Digital Research Reports

Topic Modelling of Research in the Arts and Humanities

An analysis of AHRC grant applications

Hélène Draux and Martin Szomszor

NOVEMBER 2017

DIGITAL science

Arts & Humanities Research Council

About Digital Science

Digital Science is a technology company serving the needs of scientific and research communities at key points along the full cycle of research. We invest in, nurture and support innovative businesses and technologies that make all parts of the research process more open, efficient and effective. We believe that together, we can change research for good.
Visit www.digital-science.com

About Consultancy

Our consultancy team delivers custom reporting and analysis to help you make better decisions faster. With in-depth knowledge of the historical and current research ecosystem, our unique perspective helps get the most value from data on the research lifecycle. Our team of data scientists are experts in using innovative analytical techniques to develop revealing visualisations and powerful insights. We understand the changing research landscape, and we can help you develop an evidence base on which to build the best research management and policy decisions.

About the authors

Hélène Draux is Research Data Scientist at Digital Science. She has an academic background in social geography, with experience in data visualisation and data science. Her work has focused on facilitating the inclusion of the general public through interactive visualisations, to communicate and collect data. She has worked in the UK and Denmark, and participated in several national and European-funded research projects. She has also worked in R&D in the private sector, where she beta-tested the use of phone applications for place discovery.

iD http://orcid.org/0000-0001-8837-168X

Martin Szomszor is Consultant Data Scientist at Digital Science and was founder of the Global Research Identifier Database (GRID). Martin applies his knowledge of machine learning, data integration, and visualisation techniques to analysis of the research lifecycle. He was Deputy Head of Centre at the City eHealth Research Centre (2009-2011), Chair of the 4th International Conference on Electronic Healthcare for C21, and a Research Fellow at the University of Southampton (2006-2009) working on linked data, semantic web, and social network projects.

iD http://orcid.org/0000-0003-0347-3527

This report has been published by Digital Science which is part of the Holtzbrinck Publishing Group, a global media company dedicated to science and education.

The Campus, 4 Crinan Street, London N1 9XW
consultancy@digital-science.com
Copyright© 2017 Digital Science

ISBN 978-0-9956245-6-6

Foreword

This report offers an overview of some novel methods of analysing research funding data. The UK Arts and Humanities Research Council (AHRC) has worked with Digital Science Consultancy on an exploratory project using innovative techniques to offer new perspectives of the AHRC's competitively funded research. It has unlocked rich information within the funding data, not through the existing discipline taxonomies but through analysis of the content of application summaries. This has uncovered changes across time and location, and at varying levels of detail. It's the first time topic modelling has been applied on a national funding dataset of leading, peer reviewed research grants activity within the arts and humanities.

Typically, with the exception of Research Assessment Exercises or the Research Excellence Framework (REF), descriptions of research activity and performance have concentrated on the fields of science and technology. Larger-scale public funding for the arts and humanities in the UK began only in 1998 with the creation of AHRB which preceded the AHRC. Limited data availability has also been a barrier to the types of methodologies used. Unlike the natural sciences, and increasingly prioritised among engineering and some of the social sciences through the 1990s (Adams & Gurney 2015), journal articles are not the key research dissemination medium for arts and humanities research. With little indexed data, the arts and humanities continued to be marginalised in national audits such as Elsevier's annual international comparator reports (Elsevier 2013).

We are now seeing new methods, and more inclusive databases being developed which can begin to change this skewed emphasis. This is particularly valuable as the arts and humanities is increasingly recognised for their wider contributions to the knowledge, creative and cultural economies. From the recognition of disciplines like design as critical contributors to competitive value in technology - e.g. the Apple iPhone - to the role of Philosophy underpinning software development. Disciplines such as History, Museum Studies, Classics, Literature, Archaeology, the Performing and Visual Arts generate significant cultural revenue and wider benefits.

Our ability to use novel techniques for data analyses has been transformed by changes in computing power and data availability. These digital transformations, including the advances in digital humanities, allow us to better tackle entire publication corpuses including books, their chapters, conference series, grey literature, public reports and other media. We no longer need to rely solely on curated citation lists, nor do we have to solely depend on predetermined metadata such as titles and keywords, but can instead draw on full textual content.

This offers exciting new possibilities but challenges remain. Recognising this will allow us to find ways of framing topic modelling analyses and interpreting the findings appropriately. We identified the following as important considerations when interpreting the data. First, recognising that some words or phrases might have multiple meanings across disciplines. Context, therefore, is key; analyses such as these require interpretation by subject experts.

Second, proposal summaries were used. As applicants are asked to write summaries for a lay audience they are less likely to use domain-specific terminology. Third, there was considerable interest in highlighting hotspots of research activity, using a set of words or phrases. However, in the face of these contextual challenges, the words of the Astronomer Royal, Lord Rees that "the absence of evidence is not evidence of absence" rings true!

Notwithstanding the challenges, this report shows that topic modelling, used appropriately, helps us to: uncover potential links within the data; offer a tantalising glimpse of the breadth of topics and concepts of interest; and, demonstrate a possible text-focused approach to understanding the nature of interdisciplinary research.

"We no longer need to rely solely on curated citation lists, nor do we have to solely depend on predetermined metadata such as titles and keywords, but can instead draw on full textual content."

Sumi David and Sarah Wingrove
Strategy and Development Manager
Arts and Humanities Research Council

Introduction

"*Topic modelling, by contrast, is a bottom-up approach driven by the material that is available*"

This report presents a series of visualisations of a topic model - a landscape analysis - based on processing the title and abstract text of ten years of grant applications to the Arts & Humanities Research Council (AHRC). This kind of analysis, building a picture of common topics using researchers' own descriptions of their work, has only recently become possible. The methodology is relatively complex, somewhat value-laden and not entirely settled, so we have summarised this at the end of the report and listed key references for those interested.

The key purpose of this report is to show what can now be done with topic models, to illustrate some of the different 'landscape features' that the contrasting visualisations reveal about arts and humanities research, and to provide a glimpse of material that have yet to be fully explored but - linked to expert and experienced peer review - may come to be a valuable additional tool in research management and policy.

This analysis is about grants, but it could have been about other parts of the huge diversity of text that describes research activity: grant proposals, clinical trials, journal articles, books, reports, and patents. The challenge is that any large corpus of documents is difficult to understand as a single entity (which is why bibliometrics reduces research articles to rather superficial numbers). Large text stores, such as libraries, have long been organised via standard classifications to help search and discovery and to understand what the document set represents. Classification follows a top-down approach and to do this you need to understand the entire library: first, to create a classification: then, to assign items.

The fields of research used to structure the millions of articles published every year are also created top-down, They are based on journals; they are rather inflexible; and they do not capture the changing reality. For example, where do we find new fields of research? How fast can they be introduced into formal classification? If soil science includes biogeochemistry, plant nutrition, ecology, genomics, pedology, then how are multi- and inter-disciplinary research captured?

Topic modelling, by contrast, is a bottom-up approach driven by the material that is available. It uses computational methods that analyse the text of each document to create a novel and specific classification (topics). It then assigns each document to one or more topics. The statistical method uses word frequency, and is entirely adaptable to new sets of documents used for each new purpose.

Topic modelling, as a data-driven approach, can model millions of documents. Documents belong to multiple topics, and a significant proportion of the documents (though not always all) identified in a topic about 'music' and 'Wales' will actually be about 'Wales and music'. The granularity of the classification depends on the number of topics selected.

Topic modelling is a promising approach that can capture trends in research production. It can map documents to time patterns and spatial distribution. Potentially, combined with expert interpretation, it could leverage information to create insights on emerging and branching topics. Digital Science uses this powerful tool to study publications, grants, case studies and other documents and provide understanding of topic evolution, clusters and trends.

Presentation of the Topics

"There is a dynamic transition reflecting intellectual development."

