Which types of online resource support US patent claims?[1]

Cristina I. Font-Julián[1]
José-Antonio Ontalba-Ruipérez[2]
Enrique Orduña-Malea[3]
Mike Thelwall[4]

[1] Department of Audiovisual Communication, Documentation and History of Art, Universitat Politècnica de València, Valencia, Spain
🖂crifonju@upv.es

[2] Department of Audiovisual Communication, Documentation and History of Art, Universitat Politècnica de València, Valencia, Spain
🖂joonrui@upv.es

[3] Department of Audiovisual Communication, Documentation and History of Art, Universitat Politècnica de València, Valencia, Spain
🖂 enorma@upv.es

[4] Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of Wolverhampton, Wolverhampton, UK
🖂 m.thelwall@wlv.ac.uk

Abstract

Patents are key documents to support the commercial exploitation of inventions. Patent documents must claim inventiveness, industrial application, and novelty to be granted and may use citations and URLs to support these claims as well as to explain their ideas. Although there is much research into the citations used to support inventions, almost nothing is known about the cited URLs. This may hinder inventors and evaluators from deciding which URLs are appropriate. To investigate this issue, all 3,133,247 patents granted by the United States Patent and Trademark Office (USPTO) from 2008 to 2018 were investigated, and 2,719,705 URLs (patent outlinks) were automatically extracted using heuristics, and analyzed using link analyses techniques. A minority of patents included URLs (17.1%), with the percentage increasing over time. The inclusion of URLs differs between disciplines, with Physics (especially the subcategory Computation) having the most URLs per patent. Patents are generally embedded in the "other citations" patent section (referring to academic publications) and the "description" section (e.g., supplementary information and definitions). Online content-oriented resources (e.g., Wayback Machine, Wikipedia, YouTube), academic bibliographic databases (e.g., IEEE Xplore, Microsoft Academic, PubMed, CiteSeerX) and technological companies (e.g., IBM, Amazon, Microsoft) are often linked from USPTO patents. These findings show the broad roles that URLs can play when supporting a patent claim. Finally, in order to avoid bad practices found in the inclusion of URLs in patents, a list of recommendations to cite online resources from patents is provided.

Keywords

Informetrics; Link analysis; Patent analysis; United States Patent and Trademark Office; online resources

[1] Author version of the manuscript. Please cite as follows:
Font-Julián, C. I., Ontalba-Ruipérez, J-A., Orduña-Malea, E., & Thelwall, M. (2022). Which types of online resource support US patent claims? *Journal of Informetrics*, *16*(1), 101247.
https://doi.org/10.1016/j.joi.2021.101247

1. Introduction

According to the United States Patent and Trademark Office (USPTO), a patent for an invention is the grant of a property right to the inventor for a limited period, generally 20 years, to exclude others from making, using, offering for sale, or selling the invention in the United States, or "importing" the invention into the United States.[2]

The patent must meet three fundamental requirements to be approved: inventive (i.e., non-obvious), an industrial application, and being novel (once an invention is in the public domain, it is no longer patentable).[3] To facilitate evaluation, patent offices (responsible for granting and registering patents) have standardized patent document structures. According to the World Intellectual Property Organization (WIPO), a patent document must contain bibliographic data (patent number, title, name of the inventor, date, references to previous patents), a detailed description of the invention, and claims (which define the limits of the exploitation right).

The high commercial value of some patents has pushed researchers and professionals to extract and analyze their contents for different purposes (Breitzman & Mogee, 2002), such as technological surveillance (Lee et al., 2018), economic growth (Chang, Chen & Huang, 2012) or university-government-company partnerships (Meyer; Siniläinen; Utecht, 2003; Campbell et al., 2004). Different data analysis techniques have been applied for this, such as data mining and text mining (Aristodemou & Tietze, 2018; Van Looy & Magerman, 2019), keywords analysis (An, Kim, Mortara, & Lee, 2018) and network analysis (Yoon & Park, 2004).

Patent documents must reference prior art, for which inventors can include patent citations (citations to previous patents) and/or non-patent citations (citations to scientific publications and other materials). Statistical analyses of these citations can support the study of science-technology interactions, technological trends and technology forecasting (Meyer, 2000a; Meyer, Siniläinen & Utecht, 2003; Sharma & Tripathi, 2017). Thus, citations from patents form a valuable data source in their own right. Patent citation analysis can exploit patent bibliographic databases from national patent offices (e.g., The German Patent and Trade Mark Office, DPMA[4]), international patent offices (e.g., Espacenet[5]) and organizations (e.g., WIPO Patentscope[6]) or commercial indexes (e.g., Derwent World Patents Index[7]). Moreover, Google Patents[8] and Lens.org[9] provide free advanced features oriented to patent findability and patent scholarly analysis, respectively. The patent searching functionalities (Alberts et al., 2011), coverage (e.g., years, national patent offices included) and patent accessibility

[2] https://www.uspto.gov/patents/basics/general-information-patents

[3] As an illustrative example, the Dutch Patent Office originally refused the application by the Danish inventor Karl Krøyer for raising sunken vessels (NL6514306), because examiners found an old issue of the Donald Duck magazine which showed the same invention.

[4] https://www.dpma.de/

[5] https://worldwide.espacenet.com/

[6] https://patentscope.wipo.int/

[7] https://clarivate.com/derwent/solutions/derwent-world-patent-index-dwpi/

[8] https://patents.google.com/

[9] https://www.lens.org/

(abstracts or full text) offered by these databases condition and limit the statistical analysis of their patents. The services offering full text patents have made possible the analysis of all elements of patent documents, including URLs.

Inventors may also embed URLs linking to online resources in their patent proposals to identify relevant prior patents and other publications that limit the scope of their claims (Orduna-Malea, Thelwall & Kousha, 2017), to support their statements or just to provide additional information (e.g., interactive maps, definitions, images, videos) to increase the transparency of the evidence provided. These URLs can be included as part of the non-patent citations or just embedded throughout the body of the patent.

Whilst the USPTO's Manual of Patent Examining Procedure (Horwitz, Horwitz & Hershman, 2018) refers to some basic rules for including non-patent citations to prior art (e.g., Chapter 901)[10] and describes the use of social media platforms as sources of prior art (Chapter 2128),[11] no specific guidelines to include URLs (accompanying or not a reference) are provided.

It is important to understand the type of online resources that these URLs can point to, as well as to be aware about the best linking practices so that patent authors and evaluators can be guided about current appropriate uses. Links to inappropriate resources can limit the informational value of the patent or the justification of a claim, while broken links (misspelled or obsolete) can prevent the evaluator or reader from accessing information of interest. In addition, the analysis of all URLs embedded in patents can provide information about which online resources are most used.

Despite the promise of patent URLs, they have only been evaluated once before, since extracting hyperlinks from patents is still not straightforward. Most patents were originally published in paper format, being lately digitized in PDF format using optical character recognition (OCR) technologies. However, many URLs appear broken or with typographical errors, making their identification and extraction complex (Orduna-Malea, Thelwall & Kousha, 2017). Recent patents, already published in web formats by some patent offices (e.g., USPTO publishes patents in XML format) still do not mark URLs as hyperlinks.

Orduna-Malea, Thelwall and Kousha (2017) analyzed the potential use of the number of linking URLs from patents to US and UK universities as a evidence of academic technological contributions. However, a large-scale systematic analysis of patent links regardless of the cited resource is needed to better understand their roles. To partly fill this gap, the objectives of this study are twofold. First, to design and develop a method oriented to the identification and massive extraction of embedded links in patents. Second, to describe the use of all URLs in patents as online resource references, driven by the following research questions (RQs):

[10] https://www.uspto.gov/web/offices/pac/mpep/s901.html#d0e113260
[11] https://www.uspto.gov/web/offices/pac/mpep/s2128.html#d0e202564

RQ1. How frequently are URLs used in patents?
RQ2. Which application fields use patent URLs most?
RQ3. Which patent sections include URLs?
RQ4. What types of web resources are most frequently linked from patents?

To provide a systematic answer to these questions, all URLs included in a set of US patents will be analyzed and characterized.

2. Research background

Patents include references to other documents to support claims as well as to explain ideas, complex concepts and processes, to acknowledge prior discoveries, or to justify novelty. References in patents can refer to other previously granted patents (patent citations) or to other documents (non-patent citations), such as scientific journal articles (e.g., patent-journal citations). While patent citations are more oriented to create legitimacy and trustworthiness (Hammarfelt, in press), non-patent citations are seen as a representation of the output of science research (Szu-chia, 2010)

All printed publications may be used as references to other documents as long as such documents have been "disseminated or otherwise made available to the extent that persons interested and ordinarily skilled in the subject matter or art, exercising reasonable diligence, can locate it." (Horwitz, Horwitz & Hershman, 2018; pp. 2100-170), including internet publications such as discussion groups, fora, digital videos, and social media posts.

