Data in Brief 41 (2022) 107966

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Bibliometric dataset (1990–2019) concerning 35 city labels dealing with sustainable urbanism

Simon Joss[a,*], Daan Schraven[b], Martin de Jong[c]

[a] Urban Studies, School of Social and Political Sciences, University of Glasgow, 25 Bute Gardens, Glasgow G12 8RS, UK

[b] Department of Materials, Mechanics, Management and Design (3Md), Section of Infrastructure Design and Management, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Postbus 5048 GA 2600, Delft, The Netherlands

[c] Erasmus School of Law, Rotterdam School of Management, Erasmus University Rotterdam, P.O.Box 1738/ 3000 DR, Rotterdam, The Netherlands

ARTICLE INFO

Article history:
Received 10 October 2021
Revised 21 January 2022
Accepted 11 February 2022
Available online 18 February 2022

Keywords:
Bibliometrics
Scientometrics
City labels
Sustainable city
Smart city
Urban development
SDGs
Urban futures

ABSTRACT

This data article presents a tripartite dataset that formed the empirical basis for a comprehensive bibliometric analysis of the use of city labels denoting sustainable urbanism in the scientific literature (Schraven, 2021). The tripartite dataset was generated using the abstract and citation database Scopus (Elsevier). Dataset A lists 148 city labels denoting different approaches to urban planning and development. It was used to select 35 city labels that specifically address sustainable urbanism ('sustainable city', 'smart city', 'compact city' etc.). Dataset B references 11,337 journal and review articles spanning the period 1990–2019. All retrieved articles contain at least one of the 35 city labels in the title, abstract, and author keywords. This database was used to calculate the frequency of the selected city labels across time, and to analyze the co-occurrences of city labels. It was further used to calculate the future trajectory of scientific outputs using the Logistic Growth Model (LGM). Dataset C entails 22,820 author keywords extracted from across the 11,337 articles. This was used to analyze the co-occurrences of keywords with city la-

DOI of original article: 10.1016/j.jclepro.2021.125924
* Corresponding author.
E-mail address: Simon.Joss@glasgow.ac.uk (S. Joss).

https://doi.org/10.1016/j.dib.2022.107966

bels. The data article describes the methods of data collection and curation, the analysis performed, and the potential for reusing the data for further research. The comprehensiveness of the bibliometric corpus – spanning three decades and 35 city labels – lends itself to further investigation of how sustainable urban development has evolved as a topic in the scientific literature since the 1990s. Furthermore, the robust methodology developed could be adapted to other scientific repositories and, indeed, other research problems and questions.

Specifications Table

Subject	Social Sciences; Library and Information Sciences
Specific subject area	Bibliometrics; Scientometrics; sustainable urban development
Type of data	Figure (visualized data) Tables (curated data; analyzed data)
How the data were acquired	Using an initial search query for 12 city labels (for Dataset A), followed by an extended search query for 35 city labels (for Dataset B and C), the data was acquired through the Elsevier Scopus abstract and citation database.
Data format	Curated (re-ordered; filtered); analyzed
Description of data collection	The tripartite dataset consists of: • Dataset A: 148 city labels obtained from initial search query in Scopus, used to select 35 city labels based upon qualitative analysis. • Dataset B: 11,337 journal articles (author names, journal publication, year of publication, article title, abstract, author keywords) retrieved from Scopus using search query containing 35 city labels. • Dataset C: 22,820 author keywords extracted from Dataset B.
Data source location	The primary data was sourced from Scopus (Elsevier) at: www.scopus.com. To reproduce the data, use the following steps: (1) enter the system with your log in; (2) click on 'advanced document search'; (3) set the time period '1990' to '2019'; (4) place the search query for 35 city labels (see below, p.4) in 'enter search query' field; (5) click on 'search' tab; (6) retrieve result.
Data accessibility	The curated data [2], which entails datasets A, B, and C, can be accessed via the following link: doi:10.4121/13580273.v2 Repository: 4TU.ResearchData. doi:10.4121/13,580,273.v2
Related research article	D. Schraven, S Joss, M. de Jong, Past, present, future: Engagement with sustainable urban development through 35 city labels in the scientific literature 1990–2019, Journal of Cleaner Production, 292 (2021) 125,924, doi:10.1016/j.jclepro.2021.125924

