Comb and Martin
EURASIP Journal on Information Security (2024) 2024:21
https://doi.org/10.1186/s13635-024-00172-5

EURASIP Journal on Information Security

RESEARCH

Open Access

Mining digital identity insights: patent analysis using NLP

Matthew Comb[1*] and Andrew Martin[1]

Abstract

The field of digital identity innovation has grown significantly over the last 30 years, with over 6000 technology patents registered worldwide. However, many questions remain about who controls and owns our digital identity and intellectual property and, ultimately, where the future of digital identity is heading.

To investigate this further, this research mines digital identity patents and explores core themes such as identity, systems, privacy, security, and emerging fields like blockchain, financial transactions, and biometric technologies, utilizing natural language processing (NLP) methods including part-of-speech (POS) tagging, clustering, topic classification, noise reduction, and lemmatisation techniques. Finally, the research employs graph modelling and statistical analysis to discern inherent trends and forecast future developments.

The findings significantly contribute to the digital identity landscape, identifying key players, emerging trends, and technological progress. This research serves as a valuable resource for academia and industry stakeholders, aiding in strategic decision-making and investment in emerging technologies and facilitating navigation through the dynamic realm of digital identity technologies.

Keywords Digital identity, Trust, Privacy, Ecosystem, Security, Future

1 Introduction

Digital identity-defined by Kim Cameron [15] as "a set of claims made by one digital subject (e.g., a user) about itself or another digital subject"-holds great significance in today's fast-paced digital world. In modern computing systems, a digital identity is critical for safeguarding individual autonomy and enhancing digital interactions. It is vital in various domains, including financial transactions [6, 7, 30], interactions with state entities for qualifications and entitlements, and healthcare services [37, 50]. It is a fundamental component in verifying and securing personal and transactional data [8, 14].

While the world is moving towards establishing ecosystems to foster digital identity-centric products and services [1, 3, 34, 42, 60], the fundamental question of digital identity ownership-who owns and controls the standards on which society's digital identity solutions will be built-remains unresolved.

Furthermore, despite increasing consumer demand to own their personal information [40, 47, 58], companies have registered over 6000 digital identity patents, demonstrating their perceived importance in shaping the future of digital services. In particular, organisations such as Microsoft, Oracle, Mastercard, IBM, Visa, Bank of America, and Apple have all attempted to control digital-identity-related intellectual property [55] and have collectively registered a significant number of digital identity patents (see Table 2).

This research focuses on whether text mining through NLP analysis can help identify digital identity trends in commercial innovation. The study involved organising a substantial dataset of over 6000 digital identity patents

*Correspondence:
Matthew Comb
matthew.comb@linacre.ox.ac.uk
[1] Computer Science Department, University of Oxford, Parks Road, Oxford OX1 3PP, Oxfordshire, UK

through various data preparation techniques, including cleansing, noise reduction and lemmatisation, to highlight key terms and refine the dataset.

The study then applied advanced NLP methods and various statistical analysis techniques to analyse the dataset. These methods included keyword frequency analysis, basic statistical metrics, and evaluation of term significance using measures like term frequency (TF), document frequency (DF), and term frequency-inverse document frequency (TF-IDF) coefficients. Furthermore, POS tagging and K-means clustering were employed to identify essential sub-domains and topics. Finally, polynomial regression was used to predict the future significance of identified features.

In summary, this research combines historical and contemporary perspectives, thoroughly examining the commercial digital identity sector and highlighting underlying connections and patterns within the extensive collection of patents. By doing so, this research will contribute to the existing literature on digital identity and provide valuable insights to policymakers, industry practitioners, and scholars.

In the next section, we will conduct a background analysis of the patent set to identify the most active commercial entities in the digital identity field. Section 3, 'Methodology', details the data mining techniques and analytical frameworks used in our patent analysis. Section 4, 'Results and discussion', presents our findings, highlighting the trends and implications in digital identity innovations. Section 5, 'Limitations', acknowledges the constraints of our approach, providing a critical perspective on our findings. Finally, Sect. 6, 'Conclusions', synthesises our research insights, reflecting on their broader impact and suggesting future research directions in the rapidly evolving field of digital identity.

2 History

In 1948, Bell Telephone Labor Inc. filed the first 'digital identity' patent. It aimed to identify subscriber devices on a telecommunications network. Since then, 6156 additional patents referencing digital identity have been filed globally.

In this section, we group our primary patent dataset (see Sect. 3) by a patent's filing entity and examine the most active companies over the 25 years leading up to and including 2020 (see Table 1). During this period, 4047 digital identity patents were registered, contributing approximately two-thirds of the total dataset. We also provide an overview of recent digital identity patent filing activities, from the beginning of 2021 to September 2023 for which 2067 patents are registered, and compare this recent activity with the preceding 25 years. Finally, we explore the primary areas of digital identity innovation that these entities focused on during 5-year periods. This analysis provides valuable insight into the overall trend of digital identity advancement thus far.

Table 1 Companies filing the most patents containing the term "digital identity" over a 25 year period [1996–2020]

Company	Patents
Microsoft Corporation	101
Alibaba/Alipay	84
Bundesdruckerei GmbH	66
Oracle International Corporation	54
International Business Machines Corporation	54
Diebold, Incorporated	49
Cisco Technology, Inc.	38
Forcepoint, LLC	38
Mastercard International Inc.	32
Aspire Digital Technology	31
Verizon Patent and Licensing Inc.	27
Go Daddy.	26
Bank Of America Corporation	25

Source: primary patent set (see Sect. 3)

2.1 Diebold (1996–2005)

Diebold was established as a safe and lock company in 1859. When technology allowed, the company pivoted into the automated teller machine market (ATM) and, most recently, into the general election market. Today, Diebold Nixdorf, as it is now known, has an annual revenue of almost US $5 billion and provides numerous identity-related products and services to companies and governments worldwide. It is fair to say that Diebold has been one of the first companies on the frontline dealing with identity-related challenges. Consequently, it is understandable that Diebold has a long history of digital-identity-related patents.

Diebold established its first digital identity patent in 1997 [19], and its most recent patent was filed in 2012 [20]. Thirty-nine of Diebold's 49 patents are for machines involving digital identity (such as an ATM), and 37 mention banking.

Recently, Diebold Nixdorf's Scott Anderson asserted that digital identity and its implementation needs to be more easily understood-adding that replacing physical identity documents with digital ones does not represent significant progress.

Anderson stated: "A digital copy of a physical identity is not a digital identity – and is a method fraught with problems." He further asserted that building a better digital identity is about how to get specific pieces of user data to confirm identity according to the operating context in a secure way while maintaining levels of privacy.

Anderson believes establishing a credential-proofing network could be a job for banks and financial institutions with the data access and consumer trust necessary to lead that changing paradigm.

This view counters that held by John Erik Setsaas, who, at a recent webinar on distributed identity, asserted that solutions with banks at the centre are short-sighted. Furthermore, he stated that such systems have numerous challenges, including that they are not cross-border or standardised, support limited attributes, and provide a centralised attack vector [54]. However, Anderson may have a point in that companies such as Diebold are in a prime position to roll out infrastructure to banks, allowing them to participate in a broader digital identity ecosystem.

2.2 Microsoft (2001–2010)

Microsoft conducted its most significant period of digital identity innovation between 2006-2010, when it established 54 patents on digital identity (see Table 2). The table underscores the evolving nature of the sector, with newer players like Alibaba/Alipay emerging prominently. However, Microsoft's enduring involvement, as evidenced by its patent filings over the years, highlights its foundational role and continuous influence in the advancement of digital identity solutions. In addition to the 2006–2010 period, Microsoft has previously led the field with 15 patents from 2001 to 2005 and 17 patents from 2011 to 2015. Its contributions remain significant in shaping the landscape of digital identity technologies.

Given that Kim Cameron was on staff as Chief Architect of Identity at Microsoft between 1999 and 2011, he may well have been the catalyst for this busy period of innovation in which the company filed 78 of its 117 total digital identity patents.

During this period, in 2005, Cameron published 'The Laws of Identity', a landmark paper that established seven foundation laws for digital identity which would shape future research in the field. These laws are [15]:

1. *User Control and Consent*-technical identity systems must only reveal information identifying a user with the user's consent.
2. *Minimal Disclosure for a Constrained Use*-the solution that discloses the least amount of identifying information and best limits its use is the most stable long-term solution.
3. *Justifiable Parties*-digital identity systems must be designed to limit the disclosure of identifying information to parties having a necessary and justifiable place in a given identity relationship.
4. *Directed Identity*-a universal identity system must support both "omnidirectional" identifiers for use by public entities and "unidirectional" identifiers for private entities, thus facilitating discovery while preventing unnecessary release of correlation handles.
5. *Pluralism of Operators and Technologies*-a universal identity system must channel and enable the interworking of multiple identity technologies run by multiple identity providers.
6. *Human Integration*-the universal identity metasystem must define the human user as a component of the distributed system integrated through unambiguous human-machine communication mechanisms offering protection against identity attacks.
7. *Consistent Experience Across Contexts*-the unifying identity metasystem must guarantee its users a simple, consistent experience while enabling the separation of contexts through multiple operators and technologies.

Microsoft's first patent in 2002, 'Systems and Methods for Distributing Trusted Certification Authorities' [26], dealt with distributing and updating trusted certification

Table 2 Digital identity patent history over time

Period									
[1996–2000]		[2001–2005]		[2006–2010]		[2011–2015]		[2016–2020]	
Assignee	**Count**	**Assignee**	**Count**	**Assignee**	**Count**	**Assignee**	**Count**	**Assignee**	**Count**
Diebold	**29**	Microsoft	**15**	Microsoft	**54**	Bundesdruckere	**23**	Alibaba/Alipay	**82**
IBM	**8**	Diebold	**14**	Aspire Digital	**30**	Microsoft	**17**	Oracle	**52**
L. Yubin	**3**	Huawei	**8**	Bundesdruckere	**15**	Cisco	**15**	Forcepoint	**38**
American Ex.	**2**	ETRI	**6**	GoDaddy	**15**	Tyfone	**15**	IBM	**30**
N*able Tech	**2**	SAP	**4**	Alcatel	**13**	Verizone	**15**	Mastercard	**29**
Certicom	**1**	Ntt Docomo	**3**	ETRI	**12**	Zhuhai UPT	**13**	Bundesdruckerei	**28**
Critical Path	**1**	Toshiba	**3**	Beijing FCT	**12**	Go Daddy	**11**	Bank of Amer.	**25**

authorities to computer systems and users-a relevant challenge as digital identity ecosystems are constructed worldwide.

Another of Microsoft's patents in 2007, 'Authentication for a commercial transaction using a mobile module' [31], dealt with authorisation and payment of an online commercial transaction where the identity provider and the payment provider may be different network entities. Microsoft would have eight patents on digital identity transactions, but payment providers would later take over innovation in this area (see below).

Microsoft was also active in gaming. The initiative to transition gaming users to an online platform resulted in eight gaming digital identity infrastructure patents. Finally, Microsoft also had 3 patents targeting digital identity privacy and a further 4 targeting digital identity trust.

2.3 Bundesdruckerei (2011–2015)

Bundesdruckerei specialises in secure identity technologies, including products which protect sensitive data, communications, and infrastructures. The solutions are rooted in the secure identification of citizens, customers, employees, and systems in the digital and real world.

After 4 years of planning, Bundesdruckerei launched the German identity card in November 2010-and has provided Germany's ID cards and passports, residence permits, office ID cards, visas, and driving licences.

The majority of Bundesdruckerei's 66 patents in digital identity are related to its electronic identity technology, with 34 patents involving tokens and 20 patents involving attribute management [21, 32, 39].

By January 2018, more than 53 million new-generation electronic identity cards and 8 million electronic residence permits were in circulation-and, according to the Government, all 61 million German citizens (over 16 years old) would have a German national identity card by 2020. Bundesdruckerei has consistently contributed to digital identity innovation, with its first patent registered in 2008 [23]. Nineteen patents were created between 2015 and 2016, aligning with Germany's national ID card roll-out period.

2.4 Oracle corporation (2016–2020)

Oracle has been a premier database provider for a number of decades. Additionally, its Oracle Identity Manager (OIM) and Oracle Access Manager (OAM) products-along with its acquisition and integration of role management (Bridgestream/Oracle Role Manager [ORM]) and risk-based authentication (Bharosa/Oracle Adaptive Access Manager [OAAM])-have helped establish Oracle as a leader in enterprise identity [44, 45].

Oracle established its first digital identity patent (US-10063523-B2) titled 'Crafted identities' [16] in 2005, with a recent digital identity patent, 'Service Discovery for a Multi-Tenant Identity' [25] (US-2018041515-A1), filed in 2017.

In September 2001, the Liberty Alliance Project was establish as an organisation to provide standards, guidelines, and best practices for identity management [13, 66]. It grew to more than 150 organisations and Oracle Corporation was on the management board. The alliance released frameworks for federation, identity assurance, and identity web services as well as the Identity Governance Framework-which was contributed to the alliance by Oracle Corporation in February 2007 and later released to the public in July 2007. The Identity Governance Framework defined how identity-related information was used, stored, and propagated, using protocols such as LDAP, Security Assertion Markup Language, WS-Trust, and ID-WSF.

Interestingly, 52 of Oracle's 61 patents in digital identity were created during the period 2016–2020-well after the Kantara Initiative took over the work of the Liberty Alliance in 2009. The fact that 18 of the 30 patents established in 2017 were on the topic of "multi-tenancy"-and that 22 mentioned "Cloud" and 14 mentioned "Security"-may indicate that Oracle is increasing its presence in the cloud space.

2.5 Cognitive scale (2016)

Cognitive filed 17 patents involving a combination of three technologies: digital identity, blockchain, and cognitive computing. Cognitive computing describes technology platforms that are based on artificial intelligence and signal processing, encompassing one or more of the following technologies-machine learning, reasoning, NLP, speech recognition and vision (object recognition), human-computer interaction, dialog, and narrative generation [22, 33].

2.6 Bank of America (2018)

Bank of America and a group of high profile companies, including Visa, JP Morgan Chase, Symantec, and others, formed the Better Identity Coalition to encourage the United States to address remote identity proofing and identity verification solutions. The coalition established five initiatives as follows [29]:

1. Prioritise the development of next-generation remote identity proofing and verification systems.
2. Change the way the US uses social security numbers.
3. Promote and prioritise strong authentication.
4. Pursue international coordination and standardisation of identity systems.

