Creative Industries
Policy & Evidence Centre
Led by nesta

Charities speak

Mapping arts and cultural charities in England and Wales using data science

Raphael Leung
raphael.leung@nesta.org.uk

March 2020

Summary

This report analyses the activities and objectives of registered charities in England and Wales involved in 'arts, culture, heritage or science' (ACHS) with a particular focus on arts and culture. The analysis applies natural language processing and clustering techniques to the information that charities provide when they register. This allows for a more detailed understanding of what charities are doing, and what they are trying to achieve, than is available from existing classifications. The work produces an automatically-generated taxonomy of keywords used by ACHS charities. The taxonomy is then applied to create the first systematic mapping of the different activities that charities are supporting and the groups they engage with, for example, the number of performing arts charities working to engage women or specific ethnic minorities. Future work can extend these methods to map other parts of the charitable sector or build recommendation engines to allow funders and charity workers to find others promoting similar causes.

Creative Industries
Policy & Evidence Centre
Led by **nesta**

Contents

1

Introduction

Charities affect all walks of life – they care for underprivileged people; they provide spaces for communities both urban and rural; they fundraise for, and invest in, causes; they support interests both general and specialist. They range from large foundations to small meeting groups.

Charities are a core part of our society and economy. The annual income of the charities sector in England and Wales was an estimated £113.1 billion.[i] According to surveys conducted in 2018, almost 3 per cent of the total UK workforce work in the voluntary sector[ii] and one-in-five people volunteer at least once a month.[iii] Charities are also an important part of the arts and cultural sector with many venues, groups, and institutions organised on a charitable basis. In one way or another, we all interact with charities.

But how well do we understand what these charities do, and what they are trying to achieve? In this paper, we use data science methods to map the charitable sector at scale, focusing on the activities that charities advancing arts and cultural causes are undertaking and the objectives that they are trying to fulfill.

1.1 Why study charitable missions?

There is widely considered to be a growing role for charities in so-called mission-driven policies to tackle society's grand challenges. A challenge is 'a broadly defined area which a nation may identify as a priority (whether through political leadership, or the outcome of a movement in civil society).'[iv, v] The Department for Business, Energy and Industrial Strategy (BEIS) set out the first four Grand Challenges in 2019 in the Industrial Strategy – artificial intelligence and data, ageing society, clean growth, and the future of mobility.

Using data science techniques to extract and identify missions may allow policymakers to track and monitor the charitable ecosystem. If systematic data is collected over time, it may inform the development of areas for support of the sector. Mapping missions across existing data sources, like the charitable sector, can help policymakers understand how charities are engaging with particular issues.

The systematic mapping of missions also has benefits for accountability and advocacy. While grand challenges focus 'on the global trends which will transform our future,'[vi] challenges identified on the grassroots level by charities may reflect different priorities and be phrased using different language. Being able to identify missions that have emerged on the ground, from charities documentation or elsewhere, has promising applications for policymakers and those in civil society: from tracking the role that charities are playing to enabling greater transparency for causes that have been funded and are yet to be funded.

Impact investors may find that better data on missions helps with due diligence, landscaping/scoping, and impact assessment.[vii] In particular, the practice of impact investment puts emphasis on quantifying the amount of social or environmental impact that is being delivered via investments, either directly or indirectly. A data-driven approach to missions tracking will help expand the evidence toolkit for impact investors.

1.2 How can existing classifications of charities be improved?

There are two main systems that have been used to classify charities in England and Wales – one from the Charity Commission of England and Wales (CCEW) and one from the United Nations (UN) called the International Classification of Non-profit and Third Sector Organizations (ICNP/TSO). We explain these in greater depth in the Appendix.

The existing classifications are useful but have limitations. There is potential scope to make our understanding of charities more granular, flexible and timely.

1. More granular

The existing categories are broad. For example, in accordance with the official charitable purposes, all of 'arts, cultural, heritage, or science' (ACHS) is covered by one umbrella term that is not broken down further.

The categories often combine activities that are relatively distinct. For example, museums and historic sites are grouped together in ICNP/TSO.

2. More flexible

It is difficult to explore sub-categories of charitable activity. For example, we may want to know how many charities promote exercise for the elderly through dance, or the location of community centres promoting integration of refugees through sports or performance arts. The existing CCEW classification does not allow for this functionality.[1]

We show that a taxonomy created using data science techniques can help contribute to a better understanding of the sector.

3. More timely

Other granular classifications like the UN's are manually curated. Updating them requires expert knowledge which can be labour-intensive.

Such classifications can therefore be slow in reflecting societal change. For example, ICNP/TSO was updated to have more subdivision codes in December 2017, but it takes time for users to adopt updated versions.[ii] In contrast, we show that data science methods can be used to create a taxonomy both efficiently and quickly.

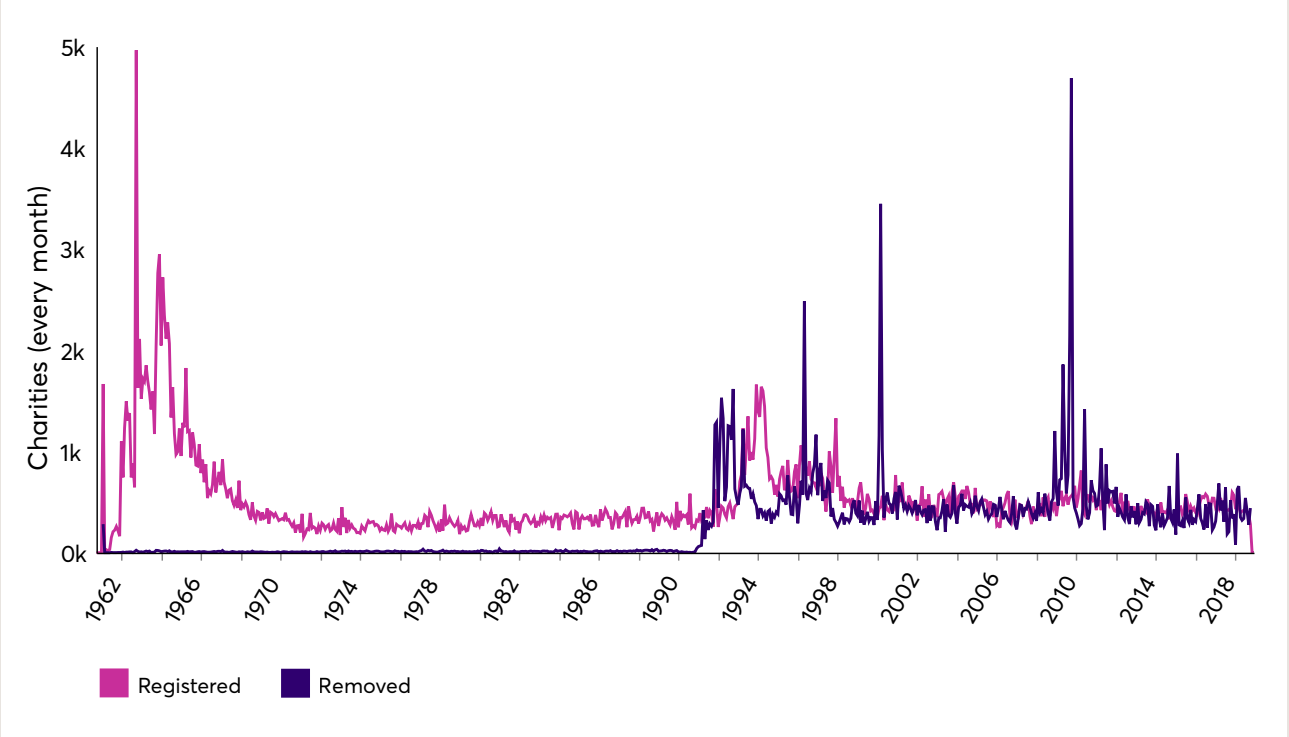
1. It is also not possible to discover charitable areas if the user does not already know what terms to search for, or have sufficient knowledge of the domain to search for related concepts that can return more useful results.

1.3 Why study change in the charitable sector?

The charitable sector is not static and undergoes constant change. For example, from 2010-2016, 469 charities were registered on average every month in England and Wales (and with an average of 425 charities removed every month as well, owing to reasons such as the charity no longer existing, it being amalgamated, or its funds being transferred).

As charitable objects are written at the point of registration, and are not edited since, we use registration and removal rates to show why registration texts are a reflection of society at various points in time. The monthly charity registration and removal rates are shown in Figure 1, with spikes representing bulk registrations and removals.[2]

Figure 1: Charities registration and removal rates in England and Wales (1962-2018)



Source: The Charity Commission for England and Wales.

A classification that is more adaptable to reflect changes should allow us to have a more up-to-date understanding of the sector, enabling emerging or declining trends in charitable work to be identified more easily.

2. There are some spikes in the data for both charity registration and removal. 4963 and 2948 charities were registered in September 1962 and November 1963 respectively. Those months were historically the highest. A variety of charities were registered in both of those months. While there were small groups of charities registered in September 1962 that shared the same name – 20 charities registered under 'Unknown Donor' and 18 under 'Fuel Allotment' – they only accounted for 0.4% and 0.3% of that month's registrations. The spikes in registration were driven by many charities using the newly implemented registration process. The historical highs for charity removals came in September 2009 and February 2000, with 4687 and 3445 charities removed respectively. 'Does not operate' was cited as the most common removal reason in both cases, accounting for 92% and 84% of removals of that month respectively. Across the entire dataset, 'Does not operate' usually only accounts for 16% of removals ('Ceased to exist' is the most common reason generally, cited by 51% of all removals from 1961 to 2018). As such, the removal spikes may be explained by the Commission removing non-operational charities from the register in batches.

