Bridging the gap between science and society: Mapping libraries' strategies for engaging in the research impact process through semantic analysis

Wang Zuorong, Sun Jiaxuan, He Shan, Deng Sanhong, and Wang Hao

Abstract

Introduction. Scientific research's societal impact increasingly relies on knowledge dissemination beyond academia. Libraries are evolving from repositories into dynamic intermediaries that bridge societal gaps by enhancing digital literacy and ensuring equitable information access. Existing literature, however, has largely overlooked such institutional intermediaries, limiting systematic understanding of their strategies.

Method. We analysed 465 library participation cases from the Research Excellence Framework using a hybrid approach combining large language models (LLMs) and BERTopic semantic analysis.

Results. Five key strategies were identified: media communication and public engagement; public dialogue and cultural presentation; artistic collaboration and live experiences; digital content creation and dissemination; and large-scale event coordination. Library contributions are concentrated in Arts and Humanities (76%), primarily within cultural (67%) and societal (22%) impact domains.

Conclusion. This study proposes an evidence-based framework elucidating the role of libraries within the research ecosystem, offering practical insights to support the societal translation of research outcomes. Future research should explore differentiated strategies for various library types and the feasibility of generalising these findings to other regions

Introduction

In the digital age, the value of scientific research depends not only on its quality, but also on how effectively it reaches and benefits communities beyond academia. However, significant barriers between researchers and the public continue to hinder this process. In this context, both public and academic libraries are evolving. They no longer function solely as repositories of knowledge but are becoming active intermediaries that facilitate knowledge sharing and connect diverse segments of society. A central objective of libraries is to enhance digital skills and ensure equitable access to information, thereby enabling broader public engagement with scientific work(Mensah Danquah et al., 2025; Rosário et al., 2025; Saffarinia, 2025).
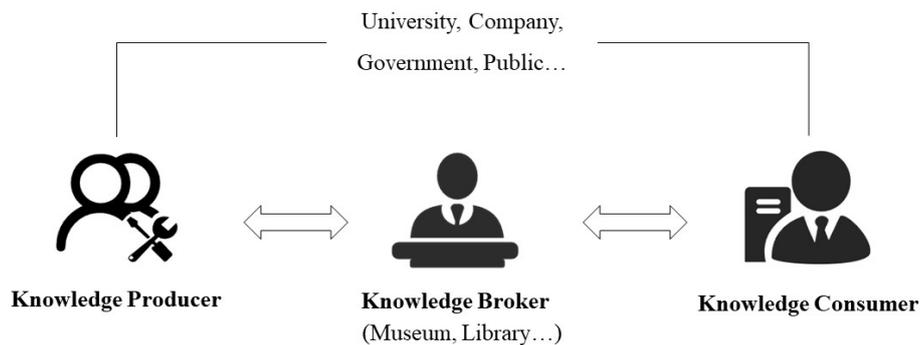
The process of generating scientific research impact is, in essence, a journey where knowledge undergoes *'production–dissemination–application,'* with different actors assuming distinct roles throughout. While existing literature on research impact often emphasises knowledge producers and consumers—such as corporations, government agencies, and universities—less attention has been paid to institutional intermediaries like libraries. This gap has limited a systematic understanding of how libraries develop and deploy multifaceted strategies to address digital inequities and enhance the reach and relevance of scientific research.

To address this, our study employed a data-driven approach. We applied modern text analysis techniques, including large language models (LLMs) and BERTopic, to examine 465 real-world case studies from the REF databases where libraries played a documented role. This methodology allowed us to impartially identify and categorise the specific strategies that libraries use to support research impact and promote digital inclusion. Overall, this study provides a practical and empirically grounded framework for understanding how libraries are adapting their strategies in the digital age.

Research on actors in the research impact generation process

The concept of scientific research impact adopted in this study aligns with the definition provided by the Research Excellence Framework (REF), which emphasises impacts beyond academia(UKRI, 2022). Analysing the historical roles of different stakeholders in the generation of scientific research's societal impact offers valuable insights: it assists research institutions in shaping future engagement strategies and provides evaluation bodies with a multidimensional understanding of how such impact is produced.

Existing research on stakeholder involvement in the research impact process has largely concentrated on four main actors: universities(Zhang et al., 2023), businesses (Bracci et al., 2022; Polidoro & Jacobs, 2024; Wang et al., 2025), government (Shin et al., 2022), and the general public (Ribeiro et al., 2024) . These groups typically function as knowledge producers or knowledge consumers. For example, universities primarily generate innovative knowledge through research, whereas government agencies often act as knowledge consumers by integrating research findings into policymaking, thereby generating policy impact (as shown in Figure 1).

