainability tool used in machine learning applications (Lundberg & Lee 2017). These values convey how much a feature has contributed to the prediction of the respective model in comparison to the average predictions of the model. Exact Shapley values for each input tag and science classification are calculated using the trained random forest model and the `SHAP` package for trees introduced in Lundberg et al. (2020). These values are the importance values used to estimate which tags capture the semantics of radio morphology.

4 TAXONOMY

4.1 Data configuration

To evaluate which data configuration (i.e. data processing parameter selection) is best, we consider the F1 scores of models trained on various configurations of the data. We grid search data across configurations including four expert thresholds, 11 similarity thresholds, and with or without lemmatization. This results in 88 configurations, which we construct and evaluate.

For the expert classification we demand that at least 20 per cent, 40 per cent, 60 per cent, or 80 per cent of the votes made on a given source agree. We call these confidence thresholds. Note that we make use of percentages. Sources which have not been classified by all experts can still have their classifications reflected in the confidence thresholds used in this search. We do not consider 100 per cent agreement amongst experts as so few classifications would survive that we could not train a model (highlighting a serious issue of the current classification scheme). The F1 scores are presented in Fig. 4, for which the statistics of the random forest models are taken over all 22 configurations (two lemmatization and 11 similarity threshold combinations) for each expert threshold.

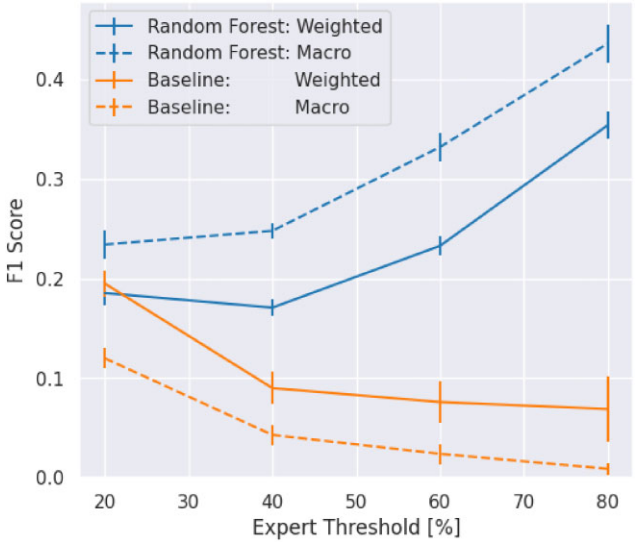

Figure 4. Aggregated F1 scores of both random forest models and baselines showing the effect of thresholding the expert classifications. The baselines are constructed by randomly predicting in accordance with the rate of positive cases per positive class. Statistics are measured over 5000 instances of the baseline estimators, and over all lemmatization and similarity threshold configurations of the random forest models.

Simply stating that the model improves as the expert threshold is increased is largely true, see Fig. 4. However, with increasing expert thresholds, the task which the model has been asked to complete becomes easier as the *noise* in the classifications is functionally reduced. The subset of data where experts have a high consensus are more likely to have clearly identifiable morphologies (reduced aleatoric uncertainty) or present with a morphological classification, which is more widely agreed upon amongst the experts (reduced epistemic uncertainty). In an attempt to capture a broader perspective on what radio morphologies are, while maintaining accuracy of the classifications, we select an expert threshold of 60 per cent for the remainder of this work.

Fig. 5 shows the performance of the 22 models for each similarity threshold and lemmatization configuration. We select the configuration with a similarity threshold of 0.80 and lemmatized inputs. This configuration results in 213 unique tags. The model achieved a weighted F1 score of 0.254 and a macro F1 score of 0.350. This is the model with the highest weighted F1 score. The highest macro F1 score is 0.352 held by both configurations with a similarity threshold of 0.6.

4.2 Tag ranking

The Shapley values are calculated for each tag provided to the model with respect to the model's outputs and the full data set. This provides us with a Shapley value for each science class and tag combination. We take the support weighted average of the Shapley values across the science cases to provide us with a descriptive Shapley value for a given tag. We normalize these values so that they sum to one across all tags. We call these values the comparative weighted Shapley values. These are presented as percentages, which reflect how much sway a given tag has over the science classification of the model.

We calculate the comparative weighted Shapley values and present the 70 most important terms in Fig. 6. To select a usable volume

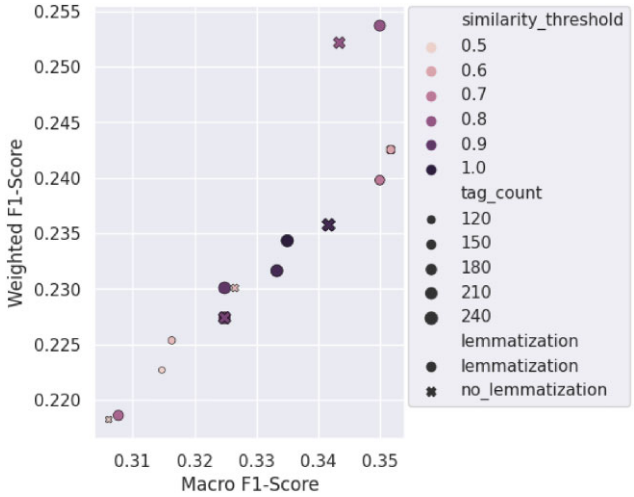

Figure 5. Grid search results across both lemmatization and the 11 similarity threshold configurations for the selected expert threshold of 60 per cent. tag_count refers to the total number of unique tags which that configuration produced, which is strongly dependent on the similarity threshold of the configuration. The selected configuration is the lemmatized version with a similarity threshold of 0.80 (expert threshold of 60 per cent).

of tags, we define the taxonomy to be the top tags required for 68 per cent of the descriptive power to be maintained (approximately 1 per cent of the comparative weighted Shapley value). In this data configuration, this results in a taxonomy of 33 tags.

Highlighting the benefits of Shapley ranking over a simpler approach such as correlations, we present an interactive graph visualization of moderately strong correlations between all combinations of both tags and science cases in Fig. 7. Importantly, the graph does not contain all 33 tags. This is because most of the top 33 tags are not strongly correlated with other terms. They are still the most impactful to the model's decision, as non-linear combinations of tags can be used to classify the science case. Their value to the non-linear classifications is captured by Shapley values.

4.3 Taxonomy adjustments

Limited to single words, the derived tags may not be an optimal selection. Suboptimal tag representations may also occur when the conjugation of a given term is relevant to the use case but lemmatization has removed it even if it would been more easily understood (e.g. 'extended' in comparison to 'extend'). Furthermore, the method only outputs single words, even if the concept the tag represents is better represented by multiple terms.

We therefore investigate each tag in turn by considering all annotations which contribute to it. We adjust tags in an attempt to optimize the taxonomy for grammatical and conceptual clarity. The adjusted tags are listed below, including descriptions of the original annotations/concepts, which a tag represents.

(i) *trace*: derives directly from numerous annotations stating 'traces host galaxy'. This is a more clear expression for what this tag represents. We therefore alter 'trace' to 'traces host galaxy'.

(ii) *disc*: derives from annotations such as 'emission from galaxy disc'. We therefore merge this with 'traces host galaxy' (originally 'trace'). This refers to the radio emission tracing the host galaxy rather than the morphology of the host.