1.4 Objectives of the report

This applies data science methods to charities' data to enhance our understanding of charities operating in the arts and culture in England and Wales. It has two specific aims:

1. To produce a detailed picture of arts and cultural charities using automated techniques, going beyond existing static classifications, and
2. To better understand what charities (say they) do and what they are trying to achieve.

Officially, as in other jurisdictions, to be registered in England and Wales charities must have 'charitable purposes' that help the public (known as being 'for public benefit'). Both are legal definitions under charity law and charities are also subject to regulations.[3] However, there are also non-registered charities (excepted or exempt charities),[4] unincorporated charitable associations, charitable trusts, charitable companies, community interest companies, and industrial and provident societies. Significant voluntary and charitable work happens outside of registered charities. In this paper, we look at registered charities only.

In this report, we restrict our analysis to a charity's 'aims and activities' and 'charitable objects', as described in text provided in two fields when a charity in England or Wales registers. For active ACHS charities where websites are available, we also include text scraped from charities' websites.

3. There are three charity regulators in the UK. The largest one, in terms of charities regulated, is the Charity Commission for England and Wales (CCEW). It started registering charities in 1961 and that register currently has over 160,000 charities. Anyone can look up the charity register on their website: https://beta.charitycommission.gov.uk
4. Certain churches, scout and guide groups, and student unions are excepted charities, whereas some universities and museums are exempt charities. See the official guidance for the definitions of 'exempt' and 'except'.[viii, iv]

2

Data

We collect official charities data from the CCEW website via web-scraping in September 2018 (n=359,245). This includes all charities ever registered, including charities that are 'linked'.[5] The charity numbers at the time of data access are:

Table 1: Data from CCEW

	Register maintained by	Data source(s)	Total entries	Active charities[6] (as a % of all entries)	ACHS charities (as a % of active charities)
England and Wales	Charity Commission for England and Wales (CCEW)	CCEW	359,245	208,057 (57.9%)	30,418 (14.6%)

For active ACHS charities where websites are listed on the register (n=19,916), we also include text scraped from their website main page.

5. Linked charities are closely connected charities that prepare only one set of aggregated annual accounts.
6. For CCEW, active was defined as all charities that were not listed as removed. This includes four other categories (up-to-date, out-of-date, recently registered and linked charities).

3

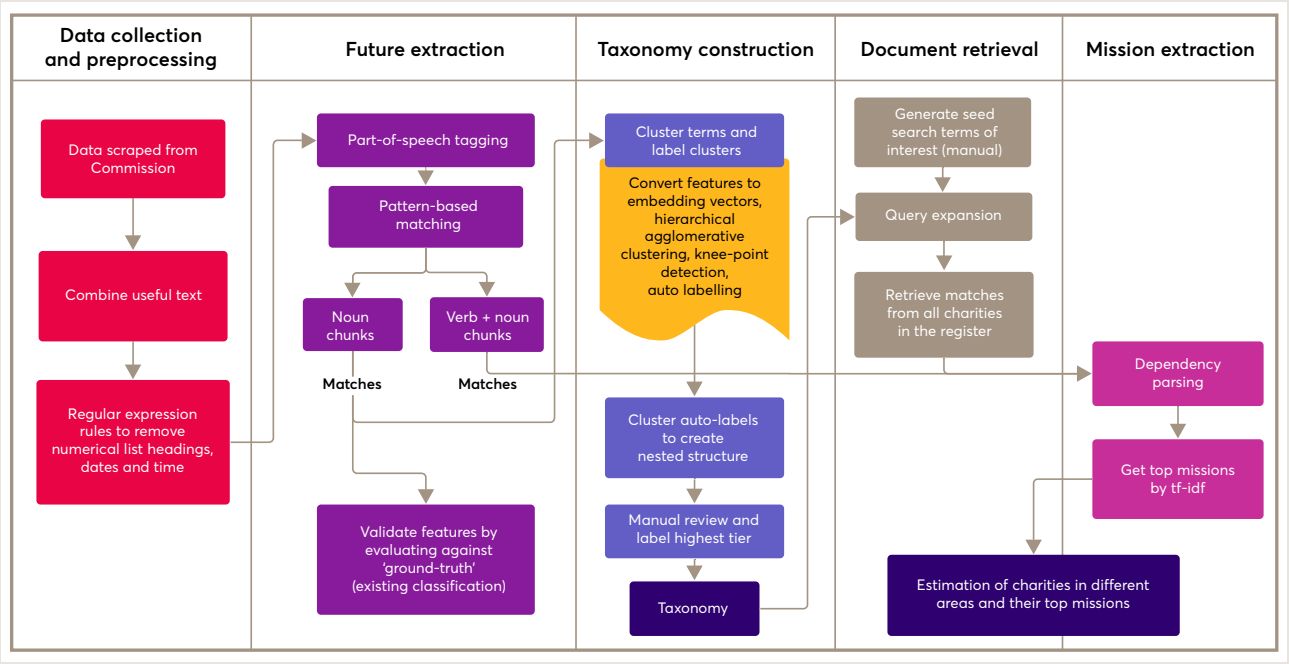
Methodology

We summarise and explain the motivation for the research methods used in this section. The analysis involves:

- Data preprocessing.
- Using part-of-speech analysis and pattern-based matching to identify relevant phrases.
- Using the existing classification to validate our features.
- Creating a taxonomy of arts and cultural charitable terms.
- Identifying relevant charities using the taxonomy.
- Using dependency parsing to extract charitable missions.

Figure 2 sets out the different stages and the different techniques, which are then discussed in more detail below.

Figure 2: Methodological pipeline



3.1 Data preprocessing

For each charity, we select two text fields - a charity's 'aims and activities' and its 'charitable objects'. While objects are more formal and sometimes contain legalese, they are the only text available for removed charities.

We carry out preprocessing to get the text into a useful format. The main reason for doing this is that certain terms are commonly used in charities' objectives, but are not relevant to understanding a charity's thematic area or mission, so we want to discard them. We take out phrases that commonly appear in charities' objectives but are not informative for our purposes, such as 'at the discretion of the trustees', 'articles of association' and 'the generality of the foregoing.'[7] We also remove the headings of numerical lists, dates and times.

3.2 Using part-of-speech analysis and pattern-based matching to identify useful phrases

Next, we programmatically parse out the most important pieces of information for each charity. The goal of this step is to extract the words and phrases that are the most informative and most able to capture the essence of each charity's self-described activities and objectives.

We use common techniques in natural language processing (NLP) to represent each charity's text as a vector. Tokens are single words or group of words that make up each sentence. In count-based models, each charity is then represented as a sequence of numbers that indicate how many times those tokens featured in each document. Tokens can be extracted with a sliding window, e.g. in a tri-gram model, we go through each three-word combination in the text information and keep recording the frequencies. Tokens frequently appearing in many documents like 'the' are weighed down. This is commonly called 'term frequency-inverse document frequency' (tf-idf). Using a sliding window the size of n-grams to extract tokens mixes up several distinct aspects of charities: it was difficult to distinguish between what charities are doing from where and how they were doing it. Therefore, we instead use part-of-speech analysis combined with pattern-based matching to identify relevant phrases.

Including part-of-speech tags helps us pick better text tokens by making use of morphology. In linguistics, morphology is the study of words, how they are formed, and their relationship to other words in the same language. English has a relatively simple morphological system. Computational linguists have created models that are trained on patterns of parts of speech in different contexts (e.g. the word following the word 'the' frequently is a noun in English) to predict and generalise across other examples. We used a pre-trained model to identify the parts of speech that are used in the charities text. The part-of-speech tagger was trained on the OntoNotes 5 version of the Penn Treebank tag set, which is a dataset of sentences annotated with syntactic or semantic sentence structures.[x]

7. To do this, we used a combination of techniques – including regular expressions, fuzzy matching with Levenshtein distances, and spell-checkers.

Obtaining tokens' parts of speech allows us to use rule-based matching, i.e. creating particular part-of-speech tag patterns to match tokens across the charities text. Part-of-speech tagging and rule-based matching allows us to use particular morphological patterns to isolate terms that answer different questions. In the charities text corpus, noun phrases may be better at capturing thematic areas, whereas verbs may shed light on the motivation of their work.

3.3 Using the existing classification to validate our features

We validated that our extracted noun phrase features were of sufficient quality by using information from existing manually-selected categories provided by the Charities Commission.

To validate that our extracted phrases are a meaningful way to identify what large numbers of charities are trying to achieve, we compared our method against official charitable purposes: the legal categories that charities choose for themselves when they register. Treating those manually-selected categories as ground truth labels, we generate subsets of charities, half of which are 'arts, culture, heritage, or science' charities and the other half from another charitable purpose (but not both).