University, Company,
Government, Public…

Knowledge Producer   Knowledge Broker   Knowledge Consumer
                       (Museum, Library…)

Figure 1. Actors in the process of generating scientific research impact.

As the pathways to research impact grow more complex, knowledge brokers—which connect producers and users of knowledge—are playing an increasingly important role(Lomas, 2007; Thompson et al., 2006; Turnhout et al., 2013). These brokers facilitate the flow and exchange of knowledge, serving as essential catalysts in the creation of impact (Hladik, 2023; Kousha et al., 2024; Meyer, 2010). Among them, libraries hold a distinctive position. They are tasked with critical functions such as disseminating scientific knowledge, improving public scientific literacy, and ensuring equitable access to information—all of which enhance the accessibility and understandability of scientific research(Gnoli, 2022; Trant, 2009). The exploration and synthesis of library engagement strategies in the scientific research impact process hold significant practical value for other knowledge brokers with similar functions, such as knowledge dissemination.
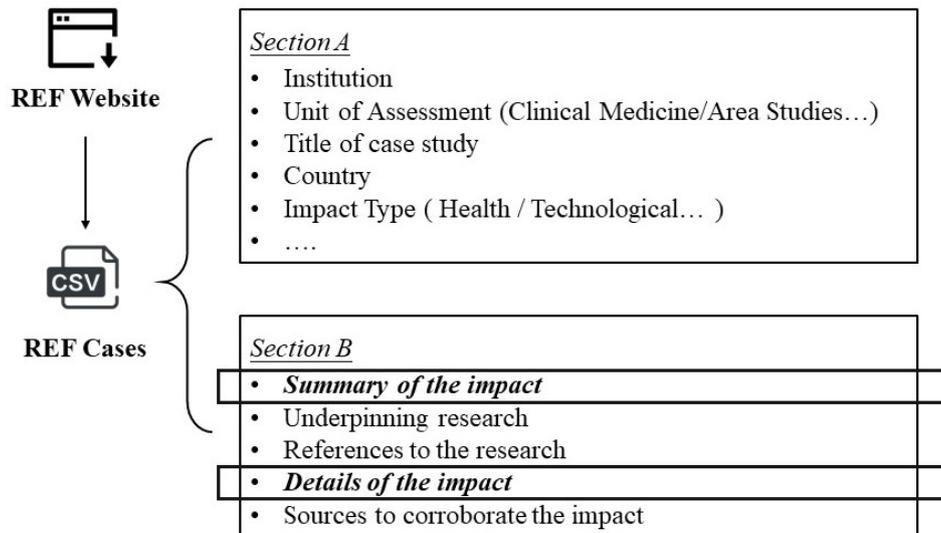
Data

Research impact case data

The Higher Education Funding Council for England (HEFCE) organises expert evaluations of universities' scientific research case studies every seven years to assess their impact, making it a long-established and well-developed research impact assessment initiative.

All research case data used in this study were obtained through an API interface from UK Research and Innovation (UKRI), including 6,637 REF2014 cases and 6,361 REF2021 cases. Each case consists of two main sections: Section A and Section B (as shown in Figure 2).

In addition to fully retaining Section A, which contains structured information related to the case, this study primarily focuses on two components in Section B—the **'*Summary of the Impact (SI)*'** and the **'*Details of the Impact (DI)*'**. **SI** should briefly state what specific impact is being described in the case study, where **DI** should provide a clear explanation of the process or means through which the research led to, underpinned or made a contribution to the impact(UKRI, 2020). for example, how it was disseminated, how it came to influence users or beneficiaries, or how it came to be exploited, taken up or applied. Both **SI** and **DI** require the submitting organisation to provide a written description, which serves as a crucial data source for the analysis in this study.

Figure 2. Structure of research case data content.

List of libraries

The library list data in this study are primarily sourced from the professional domain website Library Technology(Breeding, 1999), which maintains a global directory of libraries, with each listed library having its own dedicated webpage (as shown in Figure 3).