(iii) *bright*: refers to bright features of a presented cutout. This includes cores as well as neighbouring sources. This information is

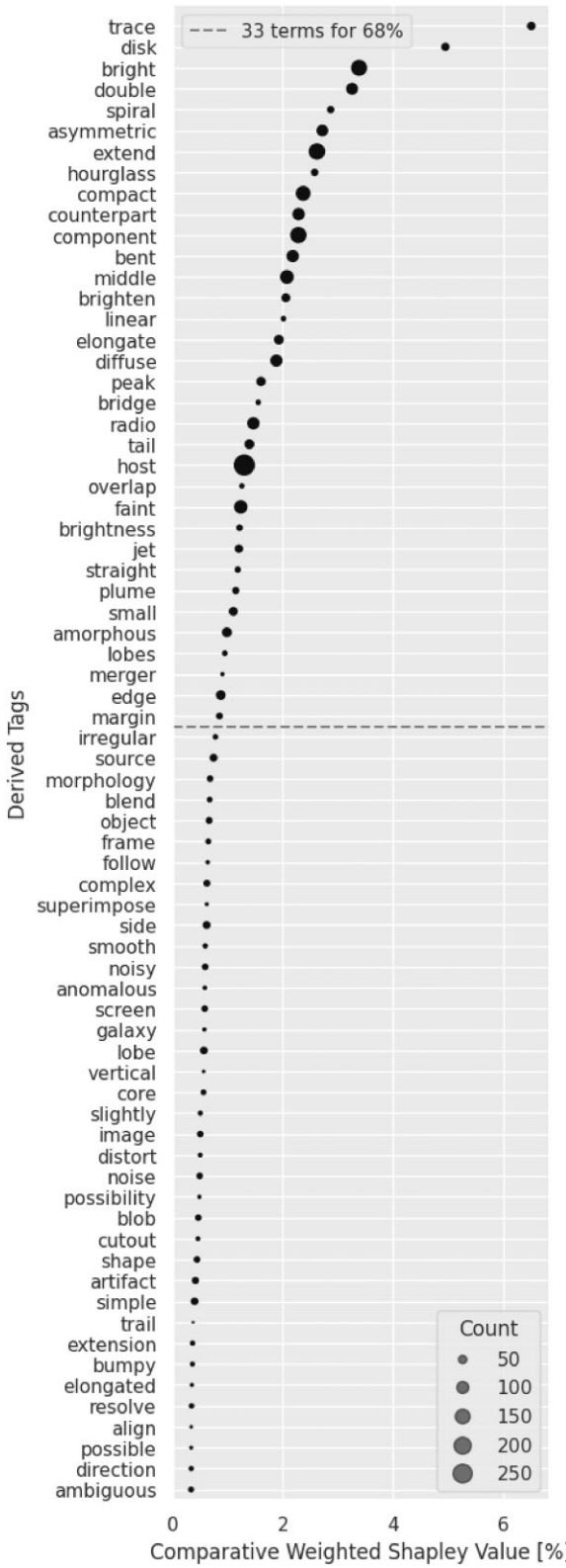

Figure 6. Top 70 tags sorted by their comparative weighted average Shapley values for the configuration selected in Section 4.1.

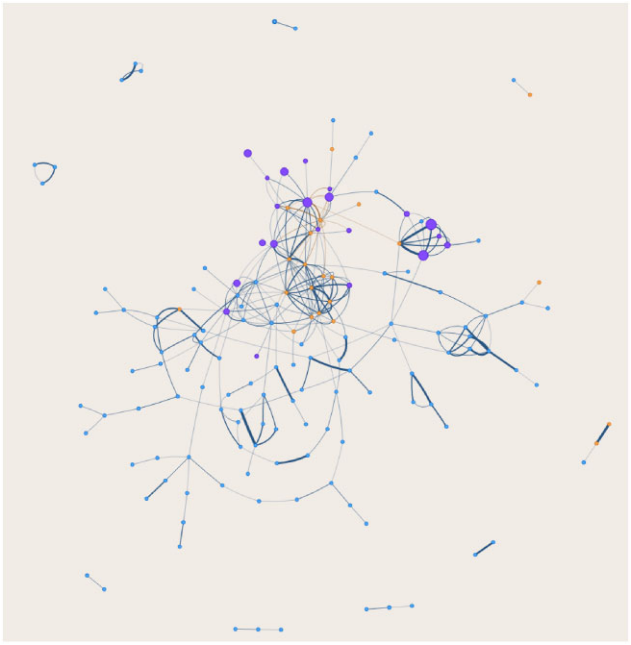

Figure 7. Still frame from the interactive graph available at https://github.com/mb010/Text2Tag/blob/main/CorrelationsGraph.html (download and open with a browser). The graph's edges are correlations above a magnitude of 0.3. Nodes are tags from the top top 33 (purple), other tags (blue), and science classes (orange). The node sizes reflect the occurrence rate of a given node across the 299 sources. The width of the edges represents the correlation between the two nodes it connects. All correlations are in the range $[-0.5, 0.9]$.

more clearly contained in the catalogues of radio component fluxes. Therefore, this tag is dropped.

(iv) *spiral*: derives from spiral galaxies being the hosts of the radio emission. This tag is changed to 'traces host galaxy' as it then contains the relevant radio morphology information.

(v) *asymmetric*: original annotations refer to asymmetric structure. To highlight the difference between asymmetric structure and brightness, we rename this tag to 'asymmetric structure'.

(vi) *extend*: grammatical adjustment to 'extended'.

(vii) *component*: refers to the number of components, which the source is composed of. This tag is dropped in favour of catalogues, which list how many separated components are assigned to a source.

(viii) *counterpart*: refers to matching emission in either optical or infrared. Host identification and 'traces host galaxy' will capture this information. Therefore, this tag is dropped.

(ix) *middle*: largely referring to presence and features of the central core of a radio galaxy. We therefore rename this tag to 'core'.

(x) *brighten*: refers to 'edge brightened' sources. We therefore clarify this by altering this tag to 'edge brightened'.

(xi) *linear*: refers to non-bent radio morphologies, which is captured in the absence of the 'bent' tag. Therefore this tag is dropped.

(xii) *elongate*: refers to elongated structures in the radio emission. This is captured by the absence of the 'bent' tag when sources are also 'extended' and is therefore dropped.

(xiii) *radio*: annotations were written commenting on the radio emission in general ways (such as the presence of a jet or how many components are visible). This information is all mapped by other tags/processes. For this reason, we drop this tag from the taxonomy.

(xiv) *overlap*: often refers to emission, which overlaps with radio contours or vice versa. It is therefore be changed to 'traces host galaxy'.

(xv) *brightness*: original annotations almost exclusively refer to 'asymmetric brightness' across components or within the structure being discussed. We therefore clarify this tag by changing it to 'asymmetric brightness'.

(xvi) *straight*: refers to the non-bent structure of the radio galaxy. We therefore drop it in favour of the absence of the tag 'bent'.

(xvii) *lobes*: we make a grammatical change to 'lobe' with the intent to make this tag less ambiguous for future users.

(xviii) *edge*: highlighting clear edges of sources, as opposed to diffuse edges. This is largely equivalent to and is merged into 'edge brightened'.

(xix) *margin*: the annotations from which this tag derives refer to the source extending beyond the margins of the cutout. This is being accounted for with updated cutouts, and is not morphologically relevant beyond the angular extent of a source, which is better presented in a catalogue format.

4.4 Semantic taxonomy

After the adjustments made to the tags in Section 4.3, we have 22 unique semantic tags. In alphabetical order, the semantic tags we propose to use for radio galaxy morphology are: amorphous, asymmetric brightness, asymmetric structure, bent, bridge, compact, core, diffuse, double, edge brightened, extended, faint, host, hourglass, jet, lobe, merger, peak, plume, small, tail, and traces host galaxy.

4.5 Effectively assigning tags

We have succeeded in deriving semantic tags for radio morphologies (see Section 4.4). However, for RGZ EMU and other citizen science approaches it is not effective to ask citizen scientists to use 22 tags. Terms would likely be ignored in a long list, and users would easily bottleneck on a small number of tags, neglecting the remainder of the taxonomy. This would be detrimental to both the science case and the user experience.

To improve the scientific results as well as the user experience, and to make the most effective use of the citizen scientists' time and energy, we consider which tags within the taxonomy can be most easily computed algorithmically at other stages of processing, e.g. 'small' is easily calculated through the angular extent of the assembly mask for a given source. We aim for 10 tags, which can be presented on a single screen to the citizen scientists. We consider each term in turn, and outline how each term might be assigned in Table 1.

