https://doi.org/10.51408/issi2025_008

Web Mining the Online Presence of Global Scientific Academies

Xiaoli Chen[1], Xuezhao Wang[2]

[1]*chenxl@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences (China)

[2]*Wangxz@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences (China)
University of Chinese Academy of Sciences (China)

Abstract

Global scientific academies have been adapting their role in fostering scientific communication and promoting science since their inception in the 15th century. Despite their prominence, the institutional norms and identities of scientific academies remain underexplored. In the digital age, their websites reflect their evolving roles, organizational priorities, and the balance between conformity and innovation. This study examines how scientific academies structure their online identities through content organization and communication strategies.

This study employs web mining techniques to analyze large-scale academy website data. It uncovers structural patterns and behavioral trends in how scientific academies present themselves online. Formal Concept Analysis (FCA) is applied to develop a unified taxonomy, enabling systematic comparisons of digital strategies across multiple academies. Using institutional theory, this study uses quantitative method to examine how academies balance conformity with differentiation in their digital presence. The research addresses two core questions: (RQ1) What content and communication patterns are adopted by global scientific academies in their online presence? And (RQ2) How do scientific academies balance imitation and innovation in their digital strategies?

The findings identify distinct website content patterns, showing how academies balance tradition and adaptation in their digital presence. Hierarchical clustering reveals three strategic approaches: (1) highly innovative academies that introduce novel digital structures, (2) conservative academies that show fragmented or underdeveloped structures, and (3) hybrid academies that combine imitation with selective innovation. The study also highlights key thematic differences in content emphasis, such as governance, scientific cooperation, and public outreach. These insights contribute to institutional theory and scholarly communication studies, revealing how scientific academies use their online presence to maintain legitimacy, engage the public, and foster international collaboration.

This study highlights common features of scientific academies' online presence, including an emphasis on membership, strategic planning, and scholarly communication to reinforce institutional legitimacy. Additionally, academies adapt their digital strategies to facilitate scientific collaboration in response to evolving societal expectations. Innovative activities include increasing transparency on the academy's decisions, achievements, budget, yearbooks, and interactive digital engagement strategies. These activities enhance public trust in scientific academies and science itself while improving communication efficiency. These findings offer guidance for scholars, academy leaders, and policymakers seeking to optimize digital engagement strategies and strengthen global scientific networks in the digital era.

Introduction

Scientific academies have long served as the cornerstone of knowledge advancement and scholarly communication since their inception in the 15th century. As technology advances and global interconnectivity increases, scientific academies

rely on their digital presence to extend influence, disseminate research, and engage with a diverse audience worldwide. Despite the increasing prominence of digital communication, little research examines how they structure their online presence and institutional identity. Additionally, there is limited understanding of how these academies balance imitation—adopting common practices—and innovation—developing unique digital strategies—in their approach to web-based communication.

Research on institutional digital presence has largely focused on universities (Lepori et al., 2014; Will & Callison, 2006), governmental organizations (Neumann et al., 2022), and research institutions (Burford, 2014; Elsayed, 2017), leaving scientific academies underexplored. Web mining has been applied to map innovation ecosystems (Kinne & Axenbeck, 2020), predict firm-level innovation (Axenbeck & Breithaupt, 2021; Kinne & Lenz, 2021), and analyze the accessibility of digital platforms (Singh et al., 2024; Alim, 2021). However, little research has specifically addressed the digital strategies of scientific academies. Unlike universities or firms, scientific academies operate at the intersection of academic prestige, policy influence, and public engagement, making their digital behavior distinct. This study utilizes prior web mining methodologies by analyzing how academies structure digital content, offering a comparative framework to assess the balance between imitation and innovation in the academies digital strategy.

This study is grounded in institutional theory, which provides a framework for understanding how scientific academies navigate community expectations and the tension between conformity and differentiation. Institutional theory explains how organizations conform to external expectations through institutional isomorphism. This process includes coercive pressures (regulatory and funding requirements), mimetic pressures (emulating successful peers), and normative pressures (adhering to professional standards and societal expectations). This theory framework provides an explanation on how scientific academies structure their online presence, influencing whether they conform to widely accepted digital taxonomies, adopt innovative approaches to distinguish themselves, or balance both strategies to maintain legitimacy while adapting to evolving scientific and societal demands. The web-based content strategies reflect their efforts to adhere to professional norms, align with stakeholder expectations, and assert their role as authoritative scientific institutions. At the same time, they face the challenge of distinguishing themselves through novel digital practices. This study uses quantitative method builds on institutional theory to analyze how scientific academies balance conformity and differentiation in their digital strategies.

To investigate these dynamics, this study utilizes web mining techniques combined with Formal Concept Analysis (FCA) to analyze the online presence of scientific academies. The hierarchical relations of web content are harvested by web mining techniques. FCA is employed to construct a unified taxonomy from the extracted hypernym-hyponym pairs. The unified taxonomy identifies patterns in content structure and content organization across these academies' websites. By quantitatively comparing these patterns, this study aims to uncover how academies engage with stakeholders, promote collaboration, and contribute to scientific

discourse on a global scale. The research is guided by two primary research questions: (RQ1) What content and communication patterns are adopted by global scientific academies in their online presence? And (RQ2) How do scientific academies balance imitation and innovation in their digital strategies?

This research makes several key contributions. First, it introduces a novel application of web mining and FCA to analyze how scientific academies structure their web content, providing a scalable and systematic method for web content taxonomy construction. Second, it advances institutional theory by exploring how academies cope with mimetic and normative pressures in shaping their digital strategies, which is reflected by their balance between imitation and innovation. Third, it provides practical insights for scholars, institutional leaders, and policymakers seeking to optimize digital engagement strategies. Understanding how academies structure their online presence can inform the development of more effective digital communication frameworks, enhance public engagement, and strengthen global scientific networks.

Related Work

The study of institutional digital strategies has gained increasing significance as institutions leverage digital platforms for communication, collaboration, and knowledge dissemination. While universities, government agencies, and firms have been extensively studied, scientific academies remain an overlooked category despite their critical role in shaping global scientific discourse. This research builds upon prior studies in institutional digital identity, web mining, and content taxonomy to assess how scientific academies structure their online presence.

Prior research has explored how institutions use digital platforms to shape institutional identity. Research has shown that institutional priorities shape online strategies across different organizations, including universities (Lepori et al., 2014; Will & Callison, 2006), government agencies (Neumann et al., 2022), and research institutions (Burford, 2014; Elsayed, 2017). Comparative studies on scientific academies (Isavand & Poormoghim, 2024) have examined regional differences but lack a global perspective on digital engagement strategies.

Studies in content organization and web architectures further demonstrate how institutions adapt their online presence to align with strategic goals (Campos et al., 2019; Karanasios et al., 2013). However, these studies primarily focus on universities and corporate entities, leaving a gap in understanding how scientific academies balance tradition and digital transformation.

Web mining has been widely applied in analyzing Organizational Structures and innovation behaviors. Researchers have used web data to map innovation ecosystems (Kinne & Axenbeck, 2020; Kinne & Lenz, 2021), predict firm-level digital strategies (Axenbeck & Breithaupt, 2021), and classify academic webpages (Kenekayoro et al., 2014, 2015). Historical web archives (Schroeder et al., 2020; Tsakalidis et al., 2021) provide insights into the evolution of institutional priorities, demonstrating how digital structures shift over time. Despite these advancements, scientific academies remain largely absent from web mining research, even though they play a crucial role in balancing scientific legitimacy, policy influence, and public

engagement. Prior methodologies have not been applied to systematically analyze how these institutions construct their digital presence.

The tension between institutional imitation and innovation is central to understanding how institutions adopt digital strategies. Institutional theory identifies coercive (regulatory), mimetic (peer-driven), and normative (professional) pressures as key factors shaping institutional behavior in digital spaces (Engelbrecht et al., 2020, 2022; Cox, 2007, 2008). Research on higher education institutions (Lepori et al., 2014) and corporate strategies (Gök et al., 2015; Thelwall, 2006) suggests that institutions often emulate established digital norms while attempting to differentiate themselves.

However, scientific academies face a unique challenge: upholding scientific authority and global credibility while adapting to national policy environments. Unlike universities, which primarily engage academic audiences, academies must also address policymakers, funding agencies, and the public. Prior research has not systematically examined how scientific academies navigate these competing demands in digital spaces. While previous studies have applied web mining, content classification, and institutional theory to universities, firms, and government agencies, no study has systematically examined the digital presence of scientific academies on a global scale. Unlike commercial enterprises, which optimize digital strategies for competitive advantage, scientific academies must balance scientific prestige, national policies, and public engagement. Furthermore, while studies on content classification and historical web evolution (Campos et al., 2019; Tsakalidis et al., 2021) provide foundational insights, they do not assess how scientific academies' digital strategies reflect their institutional missions.