Using web scraping techniques, we systematically collected library-related information from the site, including:

- *Library name*: The official name of the library

- *Address*: The library's current geographic location

- *Library type*: The classification or category of the library

- **URL**: The link to the library's information webpage

- *Update time*: The creation and most recent update time of the webpage information

This dataset provides the foundation for analysing the characteristics and participation of libraries as knowledge intermediary institutions in the context of scientific research impact.
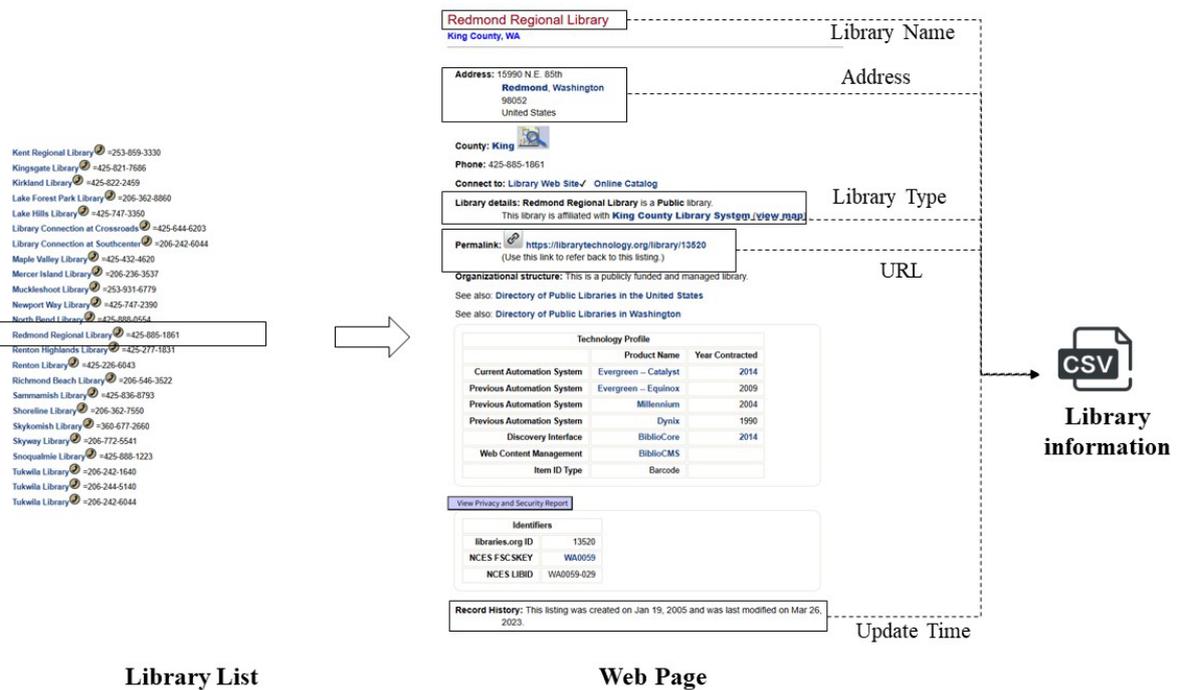
Figure 3. Composition of library information.

Research design

In real-world contexts, the process of translating scientific research into societal impact depends on the participation and interaction of multiple stakeholders (Bonaccorsi et al., 2021; Boshoff & Sefatsa, 2019; de Jong et al., 2014; SIAMPI, 2012), forming a complex network of collaborations. The Impact Case Study submissions to REF, particularly the **SI** and **DI** sections, document this network comprehensively. To accurately identify the specific contributions of libraries while preserving relevant contextual information, this study employs a hybrid methodology combining LLMs and Bertopic to extract and summarise library participation strategies. The procedure consists of two main phases: strategy extraction and strategy summarisation (as shown in Figure 4).

Strategy extraction

The official names of libraries were exactly matched against the **SI** and **DI** sections of REF cases, which describe the pathways through which research achieves impact. To maintain contextual integrity, the entire SI or DI section containing a reference to a library was retained as the corpus for subsequent analysis.

five manually annotated cases were used to refine the prompt design, ensuring alignment with the research objectives. Using the optimised prompt, all corpus texts were input into GPT-4o for strategy extraction. Each output record includes three elements:

·   **Strategy:** a concise description of the task or activity performed by the library, as identified by the model.

·   **Rationale**: an explanation provided by the model justifying why the activity was extracted as a strategy, serving as a reference for subsequent manual validation;

·   **Evidence_snippet**: the exact text segment from the source that corresponds to the described strategy.

Strategy summary

To mitigate risks of data contamination or artificial inaccuracies inherent in LLM extraction, a minimum of 10% of the outputs—corresponding to 50 cases—were randomly selected for manual verification. Two independent researchers evaluated the extractions for completeness and accuracy.