The Arts and Humanity Research Council (AHRC) supports excellent research within the UK. For this analysis we used the successful and unsuccessful applications for grants between 2005 and 2016. Analysis based on these proposal documents shows topics on which researchers focussed in this period, both as overall composition and as temporal trends.

Applications linked to any one topic may vary through time as the text referencing topics merge and new topics emerge. There is a dynamic transition reflecting intellectual development, but note that some changes in the frequency of proposals linked to a topic may reflect changing terminology, or even 'rebadging' around priority areas. That is why expert interpretation is essential.

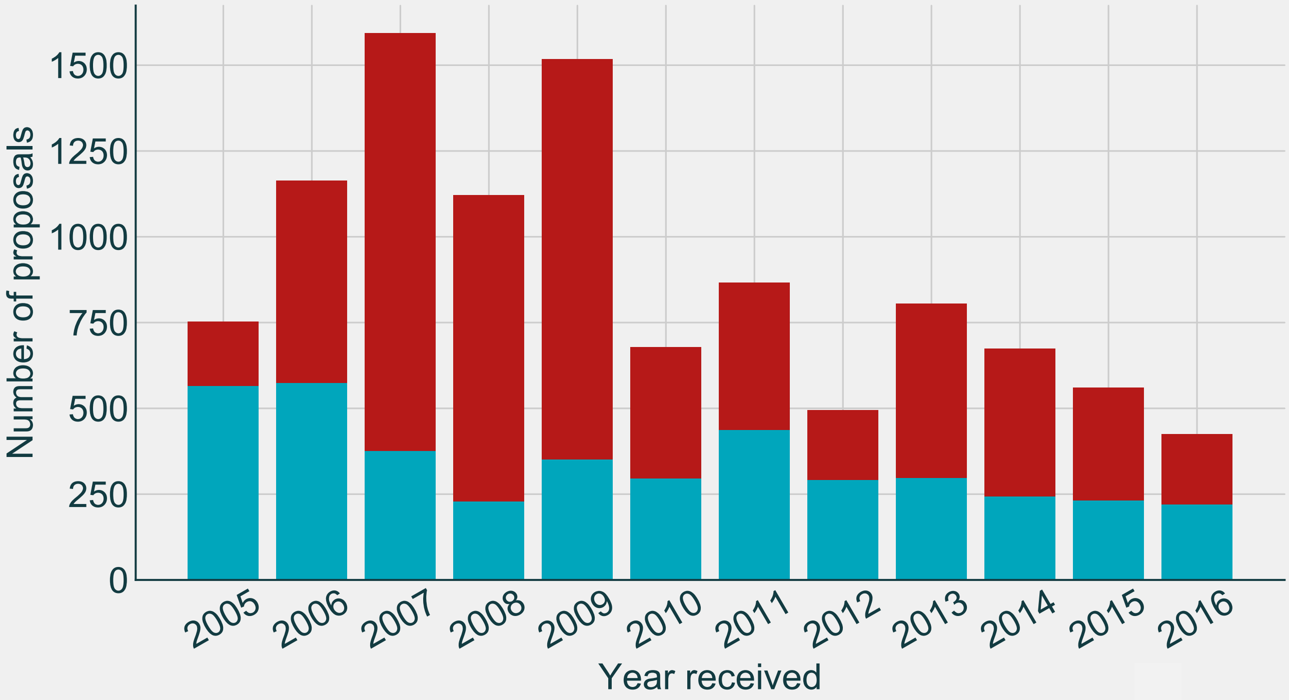


Figure 1. Number of research proposals received by AHRC per year. Successful awards are indicated in blue, unsuccessful in red. The difference in volume of funding before and after 2009 relates to changes made to funding mechanisms.

The text of each proposal contains a title (1 to 27 words, with an average of 10 words) and an abstract (21 to 842 words, with an average of 481 words).

Concepts found in AHRC text included compounds such as "five nations", "social classes", "Martin Luther King's", or "Ottoman Empire". Over 27,500 unique words were thus annotated, representing 20.5% of the vocabulary used in the combined grant texts. Wikipedia is an excellent reference set of crowd-sourced topical analysis and was used to identify these compound words via the text annotation tool DBPedia Spotlight (http://github.com/dbpedia-spotlight). This step is crucial in a meaningful topic modelling process.

"A target of 185 topics was chosen for the topic modelling process."

In this study, a target of 185 topics was chosen for the topic modelling process. Determining the operational number of topics is always a difficult task for modelling. Too small a number of topics will only provide very general groupings that are not useful for analysis. Too great a number of topics will become unwieldy and likely generate incoherent topics in which terms are not apparently related. Our approach to estimating an optimal number of topics is to generate a series of topic models across a range of values (e.g. 150-250 topics), measure the stability of topics generated (see later section on methodology), and manually evaluate those with a high stability.

A topic model defines each topic according to a weighted set of terms. Documents are assigned a weight for each topic according to the number of highly weighted terms they share, and are therefore associated with multiple topics. Typically, the top 20 most highly weighted terms are used as a topic descriptor because they are deemed sufficient to indicate the concepts modelled when they are presented to an informed expert group. Hence, in this report, we identify each topic with the top 20 terms (or the top 3 for brevity), where the first term list is the most highly weighted (or most strongly associated with the topic). For example, 'Designing Innovative Interventions with People Living with Dementia' (a 2014 research award) is assigned to multiple topics of which the highest weight is Topic 43 [dementia living carers interventions care social_care alzheimer well_being signage quality sensory healthcare stimulation needs care_homes symptoms benefit carer].

Some topics are produced by the frequent and widespread use of generic research terms, such as [concept concepts design method] and so on. Some, but not all, of these terms can be generically removed. Nonetheless, some topics can be seen to be of this nature.

Figure 2 shows the distribution of funded and unfunded applications for each topic. Above the diagonal line are topics that have proportionally more unfunded that funded applications, while below the line are topics that have more funded than unfunded applications. For example Topic 174 [concepts concept thought] had, over the 10-year period, 497 funded awards for 972 unfunded applications (34% successful), while Topic 158 [programme partner community_groups] has 539 funded for 798 unfunded applications (40% successful). There is rather little variation in success rates given the overall volume.

Some topics include many documents: for example, Topic 45 [workshop event disciplinary] which reflects methodologies; and Topic 4 [volume volumes publication] which reflects publication activity. Topics capturing big discipline groups, such as Topic 60 [writers intellectuals authors] and Topic 116 [american america united_states] also include many documents. Other topics are more specific (Topic 13 [dance dancers choreographic], or Topic 105 [eu governance european_union]) and therefore much smaller.