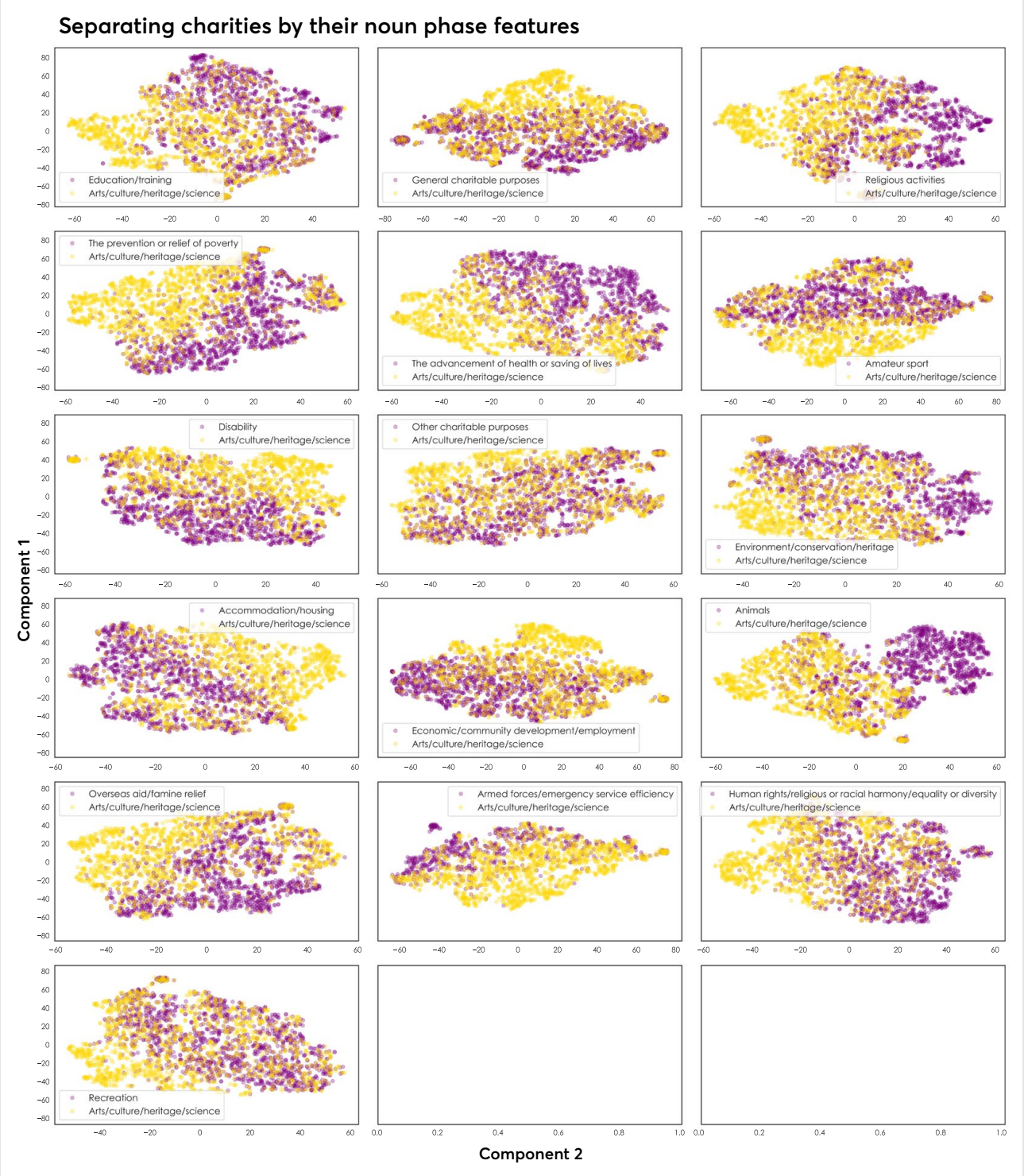
First, we created a tf-idf representation of all charities' descriptions of their activities using the extracted features.[8] Next, we converted the features to pre-trained GloVe word embedding vectors, taking the average of the token vectors for multi-word features. We multiplied the resulting matrix of token vectors by the matrix of tf-idf weights to get a tf-idf weighted representation of each document. Since these vectors live in a 300-dimensional space, which is too high to visualise, we reduce the dimensions from 300 to eight by Principal Component Analysis (PCA) and then use t-Distributed Stochastic Neighbour Embedding (t-SNE) to further reduce the dimensions to two, in order to plot the results. Each point in the resulting chart represents the text information that we are using to summarise a charity in two dimensions, with points that are closer together being charities that have text information with similar meaning.

Figure 3 shows that across all subsets of data tested, the features can be used to create visually separate clusters for charities in yellow (ACHS) and purple (another purpose). Some charities, e.g. animal charities, are very well separated from ACHS charities, meaning the words used to self-describe their activities and objectives are semantically very different for the two groups. The opposite is true for 'other charitable purposes' and 'recreation', meaning those charities use words that occupy a very similar semantic space as words used by ACHS charities. This is unsurprising as the activities of many ACHS charities relate to leisure.

Overall, this shows that using our extracted features to represent charities can successfully parse out meaningful differences at scale, which are validated by comparing to the ground-truth categories that were manually labelled when charities register.

8. Global Vectors for Word Representation (GloVe), developed by Stanford researchers, is an unsupervised learning algorithm for obtaining vector representations for words. Vectors were trained on Common Crawl, a web archive.[xi]

Figure 3: The extent to which the features can distinguish ACHS charities from other charities



3.4 Creating a taxonomy of the keywords that arts and cultural charities use to describe themselves

We include ACHS charities in England and Wales where we can find at least some official text describing their activities and goals. We used pattern-based matching, as explained above, to extract terms to build a taxonomy. The pattern used are noun chunks: these are base noun phrases or flat phrases that have a noun as their head. The noun chunk pattern picks up terms like 'suitable premises', 'elderly luncheon', and 'United Kingdom'. It also captures hyphenated phrases (like 'multi-sensory show' and 'inter-cultural community') as well as multiple adjectives (like 'periodic financial assistance', 'good mental health' and 'inclusive amateur dance'). For simplicity we refer to these extracted phrases in this report as noun phrase features.

As charities engage in a wide range of activities, creating a clustering where each charity only belongs to one group misses important nuances. In the taxonomy approach, charities that undertake multiple activities, reflected in the text information they submit when registering, are catered for more comprehensively.[9] This reflects the reality of the sector's complexity.

To identify terms that are distinct, but have a similar meaning, we clustered the extracted noun phrase features using hierarchical agglomerative clustering and estimated an optimal number of clusters to group the terms using an algorithm for knee-point detection.[xii] This generated groups of words with similar meaning, often sharing an identical word. It may be an adjective, e.g. a cluster may contain many variations of 'creative writing, creative work', etc. It may be the noun, e.g. a cluster may contain many variations of 'interactive theatre, live theatre', etc.

We generate automatic labels for these clusters of terms, using a combination of the most frequently occurring terms and their part-of-speech to produce a coherent phrase that, in most cases, summarises the cluster of phrases.

We then convert the automatically generated labels to embedding vectors for another round of clustering with the same method. The successive clustering allows us to provide structure to our taxonomy, making it easier to extract higher level meaning from, and interpret, the thousands of automatically-labelled clusters. We apply lemmatisation to the automatic labels so clusters with very similar labels are combined (e.g. 'events' and 'event' appear as event).

Finally, we review the most aggregated level of clusters and manually assign labels to the 92 clusters in the top tier. The labels attempt to summarise the majority of the terms. While the automatic labelling may miss out on important distinctions, the manual annotation process is more accurate but laborious. This produces a four-level taxonomy of keywords used by arts and cultural charities in England and Wales.

9. For example, an arts youth charity may mention terms like 'young composers' and 'musical resources', which signals that the charity may belong in clusters working with young artists and musical education respectively.

3.5 Identifying relevant charities using the taxonomy

Methods in information retrieval such as semantic matching and query expansion, as well as other implementations like knowledge graphs, allow us to pass in queries and receive matches based on similarity. To demonstrate how the taxonomy of keywords can be useful, we use a simple implementation of query expansion to search for and quantify a diverse range of over 100 terms. The search terms range from art forms ('ballet', 'poems') to adjectives ('creative', 'innovative', 'virtual'), and were manually chosen to cover a range of domains in the taxonomy.

Query expansion allows the search query to be 'enriched' by semantically similar phrases. For example, if the original search query is 'artistic', the expanded query would also include 'artistic merit', 'artistic heritage', 'musical works' and other related terms in the taxonomy. Finally, the expanded queries were used to match charities in the register that mentioned any of the matching terms.

As with most search implementations, this method only provides an estimate of charities engaging in various areas. Charities may be omitted if they use terms too dissimilar to the matching terms, and charities captured may not be exclusively focused on the area either. For example, a charity which 'helps the men and women' may not be a women's charity in the conventional sense. Still, the approach allows for estimates to be made of previously uncaptured categories.

3.6 Using dependency parsing to extract charitable missions

When charities fill out their aims and activities/charitable objects, they tend to be blending together multiple distinct questions:

- What they do (e.g. 'run a theatre').
- Why they do it (e.g. 'to promote the arts').
- Who they do it for (e.g. 'the elderly').
- How they do it (e.g. 'by running workshops').
- And where they do it (e.g. 'in Bristol and surrounding areas').

Computationally distinguishing all these different pieces of information with accuracy is a complex task. Charitable missions are often not explicitly stated, and are tied up with descriptions of activities, beneficiaries and locations. It is labour-intensive to adapt current machine reading comprehension datasets for complex reasoning.[xiii]

We therefore use dependency parsing to extract the form of the charitable mission. First, using an extended verb phrase pattern with pattern-based matching, we extract candidates of long phrases which may contain charitable missions. Predeterminers and postmodifiers[10] are included where possible to capture more complete phrases. The pattern consists of verb(s) followed by a noun phrase. We cater for consecutive verbs which are common in charities text (e.g. 'to advance, promote, and foster…'). Second, we apply dependency parsing to the extracted candidates to produce a tree showing the syntactic dependencies of its tokens. As with part-of-speech analysis, the dependency parsing uses a pre-trained model trained on OntoNotes. We identify the root and iterate through its descendents to collect mission candidates. For example, if the input is '(C) to employ, retain and pay designers and others whose services are required…', the output is 'employ designers', 'retain designers' and 'pay designers'.

While there are idiosyncratic ways of phrasing similar goals, which can be addressed by clustering, dependency parsing alone does return coherent results that can be aggregated and compared. For each of the groups of charities returned from the document retrieval, we ranked their top missions by their tf-idf score. This allows us to indicate the most popular goals among a group of charities returned from a search query. Future work can generate annotated datasets as benchmarks to evaluate the accuracy of different extraction techniques.

3.7 Limitations of our method

While our method can extract and parse charitable activities and goals at scale, there are limitations. First, the method relies on charitable missions being explicitly stated, but this is not always the case. While the motivations for charitable work usually involve tangible benefits – e.g. helping a specific group of people, cultivating awareness or a particular craft, improving wellbeing – they can be intrinsic in some cases, such as 'no reason' or 'the act is worthwhile itself'.