To balance contextual richness with thematic coherence, BERTopic modeling was applied to both the strategy and evidence snippet fields from the LLM outputs. Differential weights (w_s for strategy and w_e for evidence_snippet) were assigned during clustering to prioritise strategic intent. This approach yielded a structured overview of the targeted strategies deployed by libraries within scientific communication ecosystems.
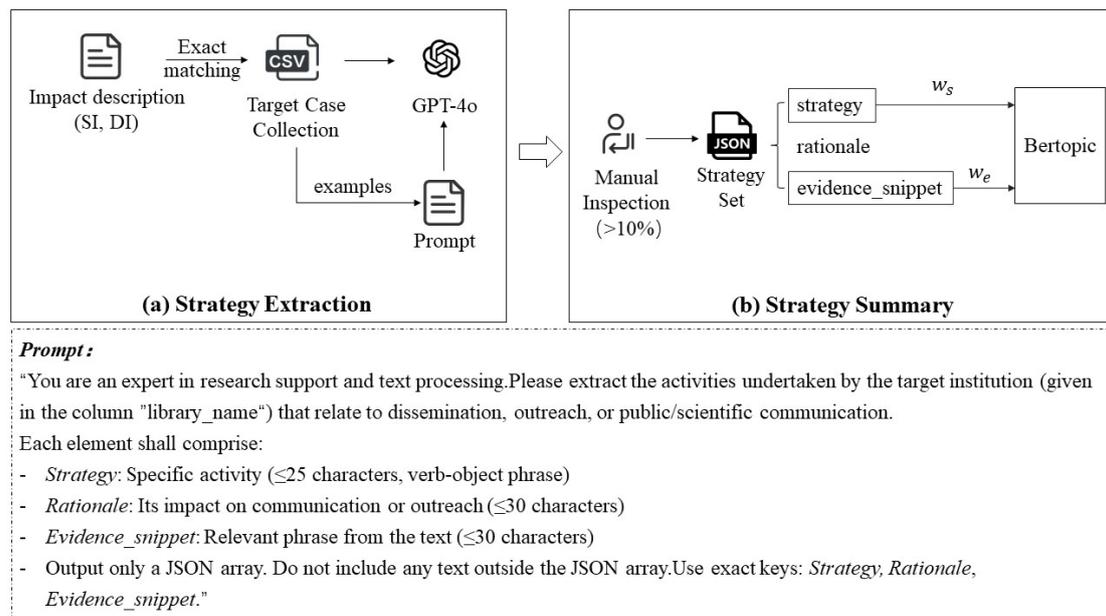

Prompt:

"You are an expert in research support and text processing. Please extract the activities undertaken by the target institution (given in the column "library_name") that relate to dissemination, outreach, or public/scientific communication.

Each element shall comprise:

- *Strategy*: Specific activity (≤25 characters, verb-object phrase)
- *Rationale*: Its impact on communication or outreach (≤30 characters)
- *Evidence_snippet*: Relevant phrase from the text (≤30 characters)
- Output only a JSON array. Do not include any text outside the JSON array. Use exact keys: *Strategy, Rationale, Evidence_snippet*."
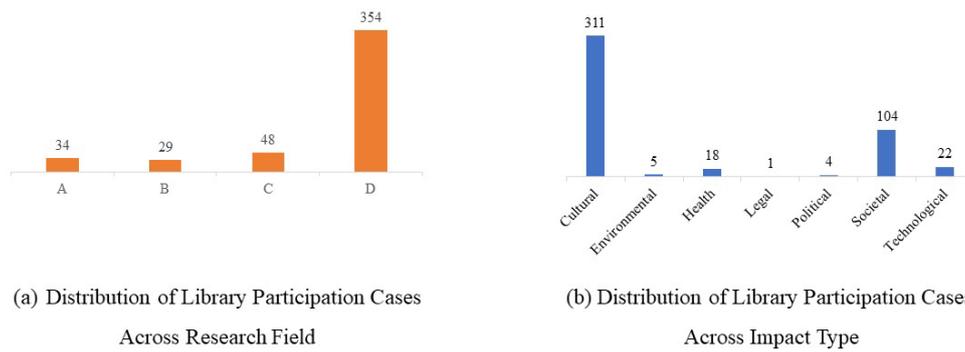
Figure 4. Research workflow diagram.

Results

Descriptive statistics

This study identified a total of 465 impact cases from the 2014 and 2021 REF datasets that explicitly mentioned library involvement. An analysis from the perspective of knowledge diffusion shows that while libraries participate in the research impact process across a wide range of fields and impact types, distinct concentration trends are evident.

In terms of the originating research fields (i.e., the starting point of knowledge production), library involvement spans various disciplines but is predominantly concentrated within **Main Panel D (Arts and Humanities)**, which accounts for 76% of all relevant cases.

