Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Short communication

300 years of British patents[☆]

Enrico Berkes [a], Matthew Lee Chen [b], Matteo Tranchero [c,*]

[a] University of Maryland, Baltimore County, United States of America
[b] Harvard University, United States of America
[c] University of Pennsylvania, United States of America

ARTICLE INFO

JEL classification:
O31
O34
N73
N43

Keywords:
Patent data
Innovation
Industrial revolution
British patent system
Patent reform

ABSTRACT

The study of innovation depends heavily on high-quality patent data. Yet, datasets containing complete patent documents focus only on recent decades, while historical patent datasets with broader temporal coverage typically lack detailed information. Therefore, our ability to leverage advances in textual analyses to study long-run innovation dynamics remains limited. To this end, we introduce a large-scale dataset of the universe of technical specifications of British patents granted between 1617–1899. Our data consists of the full specification texts alongside linked information about inventors, including their disambiguated names, occupations, and addresses. We use our data to document changes over time in total inventive activity, the geography of innovation, inventor occupations, and patent novelty and impact. Finally, we discuss use cases and avenues for subsequent research.

Resources: Dataset; GitHub

1. Introduction

Patent datasets are ubiquitous in the study of innovation (Hall et al., 2001; Jaffe, 2002). Yet, they typically focus only on recent decades, making it challenging to explore innovation dynamics and technological change over extended periods. A number of historical patent datasets have been introduced to facilitate longer-run analyses (Marco et al., 2015; Petralia et al., 2016; Kogan et al., 2017; Berkes, 2018; Sarada et al., 2019; Bergeaud and Verluise, 2024). However, these datasets provide minimal information about the inventions themselves, since they do not contain the original texts of patent specifications where technical information was reported. Recent work has increasingly applied natural language processing methods to patent texts in order to measure the direction and content of new technologies (e.g. Feng, 2020; Arts et al., 2021; Kelly et al., 2021; Kalyani, 2024; Kalyani et al., 2025). Yet, the lack of detailed technical information has limited the scope of research seeking to use these approaches in longer-run historical settings.

We create and introduce a large-scale dataset of British patent specifications covering the period 1617–1899. Our data consist of the complete technical specifications from 322,874 unique patents, representing the universe of patents published by the British Patent Office during this period. Previous research using British patent records has relied on administrative indexes created by patent clerks (Nuvolari and Tartari, 2011; Juhász et al., 2024; Rosenberger et al., 2024; Hanlon, 2025), which provided only brief abstracts and often ended by the mid-19th century. Instead, our dataset includes high-quality OCR texts of the full technical specifications written by the inventors themselves. These contain the technical information deemed necessary for specialists to reproduce each invention, which is invaluable for scholars wishing to study innovation dynamics. Our dataset combines these digitized texts with detailed inventor information automatically extracted from the front pages using a deep learning-based pipeline. We link inventors across patents and provide their full names, stated occupations, geocoded addresses, firm affiliations, application and/or sealing dates, and information on patents communicated from abroad. A distinctive feature of our data is that British patents, unlike U.S. patents, generally include inventors' occupational information which we extract with high accuracy. Moreover, we assemble information on patents communicated from abroad, providing a novel lens to study international knowledge flows.

[☆] We thank Mike Andrews and Melissa Dell for valuable comments. We are very grateful to Anthony Ellis and Siwei Ran for outstanding research assistance. We acknowledge financial support from Google Cloud for Researchers. All errors are our own.

[*] Corresponding author.
E-mail addresses: eberkes@umbc.edu (E. Berkes), matthewchen@g.harvard.edu (M.L. Chen), mtranc@wharton.upenn.edu (M. Tranchero).

https://doi.org/10.1016/j.respol.2025.105347
Received 4 February 2025; Received in revised form 22 July 2025; Accepted 8 October 2025
Available online 24 October 2025
0048-7333/© 2025 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Our dataset will facilitate long-run innovation research in several ways. First, our data contain rich information on more than 200,000 unique inventors, including details on their occupations. This will enable researchers to examine connections between human capital and frontier innovation. Second, alongside the dataset itself, we publicly release all the custom models that we fine-tuned as part of our data processing pipeline, making the dataset extensible to British patents granted after 1900. Espacenet currently provides a largely complete collection of the original scans of patent specifications published from 1900, although it has major coverage issues for prior patents.[1] Our fine-tuned models can be applied to these later texts to extract and process entities with high accuracy. Finally, our dataset is well-suited for researchers applying natural language processing to study innovation, as we release both the complete texts and tokenized words. To the best of our knowledge, this is the first historical patent dataset that readily facilitates textual analyses.

Using our data, we document several patterns in British innovation during the coverage period. We show that total patenting activity increased substantially over time, with notable jumps following major reforms of the patent system in 1852 and 1883. We also show that co-inventing became increasingly common throughout the 19th century, highlighting the long-term growth of team-based invention. Our data also document the growing importance of cross-national knowledge flows, as proxied by the patenting of inventions foreign-communicated. Examining the changing geography of innovation, we trace the emergence of northern England as an innovation powerhouse during the 18th century, followed by renewed inventive activity in southern England and London in the latter 19th century, alongside increasing inventor geographic mobility over time. We then analyze the evolving occupational composition of inventors, documenting a shift toward technical specialists such as engineers and manufacturers, and use TF–IDF analysis to demonstrate meaningful technological specialization reflected in patent language across occupational groups.

Finally, to underscore how our data can be used to construct textual measures of patent importance, we calculate patent breakthrough scores following Kelly et al. (2021). The data show a modest increase in breakthroughs during the First Industrial Revolution and more pronounced growth during the Second Industrial Revolution. We also benchmark these scores by comparing them with bibliometric patent quality measures (Nuvolari et al., 2021), showing that specialized and engineering occupations are predictive of high-breakthrough score patents, confirming earlier findings that patent reforms which lowered the cost of patenting indeed reduced the quality threshold (Nicholas, 2011), and documenting that a list of the highest breakthrough score patents in each year reveal a number of eminent historical inventors.

The remainder of the paper proceeds as follows. Section 2 provides historical background on British patent specifications. Section 3 outlines our data processing pipeline, presents a number of descriptive patterns, and provides a comparison to existing data. We conclude in Section 4 with a discussion of potential research applications and dataset limitations.

2. Background

The British patent system offers unique insight into the processes of long-run innovation and technological progress, being the longest continuously operating patent system in the world (Taylor, 1969; MacLeod, 1988). Its evolution from a series of monopoly privileges granted by the monarch to a formalized intellectual property institution coincided with Britain's industrial and economic transformation.

