Creative Industries
Policy & Evidence Centre
Led by nesta

Exploring Societal Challenges through Library Data

ISBN: 978-1-913095-56-7

India Kerle and Cath Sleeman
November 2022

Introduction

From film credits to library catalogues, the creative industries are constantly, passively generating rich textual data. Could this data be used to shed new light on societal challenges? At Nesta, we're focused on three societal challenges: A Sustainable Future, A Healthy Life and A Fairer Start.

Our Sustainable Future mission aims to accelerate the decarbonisation of household activities in the UK. In this article, we explore how library catalogue data can enrich our understanding of this mission and its evolution overtime. While we must be mindful of methodological limitations, we find that there has been both a broadening and deepening of library material related to sources of renewable energy. This supports the idea that methodologies that take advantage of untapped data emerging from creative sectors could not only help us better understand the world we live in but also reveal trends we may not otherwise see.

Beyond the renewable energy space, library catalogue data could serve many different topic areas, including Nesta's other key mission areas: A Healthy Life and A Fairer Start. For example, we could use library catalogue data to explore how the topic of health inequalities across obesity and loneliness have changed over time. More specifically, when did academic literature start to reflect the increasingly evidenced notion that this inequality is socially determined? Meanwhile, for A Fairer Start, what could we learn about literature on parenting behaviours over time? Library catalogue data does not simply reflect publications available on different topics, it also reflects the dynamic relationship between how we understand social issues and how this is reflected in academic literature.

Querying the Library Hub Discover Catalogue

We are able to access decades of library data by using an open API from Jisc's Library Hub Discover, an aggregated catalogue of over 48 million records of library holdings from over 170 institutions in the UK. The libraries represented in the dataset range from major academic institutions such as the University of Oxford Libraries, SOAS Library and the University of Reading Library to museum libraries such as the Tate and the British Museum, as well as specialist libraries such as Historical Texts, the Institute of Naval Medicine and the National Aerospace Library.

The records in Library Hub Discover contain:

1. **Bibliographic data** including the record title, author's name, publication details and subjects and;
2. **Holdings information** including the name(s) of institution(s) that hold the record(s).

An example of the bibliographic data record returned from the API is provided below:


```
{'bibliographic_data': {
'author': ['Neal, L. G.'],
'physical_description': ['ix, 227 p. :'],
'publication_details': ['Washington, D.C. : National
Aeronautics and Space Administration ; Springfield,
Va. : For sale by the National Technical Information
Service [distributor], 1971.'],
'subject': [
'Nuclear electric power generation.',
'Heat pipes.',
'Feasibility.',
'Rankine cycle.',
'Heat radiators.',
'Heat pumps.',
'Electric generators.',
'Capillary flow.',
'Electric power production.',
'Heat pipes.',
'Rankine cycle.',
'Heat Transmission.',
'Heat pumps.',
'Electric generators.'],
'title': ['Study of a heat rejection system using
capillary pumping / L.G. Neal, D.J. Wanous, and O.W.
Clausen.'],
'url': ['http://hdl.handle.net/2027/
uiug.30112106857045']}
```


fields of interest

To find the collection of records relating to renewable energy within library holdings, the open API was queried with **seven key terms**. These terms were either generic in topic or specific to different forms of renewable energy, such as solar power. A record was returned if the query term showed up anywhere within the record i.e. as part of the record's subject list or in the record's summary. The publication year was also extracted, which gave a high-level sense of when renewable energy topics were first discussed in resources contained within the catalogue. We were able to extract the publication year from publication details because the year is often contained in the string of the text. The publication year in the publication details is highlighted in yellow above.

Please note that the numbers below are subject to change as the catalogue is updated regularly. Examples of unique subjects can be found in the returned bibliographic data record above.