Regarding the types of impact achieved (i.e., the endpoint of knowledge application), cases involving libraries are primarily associated with **cultural** impacts (67%) and **societal** impacts (22%).

These findings suggest that, despite their broad engagement across the research ecosystem, libraries—as knowledge brokers—exhibit clear preferences in both disciplinary alignment and impact orientation, indicating targeted and context-specific involvement in the research impact process (as shown in Figure 5).

(a) Distribution of Library Participation Cases Across Research Field

(b) Distribution of Library Participation Cases Across Impact Type

Figure 5. Distribution of library involvement cases across research field and impact type.

Note: REF classifies research fields into four main panels：(A)Medicine, Health and Life Sciences; (B)Physical Sciences, Engineering and Mathematics; (C)Social Sciences; (D)Arts and Humanities.

Summary of strategies

During the strategy extraction phase, this study first optimised the prompt instructions using a small set of cases and then applied these prompts to guide LLM in extracting library participation strategies from the target text corpus. To evaluate the extraction performance, two researchers randomly selected 50 cases—exceeding 10% of the extracted dataset—for manual inspection and calculated precision and recall.

Given that a single case may contain multiple strategies, the evaluation was conducted at the case level rather than the individual strategy level. Specifically:

· **Recall(R)** : A case is assigned a value of 1 if all strategies present in the source text were successfully extracted, and 0 otherwise. Recall is calculated as:

$$R = \frac{\text{Number of cases with all strategies correctly extracted}}{\text{Total Number of mannually evaluated cases}}$$

· **Precision(P):** A case is assigned a value of 1 if **all extracted strategies are correctly attributable to the library** and no strategies from other institutions or fabricated by the model are included; otherwise, it is assigned 0. Precision is calculated as:

$$P = \frac{\text{Number of cases with all extracted strategies correctly attributable to the library}}{\text{Total Number of mannually evaluated cases}}$$

The manual evaluation results showed that both precision and recall exceeded 75% (precision = 84%, recall = 76%), indicating that the extraction process was sufficiently accurate and comprehensive to reflect the actual participation strategies of libraries in the research impact process. These results support proceeding to the exploratory topic modelling phase.

In the strategy summary phase, this study first employed the **Silhouette Score** to evaluate the appropriateness of different weight combinations assigned to the strategy descriptions and evidence snippets during clustering. The overall Silhouette Score for a clustering configuration is computed as the mean of $S(i)$ over all points.

For each combination, the quality of the resulting topic clusters was assessed by calculating the mean Silhouette Score across all data points. the Silhouette Score $S(i)$ for a single data point i is defined as:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where:

- $a(i)$ denotes the average distance between point i and all other points within the same cluster,

- $b(i)$ represents the minimum average distance from point i to points in any other cluster.

After evaluating multiple weight combinations, the configuration $(w_s, w_e) = (0.8, 0.2)$ was selected, as it yielded the highest Silhouette Score. This weighting scheme emphasises the semantic content of the strategy descriptions ($w_s = 0.8$) while incorporating contextual support from the evidence snippets($w_s = 0.2$). Using this optimal weighting, 16 initial thematic clusters were identified from the extracted data.

Building on the semantic distribution of the 16 initial topics (as shown in Figure 6), a second round of manual clustering was conducted to merge thematically adjacent topics. This process yielded five core strategies that capture the principal ways in which libraries participate in the generation of scientific research impact:

(1) Media outreach and public engagement (Topics 4, 9, 13)
Libraries leverage both traditional media (e.g., radio programming) and digital platforms (e.g., social media) to produce and disseminate research-related content. This strategy aims to expand public reach, share scientific narratives, and foster interactive engagement with diverse audiences at national and community levels.

(2) Public dialogue and cultural presentation (Topics 7, 11)
This approach involves organising dialogue-based activities—including public lectures, panel discussions, and forums—along with curating cultural exhibitions featuring books, music, and other multimedia content. These initiatives facilitate knowledge exchange, promote cultural awareness, and support accessible informal education.

**(3) Artistic collaboration and live experiences (Topics 6, 12)**
Libraries collaborate with artists and community partners to develop creative projects and deliver immersive experiences such as live performances, interactive exhibitions, and touring activities. These efforts aim to deepen public connection with research through participatory and sensory-rich encounters.

(4) Digital content creation and knowledge dissemination (Topics 1, 2, 10, 14)
This strategy focuses on developing, curating, and providing sustained access to digital knowledge resources. Libraries create open educational materials, research publications, digital archives, and online learning content to support self-directed learning and further discovery.