Value of the Data

- The tripartite dataset is one of the most comprehensive bibliometric repositories to date that captures essential information about sustainable urban development in the scientific literature. It spans three decades (1990–2019) and encompasses the 35 most important city labels used in the scientific and policy literatures.
- The dataset is accompanied by a robust bibliometric methodology that allows for detailed analysis of the conceptual co-evolution of city labels across consecutive time periods, including the calculation of future trends using the Logistic Growth Model (LGM).

- The comprehensive data repository is useful for researchers who are interested in exploring the broad interface between urban planning and sustainable development (including the UN's SDGs).
- Since the repository is based on a diverse set of 35 city labels that express various features of urban development goals, it can be used to examine different environmental, economic, and social aspects and their interrelationships.
- Apart from carrying out additional analysis with the existing data, the dataset can be used to expand beyond 2019 in future years.

1. Data Description

The article is accompanied by three interlinked data files deposited in the international data repository 4TU.ResearchData as follows:

Dataset A: Presented in txt format, the dataset lists 148 city labels (curated data), including a subset of 35 city labels. The document also entails curated and analyzed data in the form of qualitative evaluations of each city label based on three specified criteria (see next section). This qualitative procedure was used to select 35 city labels to generate Datasets B and C.

Dataset B: This contains curated bibliometric data (XLSX format) of the total of collected journal articles, including: authors' names; year of publication; journal publication; reference information (journal volume, issue, page numbers); article title; abstract; author keywords. See next section for the search term used to harvest the bibliometric information. There is a total of 11,337 entries. The data was collected on 6 January 2020.

Dataset C: This contains curated bibliometric data (XLSX format) of the total of author keywords extracted from the 11,337 journal articles (Dataset B). The data is presented in matrix format: 22,820 keywords x 35 city labels. This, thus, lists all keywords associated with each city label (in title, abstract, keywords) across the entire bibliometric repository (conversely, for each keyword the dataset shows the associated city labels). There is a total of 798,700 entries.

2. Experimental Design, Materials and Methods

Fig. 1 (below) illustrates the research design with its constituent components that produced the three interlocking datasets. The research design was motivated by an inquiry into the emergence and long-term evolution, within the scientific literature, of city labels addressing various aspects of sustainable urbanism (see [1] for background). A city label is defined as "a classifying phrase that succinctly expresses essential features of urban development goals" [1]. It is made up of a qualifying term preceding the word 'city', e.g., 'resilient city', 'liveable city', 'green city'. Each city label, thus, expresses a particular strategy for, and approach to, urban planning and development. Apart from the use in the scholarly literature, city labels are also frequently used in urban policy and practice (e.g., 'Vienna Smart City', 'Tianjin Ecocity') to prioritize and promote urban plans and initiatives.

Dataset A: The dataset of 148 city labels was generated using an initial list of 12 core city labels that had been identified in earlier bibliometric studies [3–5]. Using these 12 city labels, an initial harvest of journal articles was carried out in Scopus to screen for further associated city labels (in article title, abstract, and keywords). This generated 148 city labels. Each of these was analyzed qualitatively, with a view to selecting a sufficiently comprehensive yet manageable list of city labels, using the following criteria: (1) the selected city labels must conceptually relate to (aspects of) sustainable urban development; (2) they must have an established presence in the academic literature; and (3) they must resonate in policy and practice discourse (as exemplified