This study builds on these research strands by integrating insights from web mining, institutional behavior, and content taxonomy to examine the digital presence of global scientific academies. This study addresses the current research gap by applying web mining and Formal Concept Analysis (FCA) to systematically examine how scientific academies structure their digital presence. A comparative framework is proposed for assessing how academies adapt to scientific norms and policy expectations in their online representations. By integrating insights from institutional theory, web mining, and web content taxonomies, this research advances our understanding of how scientific academies construct and maintain legitimacy in the digital age.

Methodology

This study applies web mining techniques and Formal Concept Analysis (FCA) to analyze systematically the digital presence of global scientific academies. It also explores institutional digital strategies to improve theoretical understanding in this area. This methodology addresses research questions by identifying patterns of imitation and innovation in the digital communication strategies of scientific academies. The methodology consists of five distinct phases, as illustrated in Figure 1: (1) Website Data Harvesting. Extracting structured information from global scientific academy websites. (2) Hierarchical Concept Development. Identifying hypernym-hyponym relationships to model content structures. (3) Category

Development and Category Mapping. Grouping content into meaningful categories using LLMs and word embeddings. (4) Taxonomy Construction. Applying FCA and graph pruning to refine hierarchical structures. (5) Comparative Analysis. Evaluating the thematic and structural commonalities and differences among academies.

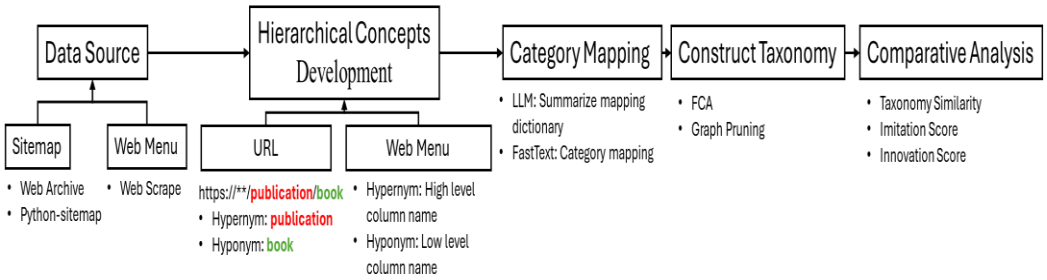

Figure 1. Framework for Web Mining and Comparative Analysis of Scientific Academies.

Website Data Harvesting

The website sitemaps provide a comprehensive structural blueprint of each academy's web presence, listing URLs that encapsulate both structural and content-related aspects. However, official sitemaps are sometimes incomplete or unavailable. The primary source of data is the sitemaps of scientific academies that are retrieved from the website archival platform[1]. As a complementary approach, automated sitemap generators like *python-sitemap*[2] are used to reclaim web pages that may be missing. However, both methods may encounter challenges due to scraping restrictions and web connection issues. To mitigate these limitations, this study uses navigation menus of scientific academies' websites to supplement data collection. Compared to sitemaps, website navigation menus provide another perspective on content organization. These menus typically highlight the key focus of institutional priorities and mission. However, this method has limitations—some websites lack a well-structured navigation website menu or offer shallow categorizations. This study combines the three data sources, and a manual check by random browsing of the website is also conducted to verify the key aspects of the website's columns are included in the data collection.

Hierarchical Concept Development

This study analyzes the hierarchical structure of website content from global scientific academies. It extracts hypernym-hyponym relationships from menu items and webpage URLs. Each URL is stripped of its domain and segmented hierarchically using the forward-slash (/) delimiter, as illustrated in Figure 1. The resulting hierarchical dictionary preserves the hypernym-hyponym relationships within the website's navigation content and webpage URLs, with higher-level menu

[1] https://web.archive.org/
[2] https://github.com/c4software/python-sitemap

items or URL relative paths (hypernyms) containing more specific subcategories (hyponyms).

To ensure cross-institutional consistency of scientific academies, this study utilizes the DeepSeek[3] Large Language Model (LLM) for language translation, normalizing terminologies across different linguistic contexts. Given that the hypernym and hyponym pairs extracted from the URL path often include acronyms and numbers, this study WordNet to retain only semantically meaningful terms. This process refines the extracted relationships and improves taxonomy accuracy.

> *LLM Instructions:*
> *You are an expert in hierarchical content classification and taxonomy development. Your task is to refine a set of extracted hypernym-hyponym pairs by identifying meaningful concepts and filtering out irrelevant terms.*
> *1. Input Format: You will receive a list of hypernym-hyponym pairs extracted from website structures.*
> *2. Objective: Identify core concepts by:*
> *- Grouping similar hyponyms under a meaningful hypernym.*
> *- Removing noisy terms, such as acronyms, numbers, and ambiguous words.*
> *- Ensuring logical consistency in hierarchical relationships.*
> *3. Output Format:*
> *- A structured JSON object where each hypernym maps to refined hyponyms.*

Categories Development and Category Mapping

After cleansing the extracted hypernym-hyponym pairs, this study establishes core concepts that form the foundation of the taxonomy. This process involves Identifying and summarizing content patterns using an LLM pipeline. These pattern words filter out irrelevant hypernyms and hyponyms to enhance dataset clarity.

For computational efficiency, the FastText model is used to compute the average word embeddings of hypernym and hyponym terms. Cosine similarity scores of these embedding vectors are mapped into the nearest normalized category embedding. To maintain classification integrity, manual verification is conducted, resolving inconsistencies and improving accuracy.

Developing Website Content Taxonomy

This study applies Formal Concept Analysis (FCA) to structure and refine hierarchical web content. FCA constructs a concept lattice, while graph pruning enhances consistency, reduces redundancy, and optimizes hierarchical relationships. Formal Concept Analysis is a well-established method for knowledge organization. It is particularly suited for this task as it enables the construction of a concept lattice,

[3] https://www.deepseek.com/

effectively capturing relationships between categories while preserving the hierarchical nature of web structures. In this study, FCA is applied to generate a formal taxonomy of web content from scientific academy websites, facilitating comparative analysis.

The formal context is represented as a binary relation $K = (G, M, I)$, where:

- Objects (G): $g_i \in G$ denotes hyponyms (specific subcategories in the taxonomy).
- Attributes(M): $m_j \in M$ denotes hypernyms (general categories representing broader concepts).
- Incidence Relation (I): A binary relation $I \subseteq G \times M$ indicating which objects belong to which attributes. The relation is represented as a binary matrix B, where:

$$B_{ij} = \begin{cases} 1, & \text{if object } g_i \text{ is associated with attribute } m_i \\ 0, & otherwise \end{cases}$$

Using this matrix representation, the attribute derivation operator A' and the object derivation operator B' could be derived:

$$A' = \{m \in M \mid \forall g \in A, (g, m) \in I\}$$
$$B' = \{m \in M \mid \forall g \in B, (g, m) \in I\}$$

Here A' is the set of all attributes shared by objects in A. B' is the set of all objects sharing the attributes in B. A formal concept is a pair (A, B), where:

$$A = B' \text{ and } B = A'$$

Here A is the extent, which means the set of all objects (hyponyms) belonging to concept B. B is the intent, which means the set of all attributes (hypernyms) that describe all objects in A. A concept lattice $L(K)$ is formed by structuring these concepts into a partially ordered set:

$$(A_1, B_1) \leq (A_2, B_2)\ if\ A_1 \subseteq A_2\ (or\ equivalently\ B_2 \subseteq B_1)$$

This implies that more general concepts are ranked higher in the lattice, while specific concepts appear lower. This is implemented by using Meet ($\wedge$) operation and Join ($\vee$). Meet ($\wedge$) operation computes the greatest lower bound of two concepts, used to identify hyponym terms:

$$(A_1, B_1) \wedge (A_2, B_2) = (A_1 \cap A_2, (A_1 \cap A_2)')$$

Join ($\vee$) operation computes the greatest least upper bound of two concepts, used to identify hypernym terms:

$$(A_1, B_1) \vee (A_2, B_2) = (B_1 \cap B_2, (B_1 \cap B_2)')$$

A key challenge of FCA is multi-parent assignments, where a single hyponym is linked to multiple hypernyms, potentially creating ambiguous or cyclic relationships. Due to the diverse structures of academy websites, the extracted categories often exhibit inconsistent terminology, redundancies, and overlapping concepts. To further

refine the extracted hierarchical taxonomy, this study applies a graph pruning method to ensure hierarchical consistency, eliminate conflicts, and resolve structural inconsistencies.