Second, when put into words, charitable missions are frequently entangled with the target population and locations, as well as the activities carried out to achieve those goals. The extent to which they can be unpacked with machine learning is the key technical issue this report tackles.

Third, charitable missions evolve over time to reflect changes in priorities and circumstances. Text mining enables useful social inquiries using novel data (ranging from annual reports to websites and social media accounts). However, there may be potential biases arising from stylistic writing choices, frequency of updates, as well as data availability and retention.

10. As an example, in the phrase 'all the residents living in the area', 'all the' are the predeterminers and 'living in…' are the postmodifiers.

4

Findings

4.1 A taxonomy of keywords

The taxonomy of keywords used by charities in England and Wales that promote arts, cultural, heritage or science (ACHS) has four levels. The taxonomy unsurprisingly includes many art and cultural domains, like performing, visual, dramatic, and literary arts etc, but also include the wider number of domains where ACHS charities operate. This is evidenced by subcategories of terms in the taxonomy that concern religions, ethnic groups, disabilities, age groups, the environment, flowers and plants, the military and transportation, etc.

At the very top, there are seven main areas which break down to 92 large clusters of charitable terms, each breaking down to one or two further levels. The seven main areas of charitable terms generated by clustering the manually-assigned labels can be loosely interpreted as the:

1. People and identities charities work with.
2. Religions charities are associated with.
3. Descriptions of arts and cultural activities.
4. Descriptions about logistical operations of charities.
5. Geographies charities are associated with.
6. Formats and genres of arts and culture.
7. Descriptions about buildings and the environment.

Not all of the seven areas are relevant for each charity. The 92 manually-assigned labels are shown in Figure 12 in the Appendix.[11]

We show three examples of the taxonomy subcategories from the 92 clusters in the second level: 1) creative arts, 2) digital technology, and 3) disability.[12] Each one breaks down into 2-3 more levels. Each box is labelled with the automatically generated text, and a maximum of five noun phrase features that belong in each cluster. The accuracy of the cluster automatic labelling is not perfect: for example, in one instance, aikido classes and upholstery classes are grouped together, but visual inspection suggests they are overall sensible. To achieve highest coherence, the clustering approach can be combined with manual curation. Newer encoding techniques can also improve performance.

11. There are altogether 2747 'end clusters' which are groups of terms sufficiently similar for automatic labelling. There are 92 clusters in the second level. All of them can be broken down further. On average the 2nd level clusters have 11 child clusters that sit beneath them (number of children vary from 3-104). There're 1665 clusters in the third level. Out of those, 354 (21.3%) could be further broken down to a 4th level whereas the remaining 1311 are sufficiently broken down. There are 1436 clusters in the fourth and last level where there is the most detail.

12. The others can be viewed at https://charitiestaxonomy.azurewebsites.net/taxonomy

Figure 4: Example of the taxonomy – creative arts

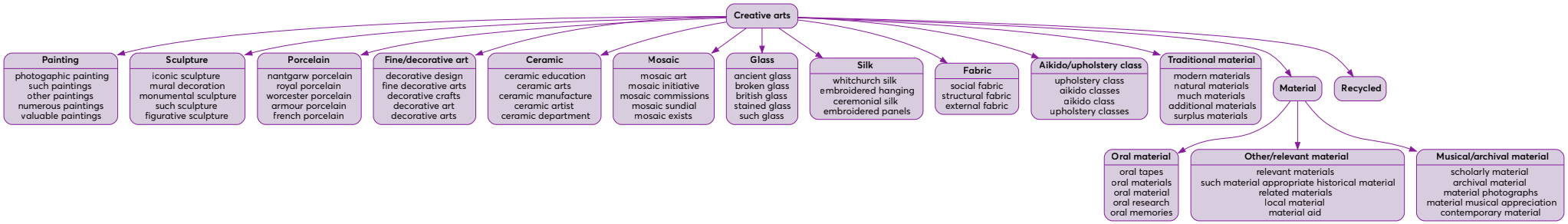


Figure 5: Example of the taxonomy – digital and technology

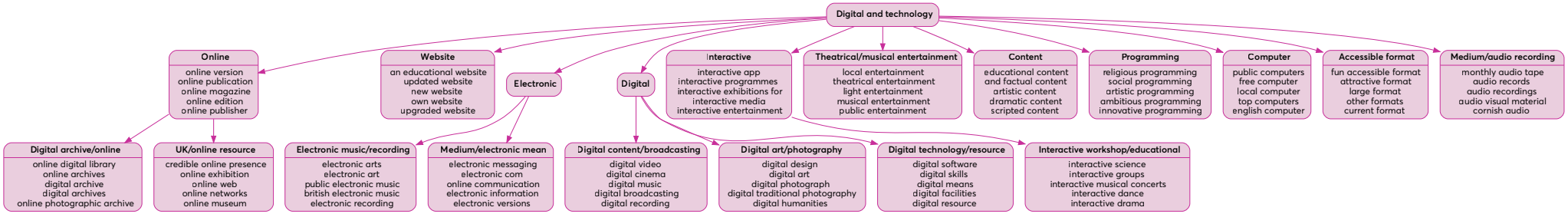
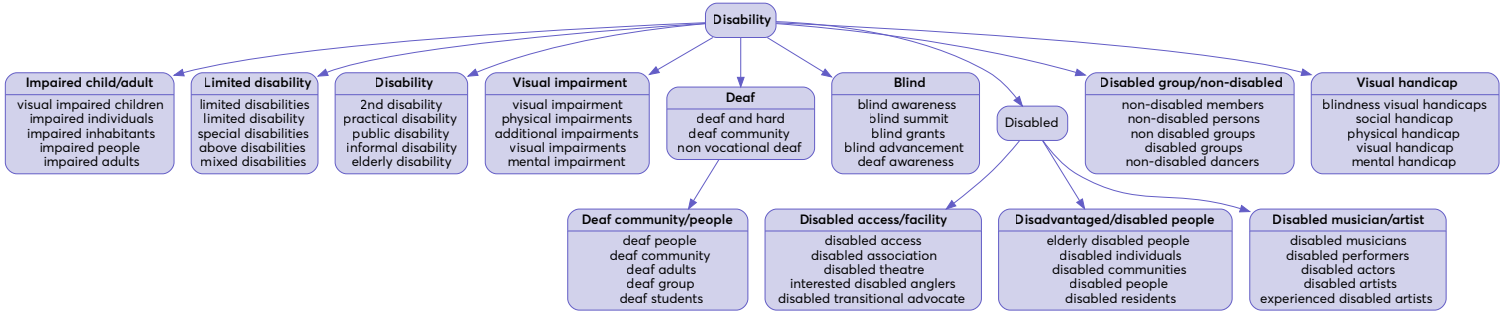


Figure 6: Example of the taxonomy – disability



4.2 Developing a vocabulary for tagging charities activities and goals

The iterative clustering creates a prototype of a vocabulary, which can be applied to index the activities and goals of charities in a semantically meaningful way. It can be helpful to think of the cluster assignments as suggested tags: if a charity mentions at least one of the terms in a cluster, the charity is tagged with the cluster label. So a charity mentioning 'young violinist' may be tagged as 'young musician', which is nested under 'young people' and may also be tagged in one of the music clusters, etc. Each tag represents a group of charitable terms that are similar enough that they can be given a name automatically.[13]

We clustered the noun phrase features (converted to word embeddings, so that semantically similar phrases would be closer to each other) successively until they can be represented in a four-tier structure which we loosely call a 'taxonomy'. We found 2747 groups of terms that are semantically similar enough and can be automatically labelled. In Figure 7, each cluster of terms is a circle, with size corresponding to an ACHS focus weight.[14] This weighting surfaces keywords that ACHS charities are more likely than non-ACHS charities to use when describing their activities and goals. A single charity will often be counted in multiple circles in this diagram.

13. 92% (n=154,974) of charities received at least 1 tag, where the base is all CCEW charities that were not removed or linked at the time of the web-scrape. On average each charity is assigned 4 tags and each tag is assigned to 55 charities.

14. If X is equal to the percentage of ACHS charities that described their activities using terms within that cluster, and Y is the percentage of all charities that described their activities using terms within that cluster, the ACHS focus weight is simply X divided by Y.

Figure 7: The taxonomy flattened and visualised, with some clusters in creative domains highlighted

4.3 Applications of the taxonomy

4.3.1 Artforms and beneficiaries

How do arts and cultural charities help people from different demographics? We used the taxonomy to identify this for the arts and cultural charitable sector in England and Wales. Using the query expansion method explained above, we constructed identified charities that were matched from specific search terms relating to artforms and demographics. The motivation for this analysis is to help answer questions like 'Are there more youth charities promoting dancing than acting?', 'are there any charities that engage in Islamic arts and crafts?', or 'is there charitable work involving dance and disabled persons?'

Figure 8: What demographics do arts and cultural charities engage with?

Among registered 'ACHS' charities in England and Wales that were active in Sept 2018.
Highest values for each subset are in red, zero matches are greyed out.