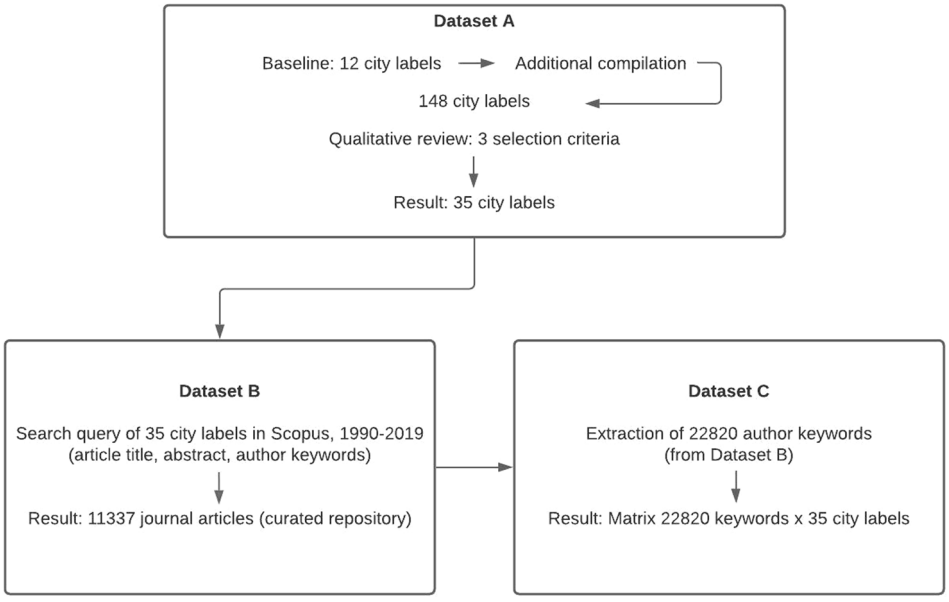

Fig. 1. Tripartite dataset A–C.

above). Dataset A includes the related qualitative information and explains the selection of 35 city labels from the overall list.

There is scope for further use of the dataset, e.g., by using the sub-set of 35 city labels for other bibliometric or webometric analyses, and carrying out bibliometric analyses with other city labels (than the 35 selected) from the list of 148.

Dataset B: Scopus was chosen as authoritative author and citation database, as it entails complete journal publication records irrespective of changing ISI status (unlike Web of Science), dating back to 1996. For the earlier period (1990–1995, which was important to include, given the foundational UN sustainable development conference, 'Earth Summit' in 1992), the Scopus records may be partially incomplete; however, this had minimal effect on the overall dataset, as the article output (<100) was dwarfed by the exponential growth in the subsequent periods. Only journal and review articles were harvested (thus excluding conference proceedings etc.), as they represent the gold standard of peer-reviewed scientific output. For each article, title, abstract, and author keywords were extracted; these encapsulate the essence of scientific findings and are typically used for bibliometric analysis. (As the Scopus keyword operator includes index keywords, its function was adjusted to author keywords only.) The following search query was used for the main data harvesting exercise, which generated a total of 11,337 articles:

> (TITLE-ABS ("biophilic city" OR "biophilic cities" OR "circular city" OR "circular cities" OR "compact city" OR "compact cities" OR "competitive city" OR "competitive cities" OR "connected city" OR "connected cities" OR "creative city" OR "creative cities" OR "digital city" OR "digital cities" OR "eco city " OR "eco cities " OR "ecological city" OR "ecological cities" OR "entrepreneurial city" OR "entrepreneurial cities" OR "experimental city" OR "experimental cities" OR "future city" OR "future cities" OR "green city" OR "green cities" OR "inclusive city" OR "inclusive cities" OR "information city" OR "information cities" OR "intelligent city" OR "intelligent cities" OR "knowledge city" OR "knowledge cities" OR "learning city" OR "learning cities" OR "liveable city" OR "liveable cities" OR " livable city" OR "livable cities" OR "low-carbon city" OR "low-carbon cities" OR "open city" OR "open cities" OR "playful city" OR "playful cities" OR "post-carbon city" OR "post-carbon cities"