To effectively resolve cyclic dependencies in hypernym-hyponym pairs, conflict cycles were detected using depth-first search (DFS). Manual evaluation was then conducted to eliminate incorrect hypernym-hyponym relationships while retaining only the most contextually appropriate ones. Multi-parent issues were addressed using a similar manual resolution process. For example, if the term "*Funding*" appeared as a subcategory under both "*Governance*" and "*Supporting Science*", the pruning process ensured its placement under "*Supporting Science*", where it aligns with funding mechanisms for scientific projects rather than administrative governance. Additionally, cyclic dependencies—such as a category incorrectly appearing as both a parent and a child (e.g., "Awards" categorized under both " *Supporting Science*" and "*Knowledge Resources*")—were detected using depth-first search (DFS) and manually resolved to preserve logical consistency in the taxonomy. The iterative manual review ensured that meaningful hierarchical relationships were maintained, preventing redundancy and ambiguity. To validate the accuracy of the final taxonomy, a manual review is conducted for a subset of academy websites, ensuring that the taxonomy aligns with real-world institutional practices.

Comparative Analysis of Global Scientific Academies

This study leverages the constructed taxonomy to examining thematic and structural differences in their digital presence. One aspect of comparison is assessing the overall scale of the websites, including the number of pages they contain, to gauge their digital footprint. Levels of URL paths are analyzed to understand how deep the content structure is, which reflects the complexity and organization of the websites. Analyzing the balance between imitation and innovation in website structures is crucial for understanding how scientific academies establish their digital presence. This study develops a methodology based on a combination of similarity analysis and unique content evaluation to quantify the extent to which websites adopt existing taxonomies, imitate peer's digital practice and introduce novel structures. Each website's hypernym-hyponym pairs were enriched by identifying and incorporating missing parent nodes from the common taxonomy to ensure structural completeness. For each site *s*, similarity to common taxonomy is assessed how closely each website adhered to the common taxonomy by computing its cosine similarity to the taxonomy. It is a balance of how many of the site's hypernym-hyponym pairs are present in the common taxonomy (Precision(s)) and how much of the taxonomy is covered by the site (Recall(s)). For site *s,* where $P_s = \{(h_k, h'_k) \mid h_k$ is a hypernym of $h'_k\}$ is the set of hypernym-hyponym pairs of site *s*. $P_{taxonomy}$ is the set of hypernym-hyponym pairs of a common taxonomy. The similarity analysis is conducted to inspect each website's similarity with the common taxonomy by performing the following method:

$$Taxonomy\ Similarity(s) = \frac{2 \times Precision(s) \times Recall(s)}{Precision(s) + Recall(s)}$$

where $Precision(s) = \frac{/P_s \cap P_{\text{taxonomy}}/}{/P_s/}$ and $Recall(s) = \frac{/P_s \cap P_{\text{taxonomy}}/}{/P_{\text{taxonomy}}/}$.

To quantify the conformity between websites, this study introduces the Imitation Score based on the average similarity to other websites. Each site *s* is represented as a binary vector v_s of length *d*, where d is the total number of unique hypernym-hyponym pairs across all websites. These pairs align with the taxonomy structure. Each entry in v_s is 1 if the corresponding hypernym-hyponym pair appears in the website, and 0 otherwise. Cosine similarity between two websites s_i and s_j is

$$cosine\_sim(s_i, s_j) = \frac{v_{s_i} \cdot v_{s_j}}{|v_{s_i}||v_{s_j}|}$$

The imitation score for website *s* is its average cosine similarity with all other websites

$$Imitation\ Score(s_i) = \frac{1}{N-1} \sum_{s' \in S, s' \neq s} cosine\_sim(s, s')$$

where *N* is the total number of websites.

To measure the uniqueness of a website's structure, this study computes an Innovation Score by comparing its hypernym-hyponym pairs with those of other websites. The Innovation Score for a website s is defined as the average number of unique hypernym-hyponym pairs it has compared to all other websites. Each website s_i is represented as a set of hypernym-hyponym pairs P_{s_i}. The uniqueness of s_i is determined by counting the number of pairs that do not exist in any other website s_j, where $j \neq i$.

$$Innovation\ Score(s_i) = \frac{1}{N-1} \sum_{s' \in S, s' \neq s} \left| P_{s_i} \setminus P_{s_j} \right|$$

Where $P_{s_i} = \{(h_k, h'_k) \mid h_k \text{ is a hypernym of } h'_k\}$ is a set of hypernym-hyponym pairs of website s_i. $P_{s_i} \setminus P_{s_j}$ denotes the set difference, capturing pairs that exist in s_i but not in s_j. The summation iterates over all other websites s_j, averaging the unique pairs. To ensure comparability between the Imitation Scores and the Innovation Scores, this study applies Min-Max Scaling for both the Imitation Scores and the Innovation Scores.

To further explore how academy websites differentiation, this study introduces a Distinctiveness Score to identify the most unique hypernym-hyponym pairs in each cluster. Given a set of clusters $C = \{C_1, C_2, \cdots, C_m\}$, each website s_i is assigned to a cluster c_j through hierarchical clustering:

$$f: S \to C, \quad f(s_i) = c_j$$

For each cluster c_j, this study aggregates all pairs from its member websites

$$P_{c_j} = \bigcup_{s_i \in c_j} P_{s_i}$$

The cluster-level frequency of a pair *(h, h')* is computed as

$$\text{count}_{c_j}(h, h') = \sum_{s_i \in c_j} 1\left((h, h') \in P_{s_i}\right)$$

The global frequency of a pair across all websites is:

$$\text{global\_count}(h, h') = \sum_{s_i \in S} 1\left((h, h') \in P_{s_i}\right)$$

The relative frequency of a pair *(h, h')* in cluster c_j is given by

$$\text{relative\_freq}_{c_j}(h, h') = \frac{\text{count}_{c_j}(h, h')}{\sum_{(h,h') \in P_{c_j}} \text{count}_{c_j}(h, h')}$$

The global probability of a pair appearing in the entire dataset is

$$P(h, h') = \frac{\text{global\_count}(h, h')}{\sum_{(h,h') \in P} \text{global\_count}(h, h')}$$

The distinctiveness score of a pair *(h, h')* in cluster c_j is

$$\text{distinctiveness}_{c_j}(h, h') = \frac{\text{relative\_freq}_{c_j}(h, h')}{P(h, h')}$$

Statistical methods were used to validate the effectiveness of the cluster partition by distinguishing the imitation score and innovation score. Statistical methods were also applied to test if the identified the most distinctive hypernym-hyponyms and the least distinctive hypernym-hyponyms are significant in different types of scientific academies.

This study integrates Formal Concept Analysis (FCA), graph pruning, and manual verification to construct a reliable and accurate taxonomy, serving as the knowledge backbone for understanding the website content of scientific academies. To quantitatively assess digital strategies, Taxonomy Similarity, Imitation Score, and Innovation Score were developed to measure the extent to which academies adopt common practices, conform to established norms, and differentiate their digital presence. Additionally, the Distinctiveness Score was introduced to identify both the most unique and the most standardized content, providing insights into the balance between conformity and differentiation in the digital strategies of scientific academies.

Result and Analysis

The results of this study are organized into three main sections. The first section, Data Description, provides an overview of the dataset, detailing the structural and institutional patterns of scientific academy websites. The second section, Taxonomy of Scientific Academies' Web Content Organization, presents the taxonomy derived from Formal Concept Analysis (FCA) and graph pruning, demonstrating how these academies define their digital identities. The final section, Comparative Analysis of Digital Presence Across Scientific Academies, explores the balance between

imitation and innovation, revealing how different academies strategically position themselves within the global scientific community.

Data Description

This study utilizes the dataset of global scientific academies (Chen, 2024), focusing on a subset of 112 national scientific academies dedicated to the natural sciences and excluding those centered on medical and engineering disciplines. The sitemap and navigation menu data spans June to August 2024. After parsing and cleaning the datasets, and removing duplicate webpage entries and external links, 13,122,124 URLs from the sitemaps and 9,953 URLs from the navigation menus were retained for further analysis. These URLs were then analyzed using the taxonomy induction method outlined in the methodology section, which incorporates Formal Concept Analysis (FCA) and graph pruning. Through this process, 2,781 hypernym-hyponym pairs across the 112 websites were identified for content exploration and comparative analysis.