The tags which we believe are least easily computed will benefit the most from citizen scientist input. These are the tags, which are presented as 'proposed for tagging' in Table 1. These 10 tags are those which the RGZ EMU project will present to its citizen scientist volunteers.

4.6 RGZ EMU early feedback

While working towards our final release of RGZ EMU, we asked a small group of 16 testers who have never worked on radio galaxy studies before (eight from China, seven from Pakistan, and one from Germany) for feedback on an early version of the tags terms provided by a beta version of the pipeline presented in this work. The tags presented to the testers were: bent, bridge, complex, diffuse, distorted, elongated, hourglass, jet, plume, and tail.

Table 1. A summary of the suggestions on how the tags of the final adjusted taxonomy are to be assigned. 'Assembly mask' refers to a mask derived from the source assembly process where multiple source components are grouped as (likely) belonging to a single source. 'Tagging' refers to citizen science support, or calibrated and trained machine learning models.

Proposed for algorithmic assignment	
Asymmetric brightness	Integrated flux ratio between source sections.
Asymmetric structure	Symmetric components around host.
Compact	Angular extent of the components.
Diffuse	Proportion of assembly mask with emission.
Double	A 'component' number of two.
Edge brightened	Relative radial brightness distribution.
Extended	Angular extent of the source.
Faint	Integrated relative flux.
Host	Whether or not a host is identifiable.
Peak	Peak within the assembly mask.
Small	Angular extent of assembly mask.
Traces host galaxy	Assembly mask and host emission correlation.
Proposed for tagging	
Amorphous, bent, bridge, core, hourglass, jet, lobe, merger, plume, and tail	

In general, our testers found most provided tags self-explanatory. The main concern which the testers raised, was around the definition of three words they were not very familiar with: 'plume', 'tail', and 'elongated'. We believe there are two main contributions to this phenomena:

(i) Our testers were all non-native English speakers, which is likely to explain their struggle with the meaning of 'plume'.
(ii) The testers showed differences in their thinking around terms. For example, this included describing 'elongated' as 'extended', 'tail of a comet', 'oval shape', or a 'jet-like structure'.

To address these concerns, the final workflow will contain examples and conceptual definitions (see Appendix A), which users can reference for guidance. Furthermore, the RGZ EMU team is considering the translation of the tags into multiple languages, where issues such as this may be less relevant.

5 RADIO ASTRONOMY CHALLENGES AND SEMANTIC MORPHOLOGIES

The semantic taxonomy derived in this work (Section 4.4) is expected to be most useful as a tool by which astronomers can select samples of radio galaxies from source catalogues in a flexible manner. Assuming each of the tags is present or not, we can estimate how many populations can be selected. For the full taxonomy of 22 tags, $2^{22} = 4194\,304$ populations can be selected ($2^{10} = 1024$ for the 10 tags that RGZ EMU citizen scientists will use; see Section 4.5).

In practice, the number of populations that the tags map may be quite different. For example, the binary estimate presented here does not consider the use of other catalogued features, such as flux, spectral index, or redshift. It also does not take into account that certain tags may be fundamentally correlated. Additionally, one might expect that given enough data, catalogues containing vote fractions for each tag could enable uncertainty and strength estimates, i.e. how 'bent' a source might be could be approximated by the fraction of citizen scientists which return the tag for that source.

To demonstrate the utility of such semantically selected samples, we here synthesize a catalogue and perform some example selections. To synthesize our small data set into a catalogue, we estimate the tags for this catalogue by considering a source to have been assigned a given semantic tag if at least one of its annotations maps on to one of the tags in our final taxonomy. In this way, we treat our source annotations as a tagged catalogue.[7] Future catalogues will likely improve upon this synthesized catalogue through multiple individuals making direct use of available semantic tags.

We use this pseudo-catalogue to demonstrate the impact that catalogues using a semantic taxonomy can have by considering two practical use cases. First, we demonstrate the recovery of an existing population of radio galaxies in Section 5.1. We then highlight our ability to find morphological outliers in Section 5.2.

5.1 Detecting traditional populations

We demonstrate how traditional populations can be recovered by recovering star-forming galaxies. We do this by considering sources tagged with *traces host galaxy*. In practice, we query our data for sources, which were originally tagged with 'trace'. This is the closest proxy to 'traces host galaxy' tag that we can produce with our current data (see Sections 4.2 and 4.3).

By simply considering sources with the 'trace' tag, we identify 38 objects. These are listed along side their respective expert SFG classification and estimated tags in Table 2. This simple approach recovers 33 of the 45 sources labelled as SFGs by our experts (with at least 60 per cent expert agreement).

Five sources with the 'trace' tag were not classified as SFGs in our expert classification scheme, i.e. 34–38 in Table 2. Images of these sources are presented in Fig. 8, where their radio contours are shown overlaid on optical data from the Dark Energy Survey (DES; Abbott et al. 2018). Combining the deeper and higher resolution DES optical data into an RGB image aids visual interpretation compared to using DSS greyscale images. Now the primary choice for the EMU Zoo project, DES data were not initially used due to concerns about accessibility and coverage. Each of these sources is discussed individually below with respect to the optical morphology catalogues presented in Walmsley et al. (in preparation) made with the Zoobot[8] package (as described in Walmsley et al. 2022), where the percentage of people who would have answered with a given feature is stated behind each catalogued feature.

Source 34 is a smooth (78 per cent) cigar-shaped (70 per cent) galaxy (could be an edge-on galaxy). Source 35 is a featured (82 per cent) face-on (98 per cent; not edge-on) spiral (73 per cent) galaxy without a bar (70 per cent). Due to selection cuts, Source 36 was not included in the Walmsley et al. (in preparation) catalogues; however, the radio emission is expected to stem from at least two small galaxies bounded by the contours in the image. Source 37 is a smooth (67 per cent) round (59 per cent) galaxy. Source 38 is a featured (91 per cent) face-on (98 per cent; not edge-on) spiral (98 per cent) galaxy. It does not have a bar (71 per cent) but has a small bulge (80 per cent) and tightly wound spiral arms (83 per cent).

Consequently, of the five sources initially not classified as SFGs, when reconsidered with the deeper DES (Abbott et al. 2018) images and optical morphology catalogues, it is clear that at least two sources (35 and 38) are star-forming spiral galaxies. It is likely that the experts simply did not feel confident in classifying these sources as SFG with the limited resolution and sensitivity of the DSS data.

[7] Available at: https://github.com/mb010/Text2Tag/blob/main/data/mock_catalogue.csv
[8] https://github.com/mwalmsley/zoobot

Table 2. Sources selected for the *traces host galaxy* tag ('trace' in practice; see Section 5.1). Using an expert threshold of 60 per cent (see Section 4.1) we confirm whether or not the sampled sources are SFGs. For each source, the tags are listed in alphabetical order.