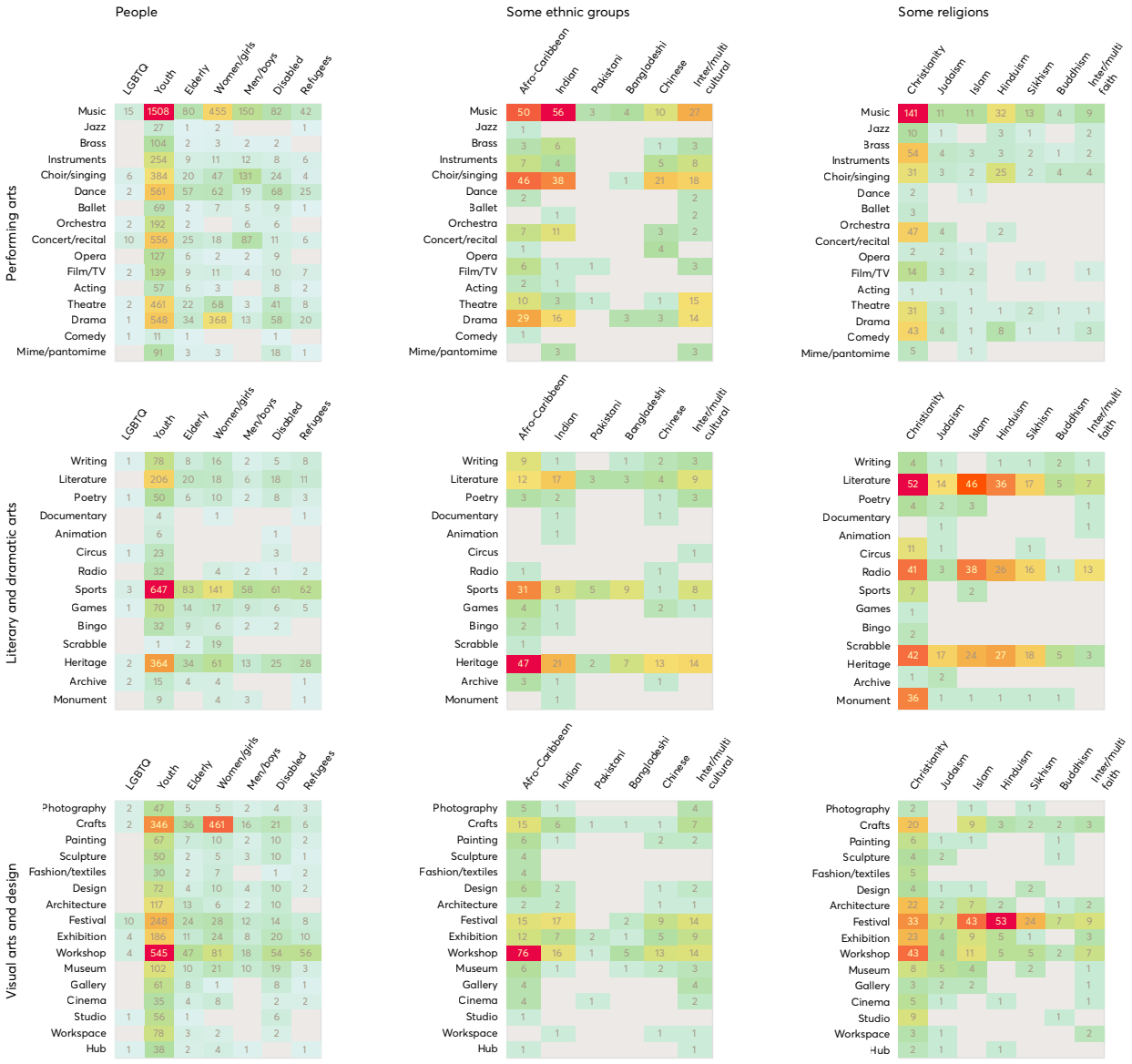


Figure 8 shows how keywords about people, faiths, and ethnicities interrelate with keywords about arts and culture, via a series of heatmaps with the number of charities labelled. Only active ACHS charities are shown. The cell colouring reflects the number of charities involved, with a red indicating a greater number than green, and grey cells indicating that there are no charities which mention at least one term in either bucket of terms, e.g. there are no charities mentioning Buddhism and opera.

Arts and cultural charities operate across a wide range of disciplines (e.g. performing arts, literary/dramatic arts, visual arts and design, sport) helping a diverse range of demographic groups. For example, the extracted results show 368 active charities mentioning women/girls and drama, 68 charities mentioning disability and dance, eight charities that mention refugees and theatre, and six charities that mention LGBTQ and choir/singing and six that mention Christianity and paintings.

Also, there are over 1,500 active charities that mention youth and music in their description or objectives when registering. As there is research studying why there are not more young people engaged in some arts,[xiv] identifying charities working in the space can be a helpful step in finding solutions.

There are some terms that are mentioned by charities across all demographic groups: workshop, sports, music, and heritage. Terms associated with dance, drama, theatre, literature, festivals, exhibitions and crafts also had very high coverage, with at least one charity mentioning these terms across nearly all demographic groups.

Our method enables for the first time comparisons across subpopulations and between genres among charities that promote arts and culture. For example, there are gender differences among ACHS charities: among charities mentioning choir/singing, there are more that also mention men/boys (131) than women/girls (47), whereas the opposite is true for dance (62 for women/girls compared with 19 for men/boys). There are also multiple active Indian and Afro-Caribbean music charities and inter/multicultural theatres.

The breakdowns also enable us to identify areas relatively more and less well-covered by ACHS charities, and delve deeper to understand why. For example there are: relatively sizable groups of charities mentioning women/girls and drama (368) and crafts (461), which is partially explained by local chapters of Women's Institutes. There are more charities mentioning 'festival' along with Hinduism (53) compared with other major religions, which can indicate Hindu observances are commonly considered when charities register, whereas for 'architecture', Christianity (22) is the religion that is mentioned alongside most frequently, which can indicate a consideration for churches and buildings, and the potential role of church buildings as cultural venues.

4.3.2. Historical change

In this part of the analysis, we first verify that almost six decades of charity registration text can be used to evidence historical change. To do this we produce strip plots that visualise high-level trends, as charities across different subdomains were added to the register at various points from 1961. Using query expansion to retrieve matching charities that ever registered, the plots allow us to study charities' relative age (by registration date of charities) and relative density (by number of charities matching associated terms).

Across the broader charitable sector, we find, e.g., that ballet, mime, and opera are terms used by relatively older charities, whereas documentary, animation, and festivals are terms used by relatively more recent charities. LGBTQ+, refugees and specific ethnic minority groups are relatively recent beneficiaries, whereas men/boys are relatively older beneficiaries, and the data also show a shift in language from charities using terms relating to handicapped to disabled. These verification checks are presented in the Appendix.

Second, we show how the most 'strongly arts and cultural' phrases vary for each decade. Using f-regression feature selection (a univariate linear regression test), we test how effectively each of the noun phrase features within the taxonomy predicts whether it identifies an ACHS charity in each of the six decades. The noun phrase features are ranked according to the significance of the regression parameter, with the term at the top the term most likely to be a predictor of ACHS status. We perform the analysis separately for ACHS charities that are more narrowly focussed (defined as having ACHS and at most one other charitable purpose) and those that have a broad remit (having three to eight charitable purposes including ACHS).

Figure 9 shows an interesting range of terms that were popular for each decade, from handbell ringing in the 1980s to literary festivals in 2010s. For ACHS charities that also work on other domains, e.g. health or community development, the more strongly arts and cultural phrases for each decade evolve from memorial halls and indoor bowls in the 1960s to male choral groups in the 1980s and mentions of older people in the 1990s and 2000s.

Figure 9: Noun phrases most strongly associated with ACHS over the decades

Narrow focus (max 2 purposes)					
1960	1970	1980	1990	2000	2010
Permanent theatre	Public promotion	Public entertainment	Professional adjudicators	Public education	Public education
Professional standards	Public library	High artistic merit	Educational drama	Musical instruments	Public history
Good design	Educational projects	Annual pantomime	Public works	Related arts	Understanding enjoyment
Public advancement	Highest quality	Choral repertoire	Creative projects	Cultural activities	Educational plays
Musical students	Amateur productions	Live orchestra	French language	Educational plays	Literary festival
Worldwide public	Carol concert	Classical choral concerts	Educational cultural activities	Public study	Musical instruments
Diverse groups	Creative art	Annual concert	Regular concerts	Professional recitals	Related arts
Affordable theatre	Good music	Handbell ringing	Annual competitive music	Professional performers	Public exhibition
Diverse range	Non members	Public stage	Public exhibitions	Public works	Cultural events
Common fellowship	Foster research	Handbell tune	Musical organisations	Understanding enjoyment	Public display

Broad focus (3-8 purposes)					
1960	1970	1980	1990	2000	2010
Local community	Young agriculture	Male choral	Older people	Best contribution	Stained glass
Memorial hall	Promote education	Social activities	Educational opportunities	Common good	Industrial heritage
Physical mental recreation	Local clubs	Cultural societies	Young agriculture	Good citizenship	Architectural importance
Mental recreation	Cultural societies	Spiritual wellbeing	New skills	Useful results	Public heritage
Indoor bowls	Open days	Social sporting	Monthly meeting	Effective relationships	Understanding enjoyment
Social intercourse	Social fundraising	Choral work	Primary school	Recreational physical activity	Special facilities
Social activities	Local broadcasting	Primary school	Young women	Chinese community	Public performing
Local clubs	Recreational leisure	National trust	Scottish country	Older people	20th century
Political opinions	Voluntary groups	Regular basis	Wider world	Weihai lishi	Highest standard
Main hall	Reasonable recreation	Widespread performance	Traditional Scottish country	Wider community	Creative performing

4.3.3. Charitable missions

Many charities in the arts and culture sector aim to advance similar goals. We explore if we can meaningfully extract and aggregate charitable missions. While there is some noise from the dependency parsing partially retained, in most cases, many of the top missions are sensible. Domain experts may be able to validate and explain the popularity of certain extracted missions. This may have useful applications for charity workers, funders, and researchers.

We find, for example, that

- For charities mentioning 'crafts' and associated terms, one of the top missions is to 'advance the education of young members of the public'.
- For charities mentioning 'LGBTQ' and associated terms, one of the top missions is around 'eliminate discrimination' and for 'refugees', it is 'adapt within a new community'.
- For charities mentioning 'monuments' and associated terms, one of the top missions is around 'reconstructing churches'.
- For charities mentioning 'diversity' and associated terms, one of the top missions is around 'conducting research on equality and diversity issues'.
- For charities mentioning 'sustainable' and associated terms, one of the top missions is 'achieving economic growth and regeneration'.
- For charities mentioning 'radio' and associated terms, one of the top missions is 'providing a local broadcasting service for hospitals'.