(5) Major events and public programming (Topics 0, 3, 5, 8, 15)
Libraries coordinate and host large-scale public events—including international conferences, professional workshops, virtual and physical exhibitions, and invited speaker series—designed to engage domestic and global audiences while enhancing institutional visibility and research impact.
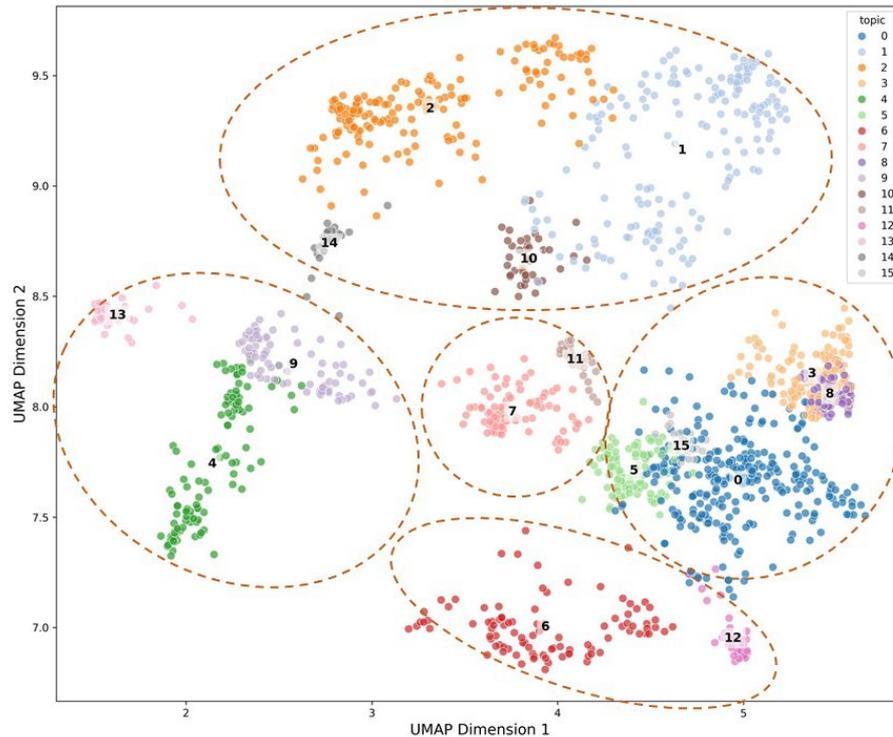
Figure 6. Distribution of library participation strategies in semantic space (Circles indicate the results of secondary manual clustering).

Conclusion and discussion

Main findings

This study identified 465 real-world cases of library engagement through precise matching of library entities within publicly available research impact databases. Analysis shows that while libraries participate across various research fields and contribute to multiple impact dimensions, their involvement is particularly concentrated in the *arts and humanities* domain, with *cultural and social impact* being the most prominent outcome.

Using an integrated LLM and BERTopic approach, this study extracted and synthesised textual descriptions of library contributions, leading to the identification of five key strategies through which libraries facilitate the translation of scientific research into societal benefits:(1)Media Outreach and Public Engagement; (2) Public Dialogue and Cultural Presentation; (3)Artistic Collaboration and Live Experiences; (4) Digital Content Creation and Knowledge Dissemination; (5) Major Events and Public Programming.

By systematically summarising practical experiences, this study provides a valuable reference for libraries seeking to enhance their role in fostering diverse societal impacts from scientific research. These strategies help channel scientific knowledge to different types of knowledge consumers, thereby contributing to the enhancement of public cultural literacy.

Future research directions

As representative knowledge intermediary institutions, libraries play a crucial role in translating scientific research into broader societal impact. This study used semantic analysis to identify specific participation strategies adopted by UK libraries within this process.

However, due to word-length limitations in the REF case data, the corpus available for analysing strategies across different types of libraries was relatively constrained. Future research could be extended or deepened in the following directions. First, building on the disciplinary fields and impact dimensions in which libraries are most actively involved—as identified through the statistical analysis in this study—future work could conduct more in-depth examinations of specific libraries by integrating multi-source data, such as publicly available online materials or offline interviews. Second, the generalisability of the findings could be examined from a regionally differentiated perspective, expanding case data based on geographical distribution to enhance the practical applicability of the conclusions.