Non-patent citations have been used to identify and assess the relationship between science (i.e., basic research) and industry (i.e., technology) through technological indicators (Narin, Hamilton & Olivastro, 1997; Carpenter, Cooper & Narin, 1980; Schmoch, 1993). However, the legal and economic implications of patent applications make citing motivations in these documents different from those in academic publications (Thelwall & Kousha, 2015). Consequently, patent citations do not necessarily reflect knowledge flows from research to industry (Meyer, 2000b; Alcácer & Gittelman, 2006; Alcácer, Gittelman, & Sampat, 2009; Roach & Cohen, 2013) or technological innovation (Jaffe, Trajtenberg, & Fogarty, 2000).

References can be inserted either by the applicant or assignee (found throughout in the body of the text) or the examiner (often found on the front page, or on a separate data sheet). The roles and motivations of applicants and examiners affect their citation behaviors differently. Examiners are supposed to be neutral and focus on patent law, whereas applicants must persuade the examiners (Latour, 1987), which involves making stronger claims (stacking) whilst delimiting the borders of the invention accurately (fencing) (Rip, 1986). On this basis, examiners seem to be more inclined to choose references satisfying legal requirements, as they should focus on the claims made and are not obliged to cite additional sources (Oppenheim, 2000). Thus, examiners tend to cite other patents (Bryan, Ozcan & Sampat, 2020), reference books and abstracting journals. In contrast, applicants' citing behavior is closer to that of researchers (Collins & Wyatt, 1988; Oppenheim, 2000), more frequently citing academic journals and background information, which in turn could lead to increased use of online resources.

There are also other differences between authors and inventors when citing (Meyer, 2000b). In general, authors' assumptions are wider (at the research community level) than those of the inventors (just at the level of the claims of the invention). In contrast, authors' risks when adding new references is low, while inventors' risks is high, as claims made in relation to the literature could be challenged in court (Hammarfelf, 2021). Citations in patents are also dependent on the specific patent legislation of a country or region, which may result in huge differences not only in the number of references given but also in the kind of materials cited (Meyer, 2020b). There is also a strong tendency for inventors to cite articles authored in their own country, at prestigious universities and laboratories (Kousha, & Thelwall, 2017)

Although there is much research into the citations used to support inventions, almost nothing is known about the cited URLs referenced by inventors (patent-URL citations). While these URLs can sometimes just be part of journal articles' bibliographic records, they can also be used to mention non-traditional research objects (e.g., programs, data) or even non-academic online resources (e.g., maps). The systematic and massive analysis of URLs cited by inventors can thus enhance our understanding about the citing motivations of both applicants and examiners on the one hand and help them from deciding which URLs are most appropriate in future applications on the other hand.

3. Methods

All 3,133,247 patents granted by the United States Patent and Trademark Office (USPTO) from 2008 to 2018 were analyzed. This period is long enough to show changes over time and the final year was complete at the time when the data collection began in 2019. USPTO was chosen because it oversees one of the world's largest collections of patents and makes available various informational resources and databases that allow comprehensive analysis of granted patents (Kim & Lee, 2015). While the China National Intellectual Property Administration (CNIPA) has more patent requests, it was not used due to language constraints and lower accessibility.

The USPTO bulk download feature (Patent Grant Full Text Data - No Images) was used to obtain a full text copy of all its patents from 2008 to 2018 in XML format.[12] The first author developed a program to extract the URLs from each XML file. The code extracted URLs only from the XML tagged sections where hyperlinks are normally embedded (<abstract>, <othercit>, description>, and <claims>) to avoid scanning each document in its entirety, saving processing time. In addition, the tag containing the International Patent Classification (<classification-ipcr>) was extracted to allow further

[12] https://bulkdata.uspto.gov/

category analyses.[13] Although a patent may belong to more than one category, only the main category was analyzed for simplicity.[14]

The XML files did not use XLink or any other standard tag to mark-up their hyperlinks, so regular expressions (RegEx) were used to extract the URLs. To maximize URL findability, two RegEx were designed. The first is based on internet protocols (finding URLs beginning with the HTTP, HTTPS, or FTP protocols), and the second is based on top-level domains (finding URLs containing specified Top-Level Domains [TLDs]). See Appendix A for the RegEx.

The domain-based RegEx extracted 2,681,314 URLs, while the protocol-based RegEx extracted 59,320 URLs. However, several errors in the XML files were found when executing both RegEx functions, such as links missing elements (e.g., 'mit edu'), protocols wrongly written (e.g., "http:///", "wwww."), typographic errors ("mil.cdu" instead of "mit.edu") or alternative transcriptions (e.g., "4Ctep(dot)cancer(dot)gov(slash)" instead of "4Ctep.cancer.gov/").

Given the relative low volume of URLs from the protocol-based RegEx, all were manually curated, and broken URLs were deleted. This was repeated for the domain-based RegEx applied to patents granted in 2008, giving an error percentage of 5.82%. This is an estimate for the overall error rate of the protocol-based RegEx, which seems low enough to be acceptable, given that it is impractical to manually check the remaining URLs.[15]

A total of 2,719,705 URLs (hereafter referred to as patent outlinks) were obtained by this above process. The websites of these outlinks were then categorized. An entity-based classification scheme (company, service, organization, university, government, and media) was used, covering the main types of organizational entities (Table 1).

Table 1
Website owner entity types for the URL categorization.

Entity	Scope
University	Public or private universities or higher education institutions. For example, Harvard University (harvard.edu).
Government & Public Administrations	Governmental bodies, including research institutions. For example, US National Institutes of Health (nih.gov).
Organization	Organizations, associations, federations or any other institution, excluding organizations specified in other entity descriptions.
Company	The website belongs to public or private companies. For example, Nike (nike.com).
Media	Public or private companies dedicated both to mass media

[13] The field categories are as follows: A - Human necessities; B - Performing operations; transporting; C - Chemistry; metallurgy; D - Textiles; paper; E - Fixed constructions; F - Mechanical engineering; lighting; heating; weapons; blasting engines or pumps; G - Physics; H – Electricity.
[14] https://www.wipo.int/edocs/pubdocs/en/wipo_guide_ipc_2018.pdf
[15] The percentage error over all the years is estimated to be even lower, as errors in the protocol-based RegEx were scarcer for recent years.

	communication and academic publishing. For example, the New York Times (nytimes.com) and Elsevier (elsevier.com), respectively.
Product/Service	The website is dedicated to a specific product or service, regardless of the owner. For example, Science Direct (sciencedirect.com) from Elsevier.

All websites receiving at least 1,000 links from patents (200 websites) were categorized. This corpus represents 943,403 links (34.7%), to the most popular online resources linked from patents. The first author visited each website to assign it an entity type. The second and third author replicated the process to check robustness. Considering the three coders, a Krippendorff's alpha (nominal) value of 0.73 was obtained, which is considered strong enough to validate the results (Krippendorff, 2018).

4. Results

RQ1. How frequently are URLs used in patents?

The percentage of patents embedding at least one outlink (hereafter linking patents) is low (17.1%) but increased from 11.25% in 2008 (20,837 linking patents) to 20.68% in 2018 (70,541 linking patents) (Figure 1). The number of outlinks per patent has also increased over time (Table 2).