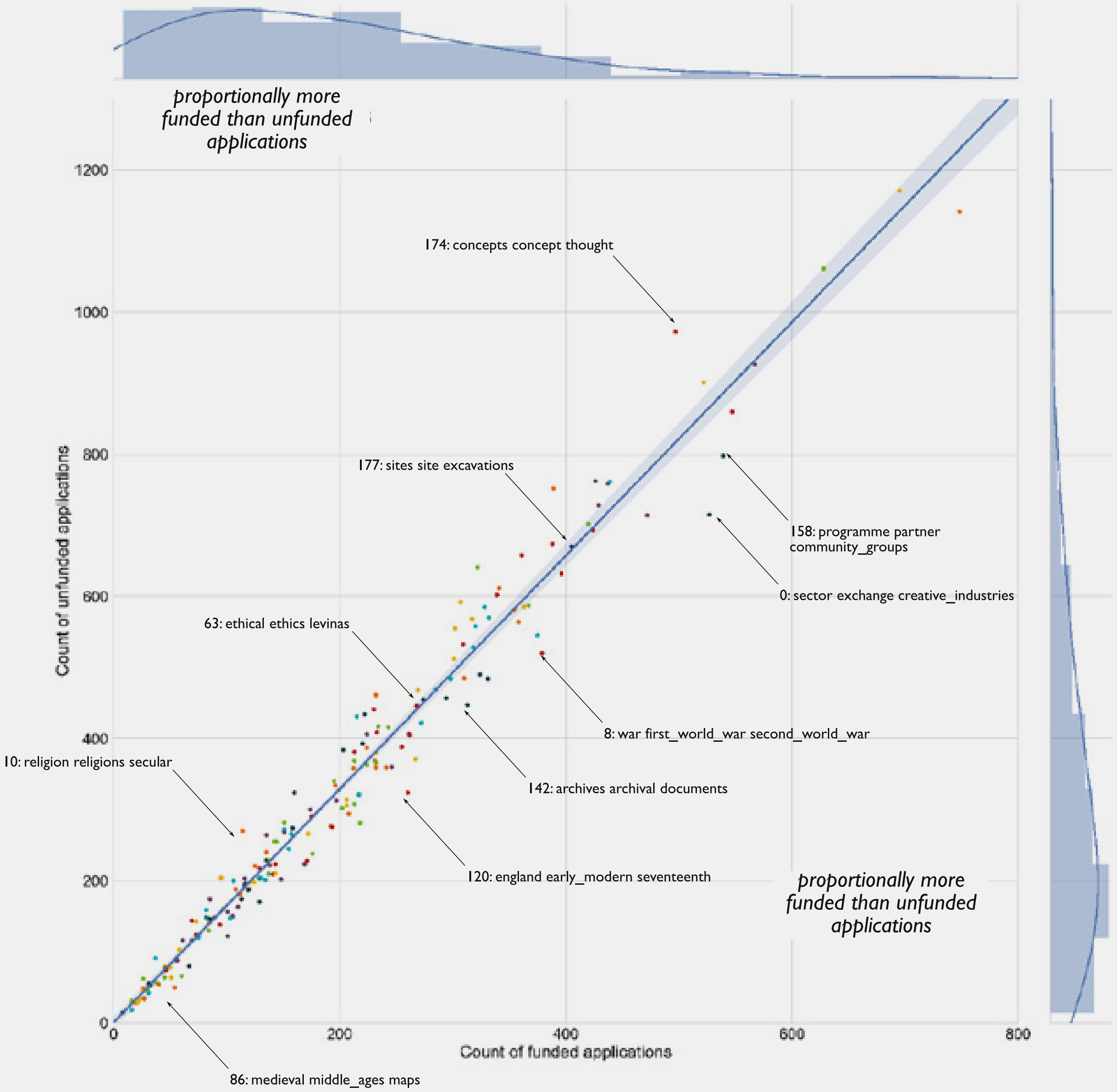


Figure 2. Comparison of the counts of funded and unfunded applications across topics. Each dot represents one topic.

Exemplar Topics

Each topic is constructed of many grant applications submitted over 10 years. The time series for each topic therefore describes its trajectory, which may be attributable to a change of topical interest (e.g. Topic 32 [conflict peace conflicts]) or choice of method (e.g. Topic 134 [programme partners community_groups]). In other cases, where topics correspond to general terms (e.g. Topic 79 [records record document]) or terms with multiple meanings (e.g. Topic 117 [interactive environment environments]) these variations are more difficult to interpret. For the analysis which follows, we feature topics that either rose constantly, or were rising sharply, or declined either as individual topics or as part of a coherent cluster. Note again, that some of this temporal variation may reflect changing preferences in terminology as well as underlying changes in focus. That requires further elucidation.

"Some of this temporal variation may reflect changing preferences in terminology as well as underlying changes in focus."

Steady Increase

Topic 31 *[global globalization globalisation]*

Topic 31 contains 2,363 documents. The most important terms (top 20, ordered by descending weight) are:

global globalization globalisation transnational cultures historians global_history cosmopolitanism globe networks connections western global_justice themes cosmopolitan histories transformations globally

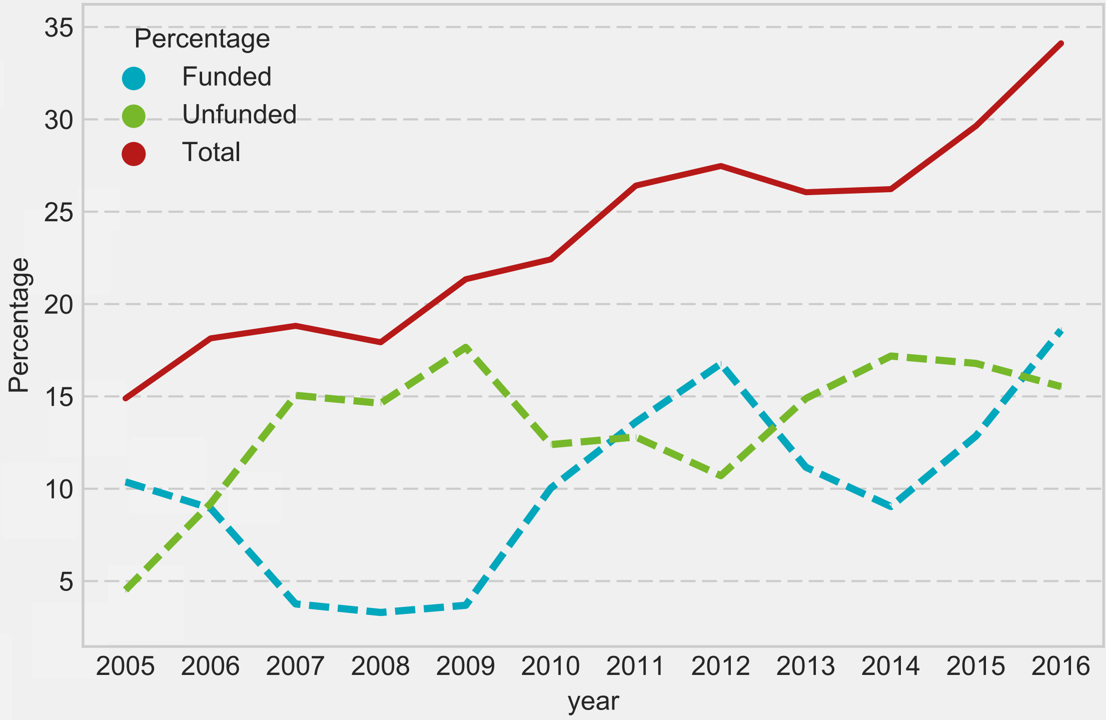
In 2005, at the start of the period analysed, 15% of all applications to AHRC were linked to this topic. This share slowly increased over ten years to one-third of applications, which may indicate a substantial increase in research interests around the topic but may also be influenced by the growing prominence of a global agenda in policy and research discourse.

Grants funded in this topic include: [AH/H034218/1] *The People's Car: A Global History of the Volkswagen Beetle*, [AH/H038477/1] *Words derived from Old Norse in Sir Gawain and the Green Knight: an etymological survey*; [AH/H039600/1] *Religion and the Origins of Modern Science*.

Fluctuation over the same period in the balance of funded and unfunded applications shows that this increase did not translate into more funded research projects. On the contrary, there was a decline in the relative number of awards assigned to this topic during the first half of the period. This may point towards the change in terminology.

The time series plot in Figure 3 (and others in the following sections) shows the number of grants classified in each topic as a percentage of the overall number of grants submitted in each year. This accounts for underlying fluctuations in submission numbers. Since multiple topics are assigned to a single grant, a sum over all percentages for each topic would exceed 100%.