Figure 1: Records in Library Hub Discover relating to sustainable energy

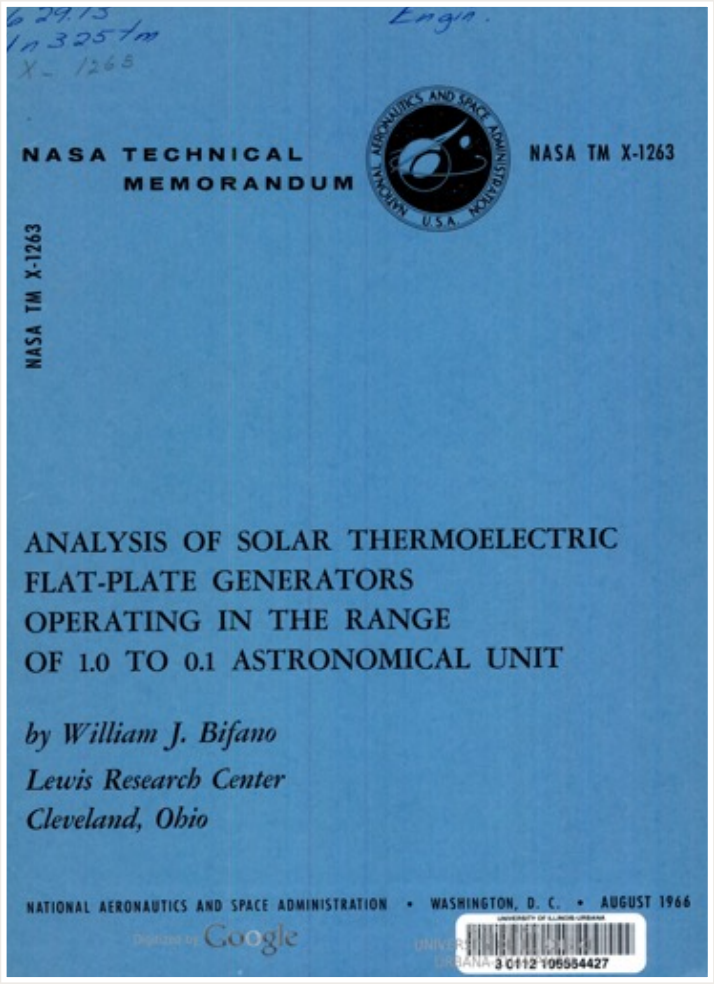
Keyword area	Queried keyword	Number of returned records	Number of subjects	Number of unique subjects	Publication year range
Heat pump	'heat pump'	773	4,731	1,975	1845 – 2022
Heat pump	'home retrofit'	5	27	24	2006 – 2019
Solar energy	'solar panel'	41	277	204	1966 – 2021
Solar energy	'solar pv'	32	880	705	1996 – 2020
Solar energy	'solar energy'	4,938	22,014	204	1966 – 2021
General	'decarbon-isation	10	87	69	2009 – 2021
General	'renewable energy'	10,637	4,731	1,975	1845 – 2022

While this gives us a reasonable sense of the renewable energy space in the Library Hub Discover catalogue, there are inevitably limitations to both the data and keyword approach. Firstly, this approach assumes that the list of keywords adequately summarises the areas of interest. Secondly, it assumes that both the list of subjects and publication details for any given book is complete and accurate. Thirdly, as we extract publication years ourselves, there will inevitably be edge cases where the four-digit number starting with 18-, 19- or 20- contained within publication details does not in fact refer to a year but to another detail. Although a manual review of publication details suggest that this is not common, it is certainly plausible. More broadly, before generalising from insights based on this catalogue, it is important to remember that libraries will have different focuses and these may have shifted overtime, which could influence the results presented below.

Renewables are older than you think

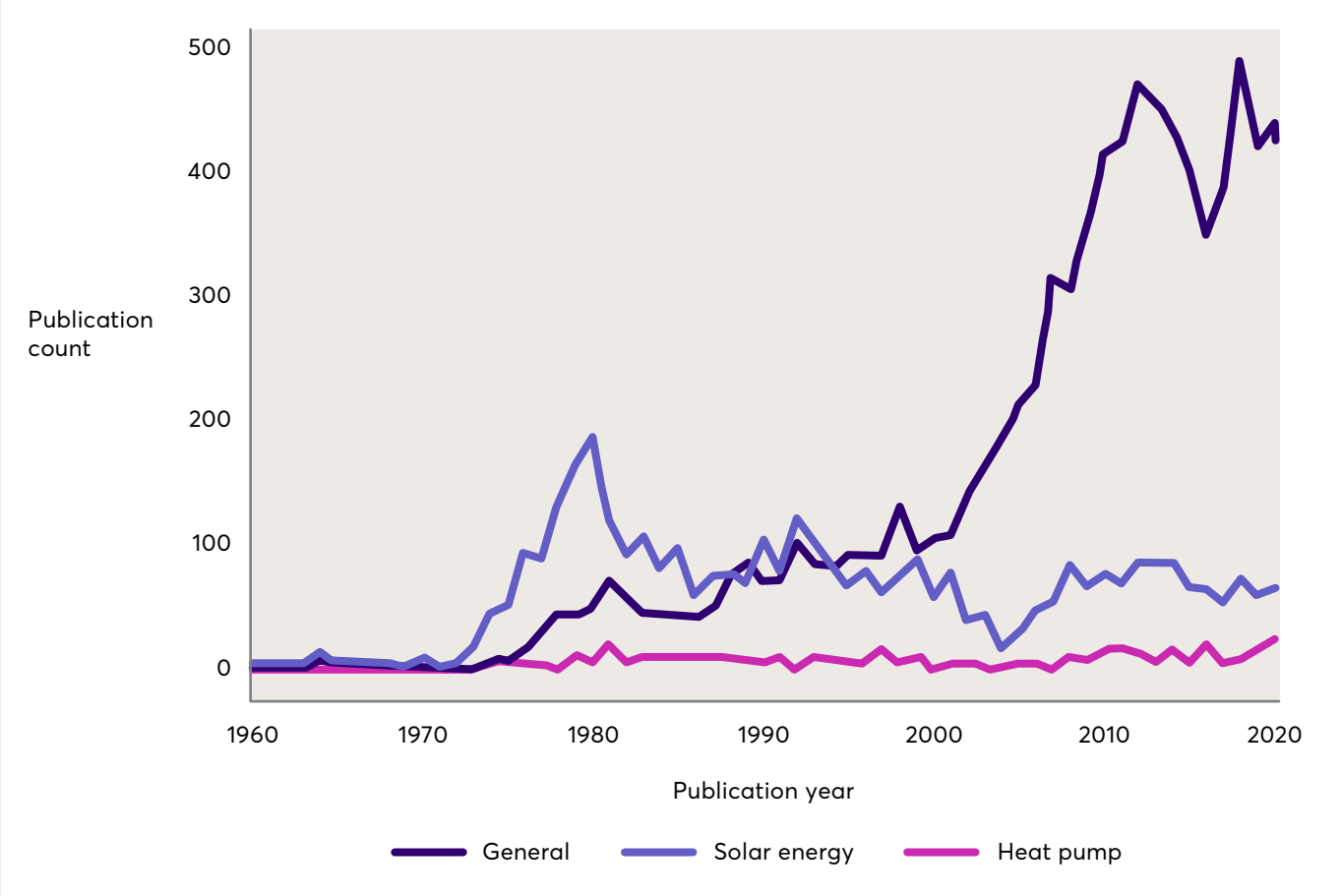
Over 10,000 records are returned from the query terms listed above. While there are over six times more records pertaining to solar energy than heat pumps and home retrofits, many of the topics across all areas have appeared in records that were published as early as the mid 19th century. For example, one of the earliest records associated with heat pumps was published in 1845 by James Booth, on new methods applied to the steam engine. Meanwhile, a record related to solar energy was found in a NASA technical memorandum, published as early as 1966.