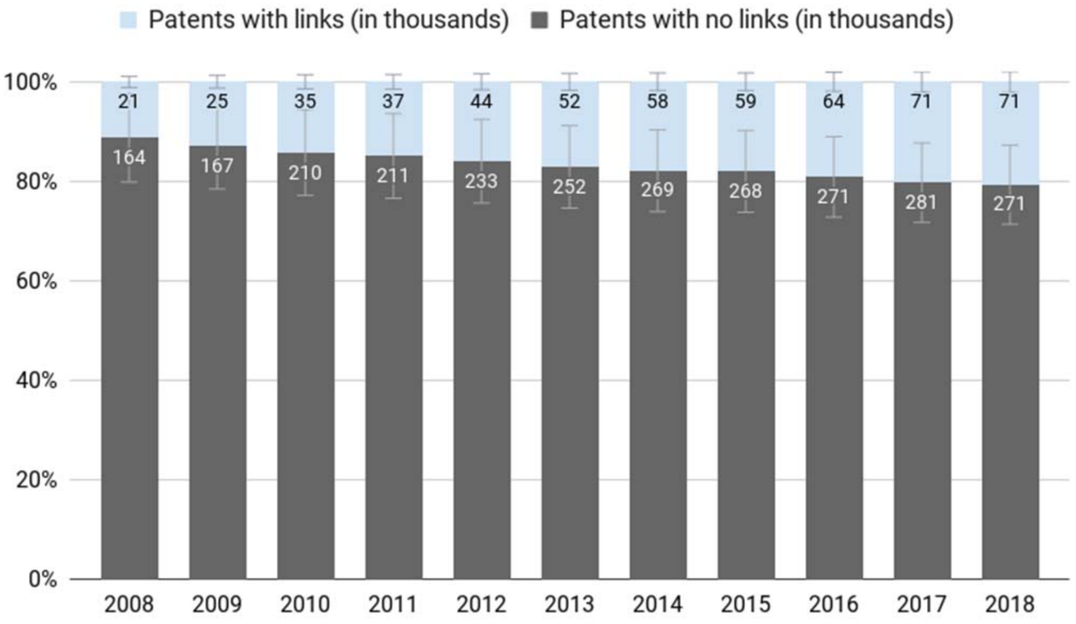

Figure 1
Evolution of the percentage of patents with and without outlinks (2008-2018).
Data source: USPTO
Note: number of patents rounded in thousands

The percentage of unique web domains per year with respect to the total number of outlinks has reduced from 28.2% to 18.3% (Table 2). This suggests an increasing concentration of linked websites (more outlinks referring to the same websites), which might be evidencing a concentration of links to well established web domains. The growth of large websites such as Archive.org, Wikipedia and YouTube, together with the use of URLs from bibliographic databases (PubMed and Microsoft Academic) accompanying non-patent citations to provide access to publications, might be explanatory factors (see below).

Table 2
Volume of URLs extracted from patents granted (2008-2018).

Year	Patents Granted	All outlinks	Rate	Unique outlinks	%
2008	185,260	88,858	0.48	25,060	28.2
2009	192,052	114,569	0.60	29,834	26.0
2010	244,599	170,958	0.70	39,659	23.2
2011	248,101	186,480	0.75	43,739	23.5
2012	277,285	218,437	0.79	49,951	22.9
2013	303,642	272,597	0.90	57,000	20.9
2014	327,014	305,854	0.94	61,787	20.2
2015	326,969	295,829	0.91	60,900	20.6
2016	334,674	323,802	0.97	64,446	19.9
2017	352,547	373,435	1.06	69,913	18.7
2018	341,104	368,886	1.08	67,689	18.3
TOTAL	**3,133,247**	**2,719,705**	**0.87**	**NA**	**NA**

Data source: USPTO.
NA: Not Available. Unique outlinks are calculated only on an annual basis.

The number of outlinks per patent shows an uneven distribution with few patents with a high number of embedded outlinks (the maximum is 1,554 URLs) and most patents having few outlinks, very approximately following a power law distribution (Figure 2).

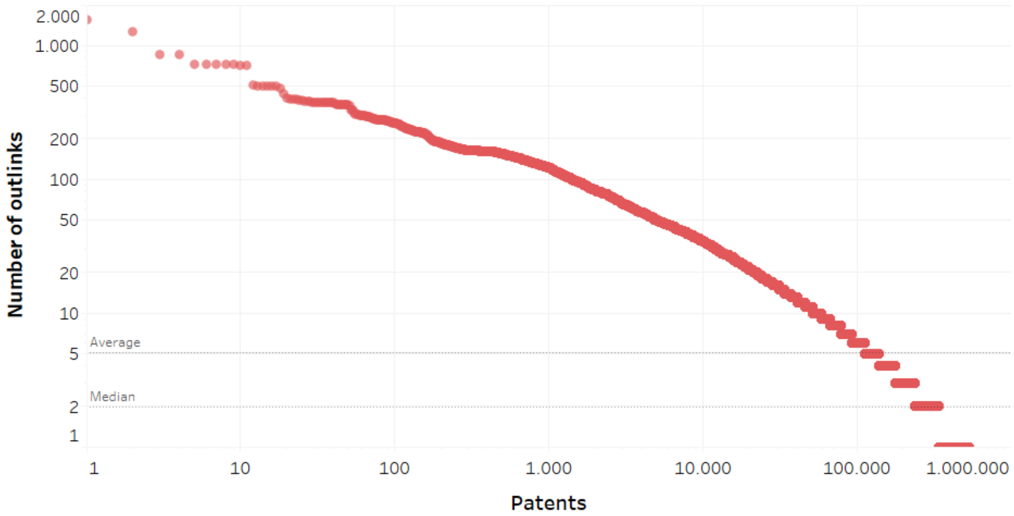

Figure 2
Distribution of outlinks according to patents (2008-2018), using a double logarithmic scale.
Data source: USPTO; powered with Tableau (https://www.tableau.com).
Note: only patents with at least 1 outlink are included.

RQ2. Which application fields use patent URLs most?

The vast majority (90.97%; 488,079) of outlinking patents had an embedded International Patent Classification - Reformed (IPCR) category, with Physics accounting for about half (51.6%). Normalizing the number of outlinks by year and by the number of linking patents per category (size), there are differences in the presence of outlinks in linking patents by IPCR category (Table 3). Physics (G), Electricity (H) and Human necessities (A) had the most outlinks per linking patent in 2018, with a slight increase over time. The number of linking patents and the total number of outlinks per IPCR category and year is available in Appendix B.

Analyzing Physics (G) in greater detail, most linking patents are from Computation (74.7%) and Measuring & Testing (10.89%). Table 4 includes a complete analysis of all Physics subcategories. Thus, the apparent dominance of Physics is misleading since Computation is a different subject.

Table 3
Outlinks per linking patent according to the patent IPCR category (2008-2018)

Year	**Outlinks / linking patent IPCR category**							
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
2008	3.3	3.4	2.7	3.4	3.6	3.1	5.3	3.9
2009	3.4	3.7	3.6	3.1	3.6	3.7	5.8	4.1
2010	3.7	3.2	3.7	2.5	3.2	3.8	6.2	4.4
2011	4.2	3.2	3.8	1.8	3.2	3.6	6.5	4.2
2012	3.9	3.5	3.5	2.1	3.0	3.7	6.3	4.2
2013	4.4	3.3	3.4	2.3	3.4	3.7	6.5	4.6
2014	4.2	3.3	3.3	2.1	3.3	3.5	6.5	5.1
2015	4.4	3.4	3.7	2.7	3.2	3.2	6.3	4.9
2016	4.7	3.3	3.8	3.1	3.0	3.4	6.1	4.8
2017	4.9	3.5	3.7	2.6	3.2	3.7	6.1	5.5
2018	4.9	3.5	3.7	3.3	3.5	3.3	6.0	5.8
Range	**1.6**	**0.1**	**1.0**	**-0.1**	**-0.1**	**0.2**	**0.7**	**1.9**

Data source: USPTO.
Note: only linking patents (those patents with at least one outlink) are considered to avoid zero-effect.
Categories: A - Human necessities; B - Performing operations; transporting; C - Chemistry; metallurgy; D - Textiles; paper; E - Fixed constructions; F - Mechanical engineering; lighting; heating; weapons; blasting engines or pumps; G - Physics; H – Electricity.

Table 4
Number of linking patents and outlinks by the G (Physics) IPCR subcategory

	2008	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**Total**
G01	1,142	378	1,591	1,703	1,787	2,058	2,270	2,293	2,578	2,907	2,765	**21,472**
G02	318	80	290	275	279	348	427	466	561	678	645	**4,367**
G03	146	76	205	223	249	225	230	192	215	218	221	**2,200**
G04	21	16	30	28	32	36	35	20	29	44	44	**335**
G05	148	52	222	237	294	341	373	380	395	479	548	**3,469**
G06	5,982	227	11,760	11,920	14,585	16,905	18,482	16,855	16,802	17,152	16,403	**147,073**
G07	27	18	61	65	69	106	175	168	208	302	295	**1,494**
G08	244	35	342	340	374	402	423	432	475	607	663	**4,337**
G09	216	48	349	358	427	536	646	596	550	551	545	**4,822**
G10	163	56	254	279	297	373	353	361	436	556	482	**3,610**
G11	166	61	227	298	351	345	373	342	399	348	322	**3,232**
G12	2	1	2	1	5	1	1	1	0	0	0	**14**
G16	0	17	1	0	0	0	0	0	0	0	40	**58**
G21	17	0	35	29	42	46	50	52	79	93	62	**505**
Total	**8,592**	**1,065**	**15,369**	**15,756**	**18,791**	**21,722**	**23,838**	**22,158**	**22,727**	**23,935**	**23,035**	**196,988**

Note: G01: Measuring; testing; G02: Optics; G03: Photography; cinematography; analogous techniques using waves other than optical waves; electrography; holography; G04: Horology; G05: Controlling; regulating; G06: Computing; calculating or counting; G07: Checking-devises; G08: Signaling; G09: Educating; cryptography; display; advertising; seals; G10: Musical instruments; acoustics; G11: Information storage; G12: Instrument details; G16: Information and communication technology [ict] specially adapted for specific application fields; G21: Nuclear physics; nuclear engineering; G99: Others.