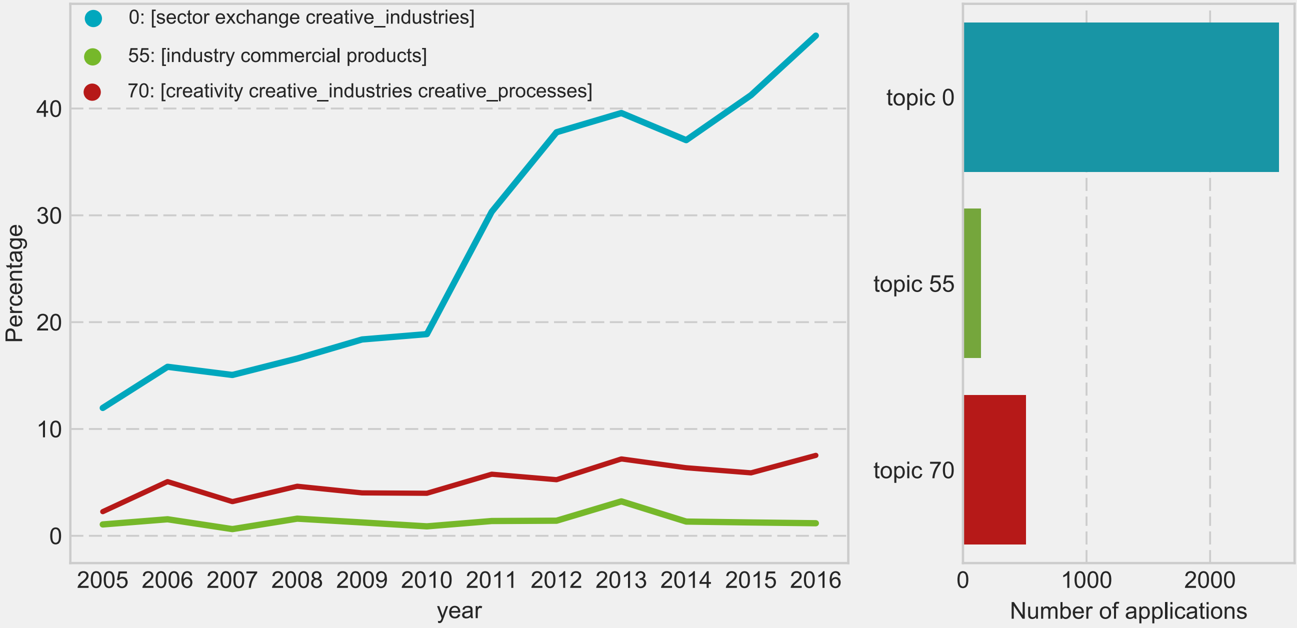
Figure 3. Time series of grants attributed to Topic 31.



Topics about Industries

Three topics were closely related to creative and commercial industries (Figure 4). These have in common a substantial increase in submissions at some point after 2010. However, only the two dealing with *creative* industries (Topic 0 [*sector exchange creative_industries*] and Topic 70 [*creativity creative_industries creative_processes*]) had an increase in success rates. Topic 55 [*industry commercial products*] does not include any creative-related term.

Figure 4. Left: Time series of grants attributed to three topics about industries. Right: Number of applications per topic.



Topic 134 *[programme partners community_groups]*

During the 10-year period, Topic 134 [*programme partners community_groups connected phase cc partnerships partnership ahrc funding build first_world_war hlf funded partner building skills co_design*] more than quadrupled in frequency as a percentage of applications submitted. This may be associated with the cross-Council AHRC-led Connected Communities programme and reflects the stimulus of directed initiatives.

"This may be associated with the cross-Council AHRC-led Connected Communities programme and reflects the stimulus of directed initiatives."

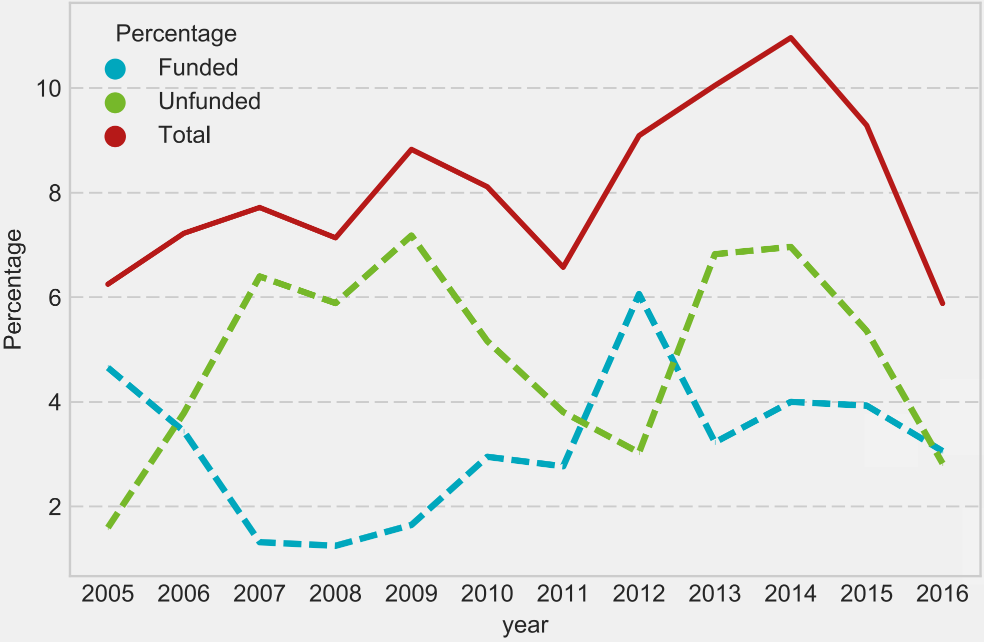


Figure 5. Time series of grants attributed to Topic 134 *[programme partners community_groups]*.

> *"The link may be to both the PaCCS (the Partnership for Conflict, Crime and Security) and the Global Challenges Research Fund."*

Abrupt Increase

A small number of topics saw a sharp increase after the year 2014. Figure 6 shows the temporal variation of applications submitted.

Topic 32 *[conflict peace conflicts*] includes funded grants such as 112869/1 *Congo's War: The Legal Dimension of a Protracted Conflict* or AH/P00492X/1 *Stories from Rwanda: Academic, Creative, Applied*. The link may be to both the PaCCS (the Partnership for Conflict, Crime and Security) and the Global Challenges Research Fund

Topic 162 [*slavery slave slaves*] include funded grants such as AH/P008690/1 *The Poetry of the Lancashire Cotton Famine (1861-65)* or AH/D500850/1 *Thinking America: Public Intellectuals and the Framing of National Identity, 1837-1909*. This topic evidently captures the complex nature of issues such as slavery which are both a direct subject for research and a framing mechanism for other projects.