OR “productive city” OR “productive cities” OR “regenerative city” OR “regenerative cities” OR “renewable city” OR “renewable cities” OR “resilient city” OR “resilient cities” OR “safe city” OR “safe cities” OR “sharing city” OR “sharing cities” OR “smart city” OR “smart cities” OR “sponge city” OR “sponge cities” OR “solar city” OR “solar cities” OR “sustainable city” OR “sustainable cities” OR “ubiquitous city” OR “ubiquitous cities” OR “virtual city” OR “virtual cities” OR “zero-carbon city” OR “zero-carbon cities”) OR AUTHKEY (“biophilic city” OR “biophilic cities” OR “circular city” OR “circular cities” OR “compact city” OR “compact cities” OR “competitive city” OR “competitive cities” OR “connected city” OR “connected cities” OR “creative city” OR “creative cities” OR “digital city” OR “digital cities” OR “eco city “ OR “eco cities “ OR “ecological city” OR “ecological cities” OR “entrepreneurial city” OR “entrepreneurial cities” OR “experimental city” OR “experimental cities” OR “future city” OR “future cities” OR “green city” OR “green cities” OR “inclusive city” OR “inclusive cities” OR “information city” OR “information cities” OR “intelligent city” OR “intelligent cities” OR “knowledge city” OR “knowledge cities” OR “learning city” OR “learning cities” OR “liveable city” OR “liveable cities” OR “ livable city” OR “livable cities” OR “low-carbon city” OR “low-carbon cities” OR “open city” OR “open cities” OR “playful city” OR “playful cities” OR “post-carbon city” OR “post-carbon cities” OR “productive city” OR “productive cities” OR “regenerative city” OR “regenerative cities” OR “renewable city” OR “renewable cities” OR “resilient city” OR “resilient cities” OR “safe city” OR “safe cities” OR “sharing city” OR “sharing cities” OR “smart city” OR “smart cities” OR “sponge city” OR “sponge cities” OR “solar city” OR “solar cities” OR “sustainable city” OR “sustainable cities” OR “ubiquitous city” OR “ubiquitous cities” OR “virtual city” OR “virtual cities” OR “zero-carbon city” OR “zero-carbon cities”)) AND DOCTYPE (ar OR re) AND PUBYEAR $>$ 1989 AND PUBYEAR $<$ 2020

The dataset was used to carry out three types of analysis:

- *Occurrences of city labels*. The frequency of a city label refers to the number of articles in which said label occurs at least once in the title, abstract, and author keywords. The frequency is a measure of the prevalence and influence of a given city label in the scientific literature. Comparing the occurrences among the 35 city labels, and across time periods, offered valuable insights into how individual city labels have fared and how the field overall has evolved.
- *Co-occurrences of city labels.* Drawing on social network analysis, the co-occurrence analysis served to identify mutual connections among the 35 city labels. It counted the number of articles containing specific combinations of two labels in the title, abstract, and keywords. By registering all instances (number of articles) where city labels are used in conjunction with other city labels, a network relationship among the 35 labels emerged. This was visualized using Pajek software’s social network analysis [6,7].
- *Future forecast.* The Logistic Growth Model (LGM) was used to predict the future trajectory of city labels in the scientific literature, by extrapolating from the occurrence rates of each city label across the 1990–2019 period [8]. LGM is a regression model based upon a set of observations between cumulative growth of the number of articles per year. The cumulative number of articles plotted over time follow a general logistic growth patterns in the shape of an S-curve.

There is scope for further analysis of the dataset, e.g., by calculating the distribution of city labels across different journals and journal classifications, and conducting a mapping of author collaboration networks.

Dataset C: The dataset was generated by counting all articles mentioning at least one city label and at least one keyword (e.g., ‘sustainable city’ and ‘planning’), and repeating this for all unique pairs of city labels and keywords. The results were stored in a large 35 $\times$ 22,820 matrix.

The dataset was used to carry out the following analysis:

Table 1
Methodological procedures for datasets A–C.