Figure 2: Example of older publications contained in the library catalogue



The rise in renewables

Figure 3: The number of publications over time across the three keyword areas related to renewables.



When we investigate the number of publications over time across the three keyword areas, we observe that records associated with 'renewable energy' ballooned after 2000. While the type of records vary, there does appear to have been a specific proliferation of bills, policy papers and academic proceedings related to the climate transition in recent years. As governments and large organisations across the world discuss interventions to tackle climate change, records across libraries reflect these growing measures. Examples of these records include a UK Summary Report on IEA Heat Pump Technology Collaboration Programme (TCP) published in 2019 and Proceedings of the International Conference on Environmental Science and Sustainable Energy published in 2017.

Did solar peak in 1980?

Curiously, the publication year associated with the greatest number of records on solar energy is 1980, after a steady rise over the 1970s. These items largely focus on the development of solar technology: a number of reports discuss both its potential and mainstream introduction, such as Solar energy: its potential contribution within the United Kingdom. Indeed, by the 1980s, solar energy in countries like the United States was readily available to citizens, and a number of federal level acts like the Solar Photovoltaic Energy Research, Development, and Demonstration Act of 1978 incentivised the roll out of solar power in homes. Again, the proliferation of library holdings reflects growing awareness around the potential of this technology.

A methodology to track renewable energy subject areas over time

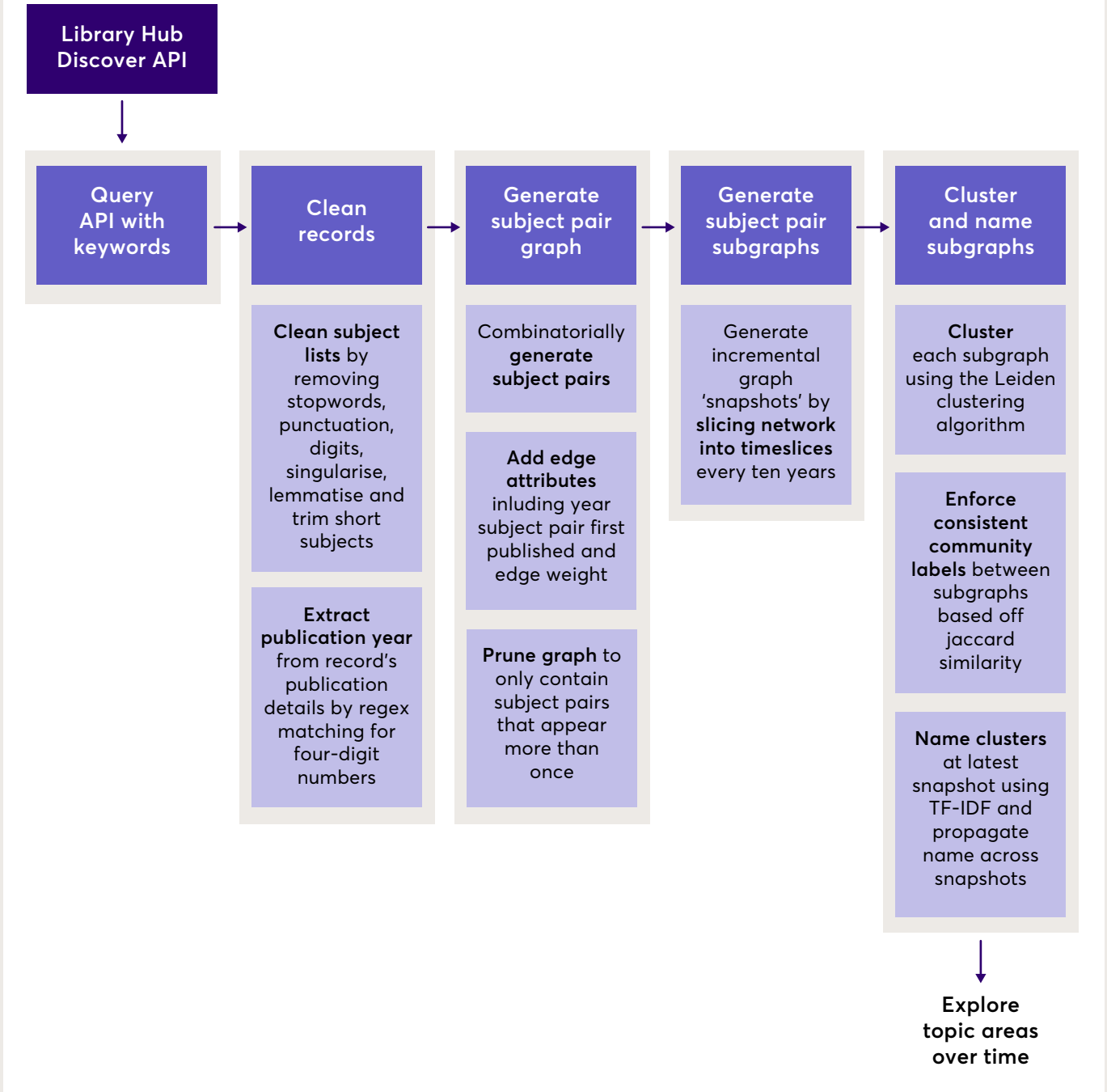
We develop a methodology to observe and track topic areas within renewable energy. This methodology allows us to spot both the emergence and 'death' of topic areas over time, from as early as 1825 to 2022.