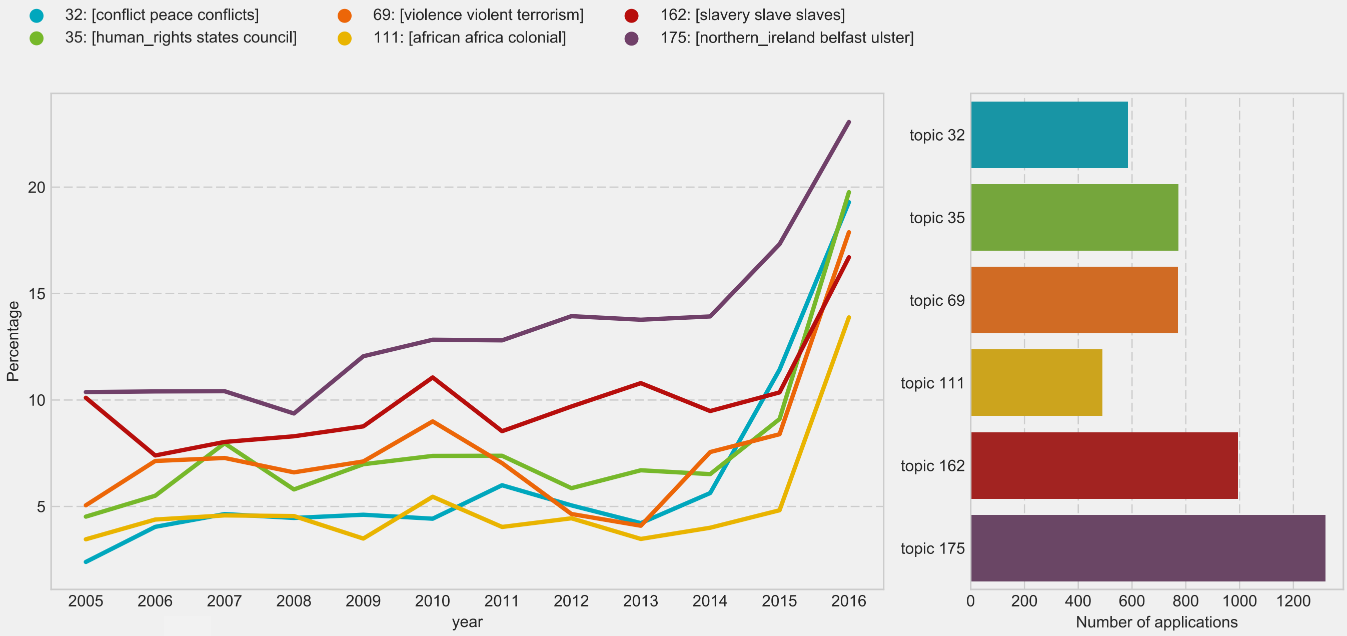


Figure 6. Left: Time series of grants attributed to six topics with an abrupt increase. Right: Number of application per topic.

Steady Decline

A few topics showed a decline in number of submissions. Topic 147 [*colour colours painting*], with 824 applications between 2005 and 2016, was one of these. At the start of the period, the number of applications nearly doubled (6.5% to 11.9%), but then progressively fell to 3% in 2016.

The terms in the topic are: [*colour colours painting techniques pigments pigment blue paintings cobalt style wool inks perception kinemacolor rp fragrance paint coloured*]. The topic included projects which cross many subjects; e.g.: AH/M005364/1: *Innovating Infographics in Public Health*, AH/H033688/1: *Looking at the overlooked: Renaissance and Early Modern Prints and Drawings from Spain*, AH/J007285/1: *Tweed: History, Culture and Design*, and AH/M005569/1: *Persons as Animals: Understanding the Animal Bases of Agency, Perceptual Knowledge and Thought*.

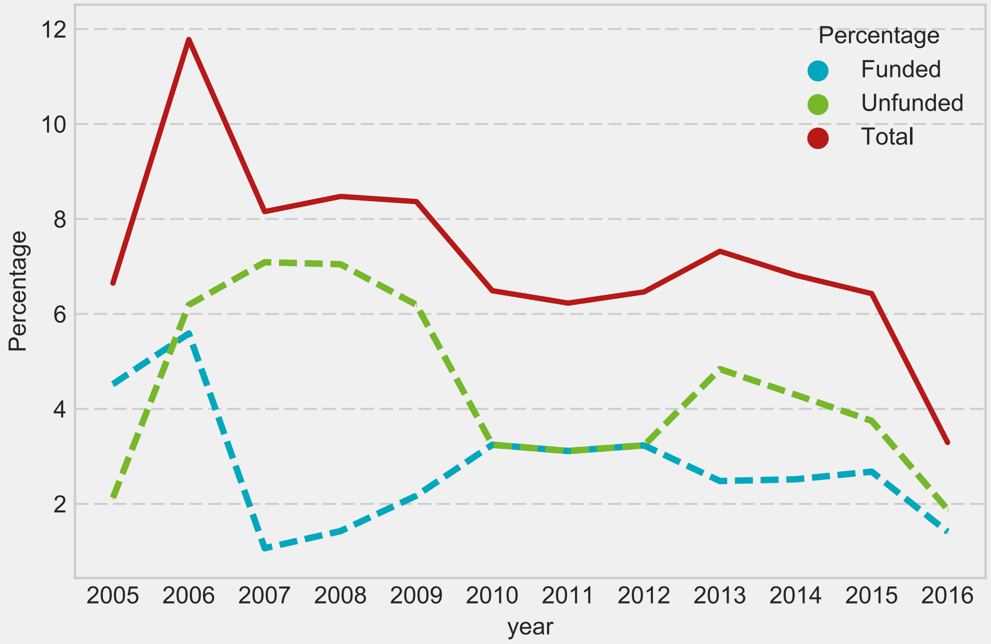


Figure 7. Time series of grants attributed to Topic 147 *[colour colours painting]*

Grants Similarity Network

Topic modelling across a large set of documents such as grant applications can, as the above examples demonstrate, track the discovered topics to reveal trends and highlight patterns obscured by the sheer scale of the overall corpus. Some of these prove to be informative but others either require more exploration or are artefacts of terminology.

Topic modelling can also be used to provide a categorical backbone on which data can be projected. One useful way to apply this is to create a network of grant applications to show the overall topical landscape. For the AHRC data, such a visualisation is presented in Figure 8.

In this diagram, each dot represents a grant application and is coloured according to the primary topic assigned (the topic with the highest weight). Edges connect similar applications based on the topics they share.

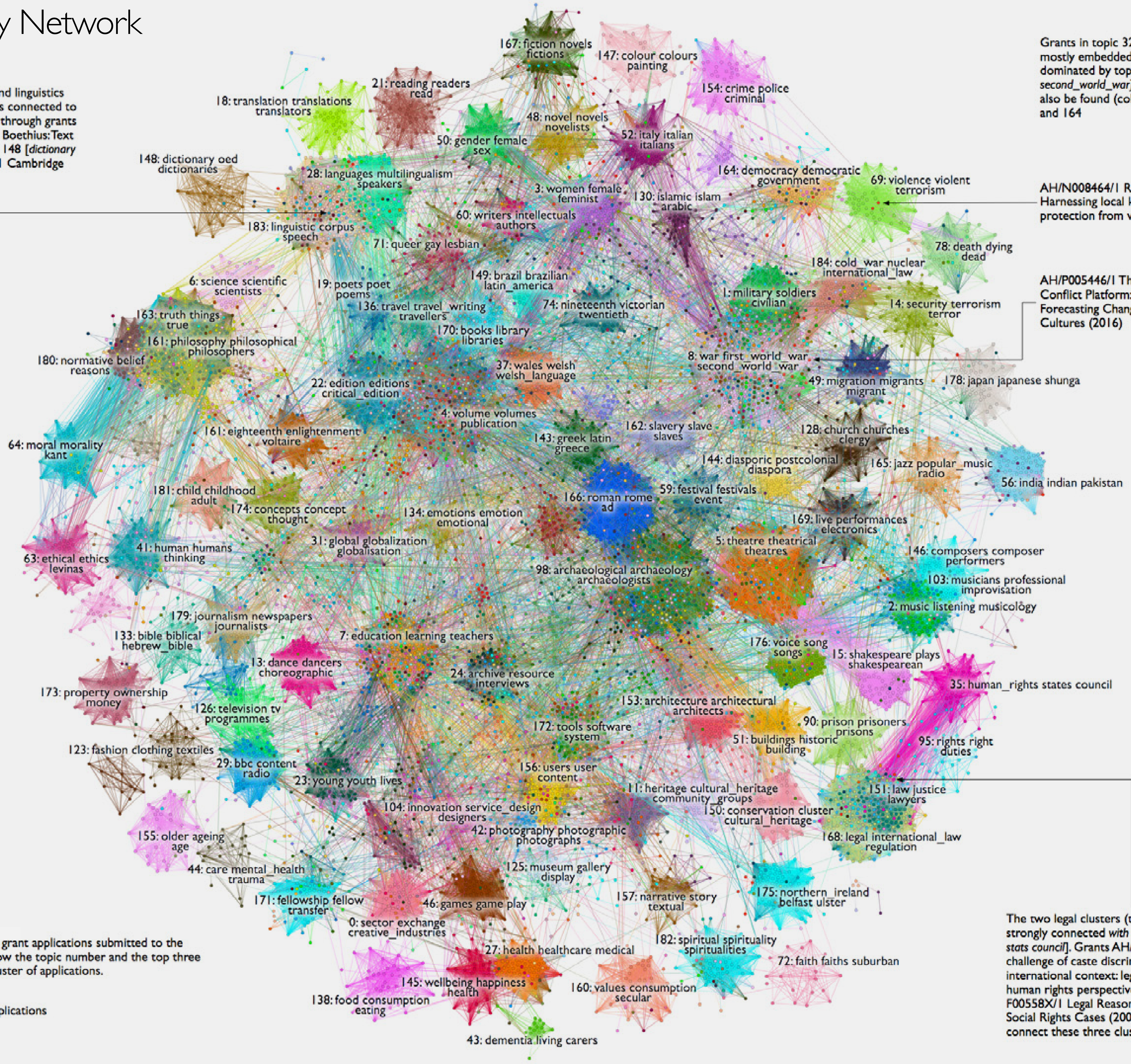
Callout boxes have been added to the figure to highlight interesting features of the network.