Acknowledgements

About the authors

Wang Zuorong is the PhD student in the School of Information Management, Nanjing University, Jiangsu, China. Her primary research interests focus on the organisation and processing of academic evaluation information, as well as the practices and roles of multiple actors within the domain of science communication. She can be contacted at wangzr@smail.nju.edu.cn

Sun Jiaxuan is the PhD student in the School of Information Management, Nanjing University, Jiangsu, China. Her primary research interests focus on the digital humanities, as well as the design and design management. She can be contacted at stella.sjx@gmail.com

He Shan is a postgraduate student at the School of Information Management, Nanjing University, Jiangsu, China. Her current research interests focus on patent complex networks, with an emphasis on exploring issues related to patent recommendation in the context of cooperation and competition. She can be contacted at 502024140006@smail.nju.edu.cn

Deng Sanhong is professor in the School of Information Management, Nanjing University, Jiangsu, China. He received his PhD from Nanjing University, and his interests primarily encompass intelligent information processing and retrieval, information analysis, and scientific evaluation. He can be contacted at sanhong@nju.edu.cn

Wang Hao is professor in the School of Information Management, Nanjing University, Jiangsu, China. He received his PhD from Nanjing University, and his research interests primarily encompass knowledge organisation and digital humanities. He can be contacted at ywhaowang@nju.edu.cn

Reference

Bonaccorsi, A., Chiarello, F., & Fantoni, G. (2021). Impact for whom? Mapping the users of public research with lexicon-based text mining [Article]. Scientometrics, 126(2), 1745-1774. https://doi.org/10.1007/s11192-020-03803-z

Boshoff, N., & Sefatsa, M. (2019). Creating research impact through the productive interactions of an individual: An example from South African research on maritime piracy [Article]. Research Evaluation, 28(2), 145-157. https://doi.org/10.1093/reseval/rvz001

Bracci, E., Tallaki, M., Ievoli, R., & Diplotti, S. (2022). Knowledge, diffusion, and interest in blockchain technology in SMEs [Article]. Journal of Knowledge Management, 26(5), 1386-1407. https://doi.org/10.1108/JKM-02-2021-0099

Breeding, M. (1999). Library Technology Guides. https://librarytechnology.org/

de Jong, S., Barker, K., Cox, D., Sveinsdottir, T., & Van den Besselaar, P. (2014). Understanding societal impact through productive interactions: ICT research as a case. Research Evaluation, 23(2), 89-102. https://doi.org/10.1093/reseval/rvu001

Gnoli, B. H. C. (2022). Libraries, archives, and museums (LAM) Conceptual issues with focus on their convergence. Retrieved 05-10 from https://www.isko.org/cyclo/lam

Hladik, S. (2023). Examining the Roles of Science Museum Facilitators in Academic Research. Journal of Museum Education, 48(2), 153-166. https://doi.org/10.1080/10598650.2022.2142749

Kousha, K., Stuart, E., Abdoli, M., & Thelwall, M. (2024). How do museums and galleries help academics create societal impact? An analysis of the UK REF2021 impact case studies [Article]. Scientometrics, 129(12), 7759-7782, Article e0156978. https://doi.org/10.1007/s11192-024-05180-3

Lomas, J. (2007). The in-between world of knowledge brokering [Review]. British Medical Journal, 334(7585), 129-132. https://doi.org/10.1136/bmj.39038.593380.ae

Mensah Danquah, M., Ato Yankson, A., Bajamchana Bugase, E., & Yirenkyi Nyarko, C. (2025). Librarians as Guiding Lights in the Digital Darkness: A Systematic Review of Libraries' Provision of Digital Literacy Services. Ghana Library Journal, 30(2), 4-17. https://doi.org/10.4314/glj.v30i2.3

Meyer, M. (2010). The Rise of the Knowledge Broker. Science Communication, 32(1), 118-127. https://doi.org/10.1177/1075547009359797

Polidoro, F., & Jacobs, C. (2024). Knowledge diffusion in nascent industries: Asymmetries between startups and established firms in spurring inventions by other firms [Article]. Strategic Management Journal, 45(4), 807-845. https://doi.org/10.1002/smj.3568

Ribeiro, R. G. P., Dias, T. M. R., Dias, P. M., Araújo, R. F., & Costa, E. S. (2024). Social4Science Uncovering Connections and Trends in the Diffusion of Scientific Knowledge on Social Networks [Article]. Ciencia da Informacao, 53(Special issue), 180-192. https://doi.org/10.18225/ci.inf.v53i.6822