No.	Coordinates (J2000)	Expert SFG classification	Tags
1	20 h 22 min 17 s − 55°42′52″	✓	Asymmetric structure, compact, double, faint, host, peak, **traces host galaxy**
2	20 h 22 min 31s − 55°16′45″	✓	Amorphous, extended, host, **traces host galaxy**
3	20 h 23 min 00 s − 54°59′23″	✓	Amorphous, bent, compact, core, host, **traces host galaxy**
4	20 h 23 min 12 s − 53°55′43″	✓	Compact, diffuse, host, tail, **traces host galaxy**
5	20 h 26 min 02 s − 55°36′02″	✓	Bent, bridge, double, faint, host, hourglass, peak, **traces host galaxy**
6	20 h 26 min 53 s − 53°56′33″	✓	Amorphous, diffuse, host, **traces host galaxy**
7	20 h 29 min 31 s − 56°44′34″	✓	Amorphous, extended, host, **traces host galaxy**
8	20 h 29 min 44 s − 57°57′21″	✓	Bridge, core, double, extended, faint, host, peak, **traces host galaxy**
9	20 h 31 min 29 s − 53°44′17″	✓	Amorphous, extended, faint, host, peak, **traces host galaxy**
10	20 h 31 min 52 s − 53°46′30″	✓	Amorphous, compact, core, extended, host, merger, peak, **traces host galaxy**
11	20 h 32 min 02 s − 53°37′56″	✓	Amorphous, compact, extended, host, merger, peak, **traces host galaxy**
12	20 h 34 min 02 s − 52°58′50″	✓	Diffuse, extended, faint, host, **traces host galaxy**
13	20 h 35 min 33 s − 57°22′44″	✓	Amorphous, compact, extended, faint, host, **traces host galaxy**
14	20 h 36 min 11 s − 57°09′38″	✓	Compact, core, extended, faint, host, peak, **traces host galaxy**
15	20 h 38 min 19 s − 54°05′49″	✓	Host, merger, peak, small, **traces host galaxy**
16	20 h 40 min 18 s − 55°16′18″	✓	Compact, diffuse, extended, host, **traces host galaxy**
17	20 h 40 min 36 s − 53°15′53″	✓	Bent, bridge, extended, host, merger, **traces host galaxy**
18	20 h 41 min 09 s − 55°28′19″	✓	Compact, double, extended, host, **traces host galaxy**
19	20 h 43 min 10 s − 53°27′55″	✓	Compact, faint, host, peak, small, tail, **traces host galaxy**
20	20 h 43 min 41 s − 57°02′09″	✓	Diffuse, extended, host, **traces host galaxy**
21	20 h 43 min 55 s − 57°20′04″	✓	Amorphous, compact, core, extended, faint, host, merger, peak, **traces host galaxy**
22	20 h 50 min 42 s − 55°47′58″	✓	Diffuse, extended, faint, host, **traces host galaxy**
23	20 h 53 min 05 s − 56°25′08″	✓	Amorphous, compact, extended, host, **traces host galaxy**
24	20 h 53 min 43 s − 54°02′26″	✓	Compact, extended, faint, host, peak, **traces host galaxy**
25	20 h 55 min 25 s − 54°45′40″	✓	Amorphous, asymmetric structure, compact, extended, host, **traces host galaxy**
26	20 h 55 min 35 s − 55°05′54″	✓	Bent, core, diffuse, extended, host, lobe, peak, tail, **traces host galaxy**
27	20 h 59 min 43 s − 53°58′52″	✓	Amorphous, compact, extended, host, **traces host galaxy**
28	20 h 59 min 56 s − 55°33′47″	✓	Diffuse, double, edge brightened, extended, faint, host, hourglass, peak, **traces host galaxy**
29	21 h 00 min 39 s − 54°29′13″	✓	Amorphous, compact, core, host, peak, **traces host galaxy**
30	21 h 01 min 13 s − 57°14′26″	✓	Compact, extended, faint, host, peak, small, **traces host galaxy**
31	21 h 01 min 49 s − 57°56′45″	✓	Amorphous, compact, diffuse, host, **traces host galaxy**
32	21 h 02 min 10 s − 55°04′42″	✓	Core, diffuse, extended, host, peak, **traces host galaxy**
33	21 h 06 min 19 s − 56°48′27″	✓	Asymmetric brightness, compact, diffuse, extended, host, **traces host galaxy**
34	20 h 22 min 36 s − 56°16′25″	×	Asymmetric structure, compact, double, faint, host, peak, small, tail, **traces host galaxy**
35	20 h 24 min 45 s − 56°20′47″	×	Asymmetric structure, compact, faint, host, small, **traces host galaxy**
36	20 h 47 min 46 s − 56°44′04″	×	Asymmetric structure, compact, core, diffuse, extended, host, tail, **traces host galaxy**
37	20 h 58 min 37 s − 57°56′39″	×	Compact, host, **traces host galaxy**
38	20 h 59 min 51 s − 57°51′21″	×	Bent, compact, core, double, extended, faint, host, peak, small, **traces host galaxy**

12 additional sources that were classified as SFGs with 60 per cent agreement amongst our experts, but that *did not* have the 'trace' tag assigned to them are presented in Table 3. If the selection had been made on 'traces host galaxy', rather than 'trace', then these sources would have been included in Table 2. This is because 'traces host galaxy' is derived through multiple tokens such as 'counterpart' (see Section 4.3).

With a consistent use of the 'traces host galaxy' tag, the remaining eight SFG sources from Table 3 (sources 5–12) are likely to be tagged as such, as their radio emission do largely trace the respective optical host (see Appendix B).

Of the 45 sources our experts classified as SFGs, the selection of *traces host galaxy* would have recovered at least 82 per cent (33 and four from Tables 2 and 3, respectively) of these sources, making it a strong candidate to select SFG populations from future catalogues and highlighting how such a catalogue can be used.

5.2 Rare source detection

We here highlight the flexibility and practicality of our taxonomy by considering a combination of tags that a radio astronomer might consider to be abnormal. We query to find a source which (a) appears to be a merger, (b) presents bridged features, and (c) is not faint. The result is a single entry: source 17 from Table 2, shown in Fig. 9.

This source is, as expected, an unusual object requiring expert followup. It is a composite of emission from a flocculant spiral face-on 2MASS galaxy (Skrutskie et al. 2006), plus apparently associated emission to its SE with no obvious separate optical counterpart. The burned out (blue) object at the southern edge of the contours is listed as two stars in the Gaia catalogues (Gaia Collaboration et al. 2016, 2021), and has no obvious connection to the radio structure. A very careful evaluation of the chances for serendipity, and the possible physical nature of this source, are beyond the scope of this paper.

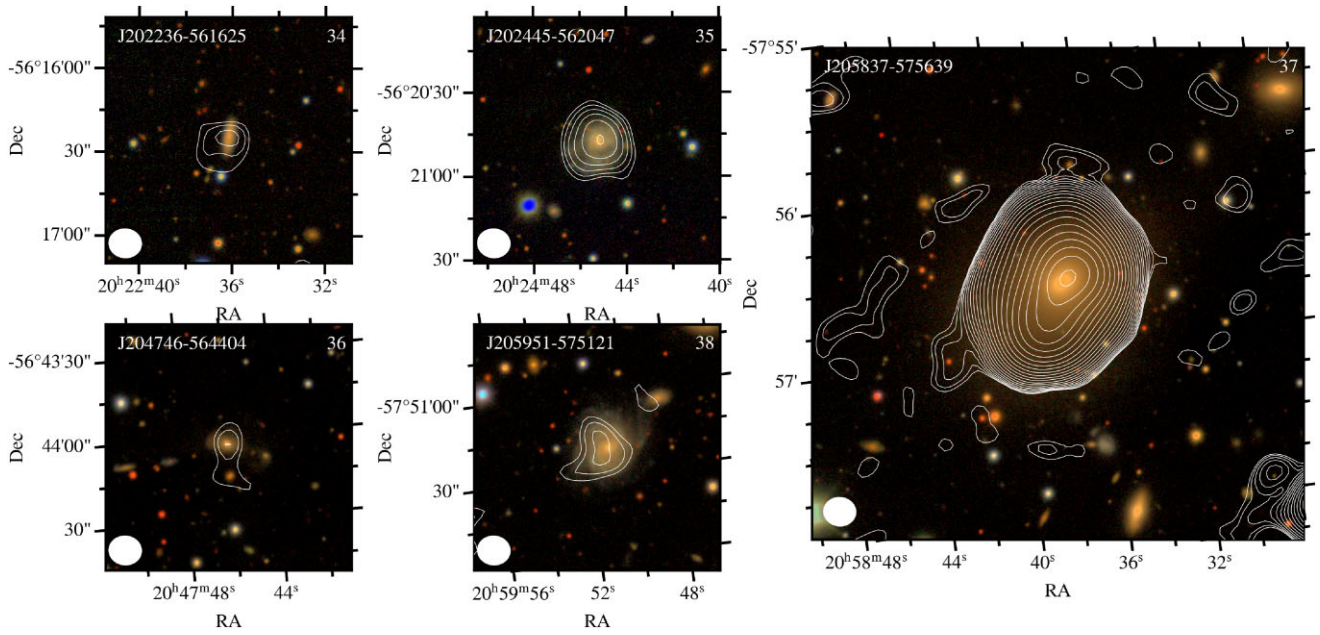

Figure 8. Sources 34–38 from Table 2 tagged with *trace* (selected as a proxy for 'traces host galaxy'), which are not classified as SFGs by experts. EMU radio brightness contours matching those in Fig. 1 with DES cutouts, combining g, r, and i band data into an RGB image following Lupton et al. (2004). Sources are annotated with their source numbers and (J2000) coordinates as presented in Table 2. Each source is shown to the same angular scale, as highlighted by the radio beam size in the bottom left-hand of each panel.