Dataset A: Selection of 35 city labels

- Check existing bibliometric studies on multiple city labels [3–5], thereby identifying 12 city labels.
- Input the 12 city labels as search query in Scopus to retrieve further city labels from author keywords of retrieved articles, resulting in 148 city labels.
- Delete any duplicate city labels; carry out qualitative review (triangulated among researchers) based on three joint criteria (derived from [3]): only select city labels that: (i) conceptually relate to (aspects of) sustainable urban development; (ii) have an established presence in the academic literature; (iii) have a presence in policy/practice discourse.

Dataset B: Compilation & analysis of 11,337 research articles

Primary data extraction

- Formulate the 35 city labels as search query: see footnote 1.
- Enter search query in Scopus, setting 1990–2019, thus retrieving 11,337 articles.
- Collect bibliometric data: (i) title; (ii) abstract; (iii) author keywords.

Temporal analysis: 1990–2019

- Arrange 5-yearly temporal incisions resulting in 6 cumulative periods: 1990–1994; 1990–1999; 1990–2004; 1990–2009; 1990–2014; 1990–2019.

Occurrence analysis of city labels

- Count all articles in database mentioning a given city label at least once (an article is counted only once even if given city label is mentioned twice or more); repeat for each of the 35 city labels.
  - Tabulate city labels from highest to lowest counts, across six cumulative time periods.
  - Draw line graph showing yearly counts 1990–2019 for all city labels; apply logarithmic scale for legibility.
  - Draw scatter plot showing relative positions (cumulative frequencies) and new entry points of 35 city labels across six time periods.

Co-occurrence analysis of city labels

- Count all articles mentioning a pair of city labels (e.g. 'sustainable city' AND 'smart city') at least once; repeat for all unique pairs of city labels; and repeat for each cumulative period.
  - Store all counts of unique pairs in 6 matrices (35 × 35 cells) representing the 6 cumulative periods.
  - In Pajek software, draw a social network graph using each of the 6 m^2atrices.
  - Use 6th matrix (1990–2019) to list 10 highest co-occurrence frequencies in ranking order.
  - Use 6th matrix (1990–2019) to list city labels co-occurring with 'sustainable city', and 'smart city', respectively, in order of strength of connection.

Future forecast of city label occurrences

- Extrapolate future trajectory of city labels from occurrence rates 1990–2019, by applying Logistic Growth Model Curve to city label occurrences as follows:
- Extract from database city label occurrences per year.
- Following General Limit Theorem, exclude city labels with <30 occurrences, thus withdrawing 10 city labels.
- For each of the 25 retained city labels, create a regression model based on occurrences between cumulative growth of articles (Y) per year (X): $y = L/(1 + e\text{^}(\text{b-kx}))$ where L represents the total estimated capacity of no. of articles that a city label could carry; b and k represent the slope of the curve which follows a natural logarithm.
- Plot no. of articles over time following general logistic growth pattern in the shape of S-curve.
- Lock the position of each of the city labels on S-curve at final complete publication year: 2019.
- Normalize S-curve to relative growth, where $L = 100\%$, then plot all locked-in city labels.
- Draw development stages 'infant', 'growth', 'mature' [8] onto S-curve.
- For each of the 25 city labels, use regression model to predict start and finish of three development stages (Zeng et al., 2019), and store predictions in matrix.
- Sort city labels by predicted longevity, from 'open city' (till 2077) to 'ubiquitous city' (till 2024), and draw stacked bar chart of 25 city labels with development stages shown.