Rosário, J., Pires, J. C., Dias, S., & Pedro, A. R. (2025). Exploring perceptions of health literacy, healthcare access, and utilisation among higher education students in Alentejo, Southern Portugal: A qualitative study. PLoS ONE, 20(6), e0326575. https://doi.org/10.1371/journal.pone.0326575

Saffarinia, M. (2025). Artificial Intelligence and Children's Human Rights. ModernTechnologies Law, 6(11), 97-113. https://doi.org/10.22133/mtlj.2024.436038.1290

Shin, S. R., Lee, J., Jung, Y. R., & Hwang, J. (2022). The diffusion of scientific discoveries in government laboratories: The role of patents filed by government scientists [Article]. Research Policy, 51(5), Article 104496. https://doi.org/10.1016/j.respol.2022.104496

SIAMPI. (2012). Case studies. http://siampi.eu/12/643.bGFuZz1FTkc.html

Thompson, G. N., Estabrooks, C. A., & Degner, L. F. (2006). Clarifying the concepts in knowledge transfer: A literature review. 53(6), 701. https://www.scopus.com/inward/record.uri?eid=2-s2.0-33644912338&doi=10.1111%2fj.1365-2648.2006.03775.x&partnerID=40&md5=7d9eea2ba94d4d1a7e395fa89c13f3ab

Trant, J. (2009). Emerging convergence? Thoughts on museums, archives, libraries, and professional training. 24(4), 387. https://www.scopus.com/inward/record.uri?eid=2-s2.0-71149111737&doi=10.1080%2f09647770903314738&partnerID=40&md5=2b2c1a34ac2a8c96ccb3eda75912ae01

Turnhout, E., Stuiver, M., Judith, J., Harms, B., & Leeuwis, C. (2013). New roles of science in society: Different repertoires of knowledge brokering. 40(3), 365. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84883143284&doi=10.1093%2fscipol%2fscs114&partnerID=40&md5=b315d997139c1e66bd147804e3112619

UKRI. (2020). Guidance on submissions (2019/01). Retrieved 11-27 from https://2021.ref.ac.uk/publications-and-reports/guidance-on-submissions-201901/index.html

UKRI. (2022). How Research England supports research excellence. Retrieved 03-31 from https://www.ukri.org/who-we-are/research-england/research-excellence/ref-impact/

Wang, S., Zhao, S., Fan, X., Zhang, B., & Shao, D. (2025). The impact of open innovation on innovation performance: the chain mediating effect of knowledge field activity and knowledge transfer [Article]. Information Technology and Management, 26(2), 139-161, Article 105767. https://doi.org/10.1007/s10799-024-00420-7

Zhang, X., Chen, S., & Wang, X. (2023). How can technology leverage university teaching & learning innovation? A longitudinal case study of diffusion of technology innovation from the knowledge creation perspective [Article]. Education and Information Technologies, 28(12), 15543-15569. https://doi.org/10.1007/s10639-023-11780-y

Appendix

Weight Combination (w_s, w_e)	Silhouette Score
(0.9, 0.1)	0.1653
(0.8, 0.2)	0.1788
(0.7, 0.3)	0.1629

Table 1. Clustering metrics Silhouette score for different weight combinations.

Note: w_s notes the weight assigned to the text of the strategy component, while w_e represents the weight assigned to the text of the evidence_snippet component.

Appendix

Topic ID	Keywords
0	exhibition, tour, host, museum, virtual, panel, public, major, digital, interactive
1	develop, educational, website, online, create, digital, launch, 20, creation, launch online
2	write, book, research, books, published, special, news, contributions, science, online
3	conduct, deliver, host, writing, creative, organise, school, conduct public, showcase, schools
4	radio, bbc, feature, media, program, participate, series, shows, programme, conduct
5	events, events host, event, events public, host, host public, public, organise public, launch, world
6	collaboration, artists, libraries, arts, plays, media, community, creative, support, museum
7	talks, discussions, public, panel, organise public, host public, host, schools, deliver, conduct public
8	public, deliver, host public, conduct public, host, conduct, series public, centre, hosting, invited
9	engagement, social, engage, media, events public, events, public, sharing, interactive, share
10	digital, access, accessible, launch, publicly, create, Britain, library, website, open
11	project, available, cultural, historical, health, display, post, music, books, making
12	live, shows, public, interactive, tour, appearances, share, promote, host, display
13	media, extensive, stories, national, reached, Wales, exhibition, conduct, news, wide
14	produce, series, production, create, content, panel, contributions, participate, participation, provide
15	conferences, conference, organise, hold, invitations, present, international, science, organise public, local

Table 2. Keywords of initial topics after first-round clustering (BERTopic).