At a high level, we:

1. **Query the API** to return records across solar energy, heat pumps and general decarbonisation.
2. Conduct **data preprocessing** to remove punctuation, digits and other 'bad' characters from subject lists. We also extract publication years from publication details by identifying four digit words starting with 18, 19 or 20.
3. **Create a network** (or graph) of subjects, where the links between each subject reflect how often two subjects co-occur with each other across records.
4. **Create incrementally larger subgraphs** based on when subjects first co-occurred on a ten year basis. The first graph is therefore the smallest and includes subjects in the first ten years of publication while the final graph is the largest and includes all subjects over time.
5. **Assign names to subject groups (or clusters)** based on subject term frequency.
6. **Propagate subject groups over time** based on how many subjects in the groups are the same.

After these six steps, we generate seven subgraphs (or smaller networks), where the largest subgraph contains all the subjects mentioned within records published between 1825 and 2022. The resulting graph at the latest time frame contained 7,835 subjects.

Figure 4: Temporal network methodology



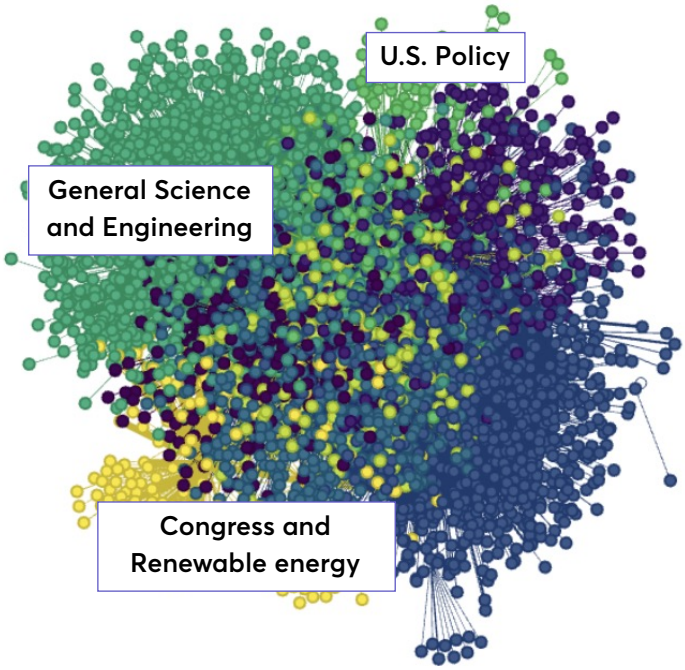
To read about the methodology in further detail, refer to the appendix.

Once the data is represented graphically, we are able to explore subject group dynamics over time, answering interesting questions such as: what can the changes in clusters over time tell us about the renewable energy space? What topic areas have remained popular in renewable energy literature? What new topics have emerged?