> *"Topic modelling can also be used to provide a categorical backbone on which data can be projected."*

Grants Similarity Network

The large cluster containing language and linguistics grant applications (topics 28 and 183) is connected to topic 22 [*edition editions critical_edition*] through grants such as AH/E505120/1 The Old English Boethius: Text and Interpretation (2006) as well topic 148 [*dictionary oed dictionaries*] through AH/D503507/1 Cambridge New Greek Lexicon Project (2005).

Grants in topic 32 [*conflict peace conflicts*] are mostly embedded within the larger cluster dominated by topic 8 [*war first_world_war second_world_war*], but many examples can also be found (coloured red) in topics 69, 49 and 164


167: fiction novels fictions
147: colour colours painting
21: reading readers read
154: crime police criminal
18: translation translations translators
48: novel novels novelists
50: gender female sex
52: italy italian italians
148: dictionary oed dictionaries
28: languages multilingualism speakers
164: democracy democratic government
69: violence violent terrorism
3: women female feminist
130: islamic islam arabic
60: writers intellectuals authors
183: linguistic corpus speech
71: queer gay lesbian
78: death dying dead
149: brazil brazilian latin_america
184: cold_war nuclear international_law
6: science scientific scientists
19: poets poet poems
1: military soldiers civilian
74: nineteenth victorian twentieth
14: security terrorism terror
136: travel travel_writing travellers
163: truth things true
170: books library libraries
161: philosophy philosophical philosophers
180: normative belief reasons
8: war first_world_war second_world_war
37: wales welsh welsh_language
49: migration migrants migrant
178: japan japanese shunga
22: edition editions critical_edition
4: volume volumes publication
162: slavery slave slaves
128: church churches clergy
161: eighteenth enlightenment voltaire
143: greek latin greece
64: moral morality kant
144: diasporic postcolonial diaspora
165: jazz popular_music radio
56: india indian pakistan
181: child childhood adult
59: festival festivals event
166: roman rome ad
174: concepts concept thought
134: emotions emotion emotional
169: live performances electronics
31: global globalization globalisation
5: theatre theatrical theatres
41: human humans thinking
146: composers composer performers
63: ethical ethics levinas
98: archaeological archaeology archaeologists
103: musicians professional improvisation
2: music listening musicology
179: journalism newspapers journalists
7: education learning teachers
133: bible biblical hebrew_bible
176: voice song songs
13: dance dancers choreographic
15: shakespeare plays shakespearean
24: archive resource interviews
35: human_rights states council
173: property ownership money
126: television tv programmes
153: architecture architectural architects
90: prison prisoners prisons
172: tools software system
51: buildings historic building
95: rights right duties
123: fashion clothing textiles
29: bbc content radio
156: users user content
23: young youth lives
11: heritage cultural_heritage community_groups
151: law justice lawyers
104: innovation service_design designers
150: conservation cluster cultural_heritage
42: photography photographic photographs
155: older ageing age
168: legal international_law regulation
125: museum gallery display
44: care mental_health trauma
175: northern_ireland belfast ulster
157: narrative story textual
171: fellowship fellow transfer
46: games game play
0: sector exchange creative_industries
182: spiritual spirituality spiritualities
27: health healthcare medical
72: faith faiths suburban
145: wellbeing happiness health
160: values consumption secular
138: food consumption eating
43: dementia living carers

AH/N008464/1 Raising Silent Voices: Harnessing local knowledge for communities' protection from violence in Myanmar (2015)

AH/P005446/1 The Changing Character of Conflict Platform: Understanding, Tracing and Forecasting Change across Time, Space and Cultures (2016)

The two legal clusters (topics 151 and 168) are strongly connected with topic 35 [*human_rights stats council*]. Grants AH/E001440/1 The challenge of caste discrimination in an international context: legal responses from a human rights perspective (2006) and AH/F00558X/1 Legal Reasoning in Economic and Social Rights Cases (2007) are examples that connect these three clusters.

Legend

Figure 8 - 10,058 funded and unfunded grant applications submitted to the AHRC between 2005-2016. Labels show the topic number and the top three topic terms over the corresponding cluster of applications.

- ● Nodes are grant applications
- ●–● Edges connect similar grant applications
- ●●● Colours denote primary topic

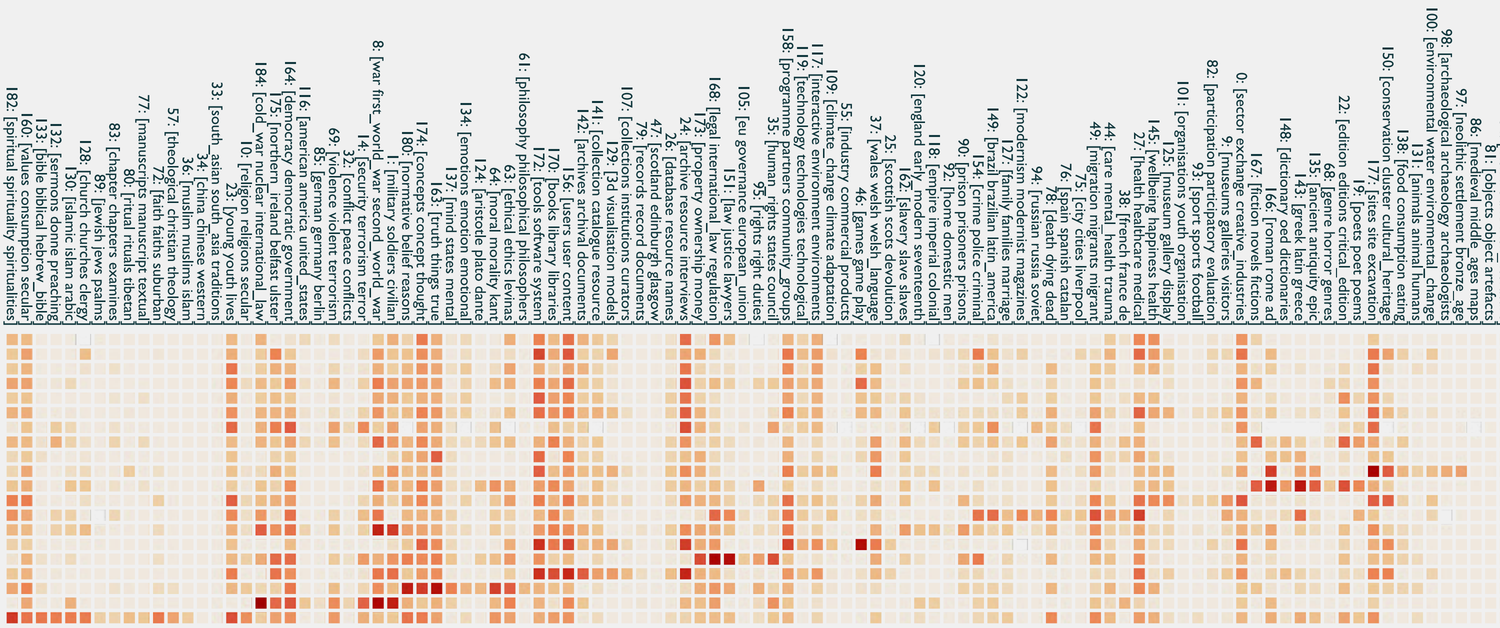


Figure 9. Heatmap of relative number of document in topics, across researcher-assigned categories