5. Educate consumers and businesses about better identity solutions.

As well as being a founder of the Better Identity Coalition, the Bank of America was also a member of the Liberty Alliance consortium which, as previously mentioned, established an open specification for identity management [52]. Bank of America's primary focus, however, has been blockchain-related digital identity, with 26 of the organisation's total 38 patents devoted to the subject. In 2018 O'Neal asserted that Bank of America had the most patents pending that involved blockchain implementation [43].

Broadly speaking, it is fair to say that user trust, transaction cost reduction, and enhanced privacy are among the most important perceived contributions from blockchain. Zheng et al. provided a list of four specific benefits from blockchain as follows [67]:

1. Decentralisation: no third-party authentication leveraged by P2P networks is key to reduce operating costs and processing bottlenecks.
2. Persistence: users in the network are constantly validating and processing transactions, making it difficult for malicious entities to tamper with the chain.
3. Anonymity: each transaction is encrypted in a way that the sender and receiver identities can only be retrieved by the parties involved in the exchange.
4. Auditability: timestamped transactions allow each participant to keep track and have visibility of the history of all transactions that have happened since the creation of the network.

2.7 Alibaba and Alipay (2019–2020)

Alibaba and Alipay have played an important part in the establishment of Chinese digital infrastructure. China has begun digitising the national ID card in an attempt to improve convenience and provide ease of access to internet-based services. Every national citizen is required to register for a Resident Identity Card upon reaching the age of 16. In fostering a digital identity ecosystem, China has collaborated with commercial entities such as Alipay with which the official digital government ID is currently integrated [28].

Alipay, a digital wallet provider, is a processor of payment transactions-and with over 500 million monthly active users, it processes more transactions than Amazon, Walmart, and eBay combined [18, 57]. Alipay is used for "restaurants, taxis, school fees, cinema tickets and even to transfer money to each other" [12].

As the focus on digital identity ecosystems shifts from identity to proof, transactions (financial and otherwise) have become a priority [6]. Consequently, digital wallet providers such as Alipay are at the forefront of the digital identity revolution. In recent years, as a result of China's ID programme or because transaction processing companies are playing catchup, a significant number of patents have been filed by companies such as Alipay, Mastercard, Visa, and Yoti-as well as a number of banking organisations such as Bank of America (as seen in Table 2). One hundred forty-four of Alibaba/Alipay's 150 total patents have been attained since the start of 2018.

Fifty-four of the patents relate to blockchain technology with digital identities-a trend shared by banking organisations such as Bank of America as similar institutions make use of blockchain in a digital-identity-based transactional world. Alipay has designed a process which establishes an intelligent transactional contract between a client and their card, the online service and invoice owner, and executes it using blockchain as the base-an approach that may become more widely used as digital identity ecosystems mature globally.

Of the remaining patents, 17 addressed traditional concerns of authentication and authorisation. 3 more patents were for digital-identity-based invoicing (all involving blockchain), 3 addressed data management, and 1 addressed communication.

2.8 Mastercard (2016-2020)

Mastercard has 62 total patents on digital identity with 9 involving payments and 8 involving authentication. The remainder are spread across a varied portfolio of technologies including contactless environments, national identity and election verification, biometrics, and tokenisation.

Mastercard recently partnered with Microsoft to address the risk of fraud and complexity and to design a service that would allow individuals to enter, control, and share their identity data on multiple devices. A key focus for Mastercard is to address the border barriers that have become apparent with other solutions as they attempt to move beyond implementation within a single country [38].

Reviewing the most active companies in digital identity patent filings offers a broad overview of industry trends and market leadership, aiding in strategic planning, innovation identification, and competitive analysis. This macro-level approach adds additional insight into the use of NLP for patent analysis. In the next section, we will outline the specific method we have used to delve deeper into the specifics of patent content, uncovering intricate patterns and developments in the digital identity field.

2.9 IBM (2016-2020)

IBM founded in 1911 and based in Armonk, New York, IBM has been a key player in the technology and consulting sectors, driving innovations in computer systems, software, and services. With a portfolio of 63 digital identity patents, IBM experienced a significant increase in patent activity between 2016 and 2020, registering 30 patents, highlighting a focus on Artificial Intelligence and blockchain.

The company's leadership in blockchain is particularly notable, with 16 digital identity patents in this area, supported by the IBM Blockchain Platform which promotes enterprise adoption of blockchain [48, 53]. IBM's patents also frequently address crucial concepts like "access," "security," "privacy," "auth," and "trust," all core terms that suggest a commitment to incorporating digital identity at a fundamental level.

2.10 Recent activity (2021-2023)

Since the beginning of 2021, substantial activity has been observed, with 2067 digital identity patents filed out of a total of 6157 in the dataset. An examination of the most active companies during this period, as outlined in Table 3, highlights the significant contribution of Chinese firms to innovation within this domain, with four of the top five companies being based in China and nine out of the top twelve as well.

The notable activity of Chinese companies in recent digital identity patent filings may be attributed to several factors, including strong government incentives for innovation, faster processing time, substantial investment in research and development, and a strategic focus on intellectual property creation to position China as a global technology leader. The rapid growth of the Chinese economy and the expansion of its tech sector also play important roles in this trend. Additionally, a broader societal emphasis on technological contribution and innovation might be playing a pivotal role in driving digital identity advancements.

Table 3 Companies filing most patents containing term "digital identity" after 2020

Company	Patents
Alibaba/Alipay	65
Industrial and Commercial Bank of China	57
Ant Blockchain Technology (Shanghai) Co.	47
Hunan University	30
Mastercard International Inc.	28
China Academy of Information and Comm.	27
NetEase (Hangzhou) Network Co.	22
China Unicom	20
Wells Fargo Bank, N.A.	18
Newland(Fukian) Public Service Co.	18
Microsoft Corporation	16
China Telecom Corporation Limited	15

Source: primary patent set (see Section 3)

While this observation is outside the scope of this article, further investigation to fully understand the underlying dynamics and implications of China's recent dominance in digital identity patent filings is warranted. Future research should explore the policies, market conditions, organisational strategies, and socio-political mechanisms, such as the social credit system, that contribute to this high level of activity.

3 Methodology

Patent archives are a key resource for mapping technological trajectories, offering insights into technological progress and future trends [41, 64]. Advanced analytical methods enable scholars and professionals to glean valuable information from these documents, thereby enriching research and development approaches [65].

In recent times, patent data mining has been improved further by using enhanced NLP approaches, which speed up the textual analysis and offer insights embedded within the structure of the text [35]. This additional process enables a more comprehensive understanding of the inherent relationships between the patents and the ability to identify emerging technological trends [62, 63].

This approach has garnered a significant following in the academic sphere, with a mounting body of work utilising text-mining approaches to delineate the historical progression of technology and to keep a vigilant eye on emerging trends (Fig. 1) [17, 62–64].

3.1 Data sources

Patents stand as a cornerstone of the intellectual property sphere, offering inventors a legally protected monopoly over the use, manufacture, and sale of their innovations for a set duration. This protection hinges on transparent disclosure, wherein the details of the invention are shared publicly.

Our investigation leverages data primarily harvested from the robust Google Patents database. This resource combines over 120 million patent publications that emanate from more than 100 patent offices worldwide. Access to the repository is streamlined through the Google Patents portal, which researchers and academics can access freely at https://patents.google.com.

3.2 Data extraction

In constructing the search strategy for the patent dataset, 'Digital Identity' was selected as the primary query, with the term encapsulated in quotation marks to filter for

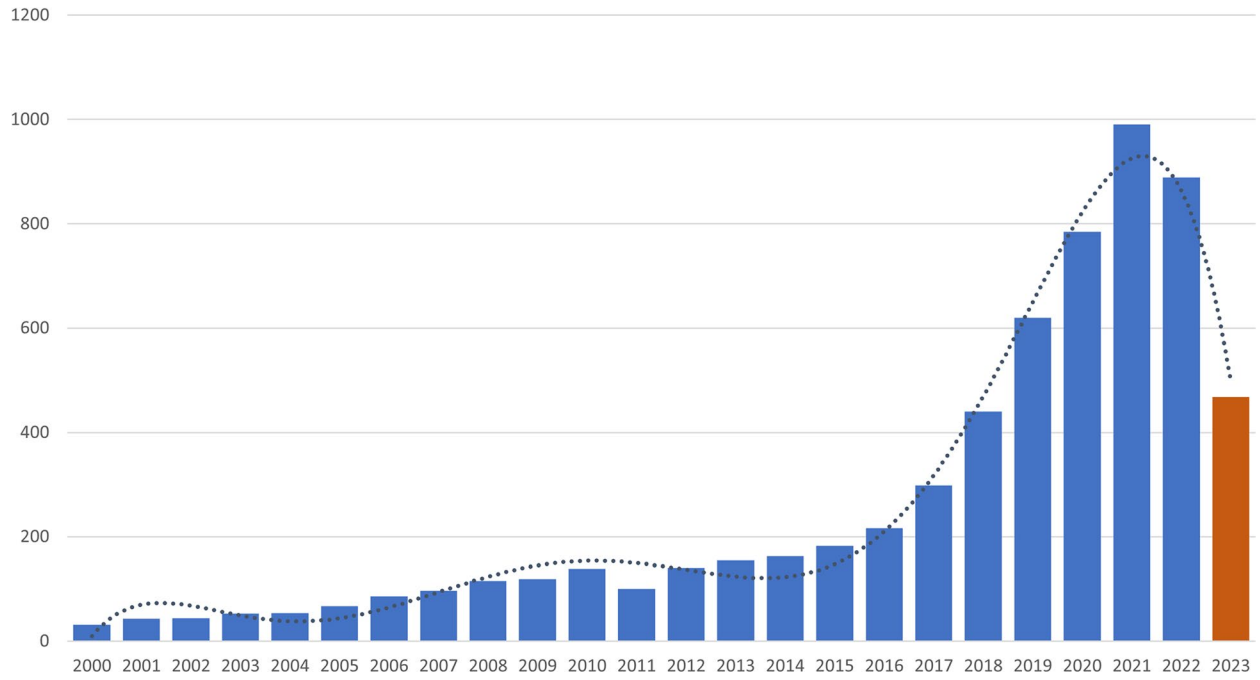

Fig. 1 Digital identity patents by year

patents explicitly containing these words consecutively. Initial considerations extended to including related domains like 'Authentication' and 'Authorization' which could be included easily given the automated pipeline, yet a preliminary review indicated their substantially higher patent volumes with many tangentially related or outside the study's digital identity framework were diluting and skewing the core results unnecessarily. Consequently, to preserve the dataset's focus and relevance, the search was confined to patents directly citing 'Digital Identity', ensuring inclusion was strictly aligned with the study's parameters.

This query was used in conjunction with the Google Patents website, as mentioned above, and resulted in the following URL:

https://patents.google.com/?q="digital+identity"

The timeline parameter was deliberately left unrestricted to encompass the entire gamut of patents accessible in the database, thereby not omitting any potential datasets from earlier periods. The result set was then downloaded using the download facility provided using the link:

"Download/Download (CSV)"

This approach yielded a substantial collection of 6157 patents pertaining to 'Digital Identity'. The resulting dataset was downloaded in CSV format, a choice dictated by the ease of manipulation and analysis this format provided.

3.3 Data sampling

The methodology adopted for sampling pursued various strategies including:

1. A comprehensive approach encompassing analysis conducted on the full sample devoid of strata.
2. Instituting a stratification spanning five-year intervals over the preceding 25 years, fostering a period-wise analytical perspective.
3. A more granular stratification predicated on yearly divisions over the most recent five years to glean nuanced insights from the data.

3.4 Pre-processing

In this stage of the study, meticulous data pre-processing steps were undertaken to ensure the robustness and reliability of the ensuing analysis. The steps involved the following:

1. *Data Cleaning*-preliminary actions were deployed to sanitise the text data to a state conducive to further processing. The detailed procedures encompassed:

   (a) Text Normalisation: Every alphabetic character was transformed to its lower case counterpart to maintain uniformity and reduce complexity.

(b) Punctuation and Numerals Elimination: All punctuation marks, special characters, and numerals were systematically removed to restrict the analysis to textual data.
(c) Redundant Successive Word Removal: Instances, where words were repeated successively, were identified and rectified by retaining a single iteration of each word.

2. *Data Reduction*-to streamline the dataset and enhance its manageability, the following measures were undertaken:

(a) Focused Retrieval: The overarching patent database was narrowed down by filtering entries explicitly associated with the term 'digital identity'.
(b) Dimensionality Reduction: An acute reduction in the dimensionality was attained by considering only the POS-tagged triples that featured 'digital identity' as either the subject or the object.

3. *Data Integration*-the assembly and management of the data incorporated the following systematic approach:

(a) Temporal Stratification: Data was stratified into discrete partitions, each spanning 5 years. Subsequently, these strata were merged to facilitate a comprehensive multi-temporal analysis.

4. *NLP Pre-processing*-leveraging NLP techniques, we prepared the data using the following procedural steps:

(a) Tokenisation: The corpus was disintegrated into individual word tokens, laying the groundwork for further NLP techniques.
(b) Lemmatisation: Words were lemmatised to their root forms to ensure grammatical consistency and to reduce the morphological variations in the dataset.
(c) Stop Word and Redundant Word Removal: To eliminate noise in the data, we removed stop words and recurrent superfluous words identified across all patent texts.
(d) POS Tagging: This involved grammatically categorising words into their respective parts of speech, thus aiding syntactic analysis.

5. *Feature Engineering*-Feature Selection and Extraction: predicated on the results from the cluster analysis (detailed below in Section 3.5), the most pertinent features were selected. Furthermore, adjacent triples related to each feature were analysed to garner additional insightful features. Data Splitting:
6. *Dataset Partitioning*-a pragmatic approach to validating the model involved segregating the data into training and test sets. The test dataset comprised the initial n records derived from the delineated strata, facilitating both the training and validation of the model through a pragmatic lens.