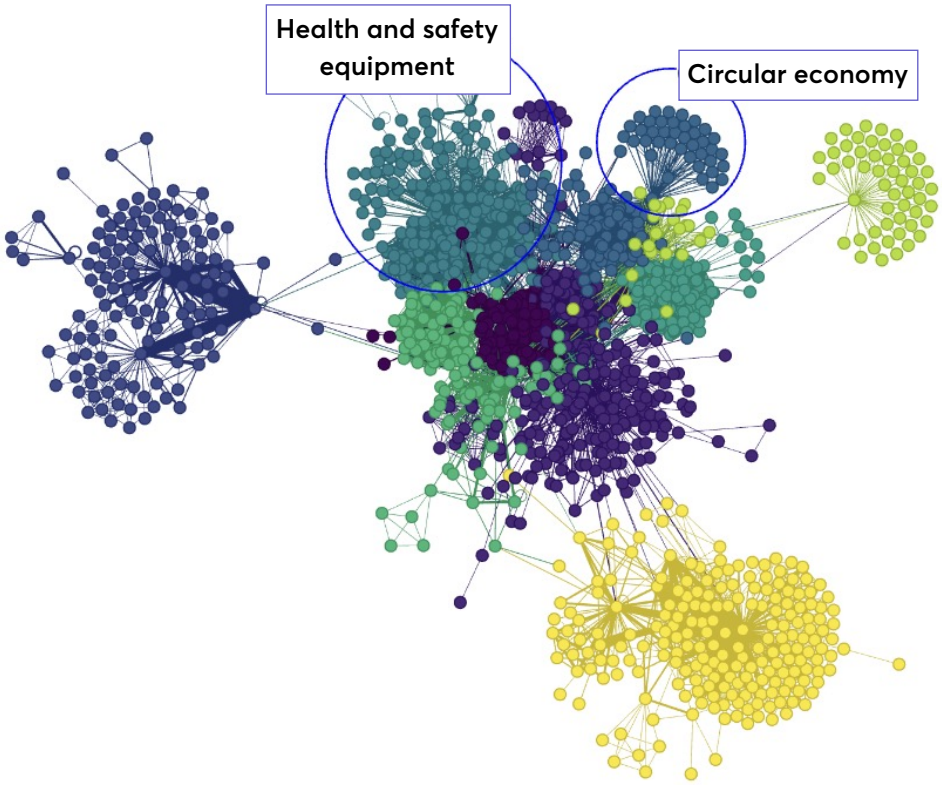
The literature on renewables has broadened and deepened

The Library Hub Discover records reveal that there has been both a broadening and deepening of the renewable energy space over time. At the earliest timestamp (1965 and before), there are **11** clusters of subjects, and these poorly clustered subjects primarily relate to energy, policy and science. By the latest timestamp (which includes all items published up until 2022), the number and specificity of clusters has proliferated to more than 188. Moreover, the mix of topics now spans beyond science and includes the likes of the circular economy and regional policies.

Figure 5: Networks of subjects from records published between 1825 and 1965 (upper network) and between 1825 and 2022 (lower network)



Full graph at the earliest timestamp (up to 1965), highlighting three clusters



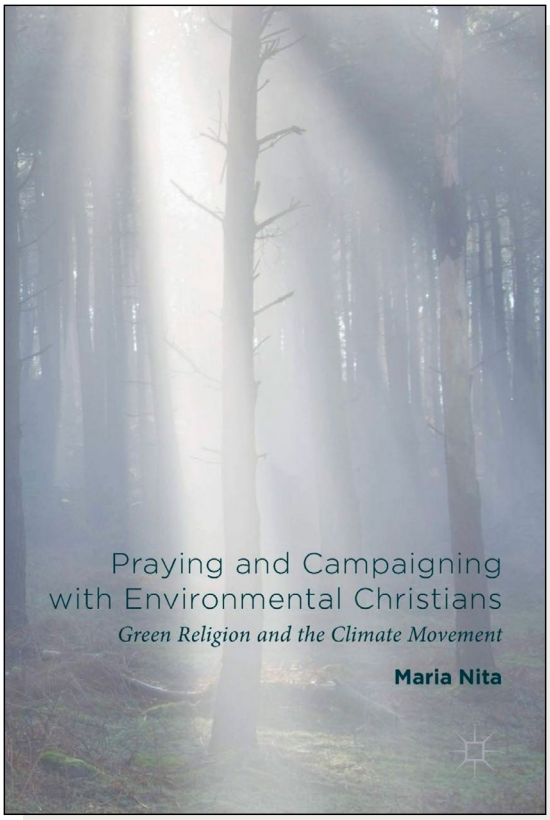
Subgraph of ten large clusters at the latest timestamp (up to 2022), highlighting two clusters

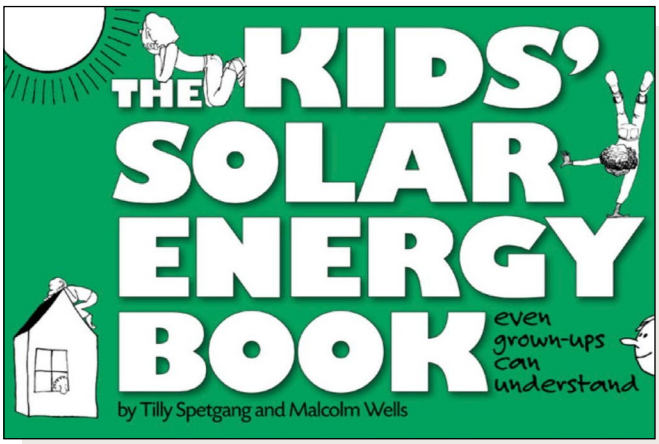
Note: Each point represents a subject that was mentioned in an item related to the renewable energy space.