Table 3. Sources which had expert SFG classifications (above 60 per cent agreement) but were not selected through 'trace' as described in Section 5.1 and used for Table 2.

No.	Coordinates (J2000)	Tags
1	20 h 32 min 03 s − 53°44′35″	Amorphous, compact, core, extended, host, **traces host galaxy**
2	20 h 34 min 13 s − 54°01′30″	Asymmetric structure, compact, core, faint, host, tail, **traces host galaxy**
3	20 h 36 min 30 s − 57°07′19″	Asymmetric structure, compact, extended, host, small, **traces host galaxy**
4	20 h 44 min 36 s − 57°37′29″	Asymmetric structure, compact, extended, faint, host, small, tail, **traces host galaxy**
5	20 h 32 min 59 s − 55°38′04″	Compact, host, lobe
6	20 h 33 min 34 s − 54°31′22″	Amorphous, compact, extended, host, merger, peak
7	20 h 34 min 49 s − 54°12′39″	Amorphous, compact, diffuse, double, extended, host, hourglass
8	20 h 41 min 41 s − 56°10′21″	Amorphous, compact, double, faint, host, merger, peak
9	20 h 46 min 26 s − 54°00′51″	Diffuse, double, host
10	20 h 55 min 12 s − 54°31′25″	Amorphous, compact, extended, host, merger, small
11	20 h 56 min 44 s − 56°37′48″	Compact, diffuse, double, extended, host, merger
12	21 h 02 min 57 s − 54°29′35″	Amorphous, asymmetric structure, bent, core, diffuse, faint, host, peak, tail

However, the control provided by these semantic tags has allowed for the selection of an unusual object worthy of further study.

6 DISCUSSION AND IMPACT

6.1 Taxonomy

The proposed semantic tags will find immediate use in the RGZ EMU citizen science project (Tang & Vardoulaki et al., in preparation). For future implementations of science tags being assigned to sources, we suggest that the community uses a hash symbol to denote the use of a tag (e.g. '#compact') as suggested in Rudnick (2021) to distinguish from traditional classification frameworks. This should prove useful to the legibility and analysis of future catalogues and works.

This is the *first step of the tagging framework* in radio astronomy morphology. The taxonomy is intentionally designed to be extensible, such that when the community decides a feature of interest is not being captured by the current version of the taxonomy it can be updated to include the appropriate tag. The presented set of semantic tags are a first step towards mapping common features of radio morphologies using plain English annotations.

The *specificity* of the tags in comparison to the current classification scheme may be a concern to some astronomers. The inconsistency with which current radio morphological classifications are defined means that the language currently in use does not have the desired specificity either – regardless of how specific a term is in an individual astronomer's mind. Furthermore, we highlight that terms are expected to be used more consistently and clearly when selected directly rather than being derived through annotations.

A *science class mapping* using the semantic taxonomy is one of the goals of the RGZ EMU team. This mapping will be constructed towards the end of the RGZ EMU project. This mapping should be able to provide the traditional classification of objects by predicting them based on the tags that citizen scientists have assigned to objects.

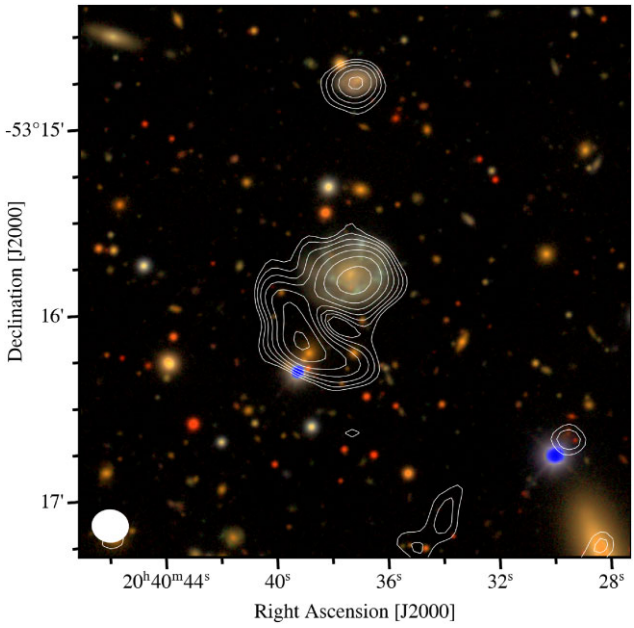

Figure 9. Rare source selected from the synthesized initial semantic tag catalogue by querying '*hourglass* \ *(amorphous* ∪ *traces host galaxy* ∪ *bent)*' (set theory notation). EMU radio brightness contours matching those in Fig. 1 ontop of a DES cutout prepared as in Fig. 8.

It will include the most common radio morphology science classes, which the community is more accustomed to. While it is hoped that the tags will span the full space of possible scientific classifications, there may be cases where the provided mapping does not cover a science case perfectly in its current form.

Regardless, the ability to combine the tags to select semantic populations, as demonstrated in Section 5, will enable feature specific population studies and sources to be omitted if they present features that are not relevant to a given science case.

6.2 Experiment and implementation

As described, the experimental set up of this work has a number of limitations, including the small size of the data set: given the degree of variation present across radio galaxy morphologies it is not possible to capture all abstract science classes of radio galaxies completely. For example, as stated in Section 3.2.3, ORCs are so rare that we could not train on them, and thus they are not taken into consideration in the weighted Shapley values. Such biases are therefore currently passed on to the derived and proposed semantic taxonomy.

Additionally, abstract science classes based primarily on morphology do not directly encode other physically relevant information. Given sufficient information, it may be more informative to derive semantic tags directly from physical parameters, e.g. active galactic nuclei accretion rates. This would encourage the derived semantic tags to carry information regarding the physics itself, rather than a proxy, e.g. abstract science classes. In practice, collecting a large enough sample with which to do this is not feasible at this time, but should be considered in future approaches and other domains.

We use a *pre-trained NLP model* in our approach to this task. This model is a limitation even though it is also a key factor in the success of our implementation and experiments. For instance, the model we used only returns individual 'tokens', and not fully grammatically correct terms or phrases. We amend this through manual inspection and adjustments, however, we recognize this is not a scalable solution, and will not be possible in all situations. We hope that future approaches will find solutions to this problem. The NLP literature is currently developing at a significant pace. We therefore urge future iterations and applications of this approach to actively re-consider which pre-trained model is used. We expect that by using the most up to date pre-trained model, future implementations will have more robust and versatile encodings of annotations and tags.

Finally, we note that the semantic radio morphology taxonomy derived in this work is inherently bound to the instrument with which the radio galaxies were imaged (ASKAP). Data from a more sensitive (e.g. SKA; Dewdney et al. 2009) or higher resolution (e.g. LOFAR long baseline Morabito et al. 2022) instrument might require additional or different semantic tags. This would be simple to implement under the proposed tagging paradigm, as it would be sufficient to supplement the existing taxonomy with the appropriate semantic tags.