3.5 Data analysis

The extensive dataset assembled from patent data underwent meticulous collation and synthesis through the use of Jupyter Notebook, renowned for its versatility in facilitating dynamic data analysis. This processing and analysis were guided by a rigorously conceived pipeline, illustrated as follows:

1. *Analysis of Word Frequencies*: The initial step involved a detailed scrutiny of word frequencies, a strategy aimed at spotlighting prominent terms and discerning potential thematic trends pervasive in the dataset.
2. *Recent Patents Analysis*: Following this, attention was pivoted to scrutinise patents filed from 2020 onwards, a manoeuvre designed to cast light on contemporary tendencies and emergent focal areas within the digital identity landscape.
3. *Word Frequencies Revisited*: This phase reincorporated further analysis of word frequencies to refine the previously gleaned insights, thereby fostering a robust comprehension of the terminologies utilised frequently in recent publications.
4. *Cluster Analysis via K-Means and LDA*: The analytical process then leveraged K-means and latent Dirichlet allocation (LDA) algorithms for cluster analysis-a tactic purposed to identify inherent groupings and delineate associative patterns in the dataset, ultimately offering a structured insight into the patent landscape.
5. *POS Tagging and Triple Extraction*: The culmination of the pipeline was a stage dedicated to intricate syntactic examination through POS tagging, paired with triple extraction, thereby enabling a deep-seated understanding of the relational dynamics and semantic structures present in the patent data.

3.6 Validation

Ensuring the reliability and accuracy of the generated data involved a series of validation steps characterised by:

1. *Test Case Development*: Devised a series of test cases to critically evaluate the correctness of data produced at each significant juncture of the pipeline, thus ensuring a high standard of data integrity.
2. *Full-Scale Data Sample Processing*: Undertook a complete processing of select data samples to identify and rectify potential issues in the pipeline, thereby safeguarding against errors in the later stages of analysis.
3. *Coherence Testing*: Conducted exhaustive coherence testing on cluster-ready datasets to affirm the optimal number of clusters. Coherence testing was conducted with cluster numbers from one and 20, and the number with the highest coherence metric was used to output cluster groupings.
4. *Cross-Validation*: Applied cross-validation techniques between stratified sample datasets. This ensured the model was proficient at generalising new unseen data.
5. *Expert Review*: Enlisted domain experts to review the analytical results and provide feedback. This fostered a refinement of the analysis based on expert insights.
6. *Sensitivity Analyses*: Conducted sensitivity analyses to evaluate the robustness of the analysis results by assessing how variations in input parameters affected the outputs.
7. *Comparison with Benchmark Datasets*: Aligned the results with benchmark datasets or established truths in the field to discern the accuracy and reliability of the findings.
8. *Error Analysis*: Built error validation steps into the pipeline to detect and log inconsistencies encountered. This rigorous analysis provided a conduit to categorise and understand potential sources of errors, fostering enhancements in the predictive accuracy of the model.

3.7 Tools and software

In this research, we crafted a data mining pipeline to analyse patent data through a series of NLP techniques. Leveraging the synergies of Visual Studio (utilising C#) and Python, we employed a variety of supplementary modules, each serving a distinct, pivotal role in different phases of the data mining process. Below, we detail how each tool was utilised:

1. *NumPy* [27]: This module played a crucial role in handling large arrays of numerical data intrinsic to the patent datasets. It facilitated the efficient execution of complex mathematical and logical operations, acting as the backbone for numerical data manipulation and serving as a computational engine for other Python libraries used in the research.
2. *Pandas* [36]: Utilised predominantly in the preliminary stages of data processing, Pandas was essential in cleaning, filtering, and pre-processing the structured patent data, transforming raw data into a usable format that facilitated seamless data analysis downstream.
3. *SpaCy* [29]: As a primary tool for NLP tasks, SpaCy was utilised for tokenising patent texts and performing named entity recognition, thereby structuring the unstructured text data into a format amenable for deeper analysis.
4. *Scikit-learn (sklearn)* [46]: Our pipeline leveraged Scikit-learn extensively for implementing machine learning algorithms and statistical models. Through Scikit-learn, we undertook predictive data analysis, extracting underlying patterns and trends from the patent data, which was integral in identifying relationships and gaining insights.
5. *SciPy* [61]: Employed in tandem with NumPy, SciPy enabled high-level computations such as linear algebra and optimisation that were crucial in processing and analysing large volumes of patent data.
6. *Gensim* [49]: In the latter stages of data analysis, Gensim facilitated vector space and topic modelling, enabling the identification of thematic patterns and semantic relationships within the patent texts, which were paramount in understanding the underlying topics and trends in the patent data.
7. *Natural Language Toolkit (NLTK)* [10]: This toolkit was engaged for an array of linguistic data analysis tasks, such as text classification and tokenisation, thereby aiding in the segregation and categorisation of patent data into distinct classes for more focused and detailed analysis.
8. *Py2neo* [51]: Towards the end of the data pipeline, we utilised Py2neo for interfacing with the Neo4j graph database, which was pivotal in storing, manipulating and analysing patent data in graph database formats, offering a sophisticated approach to visualise and analyse relationships and data patterns.

3.8 Summary

This methodology delineated a rigorous and multifaceted approach to mining patent data through a well-structured pipeline incorporating NLP techniques and data analysis strategies. The adopted pathway was characterised by successive stages of data preprocessing, where the raw data underwent cleaning,

reduction, and integration processes to prepare it for subsequent analytical undertakings.

Utilising an array of NLP techniques such as word frequency analysis, POS tagging and triple extraction, we facilitated a deeper exploration into the intrinsic patterns and prevalent themes in the patent dataset, thereby aiming to glean profound insights from the landscape of digital identity patents. The analytical prowess of K-Means and LDA in cluster analysis further assisted in discerning the underlying structures and relationships in the data, augmenting our understanding of the evolving trends and focal points in the patent domain.

Supplementing these efforts, we employed advanced data mining techniques, including pattern recognition and clustering, to refine the extracted information, thereby ensuring a nuanced interpretation of the dataset. The validation stage manifested a critical checkpoint in our methodology, incorporating a broad spectrum of validation techniques such as cross-validation and sensitivity analysis to ascertain the reliability and precision of our analytical outcomes.

As we transition to the results section, it is important to acknowledge the stringent validation measures put in place, including expert reviews and comparisons with benchmark datasets, which reinforced the reliability and accuracy of the forthcoming results.

4 Results and discussion

In this section, we present and interpret the findings from our comprehensive analysis of the patent dataset. This section is structured into two main subsections to provide clarity and depth to our data exploration.

Section 4.1, titled 'Feature Extraction', delves into the process of identifying and extracting features from the patents using NLP techniques. Using a longitudinal analysis of patent data across five distinct periods, we detail the methodologies employed to parse and analyse the textual content of patents, shedding light on how key terms and concepts-deemed as 'features' in this context-were systematically identified.

Following the extraction of features, Sect. 4.2, titled 'Feature Analysis,' takes a closer look at each feature and presents statistical analysis. We quantify the prevalence and significance of each term clustered with the extracted features, employing various statistical metrics to measure their impact and relevance within the dataset and then charting the trends of these terms over time. The combination of these quantitative metrics and trend analysis collectively highlights patterns, shifts, and emerging themes in the digital identity domain.

4.1 Feature extraction

The clustering analysis seen in Table 4 constructed using K-Means and LDA algorithms produced two sets of principal clusters outlining groupings of terms present across the 2020–2023 strata of patents. Of the clusters

Table 4 Key word cluster analysis [2020–2024]

Cluster	Words
Latent Dirichlet allocation (LDA) cluster analysis	
Identify management	identify, user, information, request, target, verification, certificate, service, authentication, signature
Cybersecurity	cryptographic, device, security, circuitry, pqc, detection, code, technique, mobile, risk
Computer technology	computer. program, memory, processor, storage, instruction, device, medium, method, software
Technological analysis	method, image, model, technology, analysis, information, enterprise, date, power, status
Service management	service, contract, transaction, payment, vehicle, element, provider, process, component, management
Network security	network, communication, protocol, message, security, module, card, node, architecture, wireless
Access control	device, method, access, server, data, authentication, equipment, user, control, management
User interaction	user, entity, input, time, behaviour, address, display, output, information, number
Data processing	data, block, chain, information, value, file, transaction, storage, module, processing
K-means cluster analysis	
Payment information	payment, card, transaction, credit, user, information, method, identity, data, merchant
Cryptography	cryptographic, pqc, technique, circuitry, data, communication, arrangement, mechanism, protocol, security
Network comm.	device, network, server, communication, user, mobile, data, computer, method, client
Computer technology	computer, memory, program, medium, processor, storage, instruction, method, device, data
Data processing	data, information, network, service, transaction, access, entity, process, module, communication
Identify verification	identity, information, authentication, user, verification, data, service, certificate, method, request
User authentication	user, information, data, request, service, entity, access, target, authentication, interface
Blockchain	block, method, chain, data, transaction, information, title, storage, processing, management

identified, identity management, transaction and payments, network security, and blockchain represented the intersection between the two algorithms (Table 5).

Identity management (see Sect. 4.2.2) in digital identity refers to the processes and policies for creating, maintaining. and using digital identities. It involves managing user identities, credentials, and access permissions through digital certificates and biometric verification technologies. This feature is underpinned by a broad spectrum of academic explorations that spotlight the pivotal role of sound identity management infrastructures in contemporary digital ecosystems [9].

The transaction cluster (see Sect. 4.2.3) signifies the increasingly intertwined nature of financial transactions and digital identity systems. This reflects a growing scholarly consensus on the necessity of secure and reliable identity systems in financial transactions. The evolving nature of digital transactions necessitates robust identity verification mechanisms, a topic explored in depth in recent financial literature [6, 7, 30].

The network security (see Sect. 4.2.4) cluster represents both new and customised security technologies applied within the digital identity domain-highlighted in current research focusing on advanced cryptographic measures to fortify digital identity frameworks [11].

Lastly, the blockchain cluster (see Sect. 4.2.6) relates to and reaffirms the emerging consensus regarding the advantageous integration of blockchain technologies within digital identity management solutions-a perspective steadily gaining ground in recent academic discussions [2, 8, 24, 56].

In order to validate the integrity of the identified clusters, coherence testing was conducted. With coherence values of 56% and 58% for LDA and K-means, respectively, the clusters provide a strong foundation for patent analysis.

In addition to these clusters, which were established in an unsupervised manner, we also added the biometric cluster (see Sect. 4.2.5) due to several biometric terms being present during a number of supervised data management steps. Blockchain in digital identity uses the technology's decentralisation, immutability, and transparency to manage and secure digital identities. It allows individuals to control their personal information, enhances data security, and reduces identity theft risk.

Finally, core concepts-common within the digital identity domain (see Sect. 4.2.1)-were also analysed for high-level digital identity trends. In the next section, we will analyse each of these identified features in detail.

4.2 Feature analysis

In the initial phase of this research, the objective centred around identifying and extracting thematic clusters within the patent landscape from 2020 to 2023. However,

Table 5 POS tagging key feature relationships

Identify	Count	Payment	Count	Crypto	Count	Blockchain	Count
⊢ user	**1084**	⊢ transaction	**105**	⊢ data	**121**	⊢ data	**398**
⊢ identity	**865**	⊢ user	**99**	⊢ key	**88**	⊢ information	**339**
⊢ claim	**845**	⊢ information	**61**	⊢ system	**79**	⊢ user	**323**
⊢ information	**797**	⊢ data	**59**	⊢ user	**75**	⊢ claim	**271**
⊢ data	**733**	⊢ device	**59**	⊢ claim	**73**	⊢ chain	**254**
⊢ system	**492**	⊢ blockchain	**52**	⊢ information	**69**	⊢ identity	**241**
⊢ blockchain	**475**	⊢ claim	**52**	⊢ transaction	**64**	⊢ field	**219**
⊢ verification	**451**	⊢ contract	**45**	⊢ signature	**64**	⊢ contract	**214**
⊢ time	**449**	⊢ party	**42**	⊢ device	**53**	⊢ system	**194**
⊢ request	**401**	⊢ identity	**41**	⊢ value	**52**	⊢ node	**185**
⊢ device	**394**	⊢ system	**41**	⊢ encryption	**47**	⊢ plurality	**175**
⊢ method	**387**	⊢ request	**36**	⊢ algorithm	**44**	⊢ network	**168**
⊢ network	**361**	⊢ currency	**34**	⊢ ledger	**41**	⊢ transaction	**155**
⊢ party	**339**	⊢ service	**33**	⊢ identity	**40**	⊢ request	**142**
⊢ certificate	**338**	⊢ order	**31**	⊢ operation	**35**	⊢ certificate	**139**
⊢ plurality	**335**	⊢ server	**30**	⊢ cryptographic	**34**	⊢ characteristic	**134**
⊢ order	**332**	⊢ wallet	**29**	⊢ process	**34**	⊢ aspect	**132**
⊢ number	**324**	⊢ security	**27**	⊢ party	**34**	⊢ order	**126**
⊢ signature	**317**	⊢ money	**26**	⊢ input	**33**	⊢ verification	**122**
⊢ service	**317**	⊢ bank	**26**	⊢ instance	**33**	⊢ operation	**120**
⊢ authentication	**307**	⊢ merchant	**25**	⊢ block	**31**	⊢ problem	**119**

to facilitate a more comprehensive analysis of emerging trends, the study's temporal scope was expanded retrospectively to encompass the period starting from 2010. Concurrently, to ensure the data's integrity and minimise distortion in the analytical outcomes, the study's timeframe was deliberately confined to the end of 2022. This limitation was imposed to counteract the potential anomalies introduced by the COVID-19 pandemic, which could have disproportionately influenced the nature and direction of technological advancements during this period.

The extended analysis period, spanning from 2010 to 2022, allowed for a thorough investigation into the developmental trajectories and shifts in technological emphases within the chosen domain. This study employed a robust quantitative methodology, leveraging TF, DF, and TF-IDF metrics.

4.2.1 Core components

Keyword analysis, as seen in Table 6 and validated through manual inspection, combined with POS tagging analysis (see Fig. 2), shows several core terms, which rank highly in each 5-year period since 2000. These core terms, 'User,' 'System,' 'Device,' 'Information,' and 'Data,' demonstrate varying degrees of frequency and significance in the analysed patents.