2.1. The British patent system, 1617–1900

The institutional framework of the patent system was established with the 1623 Statute of Monopolies — a piece of landmark legislation restraining the monarch's ability to grant arbitrary monopolies while establishing a system of temporary patent rights. The statute allowed temporary monopolies of 14 years to be granted, sufficient time to teach two generations of apprentices the new art (Van Dulken, 1999, p. 2).[2] The system imposed significant costs to prospective patentees, with estimates suggesting that around £100 — equivalent to the annual wages of a skilled workman — was necessary to obtain an English patent (Van Dulken, 1999, p. 24).[3] For this fee, protection was granted for the full 14-year period but extended only to England and Wales, and the disputed border town of Berwick-upon-Tweed, with Scotland and Ireland maintaining separate patent systems.

The British patent system continued without major reform for more than two centuries. In 1852, the Patent Law Amendment Act was passed in Parliament, introducing two crucial changes. First, initial filing costs were reduced by roughly three-quarters to £25 while renewal fees were introduced at years 3 and 7, imposing new costs for those patentees seeking to maintain the patent for a full term of 14 years. Second, the reform established the British Patent Office, unifying the previously distinct patent systems of England and Wales, Ireland, and Scotland. The reform significantly increased patent grants and the number of new inventors.

In 1883, an additional reform, the Patents, Designs, and Trade Marks Acts, introduced further measures to modernize the patent system. It established the Comptroller General's office that instituted patent examinations to probe the accuracy of technical information provided by the patentee. However, this did not constitute a formal examination of novelty in the modern sense.[4] The reform also reduced initial filing fees to only £4 (Van Dulken, 1999, p. 5), greatly increasing patent accessibility and further raising patent volumes (Nicholas, 2011).

2.2. Technical specifications

Patent specifications served as detailed technical documents that were required to provide sufficient information for a specialist to reproduce the invention (see the ruling of Boulton & Watt v. Bull, 1795; Robinson, 1972). Beyond their legal function, these specifications became crucial vehicles for technical knowledge transfer. Available for public consultation at the Court of Chancery in London for a fee (Cox, 2020), they attracted regular attention from inventors and mechanics. Their reach extended further through technical publications such as *The Repertory of Arts* or *Mechanics Magazine*, the latter of which reached a weekly circulation of 16,000 copies (Bottomley, 2014b). By codifying technical knowledge that might otherwise have remained secret, specifications reduced informational asymmetries in technological development and helped prevent duplicative invention efforts. In an era of limited technical documentation, they emerged as a unique and current source of industrial knowledge.

The first patent with a written specification appeared in 1617, initiating a period where specification submission was voluntary. A substantive legal change came with the mandatory disclosure law of 1734, which required all patentees to submit a specification for their patent to come into force.[5] Following the establishment of the

[1] Espacenet patents can be downloaded at: https://worldwide.espacenet.com/patent/. British patent numbers use the prefix "GB".

[2] In the British patent system, there has never been any distinction between utility patents and other types of patents (e.g., design), as exists in the US system (Jaffe, 2002).

[3] By comparison, patent fees in the United States were substantially lower at $30 shortly after its patent system was introduced in 1793 and maintained at this level until 1861 (Khan, 2005).

[4] It was not until 1907 that British patents were able to be refused on the basis of novelty (Van Dulken, 1999, p. 28).

British Patent Office in 1852, inventors commonly filed provisional specifications to secure temporary protection while preparing their complete specifications (Van Dulken, 1999, p. 31). These provisional and complete specifications were typically published together in the final patent document. However, after 1884, provisional specifications went unpublished if no complete specification followed, and our dataset includes a small number of these provisional-only cases. The publication format affects specification length in predictable ways: patents containing both provisional and complete specifications are longer due to partial content duplication, while those with only provisional specification — which never gained full legal force — are typically shorter.

Appendix Section A provides an example of a technical specification. We can see that this particular patent was filed by "Henry Bessemer" and "Alfred George Bessemer", with respective addresses "Denmark Hill, in the County of Surrey" and "Oakfield, South Dulwich, in the same County". The first inventor "Henry Bessemer" has his occupation listed as a "Civil Engineer". As is typical for most patents, we see a sealing date and an application date, with the application date — 5th April 1879 — being several months before the sealing on 1st July 1879. The patent holds the title "Improvements in the Manufacture of Tin Plate and Black Plate, and in the Machinery, Apparatus, and Processes Employed in such Manufacture".

Historical British patent specifications differ from modern patents in three key institutional aspects. First, they did not include citations to prior inventions.[6] Second, rigorous novelty examinations were not introduced until the early 20th century (Van Dulken, 1999, p. 27), meaning our data lacks examination records common in modern patent documents. Third, these historical specifications had no standardized technology classification system (Van Dulken, 1999, p. 141). While the Patent Office did publish abridgments between 1855–1930 that categorized patents into 146 classes, these classifications are not publicly available and challenging to access. For a subset of years (1855–1883), these were hand-transcribed by Hanlon (2016) and can be linked to our data using the unique patent identifier.[7]

3. 300 years of British Patents

3.1. Dataset overview

The dataset we introduce in this paper is **300 Years of British Patents**, a large-scale dataset of the universe of historical patent specifications published by the British Patent Office between 1617 and 1899. The dataset contains titles, full texts of specifications, lists of tokenized words, and extracted named entities for 322,874 patents.

Our dataset is made available on HuggingFace in JSONL format to facilitate ease of access to our textual data. Entities are associated with the full texts using unique patent identifiers. Our dataset is most easily accessed by downloading via the HuggingFace `datasets` library. We provide code to convert entity data into dataframes and CSVs for easier access for researchers who do not intend to use our full text data directly. Complete documentation and a starter notebook are available on our HuggingFace page.

The dataset is accompanied by our custom fine-tuned models for text region detection, entity extraction, patent title extraction, and inventor linking, enabling researchers to extend our data to additional patent collections such as post-1900 British patents on Espacenet.

3.2. Data processing pipeline

Our data processing pipeline transforms raw patent specification scans into structured data through a number of steps, each optimized for historical patent documents.

First, we process the original specification scans from the British Intellectual Property Office using a custom-trained `YOLOv8` model (Varghese and Sambath, 2024) to identify text regions, filtering out non-text elements like diagrams, page numbers, and margin notations.[8] This text detection stage is crucial for historical documents where formatting is inconsistent and OCR quality can be compromised by extraneous elements.