RQ3. Which patent sections include URLs?

Outlinks are mainly used in the "Other Citations" section (81.75%) as part of the bibliographic records included by inventors to reference prior works. URLs in the "Description" section (18.23%) supplement and enrich the explanations given by inventors in the patent text body. Few URLs (313) appear in the "Claims" section, despite its importance in the patent document (Table 5).

Table 5
Distribution of outlinks by patent section (2008-2018)

Year	Other citations	Description	Claims	Abstract	Total
2008	69,908	18,915	32	3	88,858
2009	87,956	26,594	14	5	114,569
2010	134,918	36,018	19	3	170,958
2011	147,228	39,212	30	10	186,480
2012	174,934	43,476	5	22	218,437
2013	222,390	50,159	35	13	272,597
2014	248,676	57,143	13	22	305,854
2015	241,838	53,963	18	10	295,829
2016	266,230	57,481	90	1	323,802
2017	315,339	58,069	25	2	373,435
2018	314,053	54,798	32	3	368,886
Total	**2,223,470**	**495,828**	**313**	**94**	**2,719,705**

Data source: USPTO

RQ4. What types of web resources are most frequently linked from patents?

Links by top-level domain

256,721 unique web domains linked from patents have been identified, with .com being the most widely linked TLD (61%), followed by .org (18%) and .edu (5%). The country code top-level domains most linked are .uk (United Kingdom) and .jp (Japan) (Figure 3). Appendix C includes the full distribution of outlinks by top-level domain.

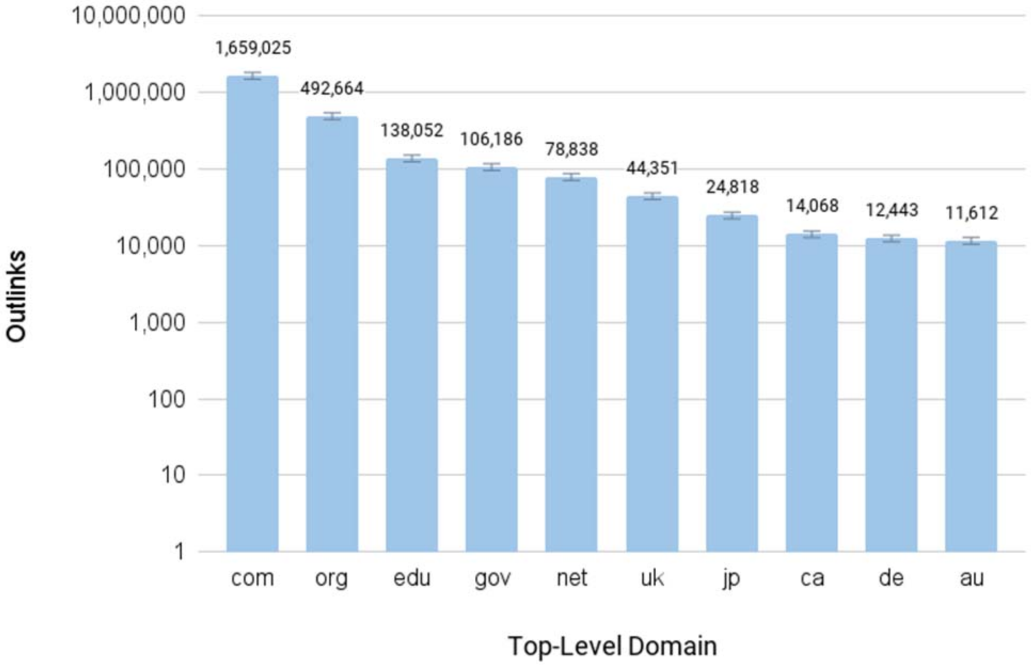

Figure 3
Principal Top-Level domains linked from patents (2008-2018).
Data source: USPTO
Note: only patents with at least 1 outlink are included.

Linked websites

The top-linked websites prominently include online content-oriented resources (e.g., Archive.org, Wikipedia, and YouTube), technological organizations (e.g., Association for Computing Machinery-ACS, The Internet Engineering Task Force-IEEE, and World Wide Web Consortium-W3C) and technological companies (Microsoft, Amazon, IBM) (Table 6). Example.com[16] (11,297 links), was excluded because it is used for URL examples rather than to link to information.[17]

Table 6
Principal Websites linked from patents (2008-2018): domain level[18]

Web domain	Domain Name	Subdomain Level 1	Subdomain Level 2	Subdomain Level 3	Total
archive.org	9,692	**87,177**	246	35	97,150
wikipedia.org*	5,306	**76,700**	179	6	82,191
nih.gov	447	5,397	**35,886**	3,102	44,832
microsoft.com	10,312	**28,199**	772	20	39,303
amazon.com	**33,846**	1,036	143	1	35,026
youtube.com	**23,317**	73	7	24	23,421

[16] Example.com is reserved web domain (it cannot be reserved by users) precisely aimed to be used as examples in texts.
[17] A doubt about the real purpose behind the use of URLs in the patent (as an example or real reference) arises for similar web domains, such as “domain.com” or “company.com”, despite their real existence.
[18] If the domain name corresponds to domain.tld, the first level corresponds to: *.domain.tld/*; second level corresponds to: x.x.domain.tld/*; and the third level corresponds to: *.*.*.domain.tld/*

google.com	10,781	**12,032**	15	16	22,844
ieee.org	642	**21,881**	69	0	22,592
gsmarena.com	**20,435**	42	0	0	20,477
w3.org	**18,500**	475	0	0	18,975
ietf.org	**10,901**	7,928	125	0	18,954
ip.com	**17,569**	258	0	0	17,827
ibm.com	5,309	**8,757**	3,569	96	17,731
clinicaltrials.gov	**15,435**	1	0	0	15,436
psu.edu	155	688	**13,257**	13	14,113
acm.org	1,396	**12,712**	1	0	14,109
3gpp.org	**9,709**	487	6	7	10,209
yahoo.com	3,644	**5,815**	558	61	10,078

Data source: USPTO

Bold values for the subdomain level with the highest value for each web domain.

* Wikipedia.com (which redirects to Wikipedia.org) receives 2,264 additional links.

There are also many links to dictionaries (e.g., merriam-webster.com, 5,984; Dictionary.com, 3,823; thefreedictionary.com, 2,096; dictionary.reference.com, 2,645). Academic and bibliographic resources appear in the internal subdomain levels, such as Pubmed (within the National Institutes of Health, nih.gov), CiteseerX (under the Pennsylvania State University, pst.edu) and Microsoft Academic (under microsoft.com). The highly linked online resources with a first-level subdomain are included in Table 7, mostly illustrating specialist resources from general sites.

Table 7

Principal contents linked from patents (2008-2018): subdomain level

Web domain	**Number of links**
web.archive.org	86,422
en.wikipedia.org	75,092
ieeexplore.ieee.org	18,607
msdn.microsoft.com	11,575
tools.ietf.org	6,998
research.microsoft.com	6,817
dx.doi.org	4,126
delivery.acm.org	4,116
portal.acm.org	3,429
dl.acm.org	3,418
java.sun.com	3,005
technet.microsoft.com	2,891
dictionary.reference.com	2,645
cs.cmu.edu	2,512
schemas.xmlsoap.org	2,197

Analyzing the 200 most linked websites (each with > 1,000 links received), services and company websites are the entities most linked from patents (44.5% of the websites are of these two entity types). There are also many media websites (43 of the 1,000 most linked) (Figure 4), dominated by CNN.com (6,322 links) and The New York Times (nytimes.com) (3,909 links).