The analysis of the 112 academies reveals significant variation in the size and organization of their web content. The number of URLs in the sitemaps varies widely, ranging from 30 to 1.5 million, with an average of 70,000 URLs per academy. Similarly, the number of items in the navigation menus ranges from 3 to 211, with an average of 40 menu items per academy. These variations indicate differing digital strategies, where some academies maintain extensive online repositories, while others prioritize streamlined, high-level navigation.

Figure 2 visualizes the depth distribution of URLs across different academies, mapping the relationship between the total number of URLs and their hierarchical depth. This analysis shows that academies with larger numbers of URLs do not necessarily structure their content deeper within the hierarchy. The lack of significant correlation, confirmed by a linear correlation analysis (p-value = 0.334), suggests that different content organization strategies influence website structure beyond mere scale. Some academies may prioritize broad, shallow hierarchies for accessibility, while others adopt deeper structures for detailed content segmentation.

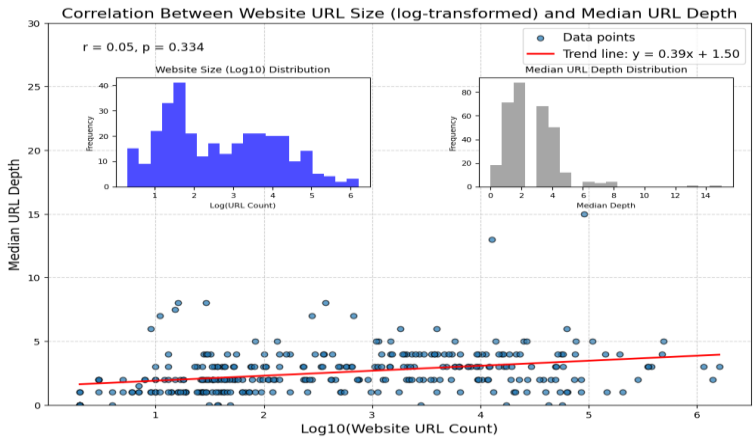

Figure 2. Correlation Between Website URL Size (Log-transformed) and Median URL Depth.

Taxonomy of Web Content

This study applies Formal Concept Analysis (FCA) and graph pruning to develop a structured taxonomy for global scientific academies' web content. The resulting classification identifies 121 unique hypernym-hyponym pairs (Figure 3). The primary categories identified through FCA are "*Governance*", "*Information*", "*Knowledge Resources*", "*Organizational Role*", "*Organizational Structure*", "*Public Outreach*", "*Scientific Cooperation*", and "*Supporting Science*". Each of the categories is further subdivided into specific subcategories that reflect the various areas of activities within these academies. These categories illuminate the institutional functions and strategic priorities of scientific academies, affirming their distinct yet overlapping roles in knowledge production, dissemination, and societal engagement. The taxonomy reveals three dominant functional categories—governance (as Learned Society archetype), public engagement (as Adviser to Society archetype), and scientific production (as Manager of Research archetype). These align with Engelbrecht et al.'s (2020) archetypes, demonstrating how academies balance internal organization, public engagement, and research leadership.

The Learned Society archetype is characterized by scientific academies as self-governing communities dedicated to fostering intellectual exchange and the advancement of knowledge. The taxonomy highlights the dominant presence of "*Organizational Structure*," "*Organizational Role*," and "*Governance*." These categories define the framework that supports scientific discourse and knowledge circulation. The legitimacy of learned societies is grounded in their ability to curate, manage, and disseminate scientific knowledge, a role further reinforced by their commitment to research documentation and public engagement.

The Adviser to Society archetype is evident in the emphasis on "*Public Outreach*", particularly in "*Science Communication*" and "*Science Advice*." These functions position scientific academies as intermediaries between researchers and the broader society. The findings suggest that academies use digital platforms to enhance scientific literacy, influence public understanding, and provide guidance on policy matters. The prominence of "*Knowledge Resources*" within this category underscores the dual responsibility of academies to engage both scientific professionals and the general public in knowledge exchange.

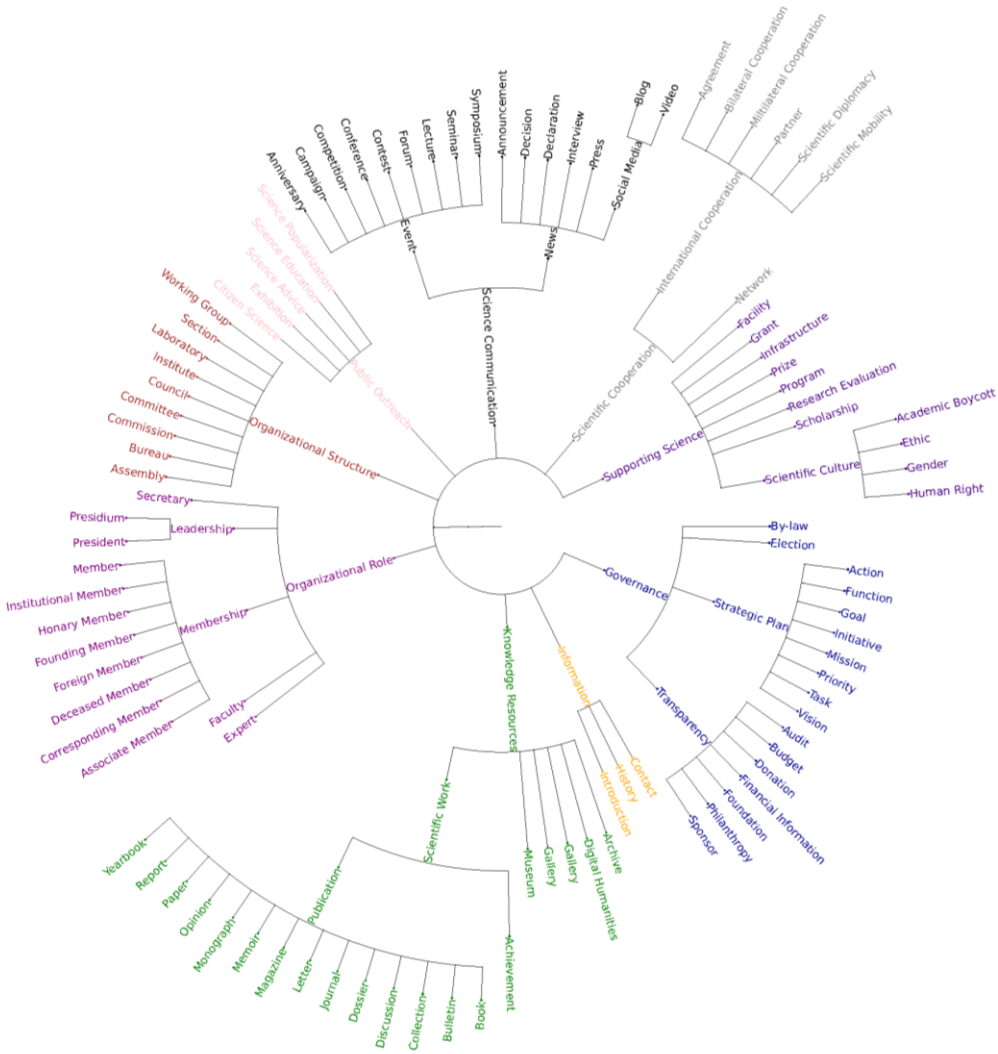

Figure 3. Hierarchical Taxonomy of Scientific Academies' Web Content.

The Manager of Research archetype extends beyond the direct management of research institutions to encompass a broader role in knowledge production and scientific excellence. The taxonomy demonstrates that "*Supporting Science*" and "*Scientific Cooperation*" are central to academy functions, signifying a strategic effort to cultivate both national and international scientific collaborations. The inclusion of "*Institution*" within the "*Organizational Structure*" of scientific academies suggests their direct involvement in knowledge creation.

Although Engelbrecht et al. (2020) primarily associated this archetype with direct research management, the taxonomy reveals that academies engage in a continuum of activities from knowledge production to dissemination. The presence of "*Knowledge Resources*" as a dominant category further illustrates that academies not only facilitate scientific research but also actively curate and preserve it. Some academies emphasize scientific recognition through awards and prizes, reinforcing

their role in advancing scientific excellence. The subcategory "*Archive*" within "*Knowledge Resources*" further highlights efforts to document and preserve national scientific and cultural heritage, reflecting a long-term commitment to maintaining and disseminating scientific knowledge.