Second, we extract and clean text from the identified regions using Google Cloud Vision OCR, concatenating text blocks based on their spatial arrangement to preserve document structure. We also provide our OCR patent texts as a list of tokenized words: we dissolve HTML and non-ASCII patterns, replace hyphens with spaces, and then apply a standard `spaCy` tokenizer (using the model `en_core_web_lg`).

Third, our pipeline extracts structured information using language models that we fine-tuned using the HuggingFace `Transformers` toolkit. We fine-tune a `XLM-RoBERTa` model (Conneau et al., 2019) on a custom dataset to identify relevant named entities (inventor names, occupations, addresses, and firms, as well as dates and foreign communicant information) and also separately patent titles.[9] We then use the OpenAI API (with the `gpt-4o` model) to link occupations, addresses, and firm affiliations to inventors, and also to geocode non-standardized historical addresses. Finally, a fine-tuned `SentenceTransformers` (Reimers and Gurevych, 2019) model — implemented with the `LinkTransformer` (Arora and Dell, 2024) framework — is used to link inventors across patent documents.[10]

Our pipeline includes validation checks at each stage, focusing on out-of-sample performance on custom test datasets that we create. The complete pipeline, available on our HuggingFace repository, is designed for extensibility and reproducibility. Researchers are able to reproduce our extraction and linking steps, and can also apply our methods to post-1900 British patents. We provide detailed documentation and model weights for each component.

Fig. 1 visually presents our data processing pipeline. The final dataset provides the full OCR texts of specifications, the lists of tokenized words, and the processed and extracted entities and patent titles. In Appendix Section A.1., we provide a more detailed breakdown of our data processing steps.

3.3. Descriptive patterns

Here, we illustrate the primary descriptive patterns to introduce and validate our dataset. First, in Fig. 2, we show that the total number of patents increased significantly throughout our coverage period,[11] with notable jumps in patenting levels following the 1852 and 1883

[5] Our dataset includes all specifications that were submitted from 1617–1734, acknowledging that voluntary submission meant not all patents during this period included specifications.

[6] While formal citations were absent from specifications themselves, patent citations did appear in legal and engineering literature. Reference indexes documenting these citations were published for patents up to 1852 (Nuvolari and Tartari, 2011; Woodcroft, 1854).

[7] The patent class data from Hanlon (2016) can be downloaded from https://walkerhanlon.com/data_resources/british_patent_classification_database.zip.

[8] Our fine-tuned model and train-val-test data splits can be found at https://huggingface.co/matthewleechen/patent_text_regions_yolov8.

[9] We make our fine-tuned entity recognition model public at https://huggingface.co/matthewleechen/patent_entities_ner, and our patent title extraction model is available at https://huggingface.co/matthewleechen/patent_titles_ner. We provide train-val-test data splits for both.

[10] Our fine-tuned model for linking inventors across patents is available at: https://huggingface.co/matthewleechen/lt-patent-inventor-linking.

[11] In the mid-17th century, there was a break in patenting activity during the Long Parliament (1640–1660) when the patent system was effectively dismantled (MacLeod, 1988). Despite a period of relatively high usage following 1660, the system became scarcely used between 1700–1720.

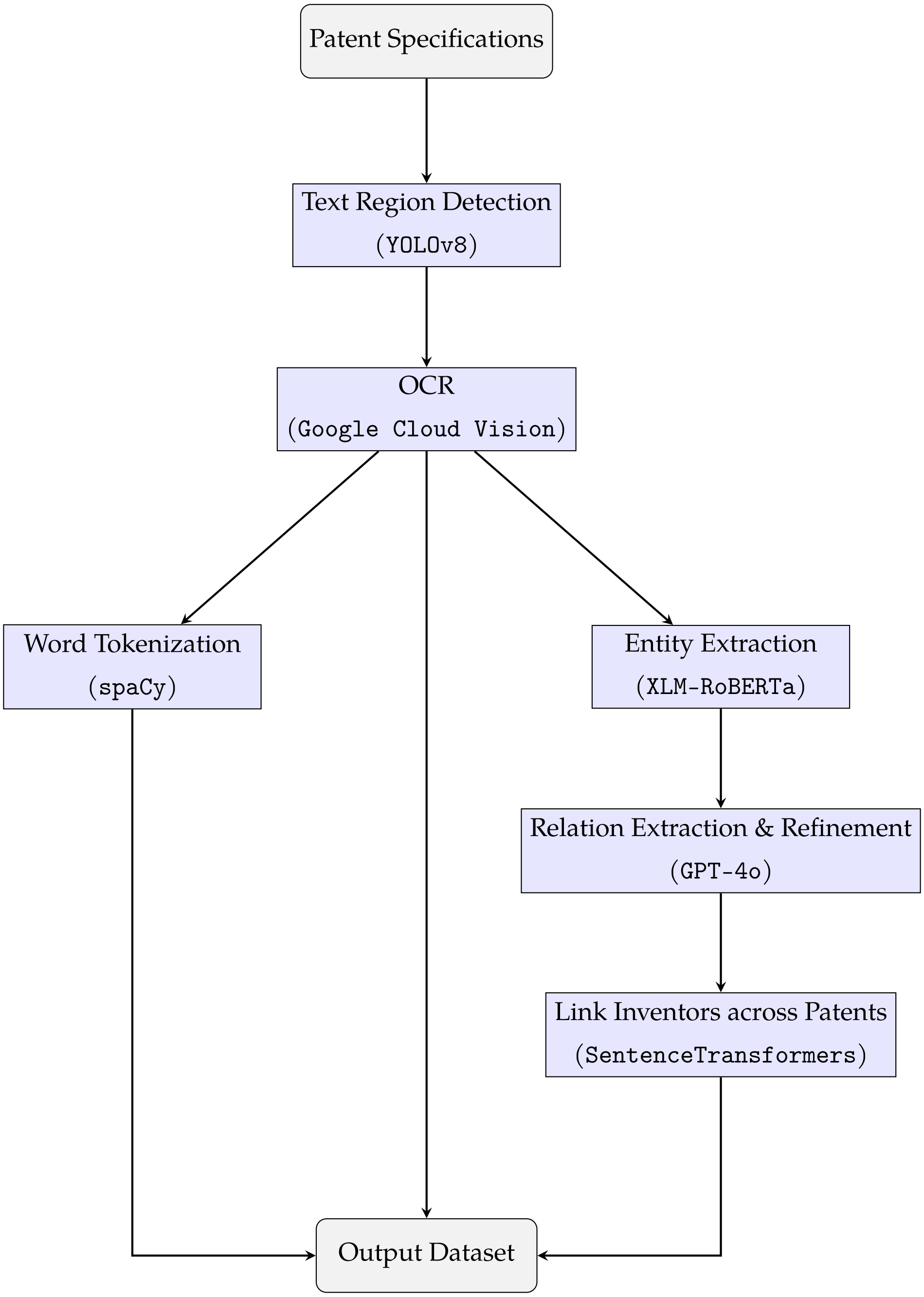
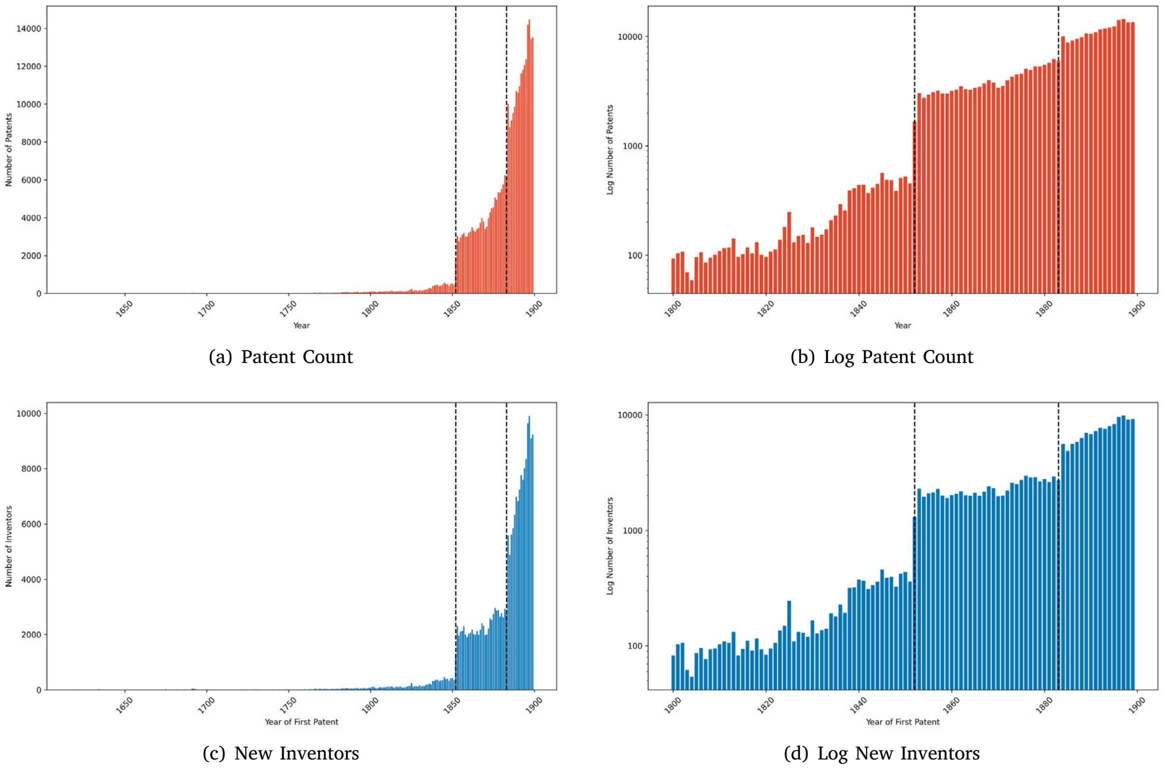