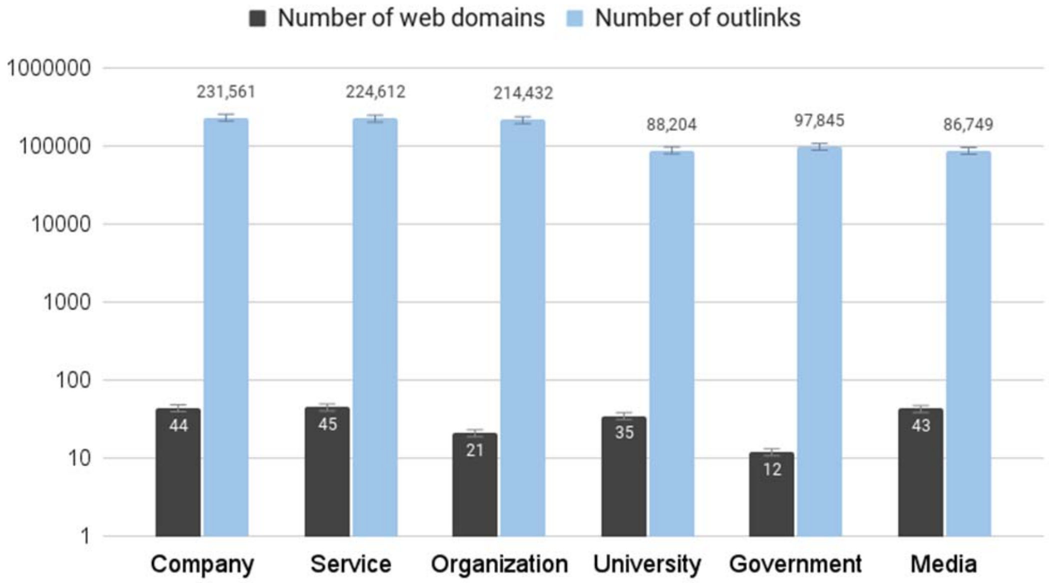

Figure 4

Principal entities linked from patents: entity-based classification

Data source: USPTO

Note: only websites with at least 1,000 outlinks received are included.

Linked files

The likely file type of linked resources was determined from their URL file name extension, ignoring HTML pages (html, php, etc.). PDF files are the most frequently used file type linked from patents (86% of all URLs with a full route found), followed by plain text files (TXT) (6%). Graphics files, especially JPG (4%), are also widely used as informational resources by inventors (Table 8).

Table 8

Principal file types linked from patents (2008-2018).

YEAR	PDF	XLS	DOC	PPT	EPUB	TXT	RTF	PNG	JPG	JPEG	Total
2008	3,811	4	72	42	0	1,116	6	8	618	1	**5,678**
2009	5,489	8	111	65	0	1,176	5	9	496	7	**7,366**
2010	9,939	9	336	165	1	1,653	5	40	611	9	**12,768**
2011	12,032	5	418	181	2	2,091	12	65	752	8	**15,566**
2012	16,985	7	516	275	0	1,648	5	73	888	14	**20,411**
2013	22,765	8	624	258	0	1,635	10	104	903	9	**26,316**
2014	26,151	16	742	281	0	1,698	10	142	1,113	7	**30,160**
2015	26,089	18	713	234	2	1,601	15	206	1,089	9	**29,976**
2016	27,748	7	676	254	0	1,113	1	220	1,286	17	**31,322**
2017	31,056	12	852	246	1	904	2	232	1,371	13	**34,689**
2018	31,058	16	759	223	0	1,007	2	218	1,102	16	**34,401**
Total	**213,123**	**110**	**5,819**	**2,224**	**6**	**15,642**	**73**	**1,317**	**10,229**	**110**	**248,653**

Data source: USPTO

Note: XLS includes XLSX; DOC includes DOCX; PPT includes PPTX.

5. Discussion

The huge numbers of URLs embedded in US patents (2,719,705) shows that they are a substantial new data source that may help enrich patent citation studies. Cited URLs are commonly used within bibliographic references (i.e., patent-publication citations). In this context the function of the databases and repositories (e.g., PubMed and CiteSeerX) is to help readers to retrieve the cited publication rather than to point to a new type of source. However, this use of URLs is not universal, and is presumably based on the preferences of the applicants and examiners.

This study has also revealed the use of URLs to reference other alternative online resources generally omitted in patent citation analyses (patent-object citations). These cited URLs can be included either as part of a reference in the “Other citations” section (see Table 9) or just embedded throughout the text (e.g., in the “Description” section) without being linked to any non-patent citation, such as the following extract of the “Description” section at US patent US20170336412A1:

> “As used herein, the term ‘fusion protein’ means a polypeptide containing a protein or a polypeptide created through the artificial joining of two or more polypeptides (see http://en.wikipedia.org/wiki/peptide)”

Table 9

Different uses for patent outlinks

Non-patent citation including outlinks	Online resource linked
Juola et al., "Learning to Translate: A Psycholinguistic Approach to the Induction of Grammars and Transfer Functions", http://citeseerx.ist.psu.edu, 1995.	Scientific publication deposited in an institutional repository
China Natural Language Open Platform (CNLOP), China Natural Language Open Platform http://www.nlp.org.cn	Specialized platform
The Free Dictionary by FARLEX http://www.thefreedictionary.com, printed Aug. 3, 2012.	Dictionary
Planetlab, 2008. [Online]. Available: http://www.planet-lab.org	Global research network organization
"Traceroute", Oct. 2008 [Online], available http://www.traceroute.org/and ping, as discussed in "ping" Oct. 2008 [Online], available http://en.wikipedia.org/wiki/Ping	2 URLs: an organization and a Wikipedia entry
Bell, "DMC data compression scheme" http://comjnl.oxfordjournals.org/content/32/1/16.abstract.	Scientific publication deposited in the journal website
Linguistic Inquiry and Word Count, http:/liwc.net	Application
Phishing corpus, http://monkey.org/7Ejose/wiki/doku.php?id=PhishingCorpus	Wiki entry
Spence, "The deceptive brain," Journal of the Royal Society of Medicine, vol. 97, No. 1, pp. 6-9, Jan. 2004. [Online]. http://www.ncbi.nlm.nih.gov/pmc/ articles/PMC1079256/pdf/0970006.pdf	Scientific publication, deposited in bibliographic database
Twitter Spam: 3 Ways Scammers are Filling Twitter With Junk, http://web.archive.org/web/20090618173995/http://mashable.com/2009/06/15/twittcr-scams/, 2009, printed Oct. 4, 2012.	Copy of website in archive.org
Goodman, et al., "The use of stylometry for email author identification: a feasibility study." http://utopia.csis.pace.edu/cs691/2007-2008/team2/docs/7.'1 EAM2-TechnicalPaper.061213-Final.pdf, Oct. 2008.	Technical paper deposited in university website

Source: all these references are included as non-patent citations in the US Patent US9292493B2.

Inventors include links to online resources with definitions, explanations and general background information related to procedures, concepts, or phenomena via Wikipedia and/or YouTube. This result reinforces the previous findings obtained by Orduna-Malea, Thelwall and Kousha (2017), who also found many links to Wikipedia and YouTube. Other online resources such as dictionaries or infotainment websites are also employed as evidence, as the following extract from US patent US20070038702A1 exemplifies:

> “The skilled addressee is well aware of how Instant Messaging (IM) works. For example, further details regarding the functionality of IM are provided at http://computer.howstuffworks.com/instant-messaging/htm”

Private companies intensively related to computer science are also highly linked from patents, presumably because many linking patents come from the computation field. A deep analysis of the complete URL route is necessary to check whether these links target specific technological contents or just point to website homepages (i.e., gratuitous links, just giving credit to institutions or centers but not referring to specific information). Further analysis is also needed to filter out those URLs used as examples, especially in programming fields, as the following extract from US patent US10621183B1 exemplifies:

```
“}, {
“rank”: 0,
“description”: “ ”,
“display_text”: “https://en.wikipedia.org/wiki/Donald_Trump”,
“url”: “https://en.wikipedia.org/wiki/Donald_Trump”,
“source_type”: “wikipedia”,
“tags”: [ ]
}],”
```

The intensive use of the Wayback Machine (web.archive.org) also evidences a logical need to have a stable, long-term reference for URLs to protect against their possible change or deletion. In fact, the USPTO’s Manual of Patent Examining Procedure explicitly mention the use of the Wayback Machine as this service stores websites as web captures, with the capture “time/date” in the form of a time stamp and the URL of the original website of capture (Horwitz, Horwitz & Hershman, 2018). This way, “prior art obtained via the Wayback Machine sets forth a prima facie case that the art was publicly accessible at the date and time provided in the time stamp”, reinforcing the authenticity, reliability or accessibility of such information. This not only explains the results obtained but also introduces a challenge, as original URLs are embedded in URLs created by the Wayback Machine, which should be decomposed in future studies to extract the original websites linked. The following extract from US patent US20160104207A1 exemplifies the use of the Wayback Machine to reference a (dated) Wikipedia entry:

> “An example of the k-means algorithm may be found under the title “k-means clustering” on Wikipedia as published on Aug. 27, 2014, herein incorporated by reference in its entirety, which may be found at: http://web.archive.org/web/20140827195754/http://en.wikipedia.org/wiki/K-means_clustering.”