The digital presence of global scientific academies is strategically structured to reflect their core missions and institutional priorities. The common taxonomy of these academies' websites reveals clear hierarchical relationships between key content categories, illustrating how they construct their institutional identity. The taxonomy highlights their role in facilitating knowledge circulation and maximizing its impact. It also identifies opportunities for public outreach, engagement, and independent advisory functions to governments. Scientific academies position themselves within a complex landscape of national and international policy, societal expectations, and intellectual networks, navigating challenges such as technological and resource disparities across institutions. Most academies do not fit neatly into a single archetype; even those within the same category may adopt distinct strategies to advance scientific excellence and promote public understanding of science. The following section will further examine the commonalities and unique characteristics of these academies' digital strategies.

Comparative Analysis of Digital Presence Across scientific academies

While global scientific academies share a common commitment to advancing scientific knowledge and assimilation knowledge, their online presences vary considerably. The Taxonomy Similarity score for the 112 scientific academies ranges from 0.2 to 0.75, with an average value of 0.42. This variation indicates differing degrees of alignment with the taxonomy developed. While some scientific academies closely follow the established taxonomy, others diverge in various ways, reflecting their unique priorities, missions, and regional contexts.

To better understand these variations, this study conducted a pairwise comparison of websites, utilizing hierarchical clustering based on Jaccard Similarity. This analysis revealed distinct groups of websites exhibiting different patterns in their hypernym-hyponym relationships. The dendrogram (tree diagram) in Figure 4 partitions the websites into three clusters, illustrating the degree of academies' web content similarity in structuring and categorization.

Figure 4 provides key insights into imitation and innovation behaviors across clusters. The left box plot in Figure 4 represents the Imitation Score for each cluster, which measures the average similarity of each website to others. Cluster 2 (green) has the lowest imitation score, meaning these websites are more unique and less similar to established patterns. This suggests a departure from conventional digital structures, possibly due to resource-limited context or underdeveloped website taxonomies. Cluster 1 (yellow) and Cluster 3 (red) have higher imitation scores, indicating stronger alignment with established conventions, implying that these websites adhere more closely to widely accepted content organization strategies.

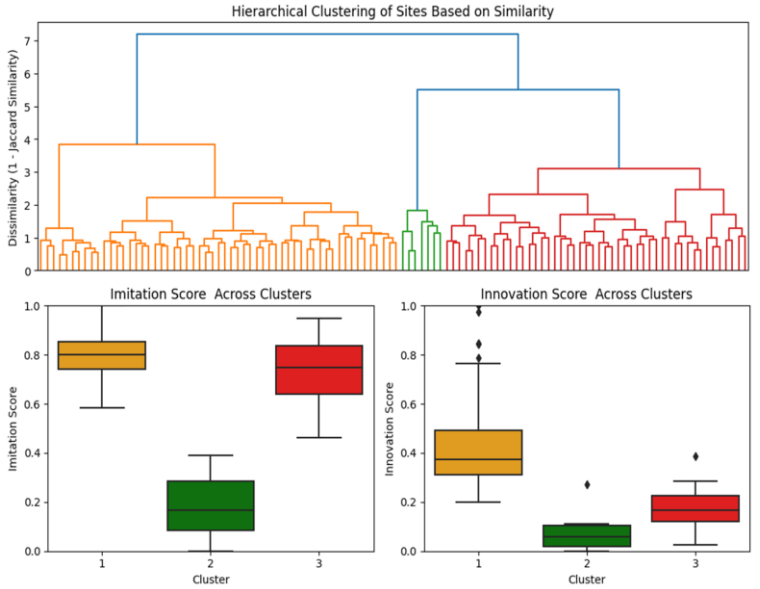

Figure 4. Hierarchical Clustering of Sites and Innovation/Imitation Scores of Website Groups.

Table 1. Descriptive Statistics for Taxonomy Similarities, Imitation Scores and Innovation Scores Across Clusters.

	Taxonomy Similarity			*Imitation Score*			*Innovation Score*		
	Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 1*	*Cluster 2*	*Cluster 3*	*Cluster 1*	*Cluster 2*	*Cluster 3*
Value	▲ High (0.51)	▼ Low (0.21)	▼ Low (0.34)	▲ High (0.80)	▼ Low (0.18)	▲ High (0.73)	▲ High (0.44)	▼ Low (0.08)	▼ Low (0.17)
Mean	0.51	0.21	0.34	0.80	0.18	0.73	0.44	0.08	0.17
Median	0.48	0.19	0.35	0.80	0.17	0.75	0.37	0.06	0.17
SD	0.09	0.07	0.06	0.10	0.15	0.13	0.19	0.09	0.07
Min	0.39	0.15	0.22	0.58	0.00	0.46	0.20	0.00	0.03
Max	0.75	0.34	0.48	1.00	0.39	0.95	1.00	0.27	0.39
Cluster size	57	7	48	57	7	48	57	7	48
Overall Avg		0.35			0.57			0.23	
ANOVA F-Statistic		94.19**			88.33**			51.93**	
Cluster 1	1			1			1		
Cluster 2	10.63**	1		10.57**	1		8.14	1	
Cluster 3	11.89**	4.79**	1	2.96**	9.19**	1	9.81**	2.33**	1

Note: *p<0.05, **p<0.01.

The ANOVA test results for Taxonomy Similarity, Imitation Score and the Innovation Score have p-value <0.01, indicating highly significant difference in the three metrics across clusters. The Pairwise t-tests results of p-values <0.01 indicate clusters have distinct imitation and innovation behaviors. The Innovation Score of Cluster 2 and Cluster 3 do not show significance with p-value above 0.05. Bootstrap resampling is conducted before statistical analysis for robustness due to small sample sizes.

The right box plot in Figure 4 presents the Innovation Score, which captures the extent to which websites introduce new hypernym-hyponym relationships. Cluster 1 (yellow) has the highest innovation score, meaning that websites in this cluster introduce more unique content structures, signifying efforts toward digital differentiation. Cluster 2 (green) has the lowest innovation score, confirming that these websites not only diverge from common patterns but also lack substantial new actions. Cluster 3 (red) demonstrates moderate innovation, balancing between adopting conventional taxonomies and integrating some novel elements. Some academies adhere closely to established frameworks, while others diverge significantly. This divergence occurs either through the introduction of new structures or fragmented content strategies.

Table 1 summarizes the Taxonomy Similarity, the Imitation Scores, and the Innovation Scores across the three identified clusters. These statistics highlight how websites align with common taxonomies, maintain structural consistency, and introduce unique elements.

Cluster 1 (yellow) websites in Table 1 exhibit high innovation but low imitation scores, indicating that they are highly innovative academies that introduce novel digital structures. This suggests that these academies take a more innovative and forward-thinking approach to structuring their digital presence. Cluster 2 (green) websites have the lowest imitation and innovation scores, reflecting conservative digital strategies. These academies exhibit fragmented or underdeveloped web structures, often lacking clear content hierarchies or comprehensive navigation systems. This pattern may reflect a lack of cohesive content strategy, potentially hindering user navigation and information retrieval. Cluster 3 (red) websites show high imitation but low-to-moderate innovation, meaning they are hybrid academies that combine imitation with selective innovation. These websites prioritize standardization, ensuring consistency in their digital frameworks while making incremental refinements.

The statistical test results confirm that the clustering approach successfully identifies meaningful distinctions. ANOVA test results for Taxonomy Similarity, Imitation Score, and Innovation Score show p-values < 0.05, indicating statistically significant differences across clusters. Pairwise t-tests further validate distinct imitation and innovation behaviors across most clusters, except for a less significant difference in innovation behaviors between Cluster 2 and Cluster 3. Bootstrap resampling is applied to enhance robustness given varying sample sizes. These statistical findings reinforce the validity of the identified clusters and their implications for digital taxonomy structuring.

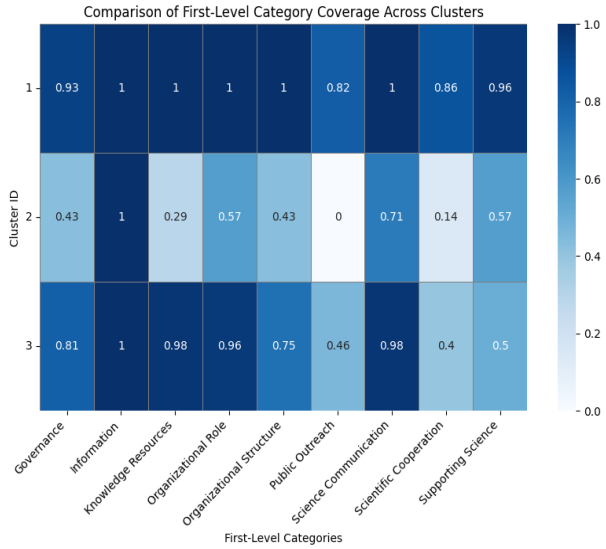

Figure 5. Comparison of First-Level Category Coverage Across Clusters.