'User' emerges as a central term with the highest frequency, underscoring its importance in the context of digital identity. This increase may be partly due to the user-centric identity paradigm established in the 2000s [59] responding to heightened global concerns about privacy and data security, where individuals have greater control over how and with whom they share their personal information on digital services.

Term	Type	μ	σ	δ	ϵ
User	TF	2.69	0.26	40.83	7.05
	DF	94.03	1.54	1.31	1.39
	TF-IDF	0.16	0.04	9.24	18.42
System	TF	2.54	0.23	-26.56	5.97
	DF	97.74	5.69	27.10	2.99
	TF-IDF	0.07	0.19	-98.07	139.95
Device	TF	2.53	0.41	54.53	10.8
	DF	96.73	5.62	26.25	2.73
	TF-IDF	0.07	0.10	-86.54	68.81
Information	TF	1.94	0.41	54.28	10.7
	DF	96.83	5.73	27.25	3.71
	TF-IDF	0.06	0.11	-91.34	113.93
Data	TF	2.52	0.74	109.29	11.3
	DF	96.31	6.20	29.96	4.04
	TF-IDF	0.08	0.12	-84.53	99.15

μ = mean %, σ = std dev %, δ = change %, ϵ = error %

Table 6 Digital identity patent key word frequencies

Period									
[2000–2004]		**[2005–2009]**		**[2010–2014]**		**[2015–2019]**		**[2020–2023]**	
Word	**Count**	**Word**	**Count**	**Word**	**Count**	**Word**	**Count**	**Word**	**Count**
system	**49**	system	**68**	device	**126**	user	**315**	data	**313**
data	**34**	user	**64**	user	**111**	data	**306**	information	**235**
method	**33**	information	**49**	system	**102**	device	**275**	user	**229**
information	**30**	device	**46**	data	**75**	system	**256**	device	**195**
user	**30**	method	**45**	information	**66**	information	**203**	system	**189**
device	**26**	data	**41**	network	**65**	identity	**176**	identity	**186**
server	**21**	network	**39**	method	**59**	method	**152**	method	**159**
service	**19**	service	**39**	service	**54**	service	**133**	service	**107**
network	**18**	identity	**36**	identity	**49**	network	**132**	network	**103**
identity	**15**	authentication	**30**	communication	**42**	authentication	**95**	block	**99**
computer	**14**	computer	**25**	module	**38**	access	**94**	chain	**87**
access	**14**	server	**24**	access	**38**	server	**94**	request	**81**
message	**14**	access	**23**	server	**35**	request	**90**	authentication	**80**
communication	**14**	communication	**23**	computer	**35**	computer	**84**	access	**72**
certificate	**12**	security	**21**	authentication	**35**	transaction	**81**	transaction	**69**
number	**11**	provider	**17**	request	**28**	communication	**80**	module	**68**
authentication	**11**	mobile	**16**	mobile	**28**	block	**78**	verification	**66**
client	**11**	request	**16**	processing	**26**	security	**61**	target	**65**
card	**10**	card	**15**	message	**24**	module	**60**	certificate	**64**
security	**10**	certificate	**14**	vehicle	**23**	entity	**59**	storage	**64**

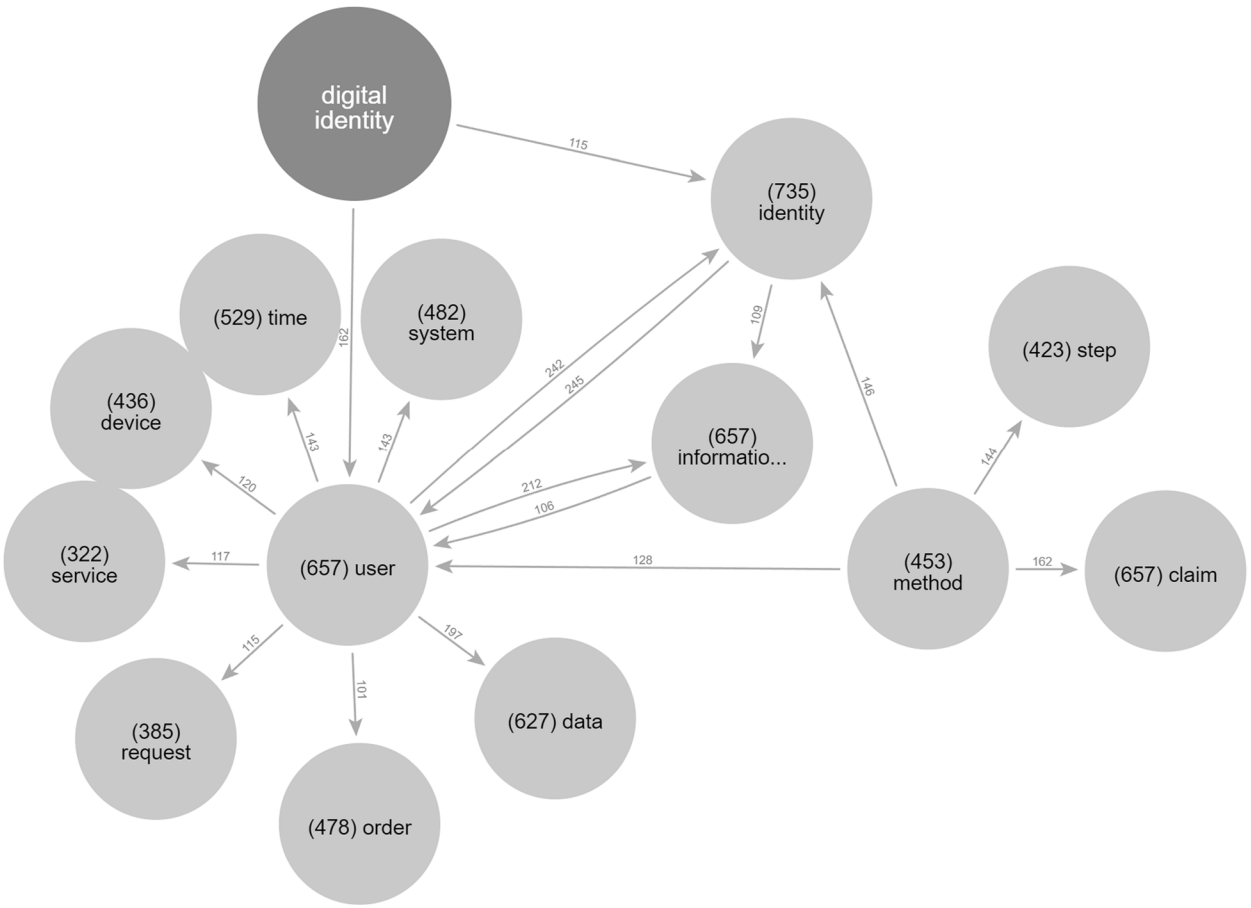

Fig. 2 Core relationships ($n > 300, r > 100$)

'Data' shows significant variability and growth in frequency and change, pointing to its evolving and increasingly crucial role in digital identity patents. Increases in 'Data' terms may indicate the evolution of digital-identity-enabling hardware to the digital identity systems themselves. In contrast, 'Information' exhibits a lower mean frequency, suggesting its more specific or limited application within this domain. The frequency of 'Information' is growing, but not as quickly as 'Data' (see Fig. 3). Collectively, these indicators provide insight into the maturity of digital identity solutions.

When considering DF, all terms display high mean values as expected, with positive movements over the analysed period, indicating widespread usage across numerous patents. Similarly, the significance values indicate the terms' very low relative importance, indicating that all terms are core and present in most digital identity patents in the dataset.

'User' has the highest significance of the core terms, while the remaining terms show high errors and negative changes in significance scores. This point indicates that while these terms are common, their distinctive significance within digital identity patents may be diminishing due to their generalised use.

Overall, subtle patterns present reflect a dynamic and evolving landscape within digital identity patents. The analysis highlights a shift towards focusing on users and data management while suggesting a potential oversaturation or generalisation of specific technical terms like 'System' and 'Device' in the patent literature.

4.2.2 Identity management

The term 'Identity' stands out for its high frequency and universal presence across documents, as indicated by its high mean values in both TF and DF (Fig. 4).

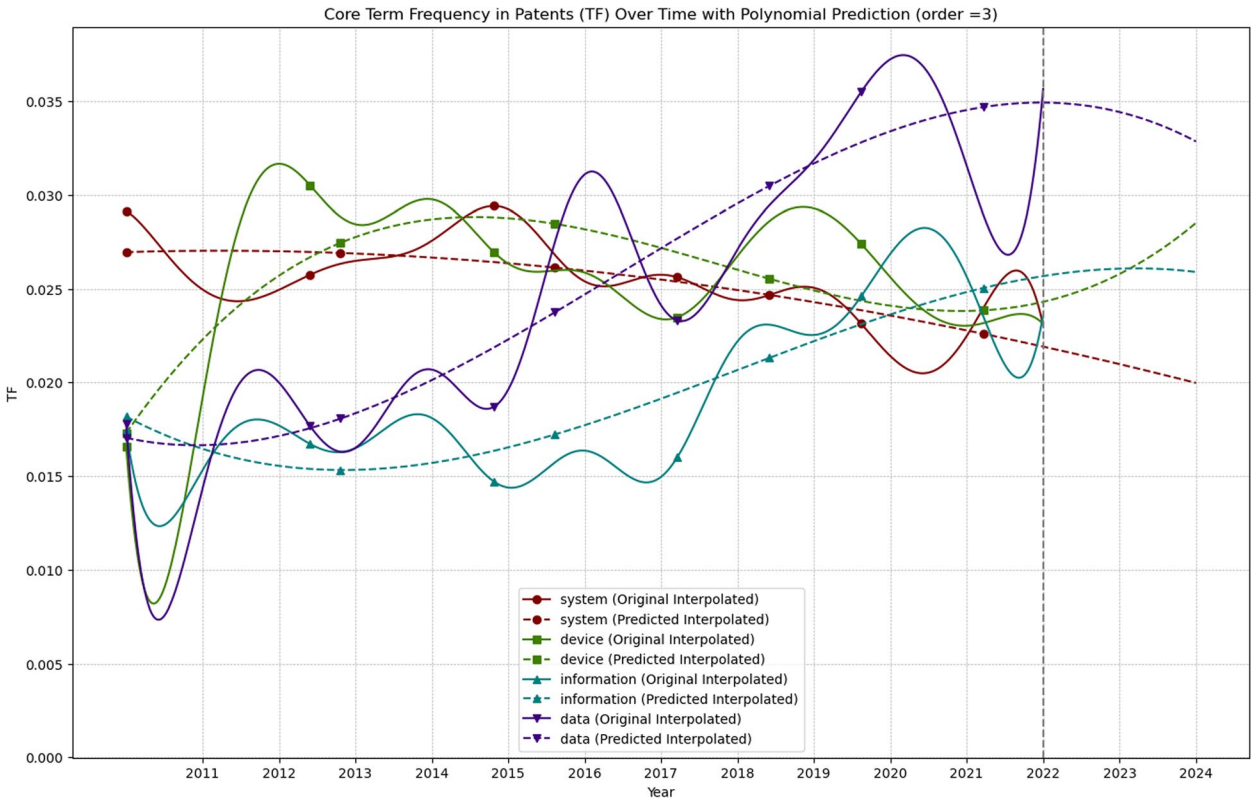

Fig. 3 Core TF in patents from 2010 to 2022

However, its ubiquity reduces its distinctiveness, as shown by its low significance scores. This suggests that while 'Identity' is a fundamental concept in digital identity patents, it is not a term that differentiates one patent from another due to its commonality. Core relationships between 'Identity' and other terms are visible in Fig. 5 and represent the most significant terms connected with identity present in the patent set.

Term	Type	μ	σ	δ	ϵ
Identity	TF	1.55	0.36	62.11	6.61
	DF	99.75	0.26	0.45	0.21
	TF-IDF	0.004	0.004	-40.57	81.89
Authenticat.	TF	0.93	0.10	-4.14	6.78
	DF	82.83	5.04	30.52	3.71
	TF-IDF	0.18	0.08	-67.67	22.62
Verification	TF	0.42	0.20	109.84	11.69
	DF	71.07	10.73	52.84	5.40
	TF-IDF	0.12	0.04	-29.61	24.68
Authorisati.	TF	0.21	0.08	30.44	36.32
	DF	58.26	5.51	31.78	5.53
	TF-IDF	0.11	0.05	-19.44	40.75

μ = mean %, σ = std dev %, δ = change %, ϵ = error %

The term 'Authentication' is notable for its consistent usage and distinctiveness compared with 'Identity' in the context of these patents. It exhibits a moderate frequency of occurrence and higher significance values, indicating that it is not only commonly used but also functions as a more specialised term within this domain. In contrast, 'Verification' displays a pattern of less frequent but diverse usage. Its significance scores reveal a moderate distinctiveness, underscoring its notable but varied presence in patents and emphasising its relevance in specific areas of the digital identity field.

'Authorisation' emerges as the least used term among those analysed, evidenced by its lowest TF mean. However, its significance scores indicate a certain level of distinctiveness. This suggests that although 'Authorisation' may not be as common as other terms, it holds significance in particular segments of digital identity patents.