Topic Heatmap

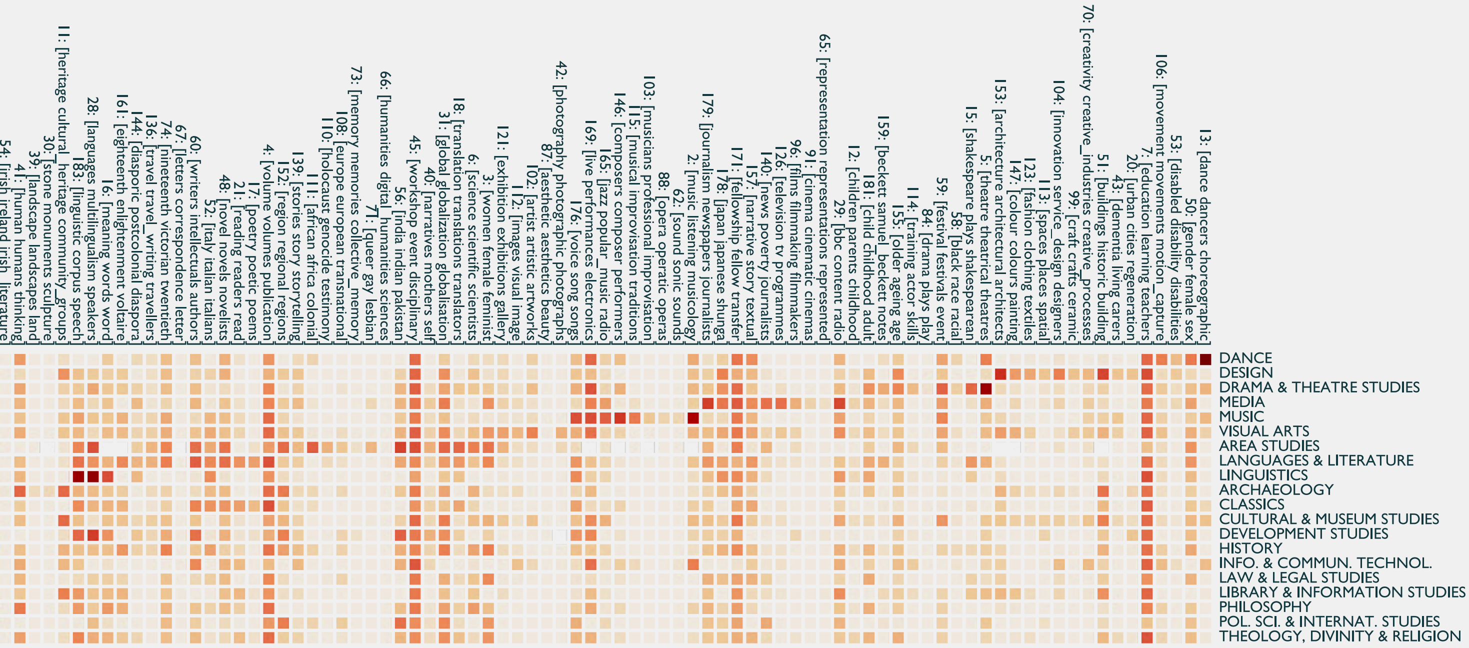
A third way of presenting the AHRC grant applications and looking for informative patterns in the topics created by topic modelling of the text is via a heatmap. The heatmap in Figure 9 makes use of additional metadata provided by applicants for each grant to compare a pre-existing categorical structure to the emergent topic model. Some topics spread across categories while others map 1-to-1.

Each AHRC grant proposal is given up to four categories, from a standard two-tier set used across Research Councils, assigned by the researcher at submission. These relate to the research type, location and subject. Following a change to the classification system in 2010, proposals also indicated the primary category. Using the 4,474 proposals that have a primary subject, a heatmap was created to show the topics with the highest weights for the 20 categories (Figure 9).

The heatmap indicates the distribution of weight for each topic: the darkest square in each column is the one with the highest weight. Although some topics are spread across multiple categories, there are a few cases where a single topic corresponds well to a single category (e.g. topic 13 [*dance dancer choreographic*]). Topics that spread across the greatest number of categories are usually about methodology (e.g. 45 [*workshop event disciplinary*]).

CLASSICS crosses over multiple topics: first (Topic 152 [*region regional regions*], Topic 4 [*volume volumes publication*], Topic 17 [*poetry poetic poems*], Topic 21 [*reading readers read*], 48 [*novel novels novelists*], and Topic 52 [*italy italian italians*]) and second (Topic 19 [*poets poet poem*], Topic 22 [*edition editions critical_edition*], Topic 68 [genre horror genres], Topic 135 [*ancient antiquity epic*], Topic 143 [*greek latin greece*], Topic 148 [*dictionary oed dictionaries*], Topic 166 [*roman rome ad*], and Topic 167 [*fiction novels fictions*]).

> *"Topics that spread across the greatest number of categories are usually about methodology."*



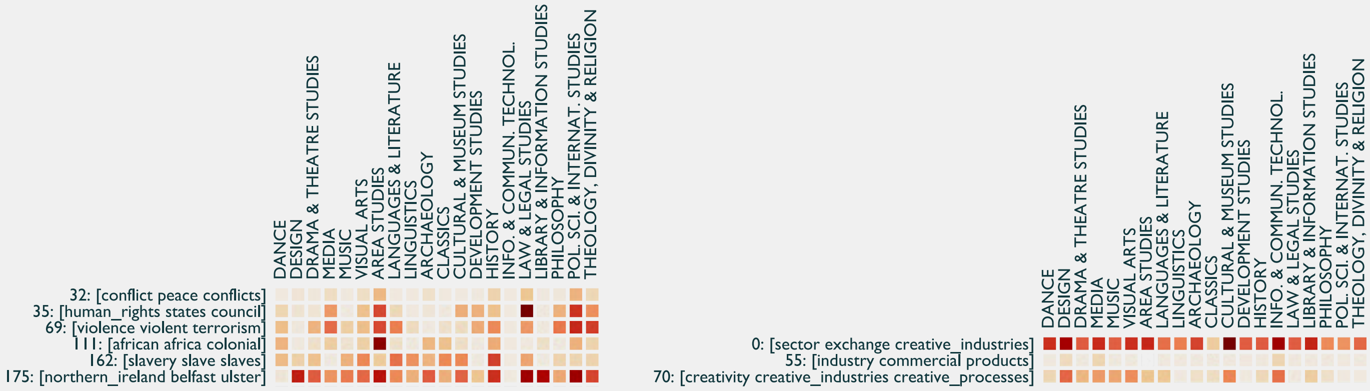


Figure 10. Extracts from Figure 9 showing topics that steadily increased (Left) and those related to industries (Right).

Figure 10 shows two extracts from the bundles of topics previously described, relating respectively to those with steady increase and sudden increase in occurrence.