Table 2. Distinctiveness and Statistical Analysis of Top Distinctiveness Pairs and Least Distinctiveness Pairs (Common Pairs) Across Clusters.

	Cluster 1			Cluster 2			Cluster 3		
	Distinct 5 Pairs	Other Pairs	Common 5 Pairs	Distinct 5 Pairs	Other Pairs	Common 5 Pairs	Distinct 5 Pairs	Other Pairs	Common 5 Pairs
Mean	1.54	1.49	0.54	11.11	1.03	0.47	2.19	1.09	0.28
Median	1.54	1.46	0.64	5.22	1.04	0.53	2.21	1.09	0.28
Std Dev	0.00	0.63	0.23	14.24	0.38	0.11	0.60	0.15	0.02
Min	1.54	0.50	0.22	3.65	0.43	0.35	1.55	0.74	0.26
Max	1.54	2.93	0.77	36.53	2.22	0.59	3.09	1.30	0.31
Hypernym-Hyponym Pairs Count	5	43	5	5	76	5	5	108	5
Distinct 5 Pairs	1			1			1		
Other Pairs	17.43**	1		1.47	1		4.93	1	
Common 5 Pairs	9.95**	-6.01**	1	1.67**	-8.66**	1	7.15**	-15.66**	1
Distinct 5 Pairs	Publication->Yearbook News->Decision Scientific Work->Achievement			Transparency->Audit Strategic Plan->Vision			Membership->Institutional Member Membership->Associate Member		

	Transparency->Budget Knowledge Resources->Museum	Membership->Corresponding Member Transparency->Financial Information Membership->Founding Member	Membership->Corresponding Member Membership->Honary Member Social Media->Blog
Common 5 Pairs	Membership->Corresponding Member Membership->Associate Member Strategic Plan->Vision Membership->Honary Member Membership->Founding Member	Knowledge Resources->Scientific Work Scientific Work->Publication Homepage->Scientific Cooperation News->Social Media Scientific Cooperation->International Cooperation	Supporting Science->Scholarship Event->Anniversary Event->Competition Publication->Memoir Organizational Structure->Assembly

Note: *p<0.05, **p<0.01.
The Pairwise t-tests results of p-values <0.01 indicate clusters have distinct imitation and innovation behaviors. The Distinct 5 Pairs and Other Pairs of Cluster 2 and Cluster 3 do not show significance with p-value above 0.05. Bootstrap Resampling is added before statistical analysis for robustness due to small sample sizes.

To gain deeper insights into how different websites cover the taxonomy categories, this study generated a heatmap (Figure 5) that visualizes the coverage of first-level categories across the 112 websites. The heatmap allows decision-makers to identify strengths and gaps in content representation. Academies can use this insight to align their digital strategies with common best practices while addressing areas of weak representation.

Cluster 1 (yellow) in Figure 5 exhibits the most comprehensive coverage across all first-level categories, with most values close to 1. Websites in this cluster consistently represent key categories, including "*Information*," "*Knowledge Resources*," "*Organizational Role*," "*Organizational Structure*," and "*Scientific Cooperation.*" This suggests that these websites follow a structured taxonomy, ensuring well-organized content and accessibility.

Cluster 2 (green) shows uneven category coverage, with "*Public Outreach*" (0.00) and "*Scientific Cooperation*" (0.14) largely absent, while "*Information*" (1.00) and "*Science Communication*" (0.71) are strongly represented. This suggests a selective emphasis on specific themes. This suggests that websites in this cluster focus on specific categories while omitting others, potentially indicating specialized or fragmented digital structures that reflect varied institutional priorities.

Cluster 3 (red) balances coverage, with high representation in "*Knowledge Resources*" (0.98), "*Organizational Structure*" (0.96), and "*Scientific Cooperation*" (0.98), while "*Public Outreach*" (0.46) and "*Supporting Science*" (0.50) are less prominent. This pattern suggests that websites in Cluster 3 generally align with common taxonomies but selectively emphasize certain content areas, striking a balance between conformity and differentiation.

These results confirm that clustering effectively differentiates websites based on their structural emphasis, highlighting distinct patterns in how scientific academies structure their online presence and the prioritization of content categories.

To further explore how specific content distinguishes scientific academies, this study applied the Distinctiveness Score (as outlined in the methodology section) to identify the most distinguishing hypernym-hyponym pairs and the least distinguishing hypernym-hyponym pairs. Table 2 presents a statistical analysis of the distinctiveness of hypernym-hyponym pairs across the three clusters, offering insights into differences in how websites structure their taxonomies. Cluster 2 exhibits the most unique structural elements, as indicated by its highest distinctiveness score (11.11) and high standard deviation (14.24). This suggests that websites in this cluster introduce the most unique structural elements. In contrast, Cluster 1 and Cluster 3 display lower distinctiveness scores (1.54 and 2.19, respectively), indicating a more moderate level of structural differentiation and stronger alignment with widely recognized taxonomies. The common pairs have significantly lower scores across all clusters (ranging from 0.28 to 0.54), confirming that frequently shared relationships follow more standardized patterns.

These findings highlight that while some academies maintain highly conventional taxonomies, others develop distinctive content structures, reflecting diverse institutional priorities and digital strategies. Pairwise t-tests confirm statistically significant differences ($p < 0.01$) between distinct and common pairs in Clusters 1

and 3, reinforcing clear structural separation. However, Cluster 2 and Cluster 3 do not show significant differences in "Other Pairs," indicating some shared taxonomy structures. These findings confirm that Cluster 2 exhibits the most structurally unique websites, while Cluster 1 and Cluster 3 balance imitation and innovation differently.
Distinct hypernym-hyponym pairs reveal unique digital strategies among scientific academies. Cluster 1 (Yellow) focuses on institutional knowledge, governance, and decision-making, emphasizing categories like "*Publication → Yearbook*" and "*Scientific Work → Achievement*" to document scholarly contributions. Cluster 2 (Green) emphasizes financial transparency and strategic vision, with categories like "*Transparency → Audit*" and "*Strategic Plan → Vision*," reflecting a focus on governance and long-term planning. Cluster 3 (Red) prioritizes digital engagement, using categories like "*Social Media → Blog*" and "*Membership → Honorary Member*" to create an interactive outreach strategy.
These distinctions illustrate how different academies adapt their digital presence based on governance models, transparency requirements, and audience engagement strategies. Common hypernym-hyponym pairs highlight shared digital structures across scientific academies. Most emphasize structured membership systems, with categories like "*Membership → Corresponding Member / Associate Member / Honorary Member / Founding Member*," reinforcing their role as academic communities. Strategic foresight and institutional direction remain central, evidenced by "*Strategic Plan → Vision*." Scientific cooperation and public communication are common priorities. Categories like "*Scientific Cooperation → International Cooperation*" and "*News → Social Media*" demonstrate the widespread use of digital platforms for knowledge dissemination and stakeholder engagement.

Discussion

The findings of this study highlight both shared and divergent patterns in how global scientific academies structure their online presence. Addressing RQ1, the taxonomy derived from Formal Concept Analysis (FCA) reveals a common framework that organizes academy websites around governance, knowledge resources, public outreach, scientific cooperation, and organizational structures. Despite this shared foundation, academies vary in how they emphasize these elements. Some prioritize structured governance and scholarly documentation, while others focus on enhancing public outreach or fostering scientific collaborations. These differences reflect the diverse roles academies play in their national and international contexts, shaping how they present their digital identities.
For RQ2, the comparative analysis of taxonomy similarity, imitation scores, and innovation scores demonstrates varying levels of adherence to standard digital frameworks. Academies in Cluster 1 exhibit high innovation but low imitation scores, indicating that they are highly innovative academies that introduce novel digital structures. In contrast, those in Cluster 2 show the lowest imitation and innovation scores, characterized by fragmented or underdeveloped digital strategies that suggest conservative digital strategies. Cluster 3 aligns closely with established taxonomies,

maintaining consistency while integrating selective innovations. These variations underscore how scientific academies navigate the balance between digital conformity and differentiation. The most distinctive hypernym-hyponym pairs reveal areas where academies differentiate themselves, such as financial transparency initiatives or interactive digital engagement strategies, while the least distinctive pairs—membership structures, strategic planning, and research collaboration—reflect widely shared priorities.