6.3 Semantic taxonomies

The proposed method, and respective task of deriving semantically meaningful tags (first outlined in Bowles et al. 2022), have the potential to impact other fields. We therefore discuss their potential impact and limitations in a domain agnostic tone.The move away from technical classes to semantic language capturing features may have a broad impact across a number of technical fields, especially where complex classes have been defined and the field has since moved beyond those initially valuable classification schemes, as is the case in radio astronomy. This could include any feature rich data product, especially where features are often repeated across classes.

The *collaborative* nature of science may be improved by the use of simplified and semantic language. Complex ideas are often shrouded in equally complex terminology, which can be highly effective when experts communicate with one another, but quickly becomes a hindrance to communicating in any other situation. Capturing features of an object using dictionary level definitions will lower the barrier to entry for established researchers who are not domain experts to study the features captured by the semantic language. This can be a significant benefit, where domain specific terminology could be an active barrier to communication in inter-disciplinary research. Additionally, the use of plain English should enable scientific collaborations within a given field, i.e. between radio astronomy domain experts and astronomers who are not experts in radio morphology.

Outreach efforts are also likely to benefit from the change to language. We hope that the simple language will reduce the barrier to entry for those who would like to become experts in the respective field. This will have direct impact on the accessibility of technical fields as a whole including communities who have not had much practice in the use of scientific language. This is in perfect alignment with the educational aspects of citizen science, which are often used to engage underprivileged communities in science with the aim to inspire and empower. The hope of citizen science outreach is that students who have seen, interacted with, and subsequently added to the international body of science feel empowered to pursue STEM subjects. Clearer language will improve engagement to support this goal.

The science in *citizen science* projects is also expected to benefit from the new language. Easily understood concepts presented by the simplified language should lead to improved usage of tags for a given source (Wald, Longo & Dobell 2016). Additionally, the reduction in training time/effort of the citizen scientists is hoped to lower the

labelling cost for projects as a whole by reducing the labelling cost for individual citizen scientists.

Deep learning and machine learning models currently learn to predict scientific classes from images (or similarly high dimensional data). Learning to encode these classes can be quite challenging as the concepts and definitions represented by these classes can be both abstract and contentious. This may be partially addressed by training models to encode a semantic taxonomy instead, as models would learn the features instead of abstracted classes. Derived semantic taxonomies presents more clearly defined concepts and may encourage models to learn a more robust feature space. This could improve the effectiveness of the encoded features of a model for various other tasks. Even in a simple use case, the more robust feature space may allow models to be more generalizable and less brittle, i.e. transferable or fine-tunable to differing data sets and tasks, which would be of immediate benefit to a number of applications.

The *anglocentric* nature of this work was alluded to previously. Although the language improvements, may benefit many populations, it does still marginalize those who do not speak English natively. The RGZ EMU team is considering a number of strategies to mitigate the effect of anglocentric labelling, including translation into multiple languages. However, we recognize that there are broader complex issues around use of language in science and recommend this as a topic of discussion in future work.

Finally, the *ethics* of deriving terms from an unstructured data set include careful consideration of the potential presence and impact of malicious agents. Caution is therefore advised when applying this process to other fields. In this work, the data set used was small enough that each annotation was inspected individually.

7 CONCLUSIONS

In this work, we derive a flexible English taxonomy for radio astronomy and the respective morphological tagging. The proposed taxonomy of 22 semantic tags is

(i) the product of experiments collecting expert classifications and plain English annotations on radio source morphologies using selected cutouts from the EMU pilot survey,
(ii) reduced to a set of 10 terms to maximize its effectiveness within citizen science projects, starting with RGZ EMU,
(iii) derived analytically through a novel method with minimal clarifying intervention.

We demonstrate the first effective use cases of the newly derived semantic morphology taxonomy. We show that using the tags we can recover

(i) known scientific morphologies, and
(ii) rare sources with abnormal morphologies.

The method which was developed, detailed, and applied in this work is domain agnostic. The method

(i) provides a framework through which plain English annotations of complex ideas can return a ranked taxonomy on a given subject,
(ii) can be applied to any scenarios where language is a barrier to future research,
(iii) can increase the accessibility of complex scientific concepts by distilling concepts into simpler English for the public, collaborators, and citizen scientists.

The potential scientific impacts, applications, and communication benefits of this method and taxonomy are discussed at length in Section 6.

ACKNOWLEDGEMENTS

MB, MW, ELA, AMS, and IVS gratefully acknowledge support from the UK Alan Turing Institute under grant reference EP/V030302/1. HT gratefully acknowledges the support from the Shuimu Tsinghua Scholar Program of Tsinghua University. EV acknowledges support by the Carl Zeiss Stiftung with the project code KODAR. ELA additionally gratefully acknowledges support from the UK Science & Technology Facilities Council (STFC) under grant reference ST/P000649/1. AD acknowledges support by the BMBF Verbundforschung under the grant 05A20STA. DL acknowledges support from the Natural Sciences and Engineering Research Council of Canada.

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

This research made use of a number of tools and software packages. Including 'Aladin sky atlas' developed at CDS, Strasbourg Observatory, France (Boch et al. 2014; Bonnarel et al. 2000), scientific color maps (Crameri 2021), SciPy (Hagberg et al. 2008), SpaCy (Honnibal and Montani 2017), Matplotlib (Hunter 2007), pandas (McKinney 2010; Pandas Development Team 2020), sci-kit learn (Pedregosa et al. 2011), and seaborn (Waskom 2021).

The Australian SKA Pathfinder is part of the Australia Telescope National Facility, which is managed by CSIRO. Operation of ASKAP is funded by the Australian Government with support from the National Collaborative Research Infrastructure Strategy. ASKAP uses the resources of the Pawsey Supercomputing Centre. Establishment of ASKAP, the Murchison Radio-astronomy Observatory, and the Pawsey Supercomputing Centre are initiatives of the Australian Government, with support from the Government of Western Australia and the Science and Industry Endowment Fund. We acknowledge the Wajarri Yamatji as the traditional owners of the Observatory site.

This project used public archival data from the DES. Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, the Center for Cosmology and Astro-Particle Physics at the Ohio State University, the Mitchell Institute for Fundamental Physics and Astronomy at Texas A&M University, Financiadora de Estudos e Projetos, Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Científico e Tecnológico and the Ministério da Ciência, Tecnologia e Inovação, the Deutsche Forschungsgemeinschaft, and the Collaborating Institutions in the DES.

The Collaborating Institutions are Argonne National Laboratory, the University of California at Santa Cruz, the University of Cambridge, Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas-Madrid, the University of Chicago, University College London, the DES-Brazil Consortium, the University of Edinburgh, the Eidgenössische Technische Hochschule (ETH) Zürich, Fermi National Accelerator Laboratory, the University of Illinois at Urbana-Champaign, the Institut de Ciències de l'Espai (IEEC/CSIC), the Institut de Física d'Altes Energies, Lawrence Berkeley National Laboratory, the Ludwig-Maximilians Universität München and the

associated Excellence Cluster Universe, the University of Michigan, the National Optical Astronomy Observatory, the University of Nottingham, The Ohio State University, the OzDES Membership Consortium, the University of Pennsylvania, the University of Portsmouth, SLAC National Accelerator Laboratory, Stanford University, the University of Sussex, and Texas A&M University.

Based in part on observations at Cerro Tololo Inter-American Observatory, National Optical Astronomy Observatory, which is operated by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation.

We thank the citizen scientists for their time and effort. We especially thank Victor Linares Sisifolibre for their substantial contributions.

DATA AVAILABILITY

All the data used in this work is publicly available. The cutouts presented to users for annotation, anonymized annotations, and anonymized expert classifications are available at https://zenodo.org/record/7254123#.Y3d5EdLP2xE. The code written for these projects are available under https://github.com/mb010/Text2Tag.