The conflict topics appear in very different categories. Their spread may indicate interdisciplinarity and/or specialisation: Topic 175 [*northern_ireland belfast ulster*] is present in most categories (even including categories such as design and therefore not directly dealing with the Irish conflict), while others are found in few categories and therefore more specialised.

Of the three 'industries' topics (see above), the topic including the largest number of projects (sector exchange creative_industries) is spread across many categories, while the topic related to 'creativity, creative_industries creative_processes) is present in only half. The topic which emerged around the words 'industry commercial products" is present mostly in MEDIA and INFO. & COMMUN. TECHNOL.

Outcomes

"*This may uncover interdisciplinary work where clusters of applications appear that link different topics, or it may identify areas of research that are relatively isolated.*"

In this report, we have demonstrated that topic modelling is a useful analytical tool that can draw a range of informative interpretations from a large text corpus, specifically grant applications.

The large network diagram (Figure 8) shows that such a model can provide an overall landscape of the content that captures many different aspects of the research. This kind of backbone can then be used as a basis for comparison, e.g. by highlighting the different areas of the network that correspond to applications from particular universities. Alternatively, this may uncover interdisciplinary work where clusters of applications appear that link different topics, or it may identify areas of research that are relatively isolated.

Topic modelling faces some challenges when applied to text about arts and humanities research because the use of domain-specific terminology differs from that in science and engineering and across different arts and humanities disciplines where the same word e.g. "medium" might be used with a different specific connotation by different disciplines. By contrast, biomedical text containing disease names, molecular entities, mathematical techniques and similar specialised terms is readily classified into research disciplines. However, although specific terminology was often not present in the AHRC grant applications that were analysed (except for linguistics and archaeology), it was found that the use of proper nouns, especially for people and places, can play an important role in characterising topics instead.

One fundamental limitation of topic modelling is that it is unable to parse sentence structure or understand language semantics. Topics are identified by words that appear frequently together in the same documents. There is no underlying language model. As a result, polysemy can create problems with interpretation because single words are used in different contexts with different meanings. To an extent, this can be mitigated through pre-processing. For example, the term resolution appeared many times in the grant applications, but was often used in conjunction with other contextual words that were joined during the pre-processing phase. This enabled the topic modelling algorithm to differentiate between uses in *conflict resolution*, *spatial resolution* and *high resolution* because these compound words were treated as single tokens. Nevertheless, other words, such as environment, are much more difficult to differentiate semantically because they are not so easily discern by the adjoining co-words.

"*When used in combination with other categorical schemes topic modelling can add nuance to analysis by showing the makeup of particular categories or show how well the content has been captured by a particular categorical scheme.*"

Crucially, for our purposes, topic modelling provides a categorical framework that is driven by the text content alone. It is not determined by pre-existing heuristic beliefs about what the content contains. When used in combination with other categorical schemes (as in Figure 9) topic modelling can add nuance to analysis by showing the makeup of particular categories or show how well the content has been captured by a particular categorical scheme.

From a funder's perspective, topic modelling may prove useful in conveying some aspects of the changing focus of grant applications. It may also offer new perspectives on aspects of the research landscapes and these can help inform policy decisions and provide more evidence for strategic decision making. It is essential that it is used only to inform and not to replace expert review mechanisms.

Methodology

Pre-processing

Text mining approaches such as Topic Modelling rely heavily in the words contained in the documents. Pre-processing, to clean and prepare the text to be ingested into Topic Modelling algorithms, is therefore an important step to create some meaningful results. Pre-processing can include the following steps:

"Pre-processing is therefore an important step to create some meaningful results."

- Removing too short text, as it would not be sufficiently long to assign topics confidently. This typically happened when the abstract was missing from the record.
- Creating compound words using a dictionary. Compound words could also be created using the text (e.g. if 'climate' is often enough followed by 'change', then 'climate change' is identified as a bigram). It is also possible to use a dictionary or encyclopaedia; for example, compounds of more than two words that have a Wikipedia entry can be considered as compounds during the pre-processing. We marked these with an underscore (e.g. climate_change) to facilitate the future processing.
- Normalisation of diacritic characters, as these are often used inconsistently across documents.
- Removing the most frequent and least frequent words. Terms that are only used once or twice in the entire corpus are not informative and only bloat the computation. Words that are used frequently (often referred to as stop words) create links between unrelated content and are therefore removed. For given corpus, than can include domain specific terminology such as 'research' or 'proposal'.

Topic Modelling

There are many different approaches to topic modelling and a wide range of software packages that provide it. Although Latent Dirichlet Analysis (Blei et al 2003) and the associated software implementation MALLET (McCallum 2002) have been used extensively, we have found Non-negative Matrix Factorisation (Dhillon and Sra 2005) to be a practical alternative that produces models with more coherent and less general topics. By using term frequency-inverse document frequency (Jones 1972) to model the appearance of words in documents (as opposed to the frequency of the word in the document), a more balanced set of topics can be produced because frequently used terms receive less importance than infrequently used terms.

Selecting the Number of Topics

The topic modelling algorithm requires to select the number of topics in which the corpus of documents should be divided into. Selecting that number is quite challenging because it depends on the data itself (size and diversity of the corpus) and the analysis intended. A few variables can be computed to help selecting a number of topics which works for the corpus and the analysis. The topic number selection is supported by:

- the distribution of documents in topics: minimum and maximum number of documents in each topic to avoid large topics and too small topics
- the stability of topics (variation of topic existence when comparing to the two adjacent smaller topic models and the two adjacent bigger topic models)

Interpreting and Labelling the Topics

> *"Topics are different from themes or categories; they represent the words that appear together in documents, regardless of their meaning."*

Topics are different from themes or categories; they represent the words that appear together in documents, regardless of their meaning. While they often bring together related terms that align well with concepts such as research discipline, location, methodology or stakeholder group, they can also reveal idiomatic or pragmatic features of the text corpus. For example, research documents such as grant applications or article abstracts will often contain non-research content such as copyright statements or phrases about the purpose of the research. These will be captured by the topic model, but can be filtered out.

Topics can be labelled for convenience, with the best results achieved using input from domain experts. However, the labelling process can lead to over-interpretation since a human will draw on background experience to infer relationships between terms that may not be present in the text.

References

Adams, J., & Gurney, K. A. (2015). Evidence for excellence: has the signal overtaken the substance?. London: Digital Science. https://www.digital-science.com/resources/digital-research-reports/digital-research-report-evidence-for-excellence-has-the-signal-overtaken-the-substance/

Blei, D. M., Ng, A. Y., Jordan, M. I. (2003) Latent Dirichlet Allocation. The Journal of Machine Learning Research 3(Jan):993-1022

Elsevier (2013) International comparative performance of the UK research base - 2011. London, UK: Department of Business, Innovation and Skills (BIS). https://www.gov.uk/government/publications/performance-of-the-uk-research-base-international-comparison-2013

McCallum, A. K. (2002) "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu

Dhillon, I. S. and Sra, S. (2005) Generalized nonnegative matrix approximations with Bregman divergences. In Proceedings of the 18th International Conference on Neural Information Processing Systems (NIPS'05), MIT Press, Cambridge, MA, USA, 283-290.
Jones, K.S. (1972) "A statistical interpretation of term specificity and its application in retrieval". Journal of Documentation, 28 (1).

Wilson, H. (1963). Speech to the Labour Party conference. In Labour Party Annual Conference Report, Scarborough (Vol. 1, pp. 139-40).

DIGITAL science

Part of the **Digital Science** family

Altmetric

BIOCRAFT

DIGITALscience Consultancy

Dimensions

figshare

GRID

ifi CLAIMS

labguru

Overleaf

Peerwith

readcube

SYMPLECTIC

TETRASCIENCE

transcriptic

digital-science.com