These trends in the patent set provide further insight into digital identity trends. 'Authentication' is well-established but shows a decreasing trend, while

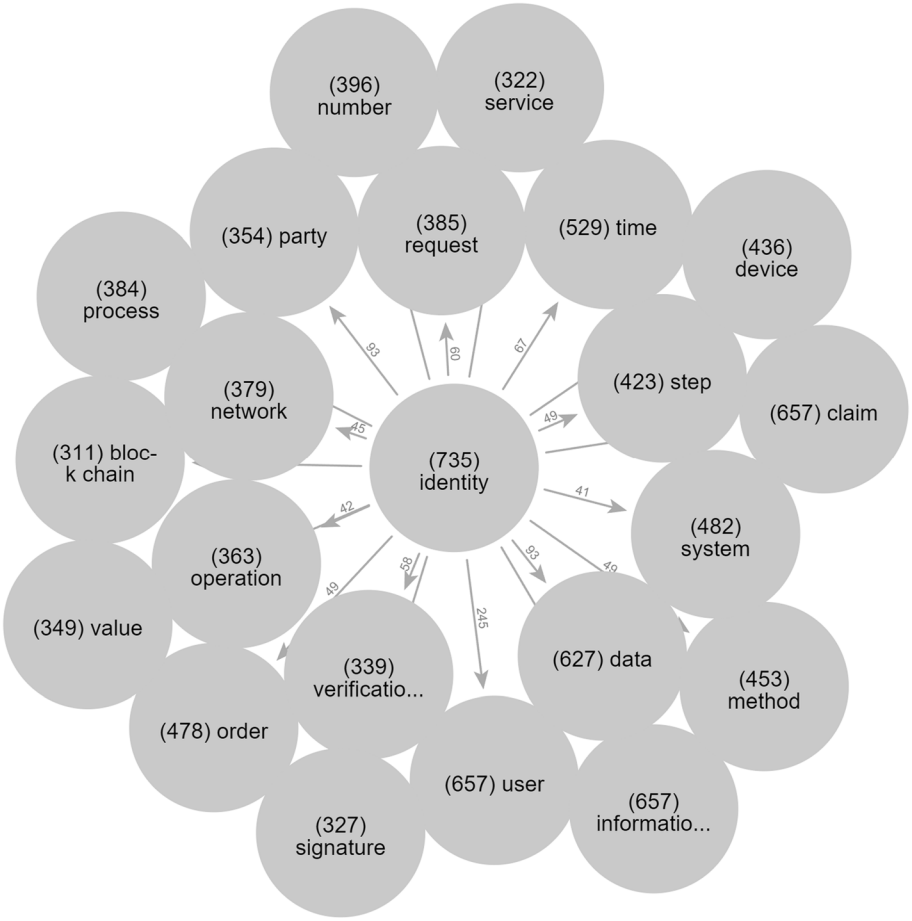

Fig. 4 Identity graph ($n > 300$, $r > 100$)

'Verification' sees the most growth, closely followed by 'Authorisation.' This suggests a shift towards credential-based digital identity solutions. This is validated by the fact that 'Authorisation' is linked to 'Authentication,' as illustrated in Fig. 5, a connection not observed with 'Verification,' which appears to be a separate process.

4.2.3 Transactions and payments

The term 'Transaction' has been present in the top 20 keyword periods since 2015 (see Table 6) and implies an increased focus on the financial aspects of digital identity verification. This is supported by the growing number of patent filings by payment firms and can be seen in current scholarly discussions [6, 7, 30]. Furthermore, K-means clustering (see Table 4) has identified a transaction feature which groups the terms 'Payment,' 'Card,' 'Transaction,' and 'Identity' together.

In this section, we omit the additional cluster terms 'Data' and 'User' as these are core terms which bridge multiple clusters and were discussed in the previous section. We also omit 'Credit' and 'Merchant' because their TF values were not statistically significant enough to warrant analysis.

Term	Type	μ	σ	δ	ϵ
Transact.	TF	0.65	0.19	23.31	19.53
	DF	60.29	10.83	53.08	8.44
	TF-IDF	0.31	0.09	-45.22	25.28
Card	TF	0.34	0.11	-54.84	9.60
	DF	69.57	8.64	-28.63	7.02
	TF-IDF	0.12	0.04	45.48	30.95
Payment	TF	0.25	0.07	-30.58	27.13
	DF	53.31	6.64	-2.21	6.93
	TF-IDF	0.15	0.05	-28.47	26.86
Wallet	TF	0.10	0.07	351.85	47.74
	DF	16.39	7.14	109.94	17.84
	TF-IDF	0.17	0.10	207.20	45.05

μ = mean %, σ = std dev %, δ = change %, ϵ = error %

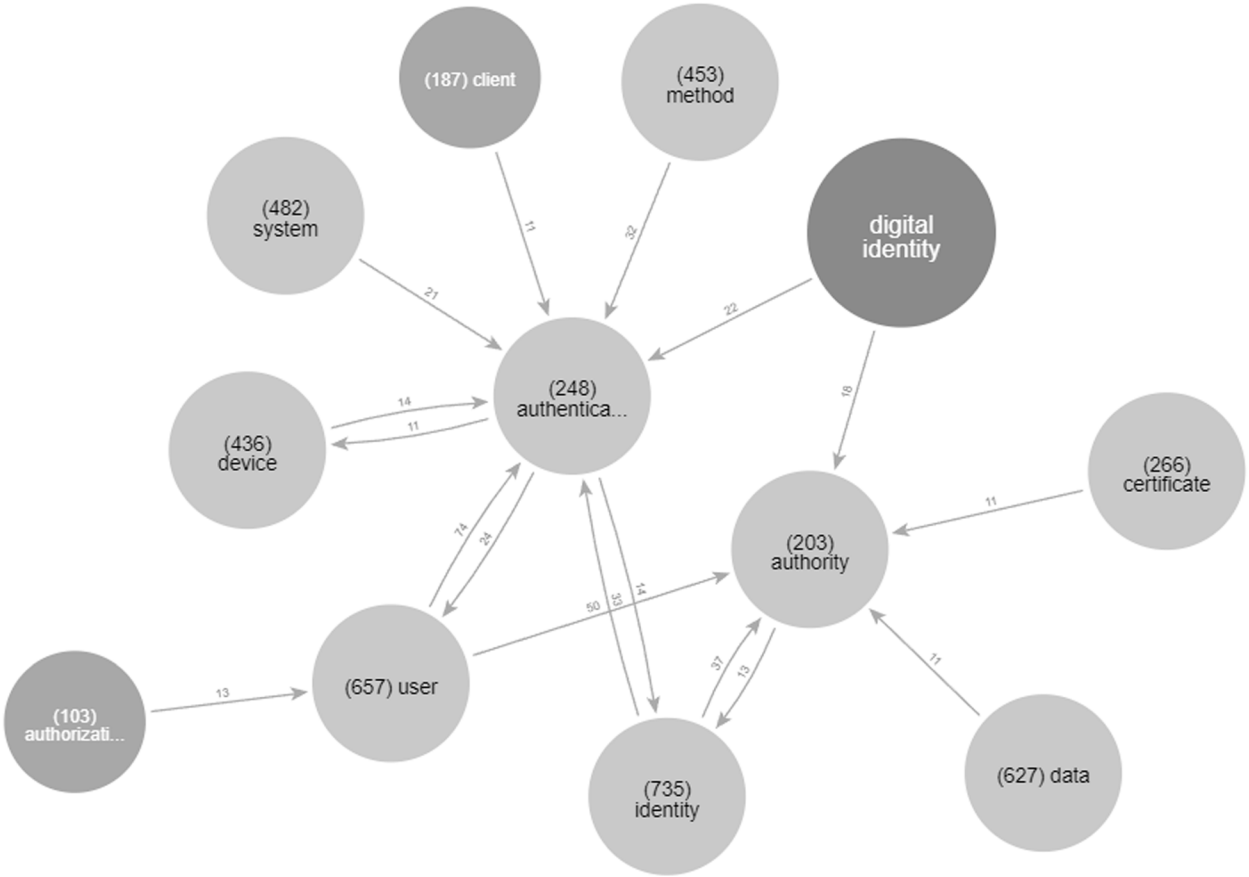

Fig. 5 Auth* relationship ($n > 100$, $r > 10$)

The term 'Transaction' exhibits a moderate usage frequency in digital identity patents, as reflected by its TF and DF values. This regular use suggests its integral role in the domain. However, the variability in its use, as indicated by the standard deviation, points to inconsistent application across various patents. Notably, TF and DF for 'Transaction' have shown an upward trend, signifying the growing importance of digital identity transactions, possibly influenced by advancements in blockchain and cryptocurrency technologies, which are integral to digital transactions and secure digital wallets. We discuss these later in Sect. 4.2.4–4.2.6.

Additionally, 'Wallet' is distinguished by a significant increase in both TF and DF metrics, and a high significance score reflecting an integration of digital identity solutions with financial technology and cybersecurity. The positive trends of both 'Transaction' and 'Wallet,' illustrated in Figs. 6 and 7 hint at this evolving landscape.

In contrast, 'Card' and 'Payment' are less frequently used than 'Transaction,' with 'Card' showing a higher, yet variable, presence in documents. Their moderately high significance scores suggest that while these terms are not as prevalent, they play a distinct role in the patents where they feature. However, the negative movement over the period for both terms indicates a possible decline in traditional card-based identity solutions.

The fact that 'Transaction' and 'Wallet' are trending up while 'Card' and 'Payment' are trending down suggests wide-ranging application of digital identity transactions extending beyond financial activities. This may indicate a significant transactional underpinning including data exchanges, social interactions within digital platforms, accessing diverse online services, and managing digital legal contracts and more. This phenomena is likely due to a movement toward credential-based identity and warrants further research.

Overall, the analysis strongly suggests a transition in innovation trends from card-based digital identity solutions, typically focused on identity verification alone, towards more integrated, credential-based digital identity solutions. This shift aligns with technological advancements and reflects a growing emphasis on user

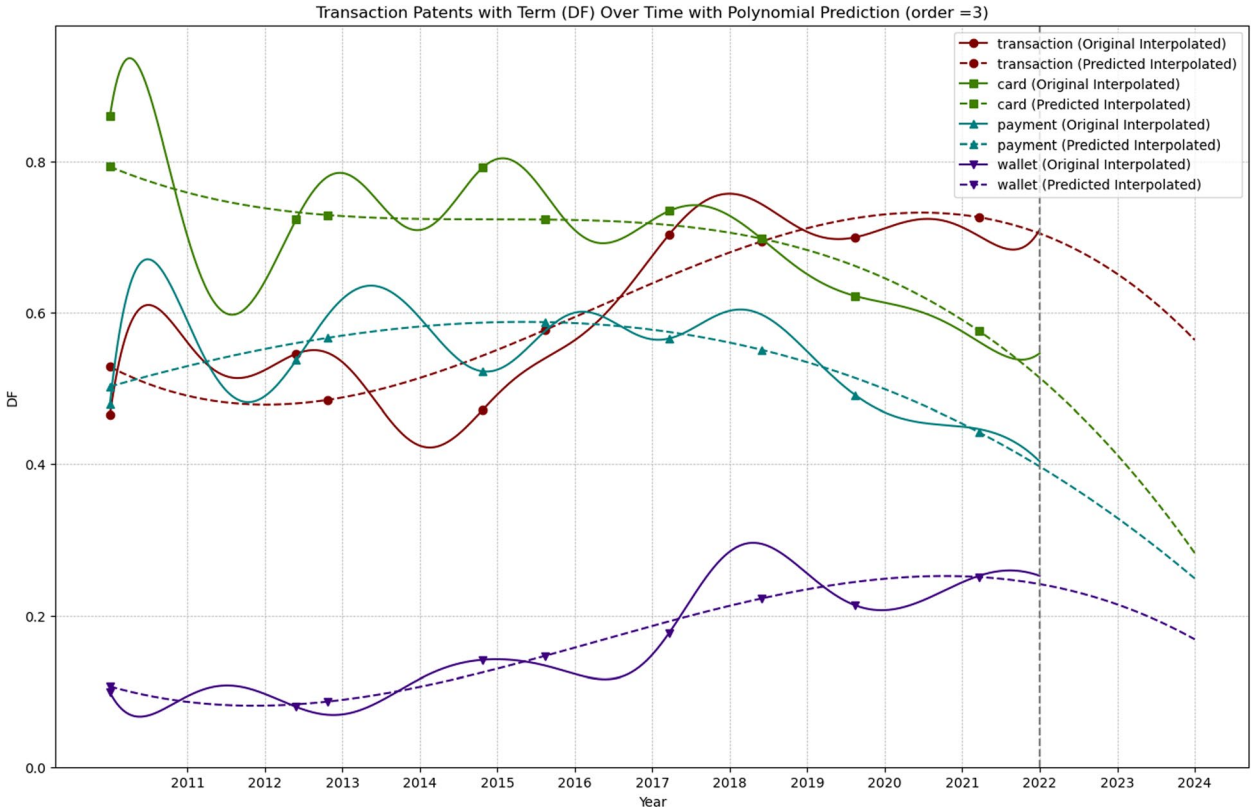

Fig. 6 Transaction DF in patents from 2010 to 2022

experience and the impact of regulatory frameworks in shaping digital identity solutions.

4.2.4 Network security

The latent Dirichlet allocation (LDA) and K-means cluster analyses in our study (see Table 4) have highlighted two security-related features within the patent set-network security and cybersecurity, with an additional feature of cryptography. The absence of a distinct, overarching security feature raises questions about the prominence of security-focused patents in digital identity.

Within the patent set, terms like 'Communication' and 'Storage' exhibit high usage frequencies, indicating their foundational role in digital identity patents. 'Communication,' despite its prevalent use, possesses a low significance score, suggesting a general rather than specialised application within the patents, possibly due to its broad applicability across various technologies. 'Storage,' while frequently mentioned, similarly lacks distinctiveness, reflecting its ubiquitous role in digital systems.