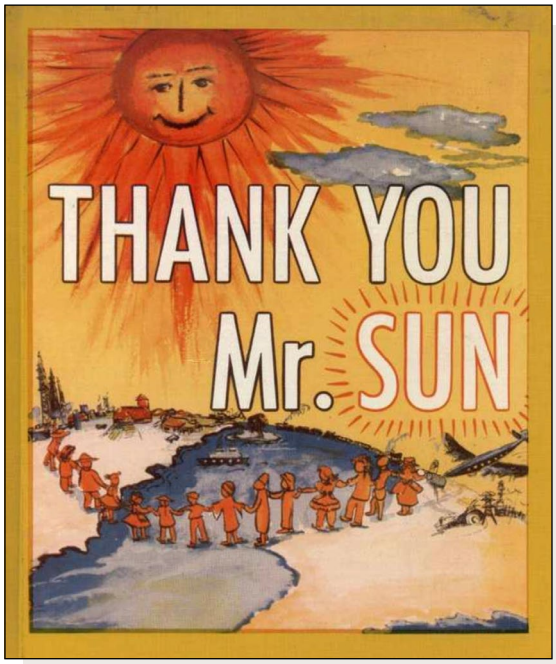
Unexpected topics

Not only do clusters beyond science emerge, clusters and records that are perhaps less expected also emerge. For example, records related to the interaction between religion and the climate crisis appear, including records such as *Praying and Campaigning with Environmental Christians*. Another surprise is the emergence of records from 'juvenile literature'. These include children's books related to renewable energy like *The children's solar energy book: even grown-ups can understand* by Tilly Spetgang and Malcolm Wells and *Thank you, Mr. Sun!*.

Figure 6: Examples of unexpected records in the network







Age old topics

The library data can also be used to identify the core subject areas that define the literature on renewables. Although we do see a proliferation of more specific, smaller clusters over time, there are **three** clusters that both persisted and grew sizeably by the final time period. These include subjects related to U.S policy, renewable energy and the economy and more general science and engineering. Given the interactions between policy, the economy and science within the field of renewable energy, there is some overlap in subject topics across the clusters.

The table below showcases the growth in these clusters and provides examples of subjects within them.

Figure 7: Persistent clusters summary statistics

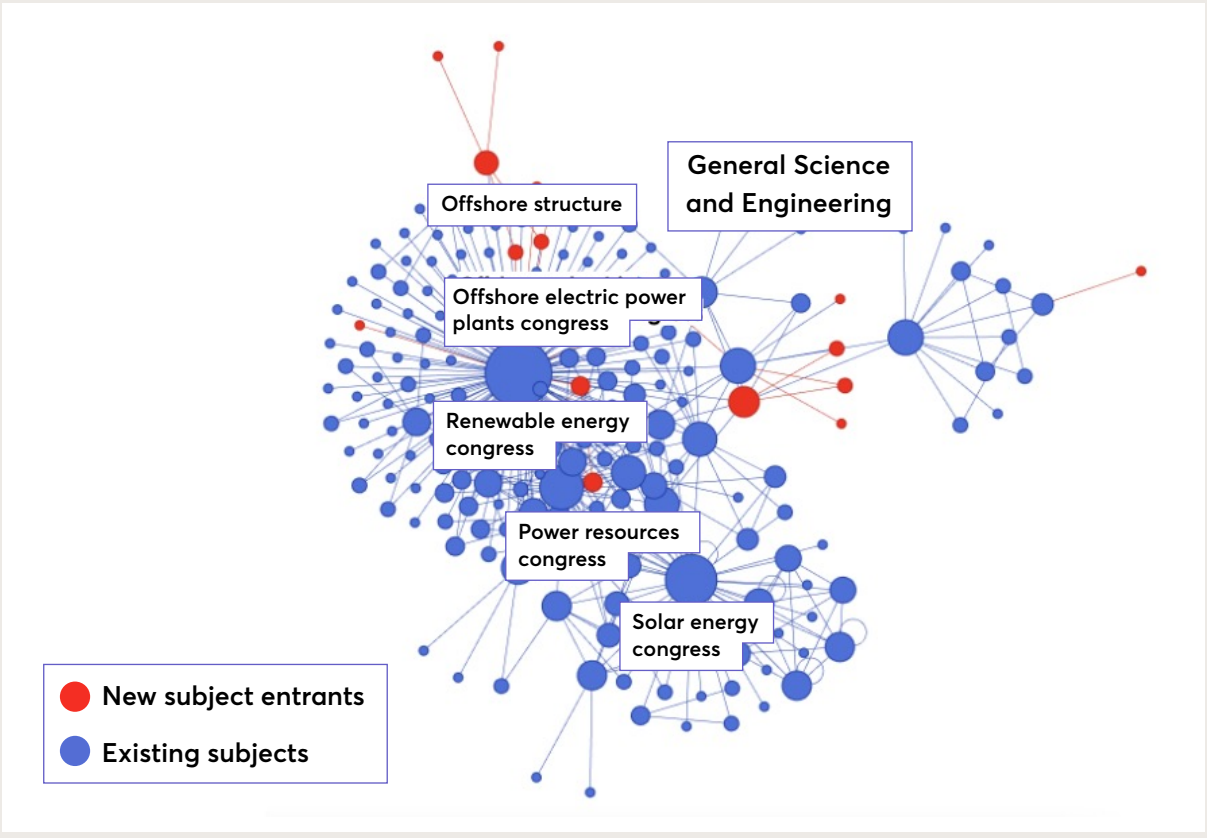
Cluster name	Subject examples	Item Examples	Number of subjects in cluster up to 1965	Number of subjects in cluster up to 2022
Renewable energy and the economy	Economy, business and finance economy renewable, renewable energy sources economic aspect, feasibility study	Feasibility analysis for sustainable technologies: an engineering-economic perspective, Careers in renewable energy, Growth isn't possible: why we need a new economic direction	408	526
Policy	Greenhouse gas mitigation international cooperation, carbon dioxide mitigation government policy united state, forest policy	Report of the United Nations Conference on New and Renewable Sources of Energy: (Nairobi, 10 to 21 August 1981), Low-Carbon Smart Cities: Tools for Climate Resilience Planning, Renewable energy from wind and solar power: law and regulation	913	2,554
Science and Engineering	Gradient descent, civil engineering, chemical engineering equipment and supply	Sustainable Energy Systems Planning, Integration and Management, Heat and Mass Transfer in Energy Systems	313	451