Fig. 1. Overview of data processing pipeline. *Note*: For text region detection, we use a fine-tuned version of YOLOv8s, available here. For entity extraction, we use fine-tuned versions of `XLM-RoBERTa-large` for inventor information extraction (here) and patent title extraction (here). For linking inventors across patents, we use a fine-tuned SentenceTransformer model, available here. The other models we use are off-the-shelf: Google Cloud Vision accessed through the API, spaCy's `en_core_web_lg`, and OpenAI's `gpt-4o` accessed via the API. The output dataset is available on our HuggingFace page here.

reforms.[12] Next, linking inventors across patents reveals that while most inventors (77%) patented only once in their lifetime, we are able to identify 47,224 inventors who patented multiple inventions (Appendix Table B.1.). Co-inventing patented technologies also became increasingly common over the 19th century (Figure B.1.). While early years are noisy due to low patent counts, there is a clear upward trend in collaborative inventions throughout the 19th century. During this period, we also see an increase in communications of inventions from abroad (Figure B.2.). While similarly noisy in early years due to low patent volumes, the share of foreign-communicated patents grew substantially after the 1852 reform, reaching approximately one-third of all patents between 1852–1883. After 1883, while the overall rise in domestic patenting reduced this share, patents originating from abroad remained a significant component of British innovation.

[12] However, while these reforms shifted the level of patenting upwards, they did not increase the underlying growth rate of new patents. The logged number of new inventors who first patented each year shows a similar pattern: reforms increased participation levels but not growth rates. These results are consistent with the findings of Nicholas (2011), who finds an uptake in patenting levels after the 1883 reform.



(a) Patent Count

(b) Log Patent Count

(c) New Inventors

(d) Log New Inventors

Fig. 2. Patents and new inventors over time. *Note*: Plot (a) shows the number of patents granted by year between 1617–1899, and plot (b) the log number between 1800–1899. Plot (c) displays the number of inventors who first patented in a given year between 1617–1899 and plot (d) the log number between 1800–1899. The patent reform years (1852, 1883) are overlaid as dashed lines.

Using our geocoded inventor addresses, we then examine the changing geography of British innovation (Fig. 3).[13] The pre-industrial period (1617–1750) shows patenting concentrated in London and key market towns like Bristol. Between 1751–1800, we observe the rise of the industrial north, with new patent clusters forming around Birmingham and Lancashire, as well as in the North East and Scotland. These northern clusters expanded significantly by the mid-19th century relative to London and southern England. However, the latter half of the 19th century saw substantial growth in patenting around London and across southern England. Inventor mobility also increased over time, as measured by the average number of different addresses per inventor throughout their patenting careers (Figure B.3.). We observe steadily increasing mobility through the 18th and 19th centuries, with a decline toward the end due to right-censoring of inventors who started patenting late in our period. The maximum distance between locations also increases over the first half of the 19th century, with later cohorts migrating further over their lifetimes.