Term	Type	μ	σ	δ	ϵ
Communic.	TF	0.88	0.14	5.93	8.63
	DF	91.51	5.70	30.18	3.34
	TF-IDF	0.08	0.54	-83.78	36.19
Certificate	TF	0.53	0.17	2.93	29.80
	DF	58.40	5.43	21.38	5.56
	TF-IDF	0.28	0.95	-31.08	33.56
Storage	TF	0.46	0.16	121.55	8.65
	DF	89.00	6.85	35.46	2.71
	TF-IDF	0.05	0.20	-64.85	24.31
Encryption	TF	0.17	0.55	41.53	18.95
	DF	55.37	4.95	24.48	3.47
	TF-IDF	0.10	0.27	0.28	20.82
Privacy	TF	0.09	0.36	154.38	25.90
	DF	47.93	6.02	47.65	6.86
	TF-IDF	0.07	0.24	53.79	31.72

μ = mean %, σ = std dev %, δ = change %, ϵ = error %

In contrast, 'Certificate' and 'Encryption' demonstrate more moderate usage but exhibit higher distinctiveness, as reflected in their significance scores. These terms, particularly relevant in emerging technologies

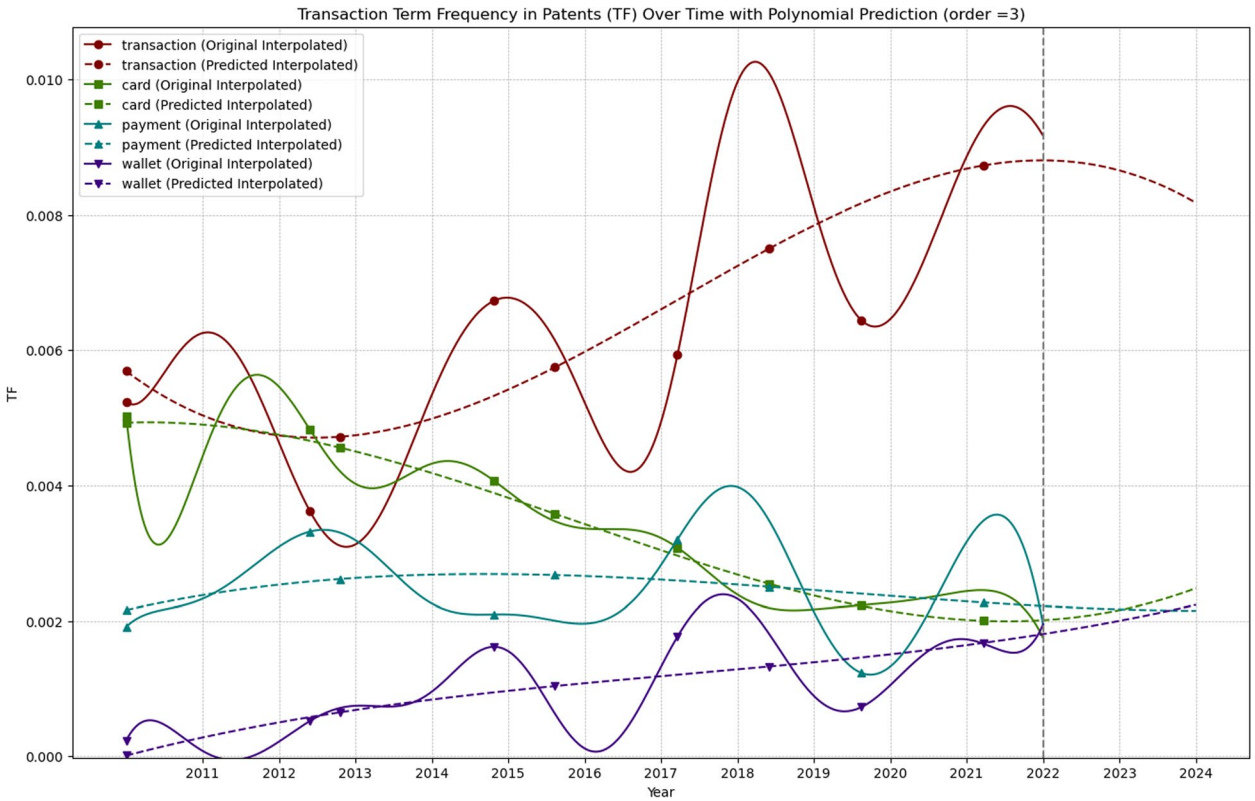

Fig. 7 Transaction TF in patents from 2010 to 2022

such as blockchain and IoT, highlight specific security applications in digital identity technologies.

'Certificate' likely refers to digital certificates, which are electronic documents used to prove the ownership of a public key. In digital identity contexts, such certificates are crucial for authentication processes, ensuring that the entities involved in a digital transaction are who they claim to be.

'Encryption,' on the other hand, is essential for protecting the confidentiality and integrity of data. In digital identity systems, encryption can be used to secure sensitive personal information from unauthorised access, making it a critical component for maintaining privacy and security. The emphasis on encryption in patents may reflect the growing concerns over data breaches and cyber threats in digital interactions (Fig. 8).

Additionally, the term 'Privacy' stands out for its lower frequency yet notable distinctiveness, suggesting its specialised application within digital identity patents. This could indicate a growing focus on privacy compliance and data protection, in line with global regulations like the General Data Protection Regulation (GDPR) [4]-a European Union regulation on information privacy-and the California Consumer Privacy Act (CCPA) [5]-a law to improve privacy and consumer rights in California, USA-signifying a focused or emerging area of innovation in the field.

Overall, these network security terms in digital identity patents present a varied landscape. Common terms like 'Communication' and 'Storage' act as foundational elements, while 'Certificate,' 'Encryption' and 'Privacy' serve more specific and evolving roles.

4.2.5 Biometric

Term	Type	μ	σ	δ	ϵ
Fingerprint	TF	0.08	0.03	-2.69	35.18
	DF	37.49	6.51	12.13	11.37
	TF-IDF	0.07	0.19	-12.97	25.21
Hand	TF	0.03	0.01	-45.11	16.05
	DF	38.25	5.37	-5.47	11.25
	TF-IDF	0.03	0.08	-42.26	16.97
Voice	TF	0.07	0.03	9.32	38.34
	DF	31.95	6.39	2.06	8.73
	TF-IDF	0.08	0.30	7.70	38.13
Face	TF	0.005	0.01	49.65	24.53
	DF	10.52	2.06	102.07	16.16
	TF-IDF	0.01	0.03	11.49	23.72
Eye	TF	0.02	0.01	-64.16	55.79
	DF	11.25	3.51	28.59	22.98
	TF-IDF	0.03	0.19	-67.68	50.54
Keystroke	TF	0.01	0.01	153.30	60.53
	DF	1.73	1.20	37.77	64.16
	TF-IDF	0.005	0.04	134.26	51.90

μ = mean %, σ = std dev %, δ = change %, ϵ = error %

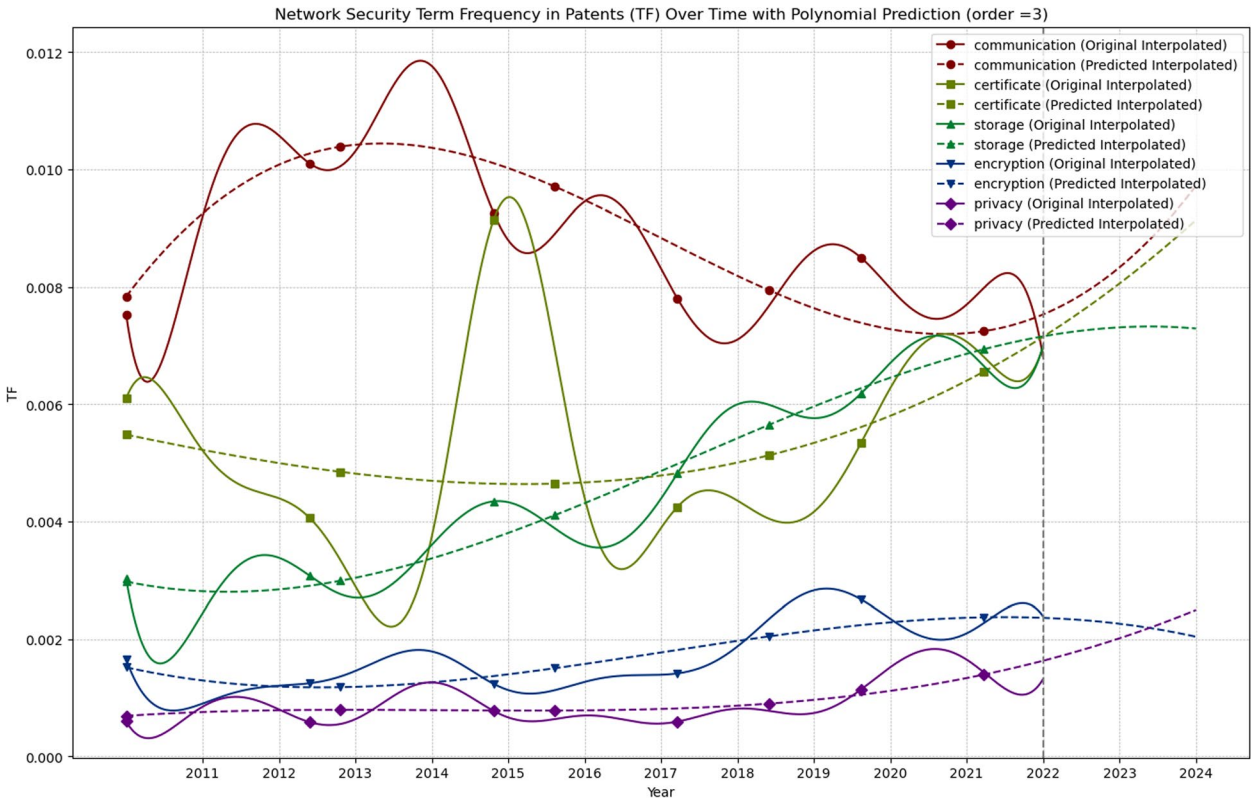

Fig. 8 Network/security TF in patents from 2010 to 2022

Biometric terms such as 'Fingerprint,' 'Hand,' and 'Voice' exhibit lower frequency in patents, with 'Fingerprint' and 'Voice' showing somewhat more distinctiveness, indicative of an established presence in the field. The prominence of 'Fingerprint' could be linked to its long-standing maturity and user-friendly nature, factors that have been shaped by user acceptance over time.

Emerging terms like 'Face' and 'Eye' display very low frequency but high change percentages, suggesting an increasing relevance in the biometric landscape. This trend might be driven by technological advancements in facial and iris recognition, which are perceived as offering enhanced security and accuracy. This shift could also be a response to growing privacy and security concerns, as well as evolving regulatory frameworks like GDPR, which emphasise data protection and privacy.

Despite the positive change over time for some terms, Fig. 9 reveals a general decline in both term and document frequency for all biometrics over the past 5 years. This decline might be due to various factors such as ethical and privacy concerns surrounding biometric data use, technological challenges and limitations in accuracy and reliability, diverse applications in different domains, and the global market trends and demand-such as the global supply chain crisis. Additionally, the impact of the COVID-19 pandemic, particularly on touch-based systems like fingerprint scanners, could have contributed to this trend, reflecting hygienic concerns. Additional research is required in this area to determine the root cause of the decline.

The term 'Keystroke,' despite its minimal current usage, is noteworthy for its significant growth, indicating an emerging focus on biometric technologies for digital identity. This increase could be attributed to the integration of keystroke dynamics with advanced technologies such as machine learning and behavioural analytics, offering innovative and non-intrusive verification methods. Again, further research is warranted.

In summary, the biometric landscape in digital identity patents is characterised by a mix of established and emerging technologies. While traditional biometrics like fingerprints and voice recognition continue to be prevalent, newer forms such as facial and eye recognition, and even keystroke dynamics, are gradually

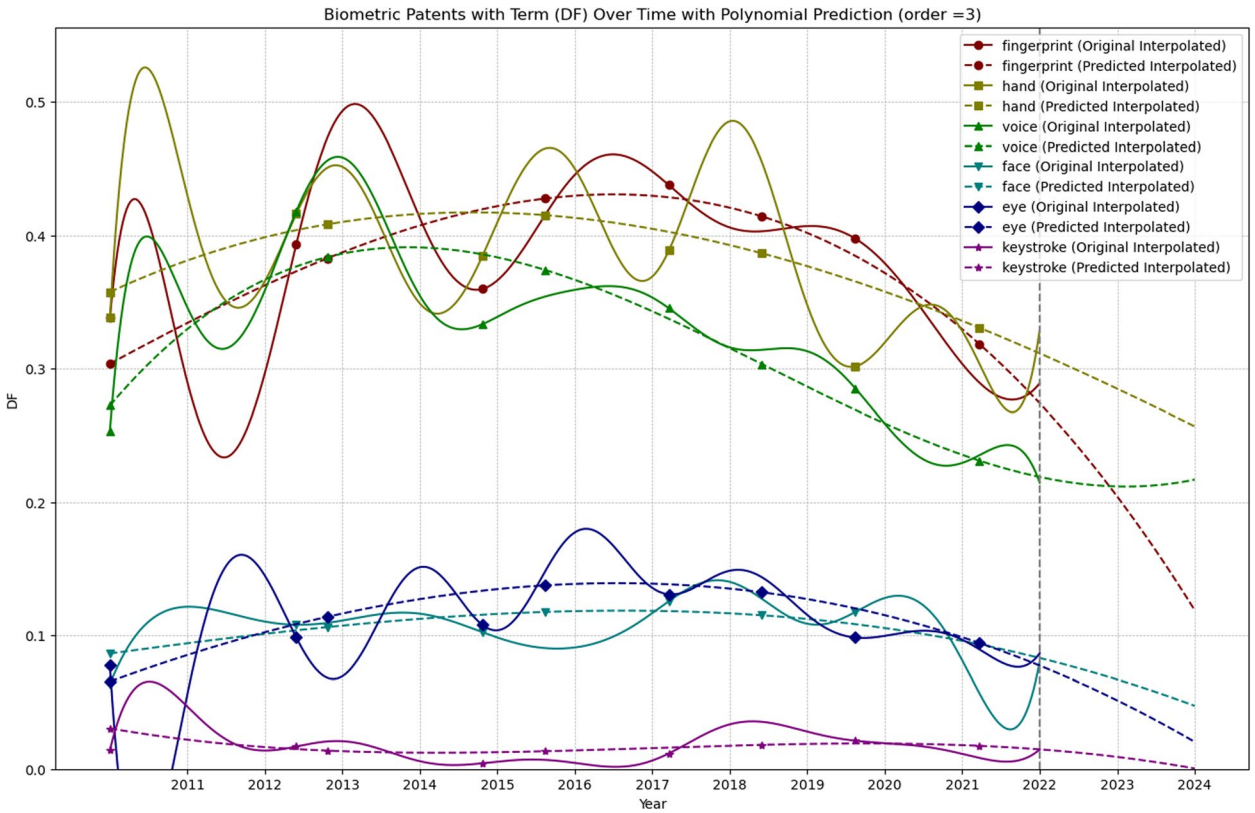

Fig. 9 Biometric DF in patents from 2010 to 2022

gaining traction. This diversity reflects the evolving nature of biometric technologies, influenced by a confluence of factors including technological advancements, regulatory changes, user acceptance, ethical considerations, and the broadening scope of biometric applications across various sectors. Declines aside, activity levels indicate a high probability that biometric technologies remain a key part of future digital identity ecosystems.