4.3.4 Web presence

In 2019, the Department for Digital, Culture, Media and Sport (DCMS) published a policy paper called 'Culture is Digital'[xv] recognising the importance of digital technologies for the sector. We use the websites listed on the Charity Commission's website, along with the taxonomy, to analyse one dimension of this – the extent to which charities have a web presence.

About two-thirds of active ACHS charities (n=19,916) list a website on the Commission's website. We visited these websites, scraping some key information from the frontpage and, where available, supplemented the data with basic information from WHOIS, a public lookup of website domain ownership. 95 per cent were valid websites, defined as having a non-expired domain with at least some relevant charities text on the front page.[15]

Figure 10 shows some arts and cultural charitable domains with relatively high and low web presence. Some groups of ACHS charities, as identified by expanded search terms with the taxonomy, have higher online presence, with charities matched by 'virtual', 'documentary', 'studio', and 'orchestra' all having over 85 per cent of functioning websites.

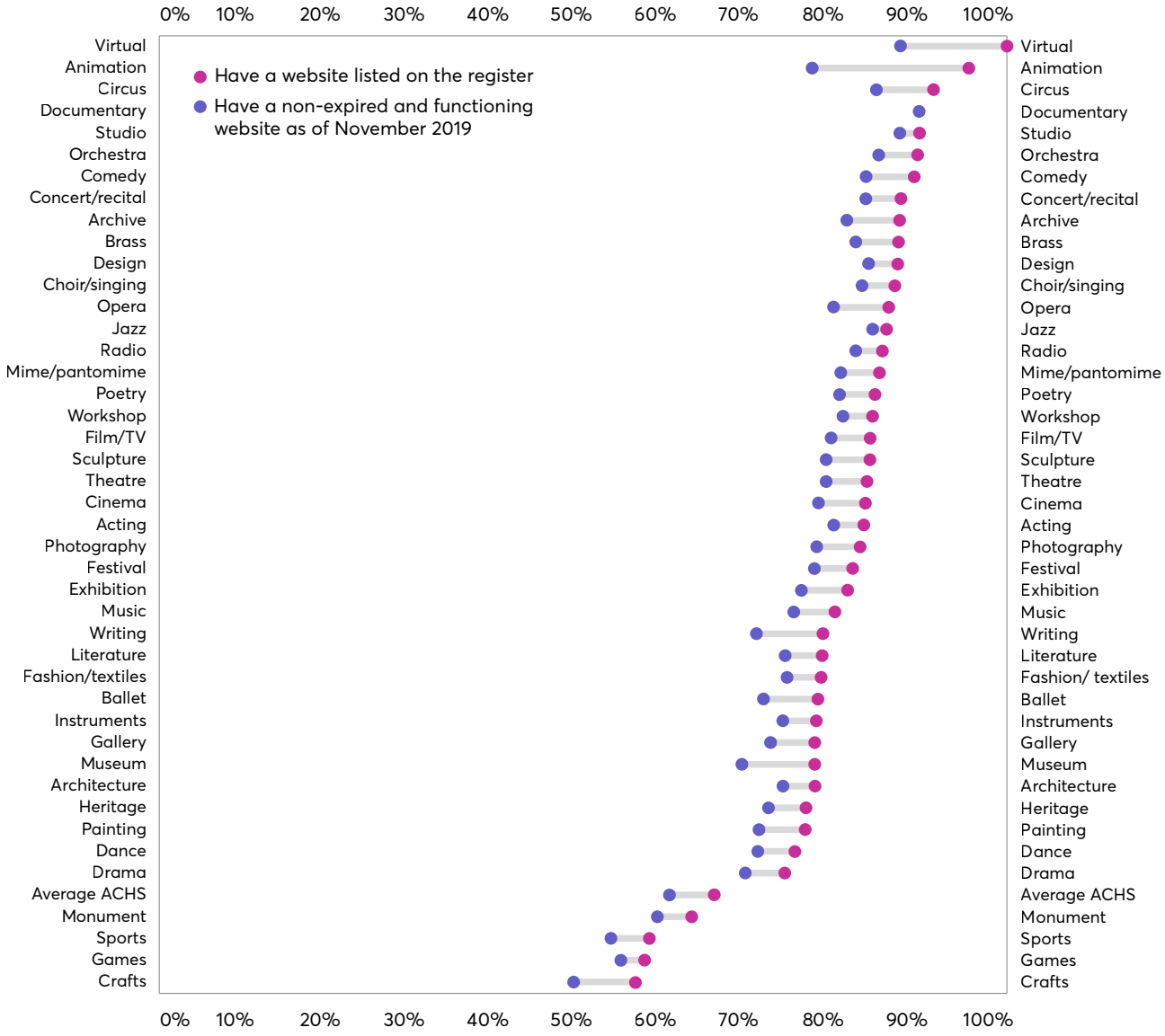
On the other hand, for charities matched by 'monuments', 'sports', 'games' and 'crafts' and their associated terms less than 60 per cent have valid websites. Their web presence is generally below the average for ACHS charities (60.4 per cent).

15. We collected the title, headings, meta tags, and all links present on the main pages of all ACHS charities that listed a website on the Charity Commission's register. Web pages displaying errors or generic messages about expired domains from registrars were counted as invalid. Very minimalist websites are still included as valid.

Figure 10: The website presence of ACHS charities

How many arts and cultural charities have a functioning website?

Proportion of active ACHS charities in England and Wales (matched by the query term)



5

Applications of this research

This paper shows that it is feasible to to use natural language processing and machine learning techniques to create a 'taxonomy' of keywords used by charities in England and Wales that advance arts, culture, heritage or science (ACHS). We use taxonomy to index charities activities and goals in a semantically meaningful way, allowing us to evidence new insights about the sector.

In the immediate term, these methods can be extended to:

- Carry out additional dimensions of mapping: for example, breaking down charities by more granular areas of focus (identified by text) alongside geography[16] and survival rates,[17] etc.
- Be updated at regular intervals, allowing us to have a live understanding of the ACHS charitable domain.
- Map other domains and break down other umbrella terms in the charitable sector.
- Track how the language used by charities to describe groups of people, causes and emerging technologies changes.
- Include alternative data sources about charities like annual reports and social media.
- Include additional data sources about voluntary organisations beyond registered charities.

In the longer term, a data science approach as outlined here can be applied to:

- Build a recommendation engine to search for similar charities or charitable causes.[18]
- Evidence how well-addressed certain goals are by charities, or how crowded certain areas are, by linking to other data sources like funding.
- Make the creation and maintenance of taxonomies of sector activity easier to help improve understanding of what the sector is doing.[19]

16. See, e.g. Corry 2020, which examines the regional breakdown of charities in England.[xvi]
17. See, e.g. Clifford 2018, which links neighborhood deprivation with charity dissolution in England.[xvii]
18. This enables funders to engage with charities that are similar to those that they fund (but who they don't engage with), and for charities to find similar organisations for collaboration and learning.
19. The NLP and machine learning techniques, like the ones described in this report, have been applied to generate tags in domains from libraries[xviii] and biotech[xix] to legal documents[xx] and regulatory codes.[xxi] One reason is that fully controlled vocabularies are expensive to produce manually, but automatically-generated tags can enrich metadata which can lend itself to outcomes like better knowledge organisation and information retrieval.

6

Glossary

Charitable objects

'Objects' describe and identify the purpose for which a charity has been set up. They are usually set out in a single clause or paragraph (the 'objects clause') when registrants write their charity's governing document. Instead of saying what the charity will do on a daily basis, the objects should accurately express all of the charity's purposes.

Charitable purpose

The Charities Act 2011 defines a charitable purpose, explicitly, as one that falls within 13 descriptions of purposes and is for the public benefit. Examples are 'the prevention or relief of poverty', 'the advancement of citizenship or community development', and 'the advancement of the arts, culture, heritage or science'.

Dependency parsing

The process of analysing the grammatical structure of a sentence, establishing relationships between 'head' words (the grammatically most important word in a phrase) and words which affect the interpretation of the head words.

Hierarchical clustering

Clusters are groups of similar objects. Hierarchical clusters are clusters with a nested structure, for example a cluster of music charities, can contain within it clusters of jazz and opera charities. Hierarchical cluster analysis is a method of cluster analysis which seeks clusters with a hierarchical structure.

Noun phrase

A noun phrase in English is a sequence of words surrounding at least one noun, e.g. 'the cupcake', 'an innovation foundation.'

Part-of-speech tagging

The process of assigning parts of speech labels, e.g. nouns, verbs, adjectives, adverbs, etc, to each word of the input text. Note the same word can have a different part of speech depending on its context (the word's relationship with adjacent and related words in a phrase, sentence, or paragraph). For example, 'I suspect that is the case' compared to 'He was a suspect in the case'.

Pre-training

Training in advance, usually refers to models trained by someone else on a dataset to solve a similar problem. For example, using pre-trained embeddings means the embedding representations we use for input words were learned separately using another algorithm.

Query expansion

A process in Information Retrieval which consists of selecting and adding terms to the user's query with the goal of returning more relevant matches or search results.

Verb phrase

A verb phrase in English consists of a verb followed by assorted other components; for example, a verb followed by a noun phrase is a kind of verb phrase.

Word embeddings

A vector representation for text where words that have close meaning have a similar representation, as prior context is 'embedded'. For example, 'knife' would be semantically close to 'fork'. The underlying idea is that 'a word is characterized by the company it keeps', which is known as the distributional hypothesis.

7

Appendix

7.1 Major existing charity classifications for charities in England and Wales

There are two main systems that have been used to classify UK charities – one from the Charity Commission of England and Wales and one from the United Nations. There is also the NTEE Classification System developed by the National Center for Charitable Statistics in the US[xxii] but it is not applied to UK charities.