We next turn our attention to the changing occupational composition of inventors (Table 1). Early periods (1617–1750) were dominated by less specialized occupations like ‘‘gentleman’’, ‘‘esquire’’, and ‘‘merchant’’, which declined over the 19th century in favor of more specialized occupations such as ‘‘manufacturer’’, ‘‘engineer’’, and ‘‘machinist’’. The occupational categories themselves evolved significantly from high-skill artisanal occupations (e.g., ‘‘watchmaker’’, ‘‘weaver’’, ‘‘hosier’’, ‘‘carpenter’’) to mass production roles (‘‘manufacturer’’, ‘‘manager’’, machinist’’). TF–IDF analysis reveals meaningful occupational specialization patterns detectable in the patent text (Table B.2.).[14] Mechanical engineers’ patents distinctively feature terms like ‘‘valve’’, ‘‘wheel’’, and ‘‘chamber’’; chemists emphasize ‘‘acid’’, ‘‘solution’’, and ‘‘gas’’; and electrical engineers focus on ‘‘current’’, ‘‘circuit’’, and ‘‘electric’’. These patterns extend previous work by Billington (2021) and Bottomley (2014a) and demonstrate how textual information enhances our understanding of the changing occupational structure of inventors.

To underscore how our data lends itself to the application of various textual measures of patent importance, we calculate patent breakthrough scores following Kelly et al. (2021).[15] We implement a backward-IDF measure that only considers prior patents when computing term importance, which helps address the temporal bias found in standard TF–IDF methods. For each patent, we compute a breakthrough score as the ratio between two measures: its cosine similarity to future patents (measured across different forward horizons of 1, 5, 10, and 20 years) and its similarity to past patents (using a 5-year backward horizon).[16][17] This ratio captures both how lexically different a patent

[13] If a patent is associated with multiple locations, one of them is chosen at random.

[14] We pool all the texts of patents corresponding to inventors with these occupations and use TF–IDF to rank the most distinctive words for each occupation, dropping a set of custom patent stopwords.

[15] We release our patent breakthrough scores publicly on HuggingFace to facilitate easier usage: https://huggingface.co/datasets/matthewleechen/300YearsOfBritishPatents_KPST.

[16] When we calculate forward and backward similarity, we drop years toward the start and end of our period for which we do not have the full horizon. For instance, when implementing a forward horizon of 10 years and a backward horizon of 5 years, we consider the years 1739–1889 (inclusive). We also impose sparsity by setting small cosine similarities below a 0.05 threshold to zero.

[17] Our implementation differs from the one implemented in Kelly et al. (2021) in that it computes the ratio of average similarities rather than sums of similarities. This aids with interpretation by centering the measure on 1,

Table 1
Top 10 occupations.

Period	Occupation	Count	Percentage	Total inventors
1617–1750	Gentleman	132	13.52%	976
	Esquire	83	8.50%	976
	Merchant	48	4.92%	976
	Knight	27	2.77%	976
	Watchmaker	7	0.72%	976
	Chymist	6	0.61%	976
	Engineer	6	0.61%	976
	Weaver	6	0.61%	976
	Captain	5	0.51%	976
	Doctor In Physick	4	0.41%	976
1751–1800	Gentleman	205	11.86%	1,728
	Esquire	87	5.03%	1,728
	Merchant	78	4.51%	1,728
	Engineer	51	2.95%	1,728
	Watchmaker	25	1.45%	1,728
	Ironmonger	21	1.22%	1,728
	Surgeon	19	1.10%	1,728
	Chymist	19	1.10%	1,728
	Hosier	15	0.87%	1,728
	Optician	14	0.81%	1,728
1801–1850	Gentleman	1,206	12.97%	9,299
	Engineer	803	8.64%	9,299
	Merchant	448	4.82%	9,299
	Esquire	418	4.50%	9,299
	Civil Engineer	282	3.03%	9,299
	Manufacturer	247	2.66%	9,299
	Chemist	138	1.48%	9,299
	Machine Maker	107	1.15%	9,299
	Machinist	86	0.92%	9,299
	Mechanic	76	0.82%	9,299
1851–1899	Engineer	21,289	11.03%	193,079
	Manufacturer	12,374	6.41%	193,079
	Gentleman	9,869	5.11%	193,079
	Merchant	5,626	2.91%	193,079
	Civil Engineer	3,257	1.69%	193,079
	Mechanical Engineer	2,434	1.26%	193,079
	Machinist	2,164	1.12%	193,079
	Mechanic	1,795	0.93%	193,079
	Manager	1,757	0.91%	193,079
	Chemist	1,667	0.86%	193,079

Note: Table displays the top 10 most frequent occupations in each of four periods: 1617–1750, 1751–1800, 1801–1850, and 1851–1899. Occupations are standardized using a series of rules: we drop plural mentions of occupations, and group together common abbreviations for 'gentleman' and 'esquire'. We then take the modal occupation of inventors (or if there does not exist a mode for a particular inventor, we assign a random occupation). We display the count (number of inventors in that period with the given occupation), the percentage of patents associated with that occupation, and the total number of inventors in the period.

is from existing technology (novelty) and how much it influenced later developments (impact). We use these breakthrough scores to identify and study important patents during the period 1734–1899, when the submission of patent specifications was mandatory in Britain. Fig. 4 shows breakthrough score percentiles over time. A modest increase emerged during the 1780s–90s, coinciding with the First Industrial Revolution, while more substantial increases occurred around the 1870s during the Second Industrial Revolution. The 1852 and 1883 reforms caused dramatic shifts reflecting changes in patent volumes.

We perform a number of additional analyses to benchmark these scores. We show that (1) they tend to correlate with existing bibliometric patent quality measures from Nuvolari et al. (2021) (Appendix Table B.3.); (2) in line with Hanlon (2025), specialized and engineering occupations predominate among high-breakthrough score patents (Appendix Table B.4.); (3) consistent with de Rassenfosse and Jaffe (2018) and Nicholas (2011), the two major patent reforms reducing the cost of patents coincided with an increase in the filing of lower-quality patents (Appendix Figure B.5.)[18]; and (4) a list of the highest breakthrough score patents in each year reveal a number of eminent historical inventors such as James Watt, Charles Wheatstone, and John Heathcoat (Appendix Table B.5).

3.4. *Comparisons to existing data*

Our dataset makes two distinct contributions relative to existing patent data. First, we provide comprehensive historical coverage of British innovation across three centuries. Second, we make available high-quality machine-readable patent texts. In this subsection, we discuss how it complements several existing sources of patent data.

Existing historical patent datasets, while valuable for studying long-run innovation, have important limitations (Andrews, 2021). These datasets are generally restricted to U.S. patents, resulting in significant blind spots in our understanding of global innovation — particularly in crucial contexts like Britain during the Industrial Revolution. Existing UK patent datasets contain only variables that were published in Patent Office indexes (Nuvolari and Tartari, 2011; Bottomley, 2014a; Hanlon, 2016; Billington, 2021; Nuvolari et al., 2021; Bergeaud and Verluise, 2024), with no information drawn from the original specifications. Moreover, they do not provide the original patent texts, preventing researchers from analyzing the technical substance of historical innovations.