The use of Wayback Machine is presumably rarely needed for cited journal articles, which are normally archived indefinitely by publishers. In fact, the obsolescence of cited URLs requires future works oriented to check the validity and utility of these links (especially if included in the patent Claims).

Applicants and examiners also include references to support patent claims. While references to scholarly publications may reflect a need to recognize prior findings (i.e., knowledge flow from science to industry), references to other online resources may reflect a need to clarify concepts and/or add relevant information.

Beyond the general informational-oriented motivations to include URLs, other reasons might also exist. An extensive body of literature have attempted to understand the underlying reasons for linking in academic contexts (Kim, 2000; Thelwall, 2003; Wilkinson et al., 2003; Bar-Ilan, 2004; 2005; Kousha & Thelwall, 2006; Stuart, Thelwall & Harries, 2007; Kenekayoro, Buckley & Thelwall, 2013). However, most results have limited to the establishment of links typologies (Chu, 2005; Bar-Ilan, 2005) and few have classified linking reasons. Among these, Kim (2000) distinguishes between scholarly, social, and technological reasons to link, while Thelwall (2003) identifies ownership (those links acknowledging authorship or co-authorship of a resource), social (those links with a primarily social reinforcement role), general navigational (those links with a general information navigation function) and gratuitous (those links that serve no communication function) reasons. Although these link taxonomies were proposed for broader environments, they can also apply for patents as well. Further research on the identification, classification and measurement of the specific reasons to insert links in patents would enhance our understanding of URLs usage.

During the analysis of the online resources linked from patents, different bad practices in the inclusion or use of URLs have been identified. Given the importance of doing this procedure properly, both for inventors (access problems to online resources may generate legal problems) and researchers (inaccurate URLs can mislead the interpretation of some results), the following best practices are proposed (Table 10).

Table 10
Best practices to embed URLs in patents

Case	Best practice
Academic URLs	References to scientific publications (e.g., journal articles, books, book chapters, working papers) should always include a URL to allow users access to this document. As the content is not likely to change, the access date is not necessary. The DOI is recommended when possible. Otherwise, thematic or institutional repository handles are also recommended, especially when academic content is not offered under Open Access. Other URLs provided by specific bibliographic databases[a] should be avoided.
Campaign URLs	URLs including UTM parameters[b] from particular analytic campaigns should be avoided. URLs should be written in the clearest and shorten form as possible.
Alternative URLs	The use of alternative transcriptions[c] for URLs should be avoided. These strings do not refer to the real online resource, and prevents automatic analyzes.

Dynamic URLs	Dynamic URLs[d] should be avoided as much as possible, as the online resource may not be reached (e.g., these URLs are long and include diacritics, augmenting the possibility of typographic errors; the database providing the contents to generate the online resource can change or disappear). When friendly URLs are not available, the use of the Wayback Machine is recommended. However, it should be pointed out that this service does not recover some complex dynamic URLs.
Example URLs	While including an example of website or endpoint, the web domain used should self-explain that it is an example (i.e., example.com) instead of using other ways[e]. Failure to do this creates non-existent URLs or refers to already existing URLs unrelated to the patent. The Internet Corporation for Assigned Names and Numbers (ICANN) reserves the “example.com” web domain for this purpose.
Gratuitous URLs	The URLs used should point to the specific resources containing the supporting information. Using the general web domain (e.g., harvard.edu) does not contribute to access directly to the information needed, unless the content mentioned is in the homepage. URLs should also not be used gratuitously or in an ambiguous or imprecise way.
Local URLs	Local URL addresses should be avoided. These URLs can disappear or exhibit access problems if the local server is not operating.[f]
Obsolete URLs	The Wayback Machine service is recommended to avoid legal problems. This will assure future accessibility as well as to evidence the access date.[g] Moreover, the use of Wayback Machine (or any other permanent URL) is also recommended when referring to online resources that can change their contents while maintaining the same URL (e.g., Wikipedia entries). This way, the access date and version used will be transparent to the user.
Short URLs	Using URL shorteners[h] can help to avoid typo errors and facilitate readability. However, these services are proprietary and could disappear (e.g., Google URL shortener was discontinued in 2019), making the online resources unreachable. Therefore, these URLs should be avoided in patents.
URL access date	When the access date cannot be embedded in the URL (e.g., Wayback Machine), it should be manually included in a reference in order to set a fixed point in time when the resource was accessed by the inventor or examiner.[i]
Walled URLs	Online resources containing supplementary information should be accessible without online registration (free or paid) when possible. URLs requiring users to manually pass a registration should be limited.[j]

[a] For example: http://ieeexplore.ieee.org/Xplore/login.jsp?url=http%3A%2F%2Fieeexplore.ieee.org%2Fiel5%2F34%2F29188%2F01316855.pdf&authDecision=-203

[b] For example: http://media.thurne.se/2016/11/BIO-SD-Flyer.pdf?utm_source=TTDKIS%20Filtrox&utm_campaign=004260d1f2-EMAIL_CAMPAIGN_2016_12_12&utm_medium=email&utm_term=0_ca7589e014-004260d1f2-217701685

[c] For example: 4Ctep(dot)cancer(dot)gov(slash)

[d] For example: http://www.burconix.com/?p=services-centrally-managed-wireless

[e] Web domain that does not exist: xx.com; web domain that does already exists: abc.com. http://web.neuro.columbia.edu/members/profiles.php?id=91

[f] For example, the following URL found at US Patent US8296173: http://localhost/UpShot/Help/User/intro.htm

[g] For example, readers cannot know which version corresponds to the following Wikipedia entry, found in US Patent US20200192567A1: www.wikipedia.org/wiki/Letter_frequency.

[h] For example: http://bit.ly/1vHVYOg.

[i] For example: CodeRun, "CodeRun Studio: A free, cross-platform browser-based IDE," <http://www.coderun.com/studio/>, 3 pages (accessed Mar. 10, 2011).

[j] For example, the following news media using paid access: https://www.telegraph.co.uk/foodanddrink/9319624/Rose-Princes-Baking-Club-raspberry-loafcakes.html

6. Conclusions

This study has revealed the degree of use of URLs in patents for the first time, along with certain descriptive aspects (e.g., in which sections they are mainly included and to which websites they are directed). The results also confirm that URLs sometimes play the same role as citations (e.g., when added to a journal article citation) and sometimes a different role (e.g., when referencing a video demonstration). Expanding patent citation analysis to include URLs may therefore enrich it and allow it to identify new patent relationships. These findings show the broad roles that URLs can play when making a patent claim, which may help inventors and evaluators decide which URLs are most appropriate and may help researchers to design new impact indicators for online resources referenced from patents. Limitations to extract and filter out URLs (from the researchers' point of view) and bad practices when including URLs in patent documents (from the inventors' and examiners' points of view) have been also discussed.

Further research is needed to delve into the underlying reasons behind the inclusion of references in patent applications, especially when URLs are included to link online resources, as well as to understand the value and utility of these links for inventors (when applying), examiners (when evaluating) and final users (when reading and searching for information), especially for computer science and technology-related fields. Likewise, exploring other patent offices to check potential geographic differences or patents already granted in other countries where specific online resources might be censored (e.g., YouTube) will be needed to gain a broader understanding of the use of online resources in patents. From the point of view of patent offices, further research is needed to better understand the effects and legal implications of the low quality or inaccessibility of online resources linked from patents. The feasibility of evaluators checking each web resource included (for which they would need protocols or more specific guides) will need to be assessed, as well as the technical support required to embed links in patents with greater ease and precision, so that they could be analyzed more easily by the scientific community.

Specifically, the next steps in this line of research will be centered on the analysis of the linked online resources at the content level (e.g., articles, images, maps, explanatory videos), and especially the role of Wikipedia and YouTube. On the other hand, a thematic analysis of the IPCR categories in greater depth is also needed, as well as to analyze the effects of link obsolescence when permanent links are not used. Similarly, it would also be useful to analyze the links that other resources provide to patents, especially from social networks (e.g., Twitter), to identify communities of interest around patents.

Finally, the proposed best practices for including URLs in patents might be useful for patent offices as a starting point to elaborate guidelines for applicants and examiners, improving the citing of online resources from patents.