From a policy perspective, scientific academies must strike a balance between standardization and differentiation in their digital strategies. Aligning with recognized taxonomies ensures clarity, institutional credibility, and interoperability, while incorporating innovative elements enhances visibility and engagement. Academies with fragmented digital structures may benefit from reassessing their web organization to improve accessibility and communication effectiveness. Strengthening public outreach, ensuring transparent governance, and supporting digital transformation initiatives—particularly for academies in regions with limited resources—can help bridge disparities in digital infrastructure. Establishing international guidelines for structuring academic web content would further enhance cohesion among global academies, fostering stronger collaboration and knowledge exchange. By refining their digital presence, scientific academies can reinforce their institutional roles, expand their public reach, and strengthen their contributions to global scientific discourse.

Conclusion

This study provides a comprehensive framework for analyzing the digital presence of global scientific academies, examining how they structure their online content and engage with stakeholders. By applying Formal Concept Analysis (FCA) and graph pruning, the research identifies both common patterns and variations in the web content taxonomy of scientific academies. The findings reveal that while academies share a foundational structure emphasizing governance, knowledge dissemination, and public engagement, they differ in the extent to which they innovate or conform to established digital frameworks. The comparative analysis of taxonomy similarity, imitation scores, and innovation scores highlights distinct strategic approaches, with some academies adhering closely to conventional taxonomies, others demonstrating fragmented or underdeveloped digital strategies, and a subset actively incorporating novel structures to enhance their digital identity. The distinctiveness analysis of hypernym-hyponym pairs further provides insights into the key areas where academies differentiate themselves, reflecting diverse institutional priorities.

This study contributes to both institutional theory and digital taxonomy research. The application of FCA advances the understanding of how scientific academies navigate the balance between standardization and differentiation in their digital strategies, shedding light on institutional isomorphism in the digital realm. Additionally, the structured web mining approach and hierarchical taxonomy construction refined methods for analyzing large-scale institutional web data, offering a scalable framework for comparative analysis. These insights have practical implications for academy leaders, policymakers, and digital strategists, providing a foundation for

developing best practices that enhance the visibility, accessibility, and interoperability of scientific academies' digital presence.

Despite its contributions, this study has certain limitations. The analysis is based solely on digital content, without accounting for offline activities and interactions that may influence an academy's broader institutional role. Additionally, while taxonomy captures structural and thematic variations, it does not measure the effectiveness of digital engagement strategies. Future research could explore the relationship between digital presence and institutional influence could further refine strategies for strengthening scientific communication and global collaboration. By continuing to refine digital strategy frameworks, this research lays the groundwork for future transformations in how scientific academies facilitate scholarly communication and contribute to the global scientific ecosystem.

References

Axenbeck, J., & Breithaupt, P. (2021). Innovation indicators based on firm websites—Which website characteristics predict firm-level innovation activity? *PloS One*, *16*(4), e0249583.

Benade, L. (2016). Learned Societies, Practitioners and their 'Professional'Societies: Grounds for developing closer links. In *Educational Philosophy and Theory* (Vol. 48, Issue 14, pp. 1395–1400). Taylor & Francis.

Bottai, C., Crosato, L., Domenech, J., Guerzoni, M., & Liberati, C. (2024). Scraping innovativeness from corporate websites: Empirical evidence on Italian manufacturing SMEs. *Technological Forecasting and Social Change*, *207*, 123597.

Burford, S. (2011). Web information architecture–a very inclusive practice. *Journal of Information Architecture*, *3*(1), 19–40.

Burford, S. (2014). A grounded theory of the practice of web information architecture in large organizations. *Journal of the Association for Information Science and Technology*, *65*(10), 2017–2034.

Campos, P. M. C., Reginato, C. C., & Almeida, J. P. A. (2019). Towards a Core Ontology for Scientific Research Activities. In G. Guizzardi, F. Gailly, & R. Suzana Pitangueira Maciel (Eds.), *Advances in Conceptual Modeling* (Vol. 11787, pp. 3–12). Springer International Publishing. https://doi.org/10.1007/978-3-030-34146-6_1

Ceci, M., & Lanotte, P. F. (2021). Closed sequential pattern mining for sitemap generation. *World Wide Web*, *24*(1), 175–203.

Chen, X.. (2024, November). *Global scientific academies Dataset* (Version V1). Science Data Bank. https://doi.org/10.57760/sciencedb.14674

Cox, A. M. (2007). Beyond information–factors in participation in networks of practice: A case study of web management in UK higher education. *Journal of Documentation*, *63*(5), 765–787.

Cox, A. M. (2008). An exploration of concepts of community through a case study of UK university web production. *Journal of Information Science*, *34*(3), 327–345.

Elsayed, A. M. (2017). Web content strategy in higher education institutions: The case of King Abdulaziz University. *Information Development*, *33*(5), 479–494. https://doi.org/10.1177/0266666916671387

Engelbrecht, J., Djurovic, M., & Reuter, T. (2020). Current tasks of academies and academia. *Cadmus*, *4*(2), 118–126.

Engelbrecht, J., & Šlaus, I. (2022). ACADEMIES OF SCIENCES IN THE CONTEMPORARY WORLD. *Trames: A Journal of the Humanities and Social Sciences*, *26*(2), 131–139.

Gloria, M. J. K., McGuinness, D. L., Luciano, J. S., & Zhang, Q. (2013). Exploration in web science: Instruments for web observatories. *Proceedings of the 22nd International Conference on World Wide Web*, 1325–1328.

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*, 653–671.

Hale, S. A., Yasseri, T., Cowls, J., Meyer, E. T., Schroeder, R., & Margetts, H. (2014). Mapping the UK webspace: Fifteen years of British universities on the web. *Proceedings of the 2014 ACM Conference on Web Science*, 62–70.

Isavand, L., & Poormoghim, H. (2024). Comparative Study of Scientific Academies between European Countries (Royal Society of Great Britain, Lincean Academy of Italy, French Scientific Academy), and Iran. *Advances in Applied Sociology*, *14*(03), 161–174. https://doi.org/10.4236/aasoci.2024.143011

Karanasios, S., Thakker, D., Lau, L., Allen, D., Dimitrova, V., & Norman, A. (2013). Making sense of digital traces: An activity theory driven ontological approach. *Journal of the American Society for Information Science and Technology*, *64*(12), 2452–2467.

Kenekayoro, P., Buckley, K., & Thelwall, M. (2014). Automatic classification of academic web page types. *Scientometrics*, *101*, 1015–1026.

Kenekayoro, P., Buckley, K., & Thelwall, M. (2015). Clustering research group website homepages. *Scientometrics*, *102*, 2023–2039.

Kinne, J., & Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics*, *125*(3), 2011–2041.

Kinne, J., & Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. *PLOS ONE*, *16*(4), e0249071. https://doi.org/10.1371/journal.pone.0249071

Krishnapuram, R., Joshi, A., Nasraoui, O., & Yi, L. (2001). Low-complexity fuzzy relational clustering algorithms for Web mining. *IEEE Transactions on Fuzzy Systems*, *9*(4), 595–607. https://doi.org/10.1109/91.940971

Late, E., Guns, R., Pölönen, J., Stojanovski, J., Urbanc, M., & Ochsner, M. (2024). Beyond borders: Examining the role of national learned societies in the social sciences and humanities. *Learned Publishing*.

Lepori, B., Aguillo, I. F., & Seeber, M. (2014). Size of web domains and interlinking behavior of higher education institutions in Europe. *Scientometrics*, *100*, 497–518.

Markus Neumann, Fridolin Linder and Bruce Desmarais. (2022). Government websites as data: A methodological pipeline with application to the websites of municipalities in the United States. *Journal of Information Technology & Politics*, *19*(4), 411–422. https://doi.org/10.1080/19331681.2021.1999880

Martínez-Torres, M. R., Toral, S. L., Palacios, B., & Barrero, F. (2012). An evolutionary factor analysis computation for mining website structures. *Expert Systems with Applications*, *39*(14), 11623–11633. https://doi.org/10.1016/j.eswa.2012.04.011

Norrby, E. (2001). The Role of Academies of Science in a Global World. *AMBIO: A Journal of the Human Environment*, *30*(2), 71–71.

Ruzza, M., Tiozzo, B., Mantovani, C., D'Este, F., & Ravarotto, L. (2017). Designing the information architecture of a complex website: A strategy based on news content and faceted classification. *International Journal of Information Management*, *37*(3), 166–176.