4.2.6 Block chain

The significant growth in the TF values for 'Block,' 'Chain,' and 'Ledger' over the past decade highlights the increasing importance of blockchain-related concepts in patent literature. This trend not only indicates a growing interest but also points to the technological convergence and interoperability of blockchain with digital identity technologies. 'Ledger' has seen the most dramatic increase, suggesting a shift towards decentralised digital identity models.

This surge could reflect broader global trends in blockchain adoption across various industries, not just in digital identity, and may also be influenced by the integration of blockchain with other emerging technologies.

The term 'Credential' shows a moderate frequency, underscoring its relevance in blockchain within digital identity patents. Its pattern and distinctiveness suggest a vital role in blockchain technology, potentially linked to self-sovereign identity (SSI) models where users control their identity. This focus on 'Credential' could be indicative of a shift towards more user-centric models in digital identity, leveraging blockchain's potential for enhancing user privacy and control.

'Block' emerges as the most frequently used term, highlighting its central role in blockchain discussions within the digital identity domain. Its high variability in usage might be driven by ongoing innovations in cryptographic security and blockchain's growing market adoption. 'Chain' and 'Ledger' show unique trends: 'Chain' with moderate frequency but high variability and distinctiveness, and 'Ledger' with lower frequency but significant distinctiveness, hinting at specialised applications in blockchain technology. These trends could also reflect technological evolution in blockchain, focusing on scalability and integration with existing digital infrastructures.

Term	Type	μ	σ	δ	ϵ
Credential	TF	0.30	0.16	196.24	39.75
	DF	49.17	6.83	54.61	8.65
	TF-IDF	0.20	0.10	77.22	40.71
Block	TF	0.58	0.38	381.74	18.70
	DF	76.09	10.64	47.85	4.59
	TF-IDF	0.11	0.37	30.97	21.84
Chain	TF	0.35	0.36	1896.20	20.87
	DF	44.87	20.51	228.82	9.69
	TF-IDF	0.16	1.08	367.75	22.00
Ledger	TF	0.10	0.11	7720.54	51.99
	DF	19.25	14.49	510.80	18.36
	TF-IDF	0.12	1.11	2800.45	37.05

μ = mean %, σ = std dev %, δ = change %, ϵ = error %

In summary, as indicated by Fig. 10, all blockchain-related terms are witnessing an increase, supported by positive changes in term and document frequencies and significance values for each term. This analysis suggests a future outlook where 'Chain' and 'Ledger' may continue to grow in importance, mirroring the evolving landscape of blockchain technology in digital identity solutions, influenced by regulatory trends, compliance requirements, and the broader integration of blockchain into diverse technological and industrial domains.

5 Lessons learned

In the 'Lessons learned' section, we reflect on the insights gained from applying NLP to analyse digital identity patents. We discuss the challenges, effective methodologies, and the importance of data preprocessing and AI integration in extracting meaningful insights. This section aims to provide guidance for future research in digital identity analysis, highlighting the key methodological takeaways and their implications for the field. The following lessons were identified:

1. *Data Preprocessing*-patent documents exhibit considerable variation in structure and frequently incorporate standardised language that may pose challenges for NLP analysis. The removal of such terminology, particularly terms irrelevant to the domain of interest, was essential for enhancing the precision of the analysis. Furthermore, meticulous cleaning of assignee information was necessitated by occasional inaccuracies in patent filings, which we resolved with the implementation of a 'partial match' filter to maintain data integrity.
2. *Strategic Data Segmentation*-employing a strategic approach to divide the dataset into defined temporal segments proved beneficial for identifying trends and shifts within the digital identity domain. Understanding the timelines associated with patent filing and processing was integral to tailoring the segmentation effectively, thereby ensuring a more accurate analysis of temporal patterns.
3. *Clustering Algorithm*-given the extensive volume of the patent dataset, selecting an appropriate clustering algorithm was critical to facilitate efficient processing and yield insightful results. LDA and K-means algorithms demonstrated efficacy, whereas Sentence Transformers, DBSCAN, and Word2Vec faced challenges managing the dataset's scale.
4. *AI Generalisation*-clustered word sets underwent further processing using OpenAI APIs to achieve generalisation, categorisation, and contextualisation. This process facilitated the identification of key features and enabled an evaluation of the congruence between the chosen clustering algorithms.
5. *Mining Architecture*-we encountered significant computational demands when conducting patent analysis combined with NLP, especially when delving into secondary and tertiary relationships. To manage the extensive analytical workload, we found that adopting a client-server architecture was beneficial. This approach allowed us to queue tasks and process them across different nodes, offering the flexibility to pause and restart the analysis without having to begin anew.

Reflecting on these lessons learned, the next section reports limitations of the study, before Sect. 7 discusses trends and conclusions drawn from the digital identity patent data, offering a comprehensive analysis of the key findings and their implications in the field.

6 Limitations of the study

This study faced limitations due to the complex nature of data mining digital identity articles. The dataset may not be entirely comprehensive due to some patents being unavailable and the updates and amendments that patent records undergo. There may also be publication and geographic biases, as the dataset may exclude potentially innovative ideas that are not patented, and may focus primarily on innovations from specific regions based on the database sources used. The study could be repeated with a wider search term-but the choice of those terms would need to be systematic and on a scientific basis.

Additionally, the analytical process had its challenges, including the difficulty of interpreting texts with high levels of semantic ambiguity through NLP. While process challenges are not a limitation of the study, the accuracy of data mining algorithms can also be constrained, potentially leading to over-fitting and limiting the model's generalisability.

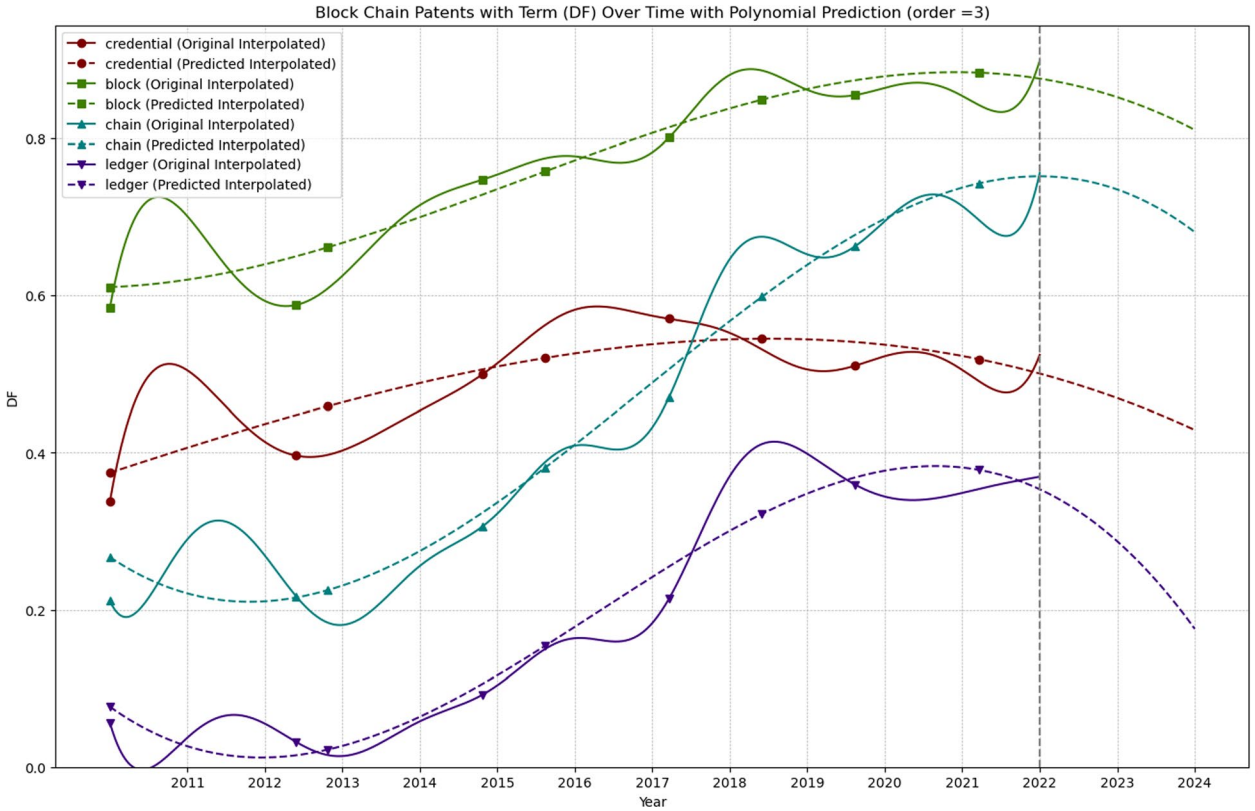

Fig. 10 Block chain DF in patents from 2010 to 2022

Finally, there is a concern regarding interpretational bias, as the parameter settings for NLP and data mining may be subjective and significantly influence the outcome. Expert validation is necessary to enhance the study's reliability, but this may introduce biases based on individual perspectives. Despite these limitations, the size and completeness of the Google patent database mitigates concerns. Future research should reduce residual limitations and build upon the findings of this study.

7 Conclusions

In our analysis of digital identity patents, quantitative methods provided objective metrics to evaluate the significance and relevance of the identified clusters. By relying on these quantitative methods and corresponding validation methods, such as coherence testing and frequency analysis, we ensured that our conclusions were grounded in measurable data, accurately reflecting trends and patterns in the digital identity patent landscape.

Additionally, the application of NLP methods in mining patents for insights is designed to eliminate bias and subjectivity and has proved invaluable. These advanced analytical techniques enabled a comprehensive understanding of high volume textual datasets, such as patent libraries. NLP provided the ability to dissect and interpret the linguistic structures within patents, revealing underlying patterns and trends that might otherwise have remained obscure, while cluster analysis further enhanced this process by grouping similar terms or concepts, allowing for a more structured analysis of the data.

Our analysis of digital identity patents specifically, revealed significant trends and shifts within the domain as follows:

1. *Core Terminology*-analysis of keywords and POS tagging has pinpointed fundamental terms, revealing significant patterns in frequency and importance. The prominence of the term 'User' suggests a shift towards a user-centric identity paradigm, highlighting an elevated focus on digital identity and associated privacy and security concerns. Additionally, the evolving usage of 'Data' and 'Information' terms indicates a maturation of digital identity solutions.
2. *Biometric*-terms indicated a mixed landscape of established and emerging technologies, with traditional biometrics like fingerprints and voice recognition being prevalent, while newer forms like facial and eye recognition were on the rise. However, a general decline in biometric terms in recent years highlighted the complex interplay of technological, ethical, and market factors.

3. *Blockchain*-the growth in blockchain-related terms such as 'Block,' 'Chain', and 'Ledger' pointed to the increasing relevance of blockchain in digital identity, potentially driven by broader global adoption trends and technological convergence. The rise of Credential in blockchain patents underscored a shift towards more user-centric, self-sovereign identity models.
4. *Financial Transactions*-rising blockchain importance also appears to be inline with advancement in digital identity transactions with evidence suggesting innovations involving blockchain enabled transactions are increasing at the expense of traditional payment methods.
5. *Privacy vs User Experience*-the analysis suggests a shift towards enhancing user experience and privacy within digital identity solutions, likely in response to increased public concern over data privacy and the demand for more user-friendly interfaces. Future solutions may thus prioritise ease of use and strong privacy protections.
6. *Regulatory Dynamics*-changes in regulatory and compliance landscapes are shaping the innovation trajectory within digital identity, with new standards and guidelines influencing patent filings and innovation focus. This regulatory evolution underscores the need for digital identity technologies to adapt, balancing innovation with compliance.

Our investigation into digital identity patents has also unveiled contributions from various sectors, highlighting the expanding role of technology firms, financial institutions and notably, the emergence of Chinese entities in the post-2021 landscape. This points to some geographical and sectoral expansion within the digital identity domain and may indicate a shift towards a more global and diverse innovation ecosystem.

Identifying key trends in digital identity technology not only highlights the field's evolution but also unlocks the door to greater interoperability. This, in turn, would make online services simpler to develop, deploy, secure, maintain, and use; however, the degree to which the identified areas in this research are moving toward inter-operability is largely unknown, showcasing the critical need for deeper exploration to maximise the trans-formative impact digital identity may have on a connected digital world.

Abbreviations

LDA	Latent Dirichlet allocation
NLP	Natural language processing
TF	Term frequency
DF	Document frequency
IDF	Inverse document frequency
POS	Part of speech
SSI	Self-sovereign identity
GDPR	General Data Protection Regulation
CCPA	California Consumer Privacy Act

Acknowledgements
The authors wish to acknowledge the contribution of Mr Michael Smith, who provided grammatical editing support for this article.

Authors' contributions
The research presented in this manuscript was primarily conducted by the first author, who was responsible for the design and execution of the study, data collection, analysis, and interpretation, as well as the drafting of the manuscript. The second author, in the capacity of a supervisor, provided overarching support and guidance throughout the research process.

Funding
A scholarship from the UK Commonwealth Scholarship Commission generously supported this research. The Commonwealth Scholarship Commission's support was purely financial and played no direct role in the study's design nor data collection, analysis, and interpretation. The findings, interpretations, and conclusions presented in this manuscript are solely the author's responsibility and do not necessarily represent the views of the UK Commonwealth Scholarship Commission. Reference: CSC CR-2019-67.

Availability of data and materials
The data central to the conclusions of this manuscript were sourced from the Google Patent Database, a publicly accessible online platform. In line with SpringerOpen's policies, the specific datasets utilized in our research are available in a machine-readable format directly from the Google Patent Database and can be accessed at https://patents.google.com.

Code availability
The source code and algorithms generated in the course of this research are publicly accessible via a GitHub repository, in line with principles of transparency and to facilitate the reproducibility of the research outcomes. This repository encompasses all requisite Jupyter notebooks and corresponding Python code essential for duplicating the computational analyses conducted in this study. Repository link: https://github.com/oxford-mc/patent-mining.
Due to the intricate nature of its operational requirements, the client/server architecture employed for the extraction and mining of clustered terms is not included in the repository at the current juncture.