Contemporary textual patent collections also face significant constraints. Recent machine learning-oriented datasets such as USPTO-2M (Li et al., 2018), BIGPATENT (Sharma et al., 2019), and the Harvard USPTO Database (Suzgun et al., 2022) are restricted to recent decades. While broader collections like Google Patents include some historical coverage, their treatment of early British patents is both incomplete and error-prone. For example, at the time of writing, Google Patents contains only about 83,000 British patents granted prior to 1900 — roughly 26% of known grants during this period — with frequent duplicates and missing specification texts. Furthermore, the OCR quality of these historical patent texts is often poor, limiting their tractability for downstream analyses.

Our dataset addresses these limitations by providing complete coverage of historical British patents up to the turn of the 20th century, with high-quality machine-readable text derived from original patent specifications. This enables researchers to leverage modern natural language processing techniques to analyze technological change over an extensive historical timespan.

4. Use cases and discussion

Our dataset enables several novel research directions that are of relevance both for understanding long-run innovation dynamics and in answering specific historical questions of interest.

First, the unique combination of textual information and granular biographical data that we extract about inventors allows researchers to study a variety of new questions. How did inventors' background characteristics condition patterns of subsequent innovation? How did patenting behavior change over time in response to economic and political shocks, and which inventor characteristics mediated this? How did inventors who moved around differ in their patenting behavior

where breakthrough scores exceeding 1 indicate patents that have a relatively higher impact. It also partly addresses the massive increase in overall patent counts over time, which could bias a measure based only on counts toward more recent years.

[18] Although this suggests that the average patent tends to decline in quality as the cost of patenting falls, the increase in the average breakthrough scores in the years following these reforms. We agree with you that this suggests an increase in patents in the very top tail of the quality distribution.

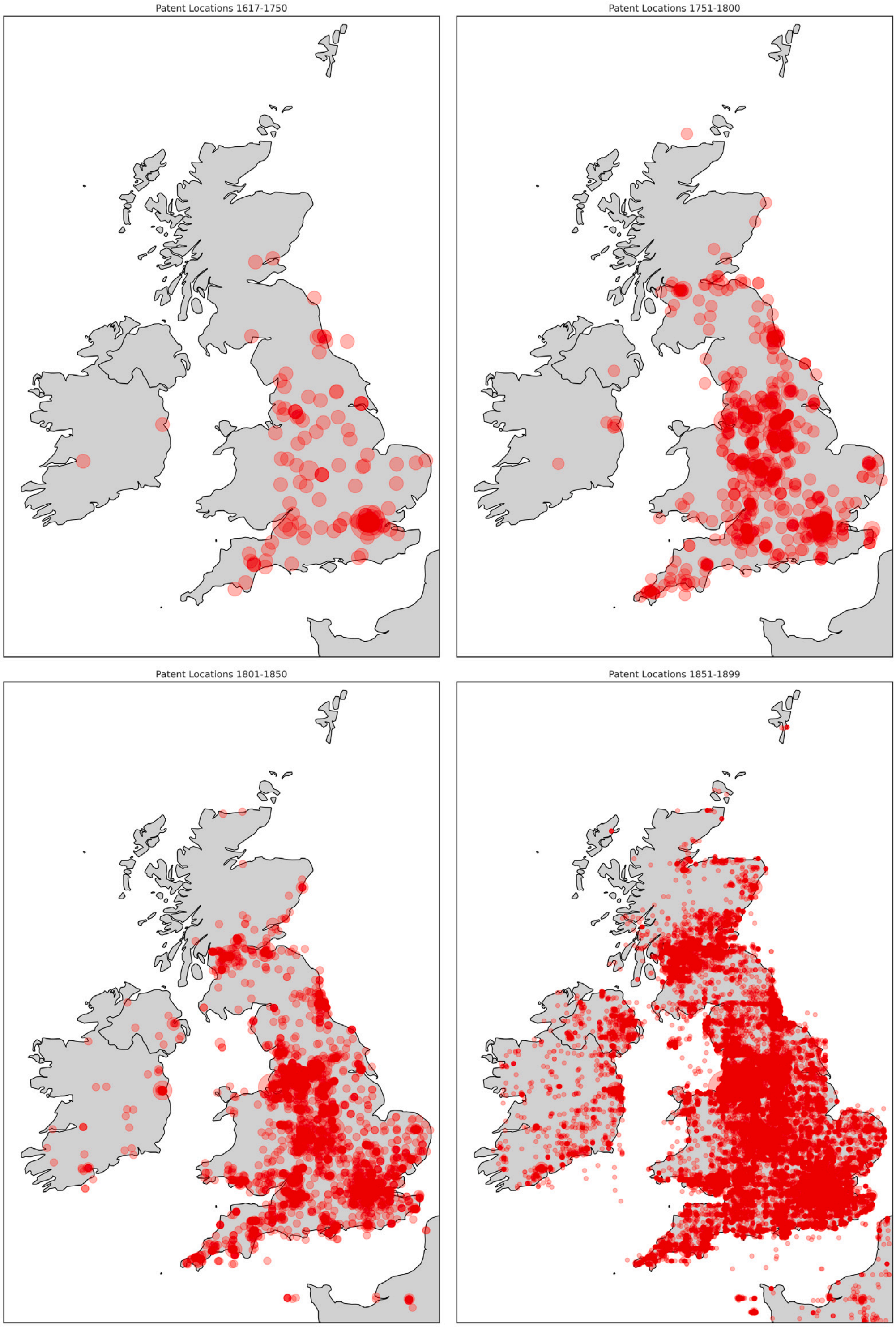

Patent Locations 1617-1750
Patent Locations 1751-1800
Patent Locations 1801-1850
Patent Locations 1851-1899

Fig. 3. UK Patent Locations, 1617–1899. *Note*: Plots show UK maps overlayed with representations of the number of patents geocoded to each longitude–latitude as a circle. For each patent, we identify the modal inventor's longitude–latitude coordinates. If there are multiple equally common locations, we randomly pick one. The size of each circle is proportional to the square root of the ratio of the number of patents at that location to the maximum number of patents observed at any single location. We use a GeoJSON file of Europe from: https://github.com/leakyMirror/map-of-europe/blob/master/GeoJSON/europe.geojson.

from those who remained in one location? Existing work that has examined the relationship between occupations and patenting behavior has been limited by the amount of textual detail available in previous datasets (Billington, 2021; Hanlon, 2025; Nuvolari et al., 2021). Moreover, the relevant Patent Office indexes ceased to be published after the mid-19th century, limiting the time horizon and — due to the massive increase in patenting activity over the 19th century — the number of observations in these studies.