Schroeder, R., Brügger, N., & Cowls, J. (2020). Historical web as a tool for analyzing social change. *Second International Handbook of Internet Research*, 489–504.

Singh, U., Divya Venkatesh, J., Muraleedharan, A., Saluja, K. S., J H, A., & Biswas, P. (2024). Accessibility Analysis of Educational Websites Using WCAG 2.0. *Digital Government: Research and Practice*, *5*(3), 1–28. https://doi.org/10.1145/3696318

Sophia Alim. (2021). Web Accessibility of the Top Research-Intensive Universities in the UK. *Sage Open*, *11*(4), 21582440211056614. https://doi.org/10.1177/21582440211056614

Sun, A., & Lim, E. (2006). Web unit-based mining of homepage relationships. *Journal of the American Society for Information Science and Technology*, *57*(3), 394–407. https://doi.org/10.1002/asi.20279

Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *Journal of the American Society for Information Science and Technology*, *57*(1), 60–68.

Tsakalidis, A., Basile, P., Bazzi, M., Cucuringu, M., & McGillivray, B. (2021). DUKweb, diachronic word representations from the UK Web Archive corpus. *Scientific Data*, *8*(1), 269. https://doi.org/10.1038/s41597-021-01047-x

Weber, M. S. (2021). Digital Data and a Multilevel Perspective of Institutions on the Web. *Proceedings of the 13th ACM Web Science Conference 2021*, 4–4.

Will, E. M., & Callison, C. (2006). Web presence of universities: Is higher education sending the right message online? *Public Relations Review*, *32*(2), 180–183. https://doi.org/10.1016/j.pubrev.2006.02.014

Yoshinaga, N., & Nobuhara, H. (2010). Formal concept analysis based web pages classification/visualization and their application to information retrieval. *2010 10th International Symposium on Communications and Information Technologies*, 153–157.

Appendices

Table 1 presents the detailed partitioning results of the hierarchical clustering of scientific academies. Cluster 1 consists of academies that prioritize innovation, introducing novel structures and diverse content categories. This cluster includes prestigious academies from G7 and other developed countries that lead in digital strategy. Cluster 2 represents academies with fragmented or less structured digital strategies, often characterized by selective content representation or weak adherence to taxonomy standards (e.g., casciences.org, mta.hu, bas.co.bw). Cluster 3 includes academies that closely follow established taxonomies, exhibiting high imitation scores and minimal structural divergence (e.g., japan-acad.go.jp, kvab.be, vast.gov.vn).

Table 1. Detail Partition Result of the Hierarchical Clustering of Scientific Academies.

Site domain	Cluster	Site domain	Cluster	Site domain	Cluster
aast.dz	1	aciencias.org.bo	2	naskr.kg	3
ria.ie	1	ais-sanmarino.org	2	dknvs.no	3
lza.lv	1	bas.co.bw	2	igd-sh.lu	3
manu.edu.mk	1	casciences.org	2	internet.hn	3
nas.gov.ua	1	mta.hu	2	japan-acad.go.jp	3
nasb.gov.by	1	zaas.org.zm	2	knasciences.or.ke	3
nasonline.org	1	zas.ac.zw	2	kvab.be	3
nast.gov.np	1			maas.edu.mm	3
oeaw.ac.at	1			nas.go.kr	3
palast.ps	1			nas.org.ng	3
pan.pl	1			rss.jo	3
paspk.org	1			nassl.org	3
rae.es	1			nast.ph	3
ras.ru	1			nauka-nanrk.kz	3
royalacademy.dk	1			rac.gov.kh	3
lincei.it	1			sav.sk	3
royalsociety.go.th	1			sci.am	3
royalsociety.org	1			science.gov.tm	3
royalsociety.org.nz	1			snas.org.sg	3
rsc-src.ca	1			unas.org.ug	3
sanu.ac.rs	1			vast.gov.vn	3
sazu.si	1			assaf.co.za	3
science.gov.az	1			avcr.cz	3
science.org.au	1			anc.cr	3
science.org.ge	1			asrt.sci.eg	3

scnat.ch	1		acfiman.org	3
taas-online.or.tz	1		abc.org.br	3
acad-ciencias.pt	1		ac.mn	3
lma.lt	1		acaciencias.org.gt	3
tuba.gov.tr	1		academiaciencias.cu	3
leopoldina.org	1		academiadeciencias.cl	3
asm.md	1		academiadecienciasrd.org	3
kva.se	1		academie-sciences.bj	3
acad.ro	1		ashak.org	3
academy.ac.il	1		academyofcyprus.cy	3
academy.uz	1		acadsci.fi	3
academyofathens.gr	1		academie.hassan2.sciences.ma	3
akad.gov.al	1		aipi.or.id	3
akadeemia.ee	1		ansts.sn	3
akademisains.gov.my	1		asduliban.org	3
anc-argentina.org.ar	1		akademia-malagasy.mg	3
anrt.tj	1		aosci.org	3
antat.ru	1		asa.gov.af	3
anubih.ba	1		ansal.bf	3
academie-sciences.fr	1		ancperu.org	3
bas.bg	1		anciu.org.uy	3
gaas-gh.org	1		amc.edu.mx	3
knaw.nl	1		cienciasdenicaragua.org	3
bas.org.bd	1			
ias.ac.ir	1			
hazu.hr	1			
insaindia.res.in	1			
eas-et.org	1			
casinapioiv.va	1			
casad.cas.cn	1			
canu.me	1			
beitalhikma.tn	1			

The heatmap in Figure 1 visually depicts the extent to which each academy covers these core content categories. This distribution highlights clear differences in digital content strategies among academies. Some institutions, particularly those in Cluster 1, exhibit comprehensive coverage across multiple categories, whereas others, especially in Cluster 2 and Cluster 3, show gaps in specific areas, such as Public Outreach and Scientific Cooperation. The clustering approach effectively groups websites with similar digital strategies, revealing distinct content structuring behaviors across institutions.

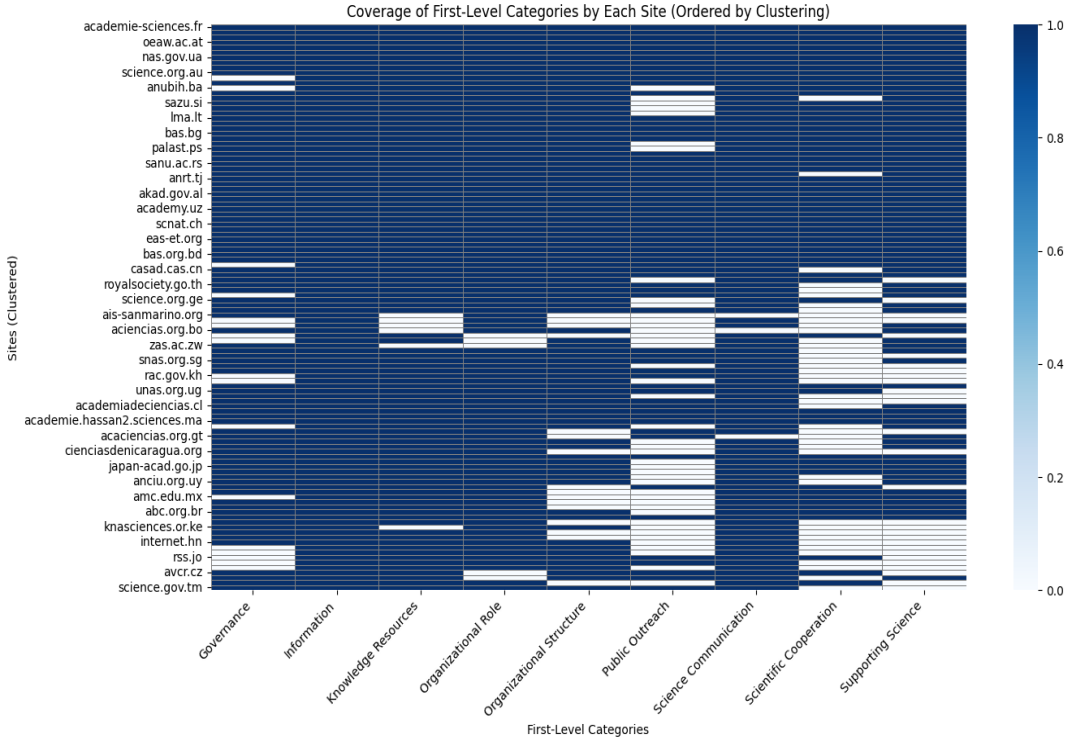

Figure 1. Comparison of First-Level Category Coverage Across Scientific Academies.