REFERENCES

Abbott T. M. C. et al., 2018, ApJS, 239, 18
Adajian T., 2022, in Zalta E. N.ed., The Stanford Encyclopedia of Philosophy, Spring 2022 edn. Metaphysics Research Lab, Stanford University, Stanford
Banfield J. K. et al., 2015, MNRAS, 453, 2326
Boch T., Fernique P., 2014, in Manset N., Forshay P., eds, ASP Conf. Ser. Vol. 485, Astronomical Data Analysis Software and Systems XXIII. Astron. Soc. Pac., San Francisco, p. 277
Bonaldi A., Bonato M., Galluzzi V., Harrison I., Massardi M., Kay S., De Zotti G., Brown M. L., 2019, MNRAS, 482, 2
Bonnarel F. et al., 2000, A&AS, 143, 33
Bowles M. et al., 2022, preprint (arXiv:2210.14760)
Crameri F., 2021, Scientific colour maps, Zenodo, available at: https://doi.org/10.5281/zenodo.5501399
Dewdney P. E., Hall P. J., Schilizzi R. T., Lazio T. J. L. W., 2009, IEEE Proc., 97, 1482
Fanaroff B. L., Riley J. M., 1974, MNRAS, 167, 31P
Gaia Collaboration, 2016, A&A, 595, A1
Gaia Collaboration, 2021, A&A, 649, A1
Grezes F. et al., 2021. preprint (arXiv:2112.00590)
Hagberg A. A., Schult D. A., Swart P. J., 2008, in Varoquaux G., Vaught T., Millman J.eds, Proc. 7th Python in Science Conference. Pasadena, p. 11
Hallinan G., Ravi V., Deep Synoptic Array Team, 2021, in American Astronomical Society Meeting Abstracts #237. p. 316.05
Hardcastle M. J., Croston J. H., 2020, New Astron. Rev., 88, 101539
Hewish A., Bell S. J., Pilkington J. D. H., Scott P. F., Collins R. A., 1968, Nature, 217, 709
Honnibal M., Montani I., 2017, spaCy 2: natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, in press
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Johnston S. et al., 2008, Exp. Astron., 22, 151
Jonas J., MeerKAT Team, 2016, in Proc. Sci., MeerKAT Science: On the Pathway to the SKA. SISSA, Trieste, PoS#001
Kapińska A. D. et al., 2017, AJ, 154, 253
Lasker B. M., Sturch C. R., McLean B. J., Russell J. L., Jenkner H., Shara M. M., 1990, Astron. J., 99, 2019
Lundberg S. M., Lee S.-I., 2017, in Guyon I., Luxburg U. V., Bengio S., Wallach H., Fergus R., Vishwanathan S., Garnett R.eds, Advances in Neural Information Processing Systems 30. Curran Associates, Inc., New York, p. 4765, http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
Lundberg S. M. et al., 2020, Nat. Mach. Intell., 2, 2522
Lupton R., Blanton M. R., Fekete G., Hogg D. W., O'Mullane W., Szalay A., Wherry N., 2004, Publ. Astron. Soc. Pac., 116, 133
McKinney W., 2010, in van der Walt S., Millman J.eds, Proc. 9th Python in Science Conference. p. 56
Menze B. H., Kelm B. M., Masuch R., Himmelreich U., Bachert P., Petrich W., Hamprecht F. A., 2009, BMC Bioinformatics, 10, 1
Mishra B. K., Kumar R., 2020, Natural Language Processing in Artificial Intelligence. CRC Press, Boca Raton
Morabito L. K. et al., 2022, A&A, 658, A1
Murphy E., ngVLA Science Advisory Council, 2020, in American Astronomical Society Meeting Abstracts #235. p. 364.01
Natal J., Ávila I., Tsukahara V. B., Pinheiro M., Maciel C. D., 2021, Entropy, 23, 1340
Norris R. P. et al., 2011, PASA, 28, 215
Norris R. P. et al., 2021a, PASA, 38, e003
Norris R. P. et al., 2021b, PASA, 38, e046
Pandas Development Team, 2020, pandas-dev/pandas: Pandas, Zenodo, available at: https://doi.org/10.5281/zenodo.3509134
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Pennington J., Socher R., Manning C. D., 2014, in Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, p. 1532, https://aclanthology.org/D14-1162
Romm J., 1991, Sixteenth Cent. J., 22, 173
Rudnick L., 2021, Galaxies, 9, 85
Von Schelling F. W. J., 1994, On the History of Modern Philosophy. Cambridge Univ. Press
Schoenmakers A. P., de Bruyn A. G., Röttgering H. J. A., van der Laan H., Kaiser C. R., 2000, MNRAS, 315, 371
Skrutskie M. F. et al., 2006, AJ, 131, 1163
Southworth G. C., 1956, Sci. Monthly, 82, 55
Thomas B., Thronson H., Buonomo A., Barbier L., 2022, Res. Notes AAS, 6, 11
Tingay S. J. et al., 2013, PASA, 30, e007
van Haarlem M. P. et al., 2013, A&A, 556, A2
Vayansky I., Kumar S. A. P., 2020, Inf. Syst., 94, 101582
Wald D. M., Longo J., Dobell A. R., 2016, Conser. Biol., 30, 562
Walmsley M. et al., 2022, MNRAS, 509, 3966
Waskom M. L., 2021, J. Open Source Softw., 6, 3021
Wolff P., Holmes K. J., 2010, Ltd WIREs Cogn Sci, 2, 253
Wright E. L. et al., 2010, AJ, 140, 1868

APPENDIX A: TAG DEFINITIONS

The below are the tags which will be used in RGZ EMU, as discussed in Section 4.5. Here, we present the same definitions of the tags as we are planning on providing to the citizen scientists:

(i) *Amorphous*: having no clearly defined shape or form.
(ii) *Bent*: curved or having an angle.
(iii) *Bridge*: a structure that connects from one side to another (not a jet; see below).
(iv) *Core*: a central part distinct from the enveloping part.
(v) *Hourglass*: shaped like an hourglass.
(vi) *Jet*: a narrow stream of material appearing to emanate from a celestial object.
(vii) *Lobe*: a roundish projecting part of something divided by a fissure/gap.
(viii) *Merger*: multiple separate things, which appear to be connected or connecting.
(ix) *Plume*: a long cloud of smoke or vapour resembling a feather as it spreads from its point of origin.

(x) *Tail*: resembling an animal's tail in its shape or position, typically extending downwards or outwards at the end of a thing.

The tags which are proposed for algorithmic assignment according to Table 1 are not defined here. They will likely be defined when respective algorithms are developed.

APPENDIX B: STAR-FORMING GALAXIES WITHOUT 'TRACES HOST GALAXY' TAG

Fig. B1 presents the sources which were classified (above 60 per cent agreement) as star-forming galaxies, but were not tagged with 'trace' in our synthetic catalogue, as described in Section 5.1. Table B1 presents the optical morphologies of the sources listed in Table 3 produced by Walmsley et al. (in preparation), and made with Zoobot[9] (initially described in Walmsley et al. 2022). Here, source numbers align with those presented in Table 3.