Declarations

Ethics approval and consent to participate
The research detailed in this manuscript did not involve human or animal subjects. However, in line with the University of Oxford's commitment to upholding the highest standards of research integrity and ethics, the study was submitted for review to the Computer Science Departmental Research Ethics Committee (DREC). Reference number CS_C1A_021_025.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 28 December 2023 Accepted: 19 June 2024
Published online: 03 July 2024

References

1. S. Abraham, Building trust: Lessons from Canada's approach to digital identity. Observer Research Foundation. ORF Issue Brief No. 367 (2020)
2. H. Alanzi, M. Alkhatib, Towards improving privacy and security of identity management systems using blockchain technology: A systematic review. Appl. Sci. (Switzerland) **12** (2022). https://doi.org/10.3390/app122312415
3. Anonymous, Discover eIDAS | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/policies/discover-eidas. Accessed 12 Mar 2024
4. Anonymous, General data protection regulation (GDPR). (2016). https://eur-lex.europa.eu/eli/reg/2016/679/oj. Accessed 12 Mar 2024

5. Anonymous, California Consumer Privacy Act (CCPA) of 2018. (2018). https://oag.ca.gov/privacy/ccpa. Accessed 12 Mar 2024
6. D.W. Arner, R.P. Buckley, D.A. Zetzsche, R. Veidt, Sustainability, fintech and financial inclusion. Eur. Bus. Organ. Law Rev. **21**, 7–35 (2020). https://doi.org/10.1007/s40804-020-00183-y
7. D.W. Arner, D.A. Zetzsche, R.P. Buckley, J.N. Barberis, The identity challenge in finance: From analogue identity to digitized identification to digital KYC utilities. SSRN Electron. J. (2018). https://doi.org/10.2139/ssrn.3224115
8. M. Aydar, S. Ayvaz, Towards a blockchain based digital identity verification, record attestation and record sharing system. Preprint **0**, 1–25 (2019). http://arxiv.org/abs/1906.09791. Accessed 12 Mar 2024
9. A. Beduschi, J. Cinnamon, J. Langford, C. Luo, D. Owen, *Building Digital Identities: The Challenges, Risks and Opportunities of Collecting Behavioural Attributes for new Digital Identity Systems*, (University of Exeter and Coelition, 2017)
10. S. Bird, E. Klein, E. Loper, *Natural language processing with Python: Analyzing text with the natural language toolkit* (O'Reilly Media, Inc., Sebastopol, 2009)
11. A. Boldyreva, V. Goyal, V. Kumart, in *Proceedings of the ACM Conference on Computer and Communications Security*. Identity-based encryption with efficient revocation. (2008), pp. 417–426. https://doi.org/10.1145/1455770.1455823.
12. R. Botsman, Who can you trust?: How technology brought us together and why it might drive us apart. (2017). https://www.penguin.com.au/books/who-can-you-trust-9780241296189. Accessed 12 Mar 2024
13. Broda, in Managing Trust in e-Health with Federated Identity Management. eHealth Workshop. Konolfingen, (2007)
14. C. Brunner, U. Gallersdörfer, F. Knirsch, D. Engel, F. Matthes, DID and VC: untangling decentralized identifiers and verifiable credentials for the web of trust. ACM Int. Conf. Proc. Ser. 61–66 (2020). https://doi.org/10.1145/3446983.3446992
15. K. Cameron, The laws of identity. Microsoft Corp. 8–11 (2005). https://doi.org/10.1126/science.22.555.206-a
16. S. Carter, L.L. Burch, D.R. Olds, Crafted identities (patent us-10063523-b2). (2005). https://patents.google.com/patent/US10063523B2/en. Accessed 12 Mar 2024
17. S. Choi, J. Yoon, K. Kim, J.Y. Lee, C.H. Kim, SAO network analysis of patents for technology trends identification: A case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells. Scientometrics **88**, 863–883 (2011). https://doi.org/10.1007/s11192-011-0420-z
18. D. Chuen, Handbook of digital currency: Bitcoin, innovation, financial instruments, and big data. (2015). https://www.sciencedirect.com/book/9780128021170/handbook-of-digital-currency. Accessed 12 Mar 2024
19. J. Drummond, D. Blackson, L. Chen, B. Cichon, M. Covert, B. Lepper, Automated banking machine apparatus and system (patent ep1672516a2). (1997). https://patents.google.com/patent/EP1672516A2/en. Accessed 12 Mar 2024
20. J. Drummond, B. Cichon, M. Smith, D. Weis, Automated banking machine that operates responsive to data bearing records (patent us008567667b2). (2013). https://patents.google.com/patent/US8567667B2/en. Accessed 12 Mar 2024
21. T. Ehrlich, D. Richter, M. Meisel, J. Anke, Self-sovereign identity als grundlage für universell einsetzbare digitale identitäten. HMD Prax. Wirtschaftsinformatik **58**, 247–270 (2021). https://doi.org/10.1365/s40702-021-00711-5
22. H.P. Enterprise, Augmented intelligence: Helping humans make smarter decisions white paper analytics and big data. www.microfocus.com. Accessed 12 Mar 2024
23. J. Fischer, F. Dietrich, M. Paeschke, Method for storing data for managing digital identity of user, involves writing data from provider computer system to token via connection to store data in token, and providing connections with connection-oriented protocol (patent de-102008042262-a1) (2008). https://patents.google.com/patent/DE102008042262A1/en. Accessed 12 Mar 2024
24. P. Gangwani, S. Joshi, H. Upadhyay, L. Lagos, Iot device identity management and blockchain for security and data integrity. Int. J. Comput. Appl. **184**, 49–55 (2023). https://doi.org/10.5120/ijca2023922529
25. L. Gupta, V. Lander, Service discovery for a multi-tenant identity and data security management cloud service (patent us-2018041515-a1). (2017). https://patents.google.com/patent/US20180041515A1/en. Accessed 12 Mar 2024
26. P.J. Hallin, J.J. Lambert, K.U. Schutz, S. Pai, Systems and methods for distributing trusted certification authorities (patent us7240194). (2002). https://patentimages.storage.googleapis.com/83/52/84/f49a0513dcf963/US7240194.pdf. Accessed 12 Mar 2024
27. C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser et al., Array programming with numpy. Nature **585**, 357–362 (2020). https://doi.org/10.1038/s41586-020-2649-2
28. F. Hersey, Alipay trials digital replacement of China's ubiquitous ID cards · technode. (2018). https://technode.com/2018/04/18/alipay-id/. Accessed 12 Mar 2024
29. M. Honnibal, I. Montani, spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear. **7**(1), 411–420 (2017)
30. W. Inambao, J. Phiri, D. Kunda, Digital identity modelling for digital financial services in Zambia. ICTACT J. Commun. Technol. **9**, 1829–1837 (2018). https://doi.org/10.21917/ijct.2018.0267
31. B.E. Johnson, C. Webster-Lam, Authentication for a commercial transaction using a mobile module (patent ep-2016543-b1). (2007). https://doi.org/10.1145/570705.570720
32. H. Kagermann, K.H. Streibich, K. Suder, Digital sovereignty. Status Quo and Perspectives. Acatech IMPULSE. (2021). Available online: https://www.acatech.de/publikation/digitale-souveraenitaet-status-quo-und-handlungsfelder/downloadpdf
33. J.E. Kelly, Computing, cognition and the future of knowing. IBM Res. Oct. **13**, 12 (2015)
34. M. Kohli, Transformation from identity stone age to digital identity. Int. J. Netw. Secur. Appl. (IJNSA) **3**, 121–136 (2011). https://doi.org/10.5121/ijnsa.2011.3309
35. F. Madani, C. Weber, The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis. World Patent Inf. **46**, 32–48 (2016)
36. W. McKinney et al., Data structures for statistical computing in python. SciPy. **445**(1), 51–56 (2010)
37. T. Mikula, R.H. Jacobsen, Identity and access management with blockchain in electronic healthcare records (2018). https://doi.org/10.1109/DSD.2018.00008
38. J. Monti, Mastercard, Microsoft to advance digital identity innovations (2022). https://www.mastercard.com/news/press/2022/april/mastercard-launches-next-generation-identity-technology-with-microsoft-to-help-more-consumers-shop-online-safely/. Accessed 12 Mar 2024
39. F. Morgner, P. Bastian, M. Fischlin, Securing transactions with the eIDAS protocols. pp. 3–18 (2016). https://doi.org/10.1007/978-3-319-45931-8_1
40. A. Mühle, A. Grüner, T. Gayvoronskaya, C. Meinel, A survey on essential components of a self-sovereign identity (2018). https://www.sciencedirect.com/science/article/pii/S1574013718301217. Accessed 12 Mar 2024
41. H. Noh, Y. Jo, S. Lee, Keyword selection and processing strategy for applying text mining to patent analysis. Expert Syst. Appl. **42**, 4348–4360 (2015). https://doi.org/10.1016/j.eswa.2015.01.050
42. Norden, The Nordic digital ecosystem actors, strategies, opportunities (2015)
43. S. O'neal, Bank of America has the most blockchain patents, but is it actually going to use them (2018). https://cointelegraph.com/news/bank-of-america-has-the-most-blockchain-patents-but-is-it-actually-going-to-use-them. Accessed 12 Mar 2024
44. Oracle, Oracle identity management 11g datasheet. pp. 1–57. (2010). https://www.oracle.com/technetwork/middleware/id-mgmt/overview/idm-ds-11g-r1-154269.pdf. Accessed 12 Mar 2024
45. Oracle, Oracle identity management 11g white paper. pp. 1–57 (2010). https://www.oracle.com/technetwork/middleware/id-mgmt/overview/oim-11gr2-business-wp-1928893.pdf. Accessed 12 Mar 2024
46. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., Scikit-learn: Machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
47. A. Preukschat, D. Reed, Self-sovereign identity (2021)
48. R, B., P, A.: Blockchain based service: A case study on ibm blockchain services and hyperledger fabric. International Journal of Case Studies in Business, IT and Education (IJCSBE) A Refereed International Journal of Srinivas University, India. Blockchain Services and Hyperledger Fabric. Int. J. Case Stud. Bus. **4**, 2581–6942 (2020). https://doi.org/10.5281/zenodo.3822411

49. R. Rehurek, P. Sojka, Gensim - python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic **3** (2011)
50. R. Rivera, J. Robledo, V. Larios, J. Avalos, How digital identity on blockchain can contribute in a smart city environment (2017). https://doi.org/10.1109/ISC2.2017.8090839
51. I. Robinson, J. Webber, E. Eifrem, Graph databases: new opportunities for connected data (O'Reilly Media, Inc., 2015)
52. G. Roussos, D. Peterson, U. Patel, Mobile identity management: An enacted view. Int. J. Electron. Commer. **8**, 81–100 (2003). https://doi.org/10.1080/10864415.2003.11044287
53. Domingo AIS, Enríquez ÁM, Digital Identity: the current state of affairs. BBVA Research. **1**, 1–46 (2018)
54. J.E. Setsaas, K. Cameron, D. Birch, Distributed identity - Should it be the way forward! EEMA (2020). https://www.eema.org/event/eema-webinar-distributed-identity-should-it-be-the-way-forward/. Accessed 12 Mar 2024
55. J. Smye, Building blocks: Conceptualizing the true socio-political potential in blockchain's facilitation of self-sovereign digital identity and decentralized organization. (2019). https://repository.library.carleton.ca/concern/etds/fj236300t. Accessed 12 Mar 2024
56. M. Takemiya, B. Vanieiev, Sora identity: Secure, digital identity on the blockchain. ieeexplore.ieee.org. (2018). https://ieeexplore.ieee.org/abstract/document/8377927/. Accessed 12 Mar 2024
57. L. Tao, A look at China's push for digital national ID cards | south china morning post. (2018). https://www.scmp.com/tech/article/2129957/look-chinas-push-national-digital-id-cards. Accessed 12 Mar 2024
58. A. Tobin, D.R.T.S. Foundation, U. 2016, The inevitable rise of self-sovereign identity. sovrin.org (2017). https://sovrin.org/wp-content/uploads/2017/06/The-Inevitable-Rise-of-Self-Sovereign-Identity.pdf. Accessed 12 Mar 2024
59. K.C. Toth, A. Anderson-Priddy, Self-sovereign digital identity: A paradigm shift for identity. IEEE Secur. Priv. **17**, 17–27 (2019). https://doi.org/10.1109/MSEC.2018.2888782
60. E. Union, Regulation (EU) no 910/2014 of the European Parliament and of the Council of 23 July 2014 on electronic identification and trust services for electronic transactions in the internal market and repealing directive 1999/93/EC (eIDAS regulation) (2014). Accessed 10 Oct 2023
61. P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski et al., Scipy 1.0: Fundamental algorithms for scientific computing in python. Nat. Methods **17**, 261–272 (2020). https://doi.org/10.1038/s41592-019-0686-2
62. X. Wang, P. Qiu, D. Zhu, L. Mitkova, M. Lei, A.L. Porter, Identification of technology development trends based on subject-action-object analysis: The case of dye-sensitized solar cells. Technol. Forecast. Soc. Chang. **98**, 24–46 (2015). https://doi.org/10.1016/j.techfore.2015.05.014
63. C.C. Wu, H.J. Leu, Examining the trends of technological development in hydrogen energy using patent co-word map analysis. Int. J. Hydrogen Energy **39**, 19262–19269 (2014). https://doi.org/10.1016/j.ijhydene.2014.05.006
64. J. Yoon, K. Kim, Identifying rapidly evolving technological trends for research and development planning using SAO-based semantic patent networks. Scientometrics **88**, 213–228 (2011). https://doi.org/10.1007/s11192-011-0383-0
65. J. Yoon, H. Park, K. Kim, Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis. Scientometrics **94**, 313–331 (2013). https://doi.org/10.1007/s11192-012-0830-6
66. S. Zeadally, F. Siddiqui, Z. Baig, A. Ibrahim, Smart healthcare: Challenges and potential solutions using Internet of Things (IoT) and big data analytics. PSU Res. Rev. **4**, 149–168 (2020). https://doi.org/10.1108/PRR-08-2019-0027
67. Z. Zheng, S. Xie, H.N. Dai, X. Chen, H. Wang, Blockchain challenges and opportunities: A survey. International journal of web and grid services. **14**(4), 352–375 (2018)

Publisher's Note