In addition, rich biographical information on inventors throughout the late 19th century makes high-fidelity census linking possible for inventors active during the Second Industrial Revolution. The recently digitized I-CeM census data (1851–1911) (Schurer et al., 2024) can be matched to our inventor records on name, occupation, and address

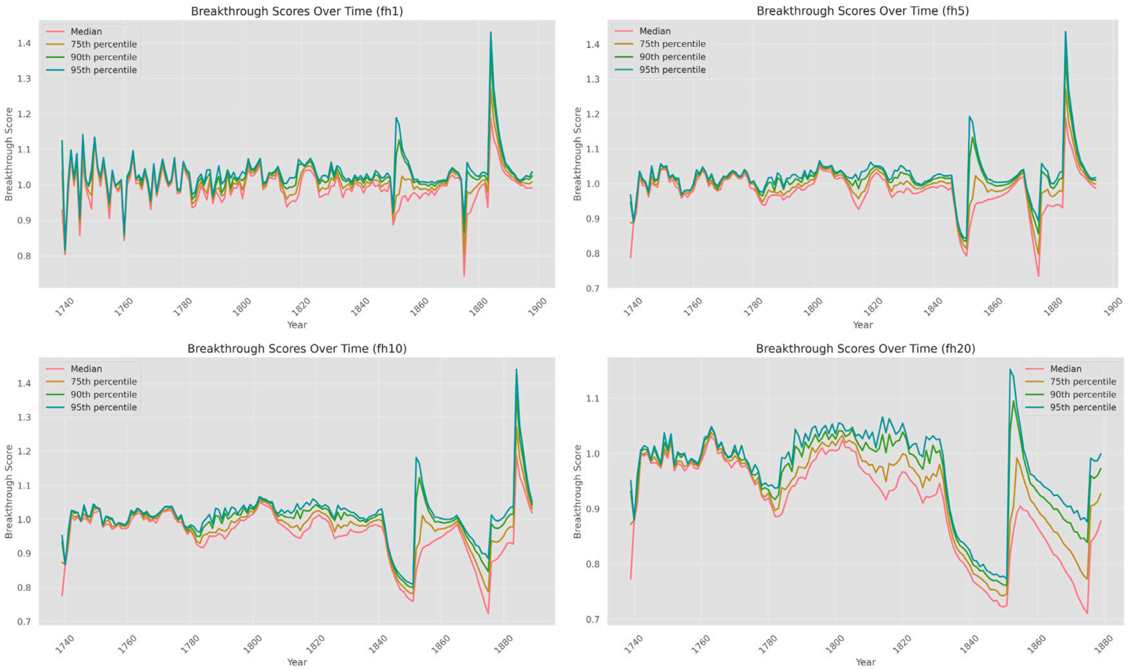


Fig. 4. Kelly et al. (2021) raw breakthrough scores. *Note*: Plots show the 50th, 75th, 90th, and 95th percentiles of raw breakthrough scores computed yearly across forward horizons (FH) of 1, 5, 10, and 20 years, and a backward horizon (BH) of 5 years.

information — a task that has proven challenging with U.S. patent data due to the lack of occupational information, making the identification of ground truth linkages difficult. The potential for accurate census linking helps to facilitate studies of inventors' family backgrounds, early life experiences, and socioeconomic trajectories. Researchers will now be able to trace how factors such as social class, fathers' occupation, and schooling shaped innovation.

The textual component of our dataset opens additional analytical possibilities. Recent advances in natural language processing, including transformer language models trained to capture patent textual similarity (e.g., Bekamiri et al., 2021; Ghosh et al., 2024) and methods for tracking novel ideas through patent text (e.g., Arts et al., 2021), can now be applied to historical patent texts at scale. This enables research that focuses on the changing language of innovation. For instance, how did technical language in patents evolve over time? How did the complexity and precision of patent language change? How did different occupational groups approach similar technical problems? Our textual data also serves as a unique source of training data to improve the performance of large language models on historical texts, especially in the technology domain. To our knowledge, this represents the first long-run patent text corpus suitable for such analyses. Alongside our data, the release of our custom fine-tuned models makes our processing steps extensible to British patent texts beyond 1900 and enables consistent analyses across extended periods.

The institutional background of the British patent system fundamentally shapes how researchers should approach this dataset. The consistent requirement for detailed technical specifications, dating back to the early 18th century, provides remarkably uniform technical information across our coverage period. However, researchers must carefully consider several important caveats. First, the patent reforms of 1852 and 1883 significantly changed the incentives and costs of patenting. These institutional changes likely affected selection into becoming an inventor, as well as what types of innovations were patented. Researchers will need to carefully consider these selection issues when comparing patents across time periods. Second, while our automated extraction methods enable improved scaling and extensibility, they introduce some degree of measurement error. Though some errors in digitization will remain, OCR quality is high, with manual inspection confirming the texts remain highly comprehensible to human readers, and we expect — to a first order — that these errors would be non-systematic. There are also errors when extracting named entities. These errors will be most severe in the earlier patents due to a lack of training examples that we can feasibly provide (since the total number of these early patents is comparatively small), as well as the use of older language that makes fine-tuning a language model more difficult. Entity extraction accuracy varies by field, and our test set evaluation scores are as follows: common elements like names (94.9% F1), occupations (93.8%), and addresses (89.9%) are reliable, while rarer elements like firm affiliations (76.2%) and foreign communications (81.8%) are noisier. For transparency, we preserve rather than correct these errors. Researchers focused on rare entity types may need to implement additional correction and validation procedures.

Finally, it is important to note that while textual analysis of patent documents offers valuable insights into innovation patterns, there are fundamental limitations to what we can learn from text alone. Patent language reflects not only technical content but also the educational background, social status, and writing conventions of inventors.[19] More fundamentally, the qualities that retrospectively define an invention as a breakthrough or historically significant may not map to observable textual features. This mismatch suggests that text-based measures will systematically misidentify breakthrough innovations—failing to recognize some genuinely transformative inventions while erroneously recognizing others.

Despite these caveats, we believe that our dataset — when these limitations are appropriately considered — offers an important new resource to facilitate research in innovation dynamics and technological change. We uniquely combine complete textual data alongside detailed inventor information over an extended time period. The fine-tuned models we provide allow researchers to reproduce and extend these analyses further. The unique long-run perspective provided by our data will help shed light on innovation, technological change, and economic growth in ways that have been challenging to explore before.