Acknowledgements

This paper has been funded by Ministerio de Economía, Industria y Competitividad under the Grant number: BES-2017-079741

References

Alberts, D., Yang, C. B., Fobare-DePonio, D., Koubek, K., Robins, S., Rodgers, M., Symmons, E., & DeMarco, D. (2011). *Introduction to patent searching*. In Lupu M., Mayer K., Tait J., Trippe A. (eds.). Current challenges in patent information retrieval (pp. 3-43). Springer, Berlin, Heidelberg. https://link.springer.com/chapter/10.1007/978-3-642-19231-9_1

Alcácer, J., Gittelman, M., & Sampat, B. (2009). Applicant and examiner citations in US patents: An overview and analysis. *Research Policy*, *38*(2), 415–427. https://doi.org/10.1016/j.respol.2008.12.001

An, J., Kim, K., Mortara, L., & Lee, S. (2018). Deriving technology intelligence from patents: Preposition-based semantic analysis. *Journal of Informetrics*, *12*(1), 217–236. https://doi.org/10.1016/j.joi.2018.01.001

Aristodemou, L., & Tietze, F. (2018). The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data. *World Patent Information*, 55, 37–51. https://doi.org/10.1016/j.wpi.2018.07.002

Bar-Ilan, J. (2004). A microscopic link analysis of academic institutions within a country-the case of Israel. *Scientometrics*, *59*(3), 391–403. https://doi.org/10.1023/b:scie.0000018540.33706.c1

Bar-Ilan, J. (2005). What do we know about links and linking? A framework for studying links in academic environments. *Information Processing & Management*, *41*(4), 973–986. https://doi.org/10.1016/j.ipm.2004.02.005

Breitzman, A. F., & Mogee, M. E. (2002). The many applications of patent analysis. *Journal of Information Science*, *28*(3), 187–205. https://doi.org/10.1177/016555150202800302

Bryan, K. A., Ozcan, Y., & Sampat, B. (2020). In-text patent citations: A user's guide. *Research Policy*, *49*(4), 1–19. https://doi.org/10.1016/j.respol.2020.103946

Campbell, E. G., Powers, J. B., Blumenthal, D., & Biles, B. (2004). Inside the triple helix: Technology transfer and commercialization in the life sciences. *Health Affairs*, *23*(1), 64–76. https://doi.org/10.1377/hlthaff.23.1.64

Carpenter, M.P., Cooper, M., & Narin, F. (1980). Linkage between basic research literature and patents. *Research Management*, 23(2), 30–35. https://doi.org/10.1080/00345334.1980.11756595

Chang, K.-C., Chen, D.-Z., & Huang, M.-H. (2012). The relationships between the patent performance and corporation performance. *Journal of Informetrics*, *6*(1), 131–139. https://doi.org/10.1016/j.joi.2011.09.001

Chu, H. (2005). Taxonomy of inlinked Web entities: What does it imply for webometric research? *Library & information science research*, *27*(1), 8–27. https://doi.org/10.1016/j.lisr.2004.09.002

Collins, P., & Wyatt, S. (1988). Citations in patents to the basic research literature. *Research Policy*, *17*(2), 65–74. https://doi.org/10.1016/0048-7333(88)90022-4

Hammarfelt, B. (in press). Linking science to technology: the "patent paper citation" and the rise of patentometrics in the 1980s. *Journal of Documentation*. https://doi.org/10.1108/JD-12-2020-0218

Horwitz, E., Horwitz, L., & Hershman, L. (2018). *Manual of Patent Examining Procedure*. *US Department of Commerce*.
https://mpep.uspto.gov/RDMS/MPEP/current

Jaffe, A., Trajtenberg, M., & Fogarty, M. (2000). *The meaning of patent citations: Report on the NBER/Case-Western Reserve Survey of Patentees*. *NBER Working Papers*, 7631.
http:// www.nber.org/papers/w7631.pdf

Kenekayoro, P., Buckley, K., & Thelwall, M. (2013). Motivation for hyperlink creation using inter-page relationships.
https://arxiv.org/pdf/1311.1082.pdf

Kim, H. J. (2000). Motivations for hyperlinking in scholarly electronic articles: A qualitative study. *Journal of the American Society for Information Science*, *51*(10), 887–899.
https://doi.org/10.1002/1097-4571(2000)51:10%3C887::aid-asi20%3E3.0.co;2-1

Kim, J., & Lee, S. (2015). Patent databases for innovation studies: A comparative analysis of USPTO, EPO, JPO and KIPO. *Technological Forecasting and Social Change*, *92*(2015), 332–345.
https://doi.org/10.1016/j.techfore.2015.01.009

Kousha, K., & Thelwall, M. (2006). Motivations for URL citations to open access library and information science articles. *Scientometrics*, *68*(3), 501–517.
https://doi.org/10.1007/s11192-006-0126-9

Kousha, K. & Thelwall, M. (2017). Patent citation analysis with Google. *Journal of the Association for Information Science and Technology*, *68*(1), 48–61.
https://doi.org/10.1002/asi.23608

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th Ed.). Thousand Oaks (CA), Sage.

Latour, B. (1987), *Science in Action: How to Follow Scientists and Engineers through Society*. Harvard University Press, Cambridge, MA.

Lee, C., Kwon, O., Kim, M., & Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. Technological Forecasting and Social Change, *127*, 291–303.
https://doi.org/10.1016/j.techfore.2017.10.002

Michel, J., & Bettels, B. (2001). Patent citation analysis: A closer look at the basic input data from patent search reports. *Scientometrics*, *51*(1), 185–201.
https://doi.org/10.1023/A:1010577030871

Meyer, M. (2000a). Does science push technology? Patents citing scientific literature. *Research Policy*, *29*(3), 409–434.
https://doi.org/10.1016/S0048-7333(99)00040-2

Meyer, M. (2000b). What is special about patent citations? Differences between scientific and patent citations. *Scientometrics*, *49*(1), 93–123.
https://link.springer.com/content/pdf/10.1023/A:1005613325648.pdf

Meyer, M., Siniläinen, T., & Utecht, J. T. (2003). Towards hybrid triple helix indicators: A study of university-related patents and a survey of academic inventors. *Scientometrics*, *58*(2), 321–350.
https://doi.org/10.1023/A:1026240727851

Narin, F., Hamilton, K.S., & Olivastro, D. (1997). The increasing linkage between US technology and public science. *Research Policy*, *26*(3), 317–330.
https://doi.org/10.1016/s0048-7333(97)00013-9

Oppenheim, C. (2000). *Do patent citations count?* In B. Cronin & H.B. Atkins (Eds.), The web of knowledge: A festschrift in honor of Eugene Garfield (pp. 405–432), ASS Monograph Series. Medford, NJ: Information Today.

Orduna-Malea, E., Thelwall, M., & Kousha, K. (2017). Web citations in patents: Evidence of technological impact? *Journal of the Association for Information Science and Technology*, *68*(8), 1967–1974.

https://doi.org/10.1002/asi.23821

Rip, A. (1986). *Mobilising resources through texts*. In Callon, Law and Rip (Eds). Mapping the Dynamics of Science and Technology (pp. 84-99). Macmillan Press, London.

Roach, M., & Cohen, W.M. (2013). Lens or prism? Patent citations as a measure of knowledge flows from public research. *Management Science*, *59*(2), 504–525. https://doi.org/10.1287/mnsc.1120.1644

Schmoch, U. (1993). Tracing the knowledge transfer from science to technology as reflected in patent indicators. *Scientometrics*, *26*(1), 193–211. https://doi.org/10.1007/bf02016800

Sharma, P., & Tripathi, R. C. (2017, December). Patent citation: A technique for measuring the knowledge flow of information and innovation. *World Patent Information*, *51*, 31–42. https://doi.org/10.1016/j.wpi.2017.11.002

Stuart, D., Thelwall, M., & Harries, G. (2007). UK academic web links and collaboration-an exploratory study. *Journal of Information Science*, *33*(2), 231–246. https://doi.org/10.1177/0165551506075326

Szu-chia, S. L. (2010). Scientific linkage of science research and technology development: A case of genetic engineering research. *Scientometrics*, *82*(1), 109-120. https://doi.org/10.1007/s11192-009-0036-8

Thelwall, M. (2003). What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information research*, *8*(3), 8–3. http://informationr.net/ir/8-3/paper151.html

Thelwall, M., & Kousha, K. (2015). Web indicators for research evaluation. Part 1: Citations and links to academic articles from the Web. *Profesional de la información*, *24*(5), 587–606. https://doi.org/10.3145/epi.2015.sep.08

Van Looy, B., & Magerman, T. (2019). *Using text mining algorithms for patent documents and publications*. In W. Glänzel, H. Moed, U. Schmoch, M. Thelwall (eds.). Springer Handbook of Science and Technology Indicators (pp. 929–956). Springer, Cham. https://doi.org/10.1007/978-3-030-02511-3_38