Classifications from the Charity Commission of England and Wales (CCEW)

The CCEW divides charities up in three ways (or classifications).

- C1: What the charity does.
- C2: Who the charity helps.
- C3: How the charity operates.

C1 mostly overlaps with charitable purposes as defined by law. According to the Charities Act 2011, charitable purposes include:[xxiii]

1. The prevention or relief of poverty.
2. The advancement of education.
3. The advancement of religion.
4. The advancement of health or the saving of lives.
5. The advancement of citizenship or community development.
6. The advancement of the arts, culture, heritage or science.
7. The advancement of amateur sport.
8. The advancement of human rights, conflict resolution or reconciliation or the promotion of religious or racial harmony or equality and diversity.
9. The advancement of environmental protection or improvement.
10. The relief of those in need, by reason of youth, age, ill-health, disability, financial hardship or other disadvantage.
11. The advancement of animal welfare.
12. The promotion of the efficiency of the armed forces of the Crown, or of the efficiency of the police, fire and rescue services or ambulance services.
13. Any other charitable purposes.

In addition, the CCEW also includes 'recreation', 'overseas aid/famine relief', 'accommodation/housing', and 'general charitable purposes' in the C1 classification. The Charity Commission for Northern Ireland (CCNI) and Office of the Scottish Charity Regulator (OSCR) use similar versions of the above.

Importantly, charities select multiple purposes when they register. In fact, only 10.9 per cent of active ACHS charities only work on that singular purpose. 40 per cent of active ACHS charities work on one to two additional charitable purposes.

The International Classification of Non-profit and Third Sector Organizations (ICNP/TSO)

There are international charity classifications that are also sometimes used. Most notably, there is a classification from the United Nations (UN).[20] Its newest version is called the International Classification of Non-profit and Third Sector Organizations (ICNP/TSO) and was last updated in December 2017.[xxvi]

The UN classification puts activities by arts and cultural non-profit organisations into a section called 'culture, communications and recreation activities.' This in turn is broken down into 'culture and arts', 'sports and recreation' and 'information and communication services'. Altogether, in the ICNP/TSO, there are ten categories that a non-profit in the arts and cultural sector can fall into, including a few that say 'not elsewhere classified'.

It is common for researchers to classify charities in England and Wales according to the original version of ICNP/TSO. Researchers started doing this in 1996 and the UN classification has remained a common way to understand the UK charitable sector.[xxvii]

Some researchers have extended the original UN classification. The National Council for Voluntary Organisations (NCVO) added new subdivision codes such as 'village halls' when such subcategories did not exist in the UN classification.[21] Sometimes the new subdivision codes were then used by researchers.[xxix] Some of the techniques have been semi-automated: e.g. researchers have used keyword searches to classify charities.[xxx]

7.2 Historical change in the broader charitable sector

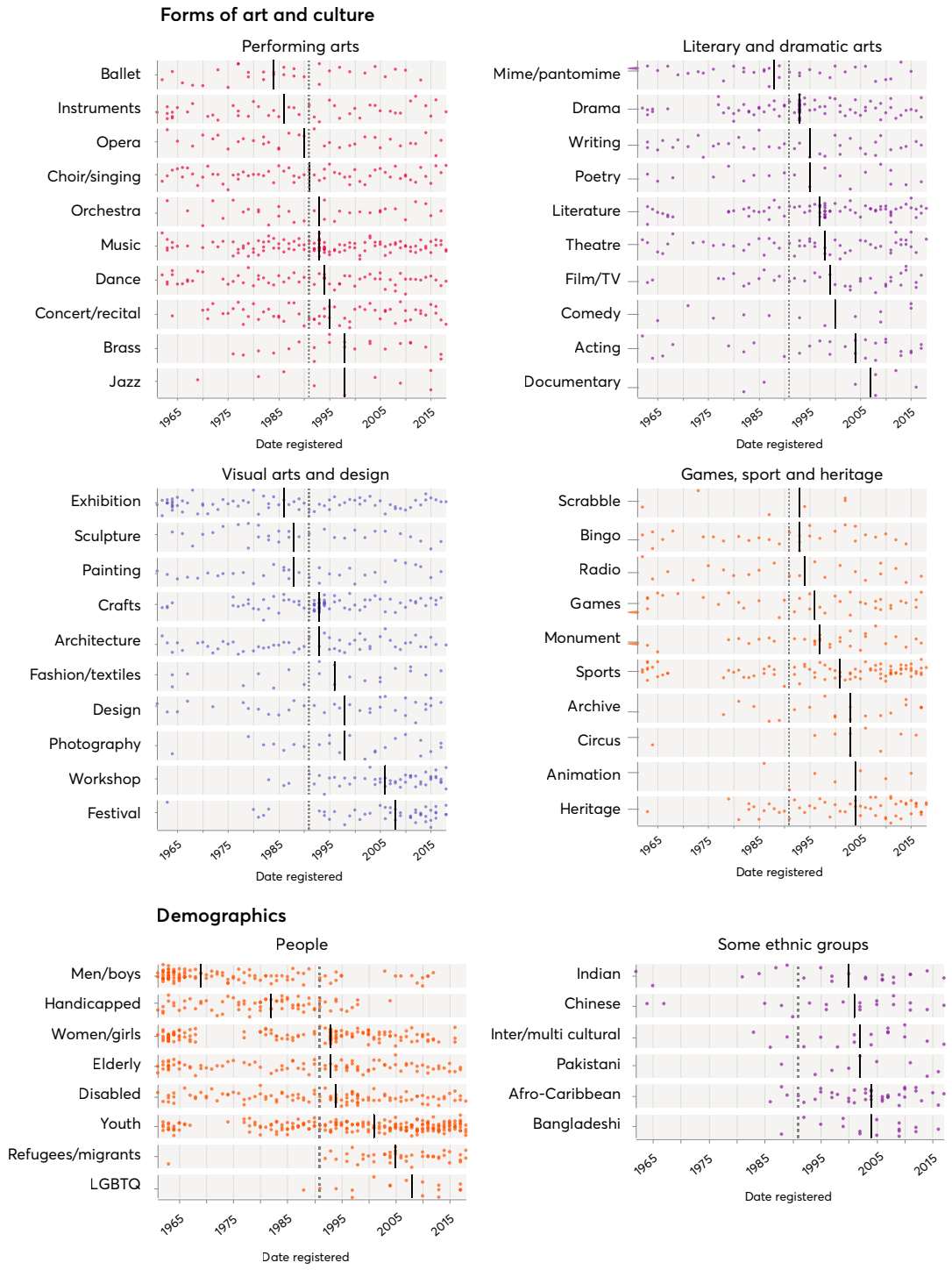
To verify that the charities data can evidence change across decades, we apply the query expansion method to over 50 search terms, with a focus on types arts and culture, as well as some demographic groups, almost all of which are currently unavailable in official charities data. All charities ever registered with CCEW are included to account for survivorship bias.

If there are ten dots on the strip, there are approximately 100 charities. A solid vertical line is drawn at the date that separates the dots on the strip in half. The dotted line indicates December 1990, which is the median registration date across the full register. For example, a strip that contains a dense area on the left but a sparse area on the right suggests areas where decreasing numbers of charities are being set up over time.

20. The United Nations Statistics Division originally introduced a non-profit classification in 2003,[xxiv, xxv] with its origins in a 1992 research paper.[xxv] This classification was eventually expanded in December 2017 to cover the 'activities of all institutional units potentially falling within the Third, or Social Economy (TSE) sector.' Many UK charities researchers use the pre-2017 original classification called the International Classification of Non-profit Organizations (ICNPO).

21. In their annual publication the Almanac, the NCVO explained how they classified organisations into categories based on the ICNPO, with examples for the subcategories they created.[xxviii]

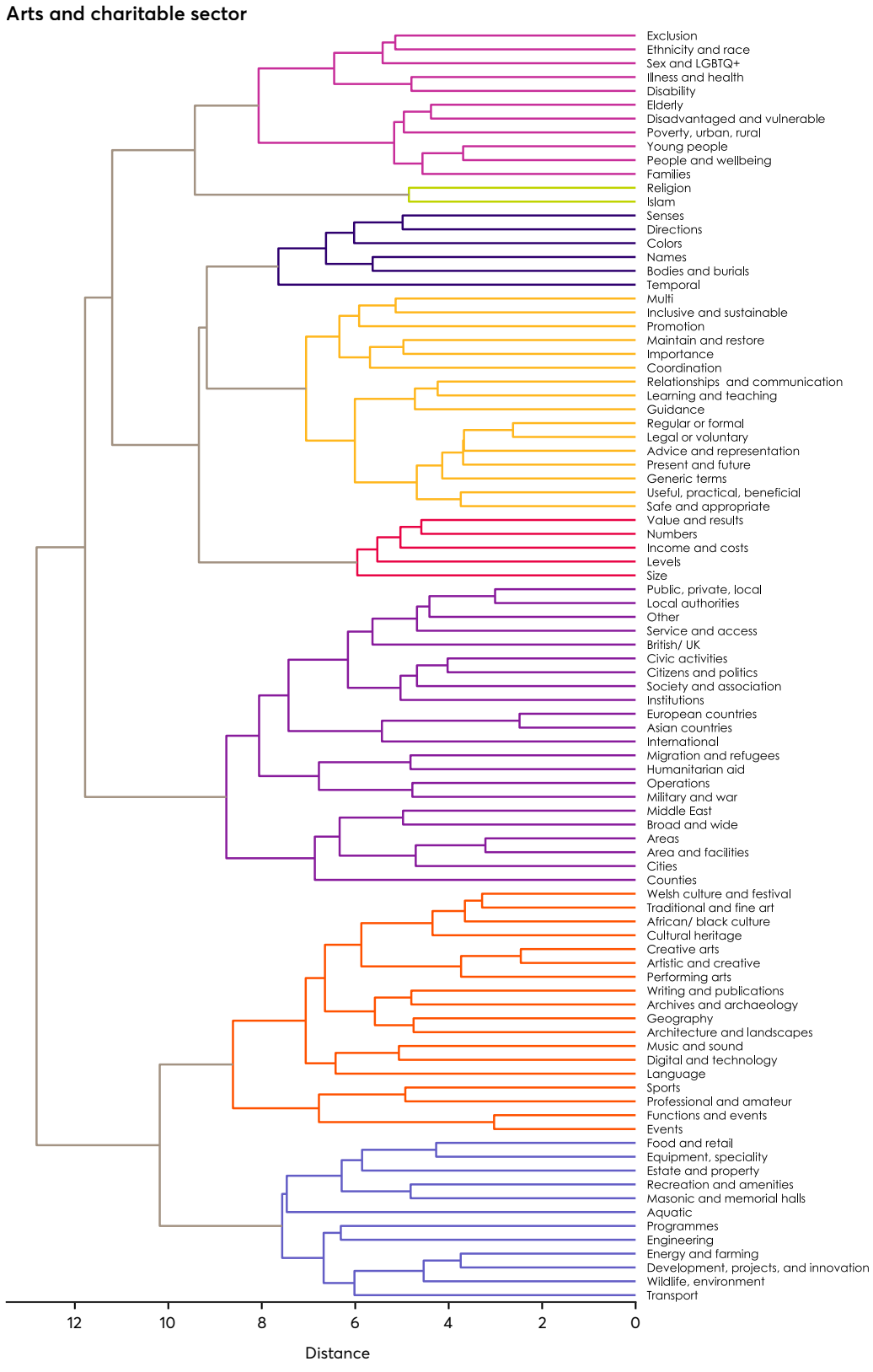
Figure 11:



The charts show a high-level trend that ballet, mime, and opera are terms used by relatively older charities, whereas documentary, animation, and festivals are terms used by relatively more recent charities. The charts also show that LGBTQ+, refugees and specific ethnic minority groups are relatively recent beneficiaries, whereas men/boys are relatively older beneficiaries, and the data also show a shift in language from charities using terms relating to handicapped to disabled.

7.3 Additional information on the taxonomy of keywords

Figure 12: 92 manually assigned labels for the taxonomy of keywords



8

References

i. National Audit Office. Regulating charities: a landscape review. Briefing for the House of Commons Public Administration Select Committee. National Audit Office; 2012 Jul.

ii. National Council for Voluntary Organisations. UK Civil Society Almanac 2019. National Council for Voluntary Organisations; 2019 Jul.

iii. Hornung L. New volunteering data out today. In: National Council for Voluntary Organisations blog [Internet]. 26 Jul 2018 [cited 22 Aug 2019]. Available: https://blogs.ncvo.org.uk/2018/07/26/new-volunteering-data-out-today

iv. Mazzucato M. Mission-oriented innovation policies: challenges and opportunities. Industrial and Corporate Change. 2018;27: 803–815.

v. Mazzucato M. A mission-oriented approach to building the entrepreneurial state. A 'Think Piece'for the Innovative UK Technology Strategy Board: London, UK. 2014. Available: https://marianamazzucato.com/wp-content/uploads/2014/11/MAZZUCATO-INNOVATE-UK.pdf

vi. Department for Business E, Strategy I. The Grand Challenge missions. Department for Business, Energy and Industrial Strategy; 2019.

vii. Ógáin EN. Impact measurement in impact investing. United Kingdom, Nesta Impact; 2015. Available: https://media.nesta.org.uk/documents/impact_measurement_in_impact_investing.pdf

viii. The Charity Commission for England and Wales. Excepted charities. In: GOV.UK Guidance [Internet]. 11 Jun 2014. Available: https://www.gov.uk/government/publications/excepted-charities/excepted-charities–2

ix. The Charity Commission for England and Wales. Exempt charities (CC23). In: GOV.UK guidance [Internet]. 9 Aug 2019. Available: https://www.gov.uk/government/publications/exempt-charities-cc23/exempt-charities

x. Explosion AI. spaCy annotations: part-of-speech tagging. Available: https://spacy.io/api/annotation#pos-en

xi. Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014. doi:10.3115/v1/d14-1162

xii. Satopaa V, Albrecht J, Irwin D, Raghavan B. Finding a 'Kneedle' in a Haystack: Detecting Knee Points in System Behavior. 2011 31st International Conference on Distributed Computing Systems Workshops. 2011. doi:10.1109/icdcsw.2011.20

xiii. Sugawara S, Stenetorp P, Inui K, Aizawa A. Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets. arXiv [cs.CL]. 2019. Available: http://arxiv.org/abs/1911.09241

xiv. Tait R, Kail A, Shea J, McLeod R, Pritchard N, Fatima A. How can we engage more young people in arts and culture? A guide to what works for funders and arts organisations. NPC; 2019 Oct. Available: https://www.thinknpc.org/resource-hub/arts

xv. Department for Digital, Culture, Media & Sport. Culture is Digital: June 2019 progress report. 2019 Mar.

xvi. Corry D. Where are England's charities? Are they in the right places and what can we do if they are not? NPC; 2020 Jan. Available: https://www.thinknpc.org/resource-hub/where-are-englands-charities

xvii. Clifford D. Neighborhood Context and Enduring Differences in the Density of Charitable Organizations: Reinforcing Dynamics of Foundation and Dissolution. Am J Sociol. 2018;123: 1535–1600.

xviii. Golub K, Hagelbäck J, Ardö A. Automatic subject classification of Swedish DDC: Impact of tuning and training data set. TPDL 2019. 2019. Available: https://nkos-eu.github.io/2019/content/NKOS2019-abstract-golub.pdf

xix. Hubain R, De Wilde M, van Hooland S. Automated SKOS Vocabulary Design for the Biopharmaceutical Industry. Cataloging & Classification Quarterly. 2016;54: 403–417.

xx. Quaresma P, Gonçalves T. Using Linguistic Information and Machine Learning Techniques to Identify Entities from Juridical Documents. In: Francesconi E, Montemagni S, Peters W, Tiscornia D, editors. Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. pp. 44–59.

xxi. Casellas N. Linked legal data: a SKOS vocabulary for the code of federal regulations. SWJ, IOS Press Journal. 2012. Available: http://www.semantic-web-journal.net/sites/default/files/swj311.pdf

xxii. Lampkin L, Romeo S, Finnin E. Introducing the Nonprofit Program Classification System: The Taxonomy We've Been Waiting for. Nonprofit and Voluntary Sector Quarterly. 2001. pp. 781–793. doi:10.1177/0899764001304009

xxiii. The Charity Commission for England and Wales. Charitable purposes: Guidance on what purposes can be charitable. The Charity Commission for England and Wales; 2013 Sep.

xxiv. United Nations Statistical Division. Handbook on Non-profit Institutions in the System of National Accounts. United Nations Publications; 2003.

xxv. Salamon LM, Anheier HK. In search of the non-profit sector II: The problem of classification. Voluntas. 1992. pp. 267–309. doi:10.1007/bf01397460

xxvi. United Nations Statistical Division. Satellite Account on Non-profit and Related Institutions and Volunteer Work. United Nations Statistical Division; 2018.

xxvii. Salamon LM, Anheier HK. The Emerging Nonprofit Sector: An Overview. Manchester University Press; 1996.

xxviii. National Council for Voluntary Organisations. Alamanac 2019: Capturing what voluntary organisations do. A reference document on the classification of subsectors (ICNPO) used for voluntary organisations. National Council for Voluntary Organisations; 2019.

xxix. Mohan J, Barnard S. Comparisons between the characteristics of charities in Scotland and those of England and Wales. Centre for Charitable Giving and Philanthropy; 2013 May.

xxx. Third Sector Research Centre. Understanding the UK third sector The work of the Third Sector Research Centre 2008–2013. Third Sector Research Centre; 2013.

Acknowledgements

We would like to thank the following people for taking time to provide useful comments and discuss charities data with us: **Margaret Bolton**, **Martin Brookes**, **Véronique Jochum**, **David Kane**, **Tris Lumley**, **Rosario Piazza**, **Mor Rubinstein**, **Lucy Smith**. Thanks also to **Andrew Mowlah** and colleagues at Arts Council England, **Harman Sagger** and colleagues at DCMS, **Nixi Cura** at the Royal Academy of Arts for hosting helpful discussions about this work. Last but not least, thanks to colleagues at both Nesta and the Creative Industries Policy and Evidence Centre: in particular **Hasan Bakhshi**, **Carrie Deacon**, **Eliza Easton**, **Trishna Nath**, **Fran Sanderson**, and **Melissa Wong**. Thanks to **Anna Zabow** for communications work and **John Davies** for extensive help with editing throughout.

The Creative Industries Policy and Evidence Centre (PEC) works to support the growth of the UK's Creative Industries through the production of independent and authoritative evidence and policy advice.

Led by Nesta and funded by the Arts and Humanities Research Council as part of the UK Government's Industrial Strategy, the Centre comprises of a consortium of universities from across the UK (Birmingham; Cardiff; Edinburgh; Glasgow; Work Foundation at Lancaster University; LSE; Manchester; Newcastle; Sussex; Ulster). The PEC works with a diverse range of industry partners including the Creative Industries Federation.

For more details visit **www.pec.ac.uk** and **@CreativePEC**

If you'd like this publication in an alternative format such as Braille, or large print, please contact us at: **enquiries@pec.ac.uk**

Creative Industries
Policy & Evidence Centre

Led by **nesta**

Creative Industries Policy and Evidence Centre (PEC)
58 Victoria Embankment
London EC4Y 0DS

+44 (0)20 7438 2500

enquiries@pec.ac.uk

@CreativePEC

www.pec.ac.uk