Birth, death, growth and decline of topics

While the topics above have remained and indeed grown over the last century, others have emerged and also died out. A cluster 'dies' when the subjects are merged with other existing clusters at a later time period. By capturing the growth and decline of topics, we can get a sense of what areas are particularly prominent in the renewable energy space and which have become obsolete.

For example, a cluster containing items on congress and renewable energy (congress-energy-power) is one example of a topic that grew. The cluster initially contained subjects that included congress and general renewable energy terms. For example, the most popular subject pair within this cluster (at its birth) was energy conservation congress and renewable energy sources congress. This pair first appeared in 1948 and occurred more than 65 times since. Over time, records emerged that related to congress and more specific types of renewable energy, like photovoltaic science, solar energy and heat pumps. In particular, new entrants include offshore structure, offshore electric power plants congress and transmutation operators congress.

Figure 8: congress-energy-power at timestamp 3 (up to 1985), where node size corresponds to subject degree

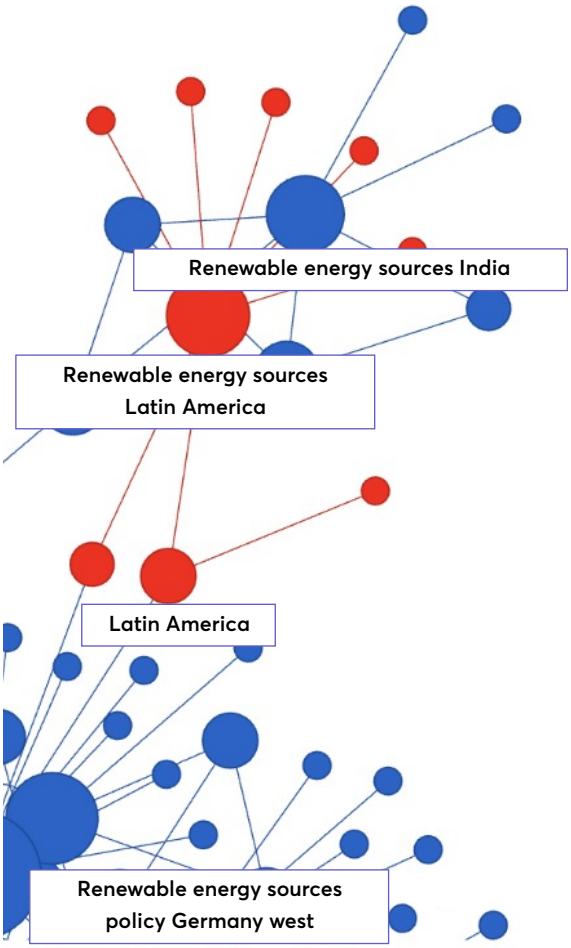


Note: In this case, degree refers to the number of times the subject was mentioned with other subjects.

From regional to global

Meanwhile, a second community of subjects that eventually died contained subjects related to renewable energy and Latin America (latin-america-energy). The small community of subjects which first appeared in 1981 included subjects like environmental economics latin america, energy policy latin america and renewable energy sources latin america. It eventually merged into a larger community related to energy policy across many different countries and geographical regions including India, Germany and the European Union.

Figure 9: great-britain-energy at timestamp 4 (up to 2005), where node size corresponds to subject degree. The red nodes represent the initial latin-america-energy cluster



Note: In this case, degree refers to the number of times the subject was mentioned with other subjects.