Wilkinson, D., Harries, G., Thelwall, M., & Price, L. (2003). Motivations for academic Web site interlinking: Evidence for the Web as a novel source of information on informal scholarly communication. *Journal of information science*, *29*(1), 49–56. https://doi.org/10.1177/016555150302900105

Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *Journal of High Technology Management Research*, *15*(1), 37–50. https://doi.org/10.1016/j.hitech.2003.09.003

Appendix A

Three-stepped Regular Expressions (RegEx) for hyperlink extraction

<table>
<tr><th colspan="2">PROTOCOL-BASED REGEX</th><th colspan="2">DOMAIN-BASED REGEX</th></tr>
<tr><th>Step</th><th>Scope</th><th>Step</th><th>Scope</th></tr>
<tr><td>(http|ftp|https)://</td><td>Searching for the protocols in the text followed by the bars '/'</td><td>[a-zA-Z0-9][a-zA-Z0-9\.-]</td><td>Any combination of letters and numbers that contains uppercase and lowercase letters, numbers and / or the hyphen symbol '-'.</td></tr>
<tr><td>([\w+?\.\w+])</td><td>Which may or may not be followed by the www protocol</td><td>*\.(ae|ai|ar|au|az|bd|be|bg|ca|cf|ch|cn|co(m)?|cz|dk|ee|es|eu|fr|ge|gr|hk|hr|hu|id|ie|il|io|ir|jp|kr|kz|lk|lt|lv|ma|me|mx|my|ng|nl|no|nz|ph|pk|pl|pt|ro|rs|ru|sa|sg|si(te)?|sk|su|tk|tr|tv|tw|ua|uk|us|uz|vn|za|info|live|net|online|org|shop|store|xyz|biz|pro|edu|gov)</td><td>Followed by a period symbol '.' And one of the possible selected TLDs.</td></tr>
<tr><td>+([a-zA-Z0-9\~\!\@\#\$\%\^\&amp;*\(\)_\-\=\+\\\/\?\.\:\;\'\,]+)</td><td>And that presents any alphanumeric combination (with or without a hyphen) that precedes a period '.' And that may or may not be followed by a slash '/' and another alphanumeric combination with / without symbols.</td><td>\b(:\d+)?(\/[-a-zA-Z0-9@:%_—-\+\.~#\?&amp;//=\$,;ºª*\\]+)?"</td><td>Which may or may not be followed by a slash symbol '/' followed by any alphanumeric combination and the symbols:</td></tr>
</table>

Note: The 92 TLDs in the TLD-based RegEx (listed in the step 2 of the formula) represent 97.7% of all internet domains.
Source: https://w3techs.com/technologies/overview/top_level_domain

Appendix B

Number of linking patents (LP) and outlinks (OL) by IPCR category

Year	A		B		C		D		E		F		G		H		ACP	AP
	LP	OL	LP	OL	LP	OL	LP	OL	LP	OL	LP	OL	LP	OL	LP	OL		
2008	2,284	7,429	1,179	4,019	1,757	4,778	38	128	241	857	419	1,298	8,599	45,562	3,623	14,294	**18,140**	**20,837**
2009	2,954	9,993	1,267	4,715	2,031	7,301	59	182	270	959	492	1,823	10,655	62,191	4,421	17,961	**22,149**	**24,624**
2010	4,953	18,140	1,910	6,154	2,741	10,264	80	197	482	1,535	672	2,556	15,372	95,909	5,991	26,107	**32,201**	**34,824**
2011	5,656	24,009	2,035	6,604	3,206	12,087	96	169	479	1,533	748	2,726	15,758	101,754	6,583	27,863	**34,561**	**36,900**
2012	7,390	29,149	2,403	8,381	3,250	11,362	119	245	610	1,830	857	3,205	18,798	118,253	8,087	33,746	**41,514**	**44,212**
2013	9,003	39,500	2,742	9,179	3,829	13,005	130	294	678	2,331	1,181	4,409	21,726	141,763	9,625	43,831	**48,914**	**51,913**
2014	10,172	43,186	3,046	9,908	4,464	14,918	138	285	823	2,690	1,217	4,309	23,838	154,605	11,415	58,352	**55,113**	**58,451**
2015	9,845	42,931	3,176	10,905	5,065	18,835	127	346	806	2,573	1,456	4,681	22,158	139,699	12,010	58,675	**54,643**	**58,846**
2016	10,232	48,309	3,342	11,024	5,343	20,159	126	389	811	2,422	1,649	5,584	22,727	138,293	12,866	61,736	**57,096**	**63,947**
2017	11,428	56,414	3,885	13,661	5,761	21,326	147	378	1,000	3,165	1,926	7,061	23,935	146,493	14,308	78,026	**62,390**	**71,469**
2018	11,045	53,881	4,046	14,028	5,904	21,998	193	629	951	3,316	2,000	6,666	23,035	137,262	14,184	81,584	**61,358**	**70,541**
Total	**84,962**	**372,941**	**29,031**	**98,578**	**43,351**	**156,033**	**1,253**	**3242**	**7,151**	**23,211**	**12,617**	**44,318**	**206,601**	**1,281,784**	**103,113**	**502,175**	**488,079**	**536,564**

Data source: USPTO.

ACP: All categorized patents; AP: all patents

Note: only linking patents (those patents with at least one outlink) are considered.

Categories: A - Human necessities; B - Performing operations; transporting; C - Chemistry; metallurgy; D - Textiles; paper; E - Fixed constructions; F - Mechanical engineering; lighting; heating; weapons; blasting engines or pumps; G - Physics; H – Electricity.

Appendix C

Distribution of outlinks according to the top-level domain (TLD)

TLD	Outlinks	TLD	Outlinks	TLD	Outlinks	TLD	Outlinks	TLD	Outlinks
com	1,659,025	sg	1,167	im	91	cr	6	vg	1
org	492,664	hr	1,071	az	91	sh	5	vc	1
edu	138,052	me	1,060	cx	85	ps	5	tt	1
gov	106,186	hu	952	online	80	lv	5	systems	1
net	78,838	tr	756	lv	78	ibm	5	sy	1
uk	44,351	xyz	698	is	77	global	5	style	1
jp	24,818	my	657	live	58	tl	4	studio	1
ca	14,068	ee	519	nu	54	ong	4	ss	1
de	12,443	mx	514	ge	53	mc	4	so	1
au	11,612	ua	499	bs	46	life	4	science	1
fr	9,319	rs	494	am	43	apple	4	sca	1
ch	7,991	store	444	ms	40	yahoo	3	reviews	1
nl	6,859	si	425	md	34	top	3	review	1
us	6,376	pk	407	kz	34	sm	3	press	1
ng	5,344	cc	405	technology	33	page	3	pink	1
info	5,287	ar	395	lu	29	ny	3	parts	1
int	5,197	br	382	mp	28	link	3	one	1
eu	5,074	ro	376	ve	26	il	3	om	1
id	4,570	sk	348	asia	21	dev	3	nf	1
be	3,910	sa	348	shop	17	dell	3	na	1
io	3,621	ae	344	by	17	ci	3	mu	1
co	3,585	ir	340	yu	16	bank	3	mn	1
no	3,486	ly	328	ni	16	tz	2	mk	1
cn	3,132	th	317	mom	15	tn	2	mit	1
dk	3,032	ai	315	et	15	tc	2	microsoft	1
kr	2,743	su	311	li	12	red	2	media	1
pl	2,739	pro	286	cy	12	re	2	je	1
ru	2,520	bg	262	as	11	pe	2	ink	1
tw	2,466	fm	257	tech	10	nyc	2	ht	1
il	2,269	site	255	ne	10	ninja	2	health	1
nz	2,211	ws	225	education	10	news	2	guru	1
se	2,210	ph	212	cm	10	museum	2	gi	1
biz	2,148	lt	199	ag	10	mg	2	fo	1
it	1,951	mobi	184	today	9	kg	2	fk	1
es	1,834	vn	161	test	9	help	2	fishing	1
tv	1,822	name	158	ba	9	guide	2	docs	1
cz	1,569	gl	156	space	8	gs	2	do	1
mil	1,565	lk	151	sc	8	final	2	digital	1
gr	1,554	to	139	la	8	energy	2	cu	1
hk	1,477	cl	131	dz	8	eh	2	clothing	1
za	1,465	tk	129	ac	8	cern	2	cg	1
pt	1,430	cf	119	st	7	bt	2	cfd	1
in	1,418	uz	102	mo	7	blue	2	cba	1
at	1,386	ma	101	google	7	al	2	cat	1
fi	1,361	bd	99	gg	7	zone	1	bike	1
ie	1,318	watch	94	bz	7	xin	1	alibaba	1