[9]https://github.com/mwalmsley/zoobot

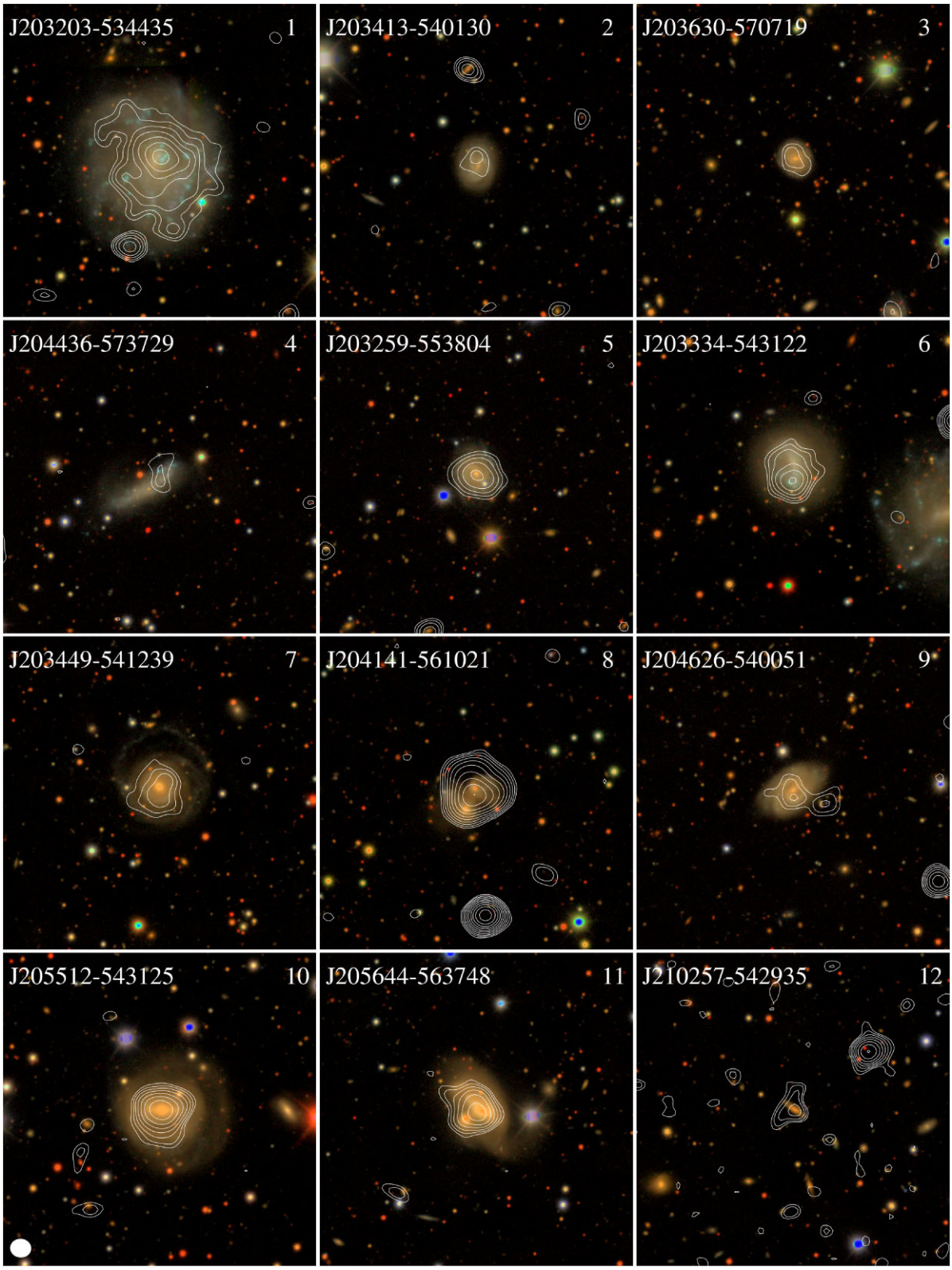

Figure B1. Composite images of SFG sources not captured by the 'trace' tag, as discussed in Section 5.1. EMU contours following Fig. 1 with optical DES RGB backgrounds as in Fig. 8. Cutout centre coordinates are presented on the image along side their respective source numbers associated with Tables 3 and B1. The radio beam size for all panels is shown in the lower left-hand of the figure; all cutouts are $3' \times 3'$.

Table B1. Optical morphologies from Walmsley et al. (in preparation) of the SFG sources not captured by the 'trace' tag, as discussed in Section 5.1 and presented in Table 3.

No.	Coordinates (J2000)	GZ morphology predictions
1	20 h 32 min 03 s -53°44′35″	Featured (74 per cent; not edge-on), face-on (98 per cent), spiral (76 per cent), no bar (59 per cent), small bulge (59 per cent)
2	20 h 34 min 13 s -54°01′30″	Featured (72 per cent), face-on (99 per cent; not edge-on), spiral (70 per cent), no bar (60 per cent), no bulge (69 per cent)
3	20 h 36 min 30 s -57°07′19″	Featured (89 per cent), face-on (97 per cent; not edge-on), spiral (98 per cent), small bulge (85 per cent), tight arms (76 per cent)
4	20 h 44 min 36 s -57°37′29″	Not in catalogue.
5	20 h 32 min 59 s -55°38′04″	Featured (87 per cent), face-on (98 per cent; not edge-on), spiral (98 per cent), no bar (77 per cent), small bulge (83 per cent), tight arms (84 per cent)
6	20 h 33 min 34 s -54°31′22″	Featured (70 per cent), face-on (95 per cent; not edge-on), no spiral (68 per cent)
7	20 h 34 min 49 s -54°12′39″	Featured (82 per cent), face-on (97 per cent; not edge-on), spiral (85 per cent), no bar (76 per cent), moderate bulge (58 per cent), tight arms (63 per cent)
8	20 h 41 min 41 s -56°10′21″	Featured (53 per cent), face-on (96 per cent; not edge-on), no spiral (77 per cent), no bar (84 per cent), merger (67 per cent)
9	20 h 46 min 26 s -54°00′51″	Featured (83 per cent), face-on (98 per cent; not edge-on), spiral (85 per cent), small bulge (57 per cent)
10	20 h 55 min 12 s -54°31′25″	Not in catalogue.
11	20 h 56 min 44 s -56°37′48″	Featured (83 per cent), face-on (96 per cent; not edge-on), spiral (92 per cent), no bar (52 per cent), moderate bulge (59 per cent), tight arms (68 per cent)
12	21h 02 min 57 s -54°29′35″	Smooth (67 per cent), cigar shaped (88 per cent)

APPENDIX C: TAG RANKINGS

To demonstrate the contributions of our tags to the classification of a given science class, we present Fig. C1, which shows the sorted most important tags for the SFG class. Equivalent plots for each class are available in our public repository: https://github.com/mb010/Text2Tag/data/AppendixB.

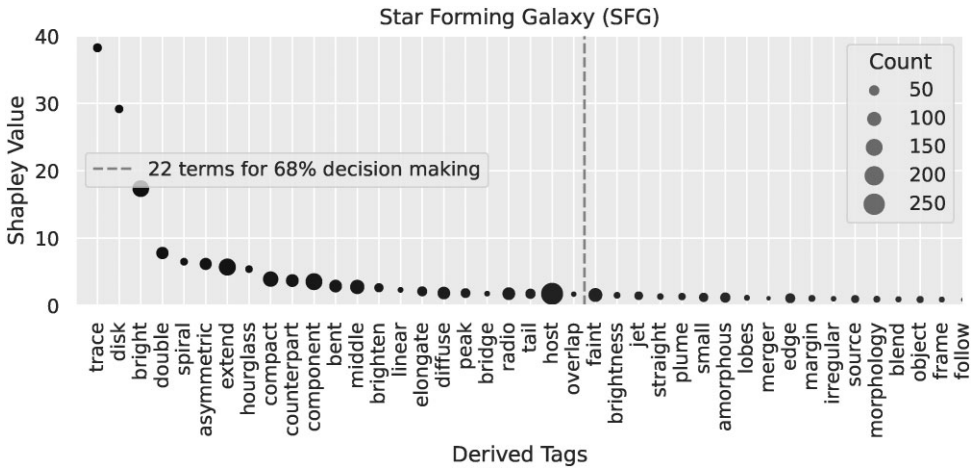

Figure C1. Class wise top Shapley value ranked tags for the SFG science class.

This paper has been typeset from a TEX/LATEX file prepared by the author.