Ultimately, the relationships between renewable energy and policy, science, engineering and technical methodologies have stood the test of time. However, and more recently, topics within renewable energy have bled into one another, become more specific and even entered unexpected areas, like in ecotheology and juvenile literature.

A spotlight on heat pumps

As government discussion of decarbonising heat in homes ramps up in the UK and public awareness of heat pumps increases, we turn to library data to investigate the evolution of this technology. We do this by creating 'heat pump' ego-graphs where the heat pump subject is the focal point of the graph and the other subjects either directly co-occur with heat pumps in records or co-occur with each other. By doing so, we can investigate which subjects tend to be mentioned alongside 'heat pumps' and how these relationships have evolved over time.

Links between heat pumps and other renewable energies

The ego-graphs from the last timestamp (which captures all records published up to and including 2022) reveal a growing interaction between heat pumps and other renewable technologies, such as research on solar assisted heat pumps or the role of heat pumps in renewable energy systems. For example, for records published up to 1965, the heat pump subject is related to general and persistent communities like scientific methodologies, science and engineering and economics and policy. By the latest timestamp (records published up to 2022), the heat pump ego graph has become associated with new subject areas, including solar energy policies, regional policy clusters and health and safety. Examples of these records include *Advanced Applications for Smart Energy Systems Considering Grid-Interactive Demand Response, Extending Permitted Development Rights for Domestic Wind Turbines, Solar Energy in Buildings* and *Refrigerating Systems and Heat pumps: Safety and environmental requirements*.

Conclusions and future directions

Decades of library data reveal that there has been a dramatic increase in academic records on renewable technologies such as solar and heat pumps, especially in the last ten years. Topics related to renewable energy have been in the literature for a surprisingly long period of time, as early as the mid 19th century. While some subject areas have grown and shrunk, others have persisted, including scientific methodologies, energy policy, science and engineering and U.S. energy policy. Meanwhile, more recent networks reveal increased specialisation both beyond and across earlier topic areas. They also show the growing interaction between different forms of renewable energy. More broadly, this article serves to demonstrate that datasets from the creative industries can serve to both confirm macrotrends and highlight unexpected or hidden topics.

There are a number of interesting directions in which to take this form of analysis. First, as the Library Discovery hub API is open, it could be used to construct a real-time monitoring system for any given topic. This system could include tracking growth in the number of publications across keywords as well as the evolution of subject areas over time. Second, given the records also return author names, prominent scholars in keywords or topic areas could be identified. Finally, a similar analysis could be conducted across a variety of topic areas, including Nesta's other key mission areas: A Healthy Life and A Fairer Start.

See the open codebase here.

Appendix

Detailed methodology

The API was queried with seven key terms mentioned in Figure 1.

After querying the API with key terms, the records were preprocessed to:

1. **Clean record subject lists** to remove stopwords, punctuation and digits and to singularise, lemmatise and trim very short subjects (under three characters);
2. **Extract publication year** from publication details by identifying four digit strings starting with 18, 19 or 20.

The preprocessed records were then used to generate a subject pair co-occurrence graph where the nodes represent subjects and the edges between subject pairs represent their co-occurrence. Edge attributes including the year the subject pair were first published and edge weight were also included. Finally, subject pairs that were only mentioned together in one book were removed from the graph. To capture the temporal element of the graph, incremental subgraphs were generated, by 'slicing' the network into time frames every ten years from 1965 (and before) onwards. This resulted in seven subgraphs, where the largest subgraph formed was the full graph at the latest time.

To capture subject clusters over time, the Leiden clustering algorithm was used to cluster each subgraph at every timeframe. To account for nodes changing communities over time, cluster membership was determined for incremental subsets of graphs based on time.

To enforce consistent cluster labels between subgraph timeframes, the labels for each subgraph were set by greedily assigning each cluster from a timeframe (t) a label corresponding with the cluster label from the timeframe before (t-1) that maximises the jaccard similarity (a measure that calculates the similarity between two sets) between the set of nodes belonging to those two clusters. This allows for the birth, death, growth and decline of clusters over time. Finally, human readable cluster names were generated using TF-IDF (a common information retrieval technique that weighs the frequency of a word or term against the inverse document frequency) of each cluster at its latest time frame. The cluster name was then propagated across all time frames. The resulting graph at the latest time frame contained 7,835 nodes.

The Creative Industries Policy and Evidence Centre (Creative 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 a consortium of universities and one joint enterprise from across the UK They are: Birmingham, Cardiff, Edinburgh, Glasgow, Work Advance, London School of Economics, Manchester, Newcastle, Sussex, and Ulster. The PEC works with a diverse range of industry partners including the Creative UK.

To find out more visit **www.pec.ac.uk** and **@CreativePEC**

The Creative Industries Policy and Evidence Centre (Creative PEC) is part of the Creative Industries Clusters Programme, which is funded by the Industrial Strategy Challenge Fund and delivered by the Arts and Humanities Research Council on behalf of UK Research and Innovation. The PEC has been awarded funding by the AHRC for an additional five years, and will have a new host organisation in 2023.

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

ISBN: 978-1-913095-56-7