Dataset C: Compilation and analysis of 22,820 author keywords x 35 city labels

Co-occurrence analysis of keywords and city labels

- Count all articles mentioning at least one city label and at least one keyword (e.g., 'sustainable city' AND 'planning'); repeat for all unique pairs of city labels and keyworks; store resulting counts in large 35 × 22,820 matrix.
- Calculate degree of centrality (co-occurrence with no. of city labels) of all keywords; rank the keywords with 15 highest degrees (cut-off at degree of centrality 10).
- Harvest and rank the 15 most frequent keywords for each city label, yielding a total of 149 keywords.
- Filter and store 149 keyword counts in 35 × 149 matrix, and draw social network graph in Pajek.
- Draw two graphs based on extracted cluster (A) 'smart'-'intelligent'-'digital'-'ubiquitous'-'future'-'creative'-'connected' and cluster (B) 'sustainable'-'low-carbon'-'liveable'-'green'-'eco'-'compact'.

- *Co-occurrences of city labels and keywords*. Author keywords encapsulate essential theoretical and empirical information and associations, chosen by authors to define and categorize their research. Analyzing their co-occurrence with city labels provided valuable insights into the conceptual association of city labels, as well as the complex network of conceptual relationships among the city labels and keywords. Co-occurrence was established by counting the number of articles which mention a given city label together with a specific keyword. Given the multitude of combinations (N = 22,820 keywords), the focus of analysis was on the 15 most frequent keywords for each city label.

There is scope for further analysis of this dataset, beyond the 15 top keywords, e.g., by examining changes in keywords associations of city labels across different time periods. This could e.g., be used to establish whether there are any significant differences in the conceptual definition of 'future city' between, say, the 1990s and the 2010s.

Table 1 details the step-by-step methodological procedures for Datasets A, B, and C. Using these procedures allows for the reproduction of the citied study results. They can be adapted to carry out further types of studies, e.g., including more recent publication years.

Ethics Statements

The research did not involve any human subjects or animal experiments. No data was collected from social media platforms.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT Author Statement

Simon Joss: Conceptualization, Methodology, Data curation, Formal analysis, Writing – original draft, Writing – review & editing; **Daan Schraven:** Conceptualization, Methodology, Formal analysis, Software, Visualization, Data curation; **Martin de Jong:** Conceptualization, Methodology, Formal analysis, Writing – review & editing.

Acknowledgments

This data research was supported with funding from the UK's Economic and Social Research Council (ESRC; Grant No. ES/S007105/1), and the Dutch Research Organisation/Chinese National Science Foundation (NWO/NSFC; Grant No. 482.19.608.).

References

[1] D. Schraven, S. Joss, M. de Jong, Past, present, future: engagement with sustainable urban development through 35 city labels in the scientific literature 1990-2019, J. Clean. Prod. 292 (2021) 125924, doi:10.1016/j.jclepro.2021.125924.
[2] D. Schraven, S. Joss, M. de Jong. Data underlying the article: past, present, future: engagement with sustainable urban development through 35 city labels in the scientific literature 1990-2019. 4TU.ResearchData. Dataset. doi:10.4121/13580273.v2.
[3] M. De Jong, S. Joss, D. Schraven, C. Zhan, M. Weijnen, Sustainable-smart-resilient-low carbon-eco-knowledge cities; making sense of a multitude of concepts promoting sustainable urbanization, J. Clean. Prod. 109 (2015) 25–38.
[4] Y. Fu, X. Zhang, Trajectory of urban sustainability concepts: a 35-year bibliometric analysis, Cities 60 (2017) 113–123.
[5] M.H. Wang, Y.S. Ho, H.Z. Fu, Global performance and development on sustainable city based on natural science and social science research: a bibliometric analysis, Sci. Total Environ. 666 (2019) 1245–1254.
[6] V. Batagelj, A. Mrvar, 2011. Pajek software, vlado.fmf.uni-lj.si/pub/networks/pajek/.

[7] T. Kamada, S. Kawai, An algorithm for drawing general undirected graphs, Inf. Process. Lett. 31 (1) (1989) 7e15.
[8] L. Zeng, Z. Li, Z. Zhao, M. Mao, Landscapes and emerging trends of virtual reality in recent 30 years: a bibliometric analysis, in: Proceedings of the IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, 2018, doi:10.1109/SmartWorld.2018.00311.