[19] Although patent agents became increasingly common throughout the 19th century, this was not ubiquitous and this concern remains even for patents filed late in our period.

CRediT authorship contribution statement

Enrico Berkes: Writing – review & editing, Writing – original draft, Validation, Investigation, Conceptualization, Resources. **Matthew Lee Chen:** Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Matteo Tranchero:** Writing – review & editing, Writing – original draft, Validation, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Matthew Lee Chen reports financial support was provided by Google Inc. Matteo Tranchero reports financial support was provided by Google Inc. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.respol.2025.105347.

Data availability

Data and codes are freely available and linked in the manuscript.

References

Andrews, M.J., 2021. Historical patent data: A practitioner's guide. J. Econ. Manag. Strat. 30 (2), 368–397.

Arora, A., Dell, M., 2024. LinkTransformer: A unified package for record linkage with transformer language models.

Arts, S., Hou, J., Gomez, J.C., 2021. Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. Res. Policy 50 (2), 104144.

Bekamiri, H., Hain, D.S., Jurowetzki, R., 2021. Hybrid model for patent classification using augmented SBERT and KNN. CoRR abs/2103.11933.

Bergeaud, A., Verluise, C., 2024. A new dataset to study a century of innovation in europe and in the US. Res. Policy 53 (1), 104903.

Berkes, E.G., 2018. Comprehensive universe of U.S. patents (CUSP): Data and facts. URL https://api.semanticscholar.org/CorpusID:231916975.

Billington, S.D., 2021. What explains patenting behaviour during britain's industrial revolution? Explor. Econ. Hist. 82, 101426.

Bottomley, S., 2014a. Patenting in England, Scotland, and Ireland during the industrial revolution, 1700–1852. Explor. Econ. Hist. 54, 48–63.

Bottomley, S., 2014b. Patents and the First Industrial Revolution in the United States, France and Britain, 1700-1850. Technical Report.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., 2019. Unsupervised cross-lingual representation learning at scale. CoRR abs/1911.02116.

Cox, G.W., 2020. Patent disclosure and England's early industrial revolution. Eur. Rev. Econ. Hist. 24 (3), 447–467.

de Rassenfosse, G., Jaffe, A.B., 2018. Are patent fees effective at weeding out low-quality patents? J. Econ. Manag. Strat. 27 (1), 134–148.

Feng, S., 2020. The proximity of ideas: An analysis of patent text using machine learning. PLoS One 15, 1–19.

Ghosh, M., Erhardt, S., Rose, M.E., Buunk, E., Harhoff, D., 2024. PaECTER: Patent-level representation learning using citation-informed transformers.

Hall, B.H., Jaffe, A.B., Trajtenberg, M., 2001. The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools. Working Paper 8498, National Bureau of Economic Research.

Hanlon, W.W., 2016. British patent technology classification database: 1855–1882. Available at: http://www.econ.ucla.edu/whanlon/.

Hanlon, W.W., 2025. The rise of the engineer: Inventing the professional inventor during the industrial revolution. Econ. J..

Jaffe, A., 2002. Patents, Citations, and Innovations: A Window on the Knowledge Economy. MIT Press.

Juhász, R., Sakabe, S., Weinstein, D., 2024. Codification, Technology Absorption, and the Globalization of the Industrial Revolution. Working Paper 32667, National Bureau of Economic Research.

Kalyani, A., 2024. The Creativity Decline: Evidence from US Patents. Working Paper 4318158, Social Science Research Network (SSRN).

Kalyani, A., Bloom, N., Carvalho, M., Hassan, T., Lerner, J., Tahoun, A., 2025. The diffusion of new technologies. Q. J. Econ. 140 (2), 1299–1365.

Kelly, B., Papanikolaou, D., Seru, A., Taddy, M., 2021. Measuring technological innovation over the long run. Am. Econ. Rev.: Insights 3 (3), 303–320.

Khan, B., 2005. The democratization of invention: Patents and copyrights in American economic development, 1790-1920. In: NBER Series on Long-Term Factors in Economic Development, Cambridge University Press.

Kogan, L., Papanikolaou, D., Seru, A., Stoffman, N., 2017. Technological innovation, resource allocation, and growth*. Q. J. Econ. 132 (2), 665–712.

Li, S., Hu, J., Cui, Y., Hu, J., 2018. DeepPatent: Patent classification with convolutional neural networks and word embedding. Sci. 117 (2), 721–744.

MacLeod, C., 1988. The development of the patent system, 1660–1800. In: Inventing the Industrial Revolution: The English Patent System, 1660–1800. Cambridge University Press, pp. 40–57.

Marco, A.C., Carley, M., Jackson, S., Myers, A.F., 2015. The USPTO historical patent data files: Two centuries of innovation. SSRN Working Paper.

Nicholas, T., 2011. Cheaper patents. Res. Policy 40 (2), 325–339.

Nuvolari, A., Tartari, V., 2011. Bennet woodcroft and the value of english patents, 1617–1841. Explor. Econ. Hist. 48 (1), 97–115.

Nuvolari, A., Tartari, V., Tranchero, M., 2021. Patterns of innovation during the industrial revolution: A reappraisal using a composite indicator of patent quality. Explor. Econ. Hist. 82, 101419.

Petralia, S., Balland, P.-A., Rigby, D., 2016. HistPat dataset.

Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Robinson, E., 1972. James Watt and the law of patents. Technol. Cult. 13 (2), 115–139.

Rosenberger, L., Hanlon, W.W., Hallmann, C., 2024. Innovation Networks in the Industrial Revolution. Working Paper 32875, National Bureau of Economic Research.

Sarada, S., Andrews, M.J., Ziebarth, N.L., 2019. Changes in the demographics of American inventors, 1870–1940. Explor. Econ. Hist. 74, 101275.

Schurer, K., Higgs, E., Limited, F., 2024. Integrated census microdata (I-CeM), 1851–1911. [Data Collection].

Sharma, E., Li, C., Wang, L., 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 2204–2213.

Suzgun, M., Melas-Kyriazi, L., Sarkar, S.K., Kominers, S.D., Shieber, S.M., 2022. The harvard USPTO patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications.

Taylor, B.M., 1969. The British patent system. I. Administration. Camb. Law J. 27 (1), 131–134.

Van Dulken, S., 1999. British Patents of Invention, 1617–1977: A Guide for Researchers. British Library.

Varghese, R., Sambath, M., 2024. YOLOv8: A novel object detection algorithm with enhanced performance and robustness. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems. ADICS, pp. 1–6.

Woodcroft, B., 1854. Chronological Index of Patents of Invention, 1617–1852, vol. 2, The Queen's Printing Office.