U.S. Patent and Trademark Office
OFFICE OF THE CHIEF ECONOMIST
OFFICE OF THE CHIEF TECHNOLOGY OFFICER
Economic Working Paper Series

USPTO Patent Prosecution Research Data: Unlocking Office Action Traits

Qiang Lu
U.S. Patent and Trademark Office

Amanda Myers
U.S. Patent and Trademark Office

Scott Beliveau
U.S. Patent and Trademark Office

November 2017

USPTO Economic Working Paper No. 2017-10
Digital Services & Big Data Project Series No. 2017-v1-1

The USPTO Office Action Research Dataset for Patents is available at:
https://bulkdata.uspto.gov/data/patent/office/actions/bigdata/2017/.

USPTO Patent Prosecution Research Data: Unlocking Office Action Traits

Qiang Lu
U.S. Patent and Trademark Office
Enterprise Advanced Analytics Branch

Amanda Myers
U.S. Patent and Trademark Office
Office of the Chief Economist

Scott Beliveau
U.S. Patent and Trademark Office
Enterprise Advanced Analytics Branch

November 2017

Abstract

Release of the United States Patent and Trademark Office (USPTO) *Office Action Research Dataset for Patents* marks the first time that comprehensive data on examiner-issued rejections are readily available to the research community. An "Office action" is a written notification to the applicant of the examiner's decision on patentability and generally discloses information, such as the grounds for a rejection, the claims affected, and the pertinent prior art. The relative inaccessibility of Office actions and the considerable effort required to obtain meaningful data therefrom has largely prevented researchers from fully exploiting this valuable information. We aim to rectify this situation by using natural language processing and machine learning techniques to systematically extract information from Office actions and construct a relational database of key data elements. This paper describes our methods and provides an overview of the main data files and variables. This data release consists of three files derived from 4.4 million Office actions mailed during the 2008 to mid-2017 period from USPTO examiners to the applicants of 2.2 million unique patent applications.

Keywords: Patents, patent examination, patent quality, patent examiners, USPTO

JEL Classification Numbers: O3, O31, O32, O34, O38

Acknowledgements: We thank Robert Kimble for excellent programing assistance. We also thank Andrew Toole, Thomas Beach, Jesse Frumkin, Asrat Tesfayesus, Peter Mehravari, and Pinchus Laufer for valuable comments and suggestions.

The USPTO *Office Action Research Dataset for Patents* is available at: https://bulkdata.uspto.gov/data/patent/office/actions/bigdata/2017/.

I. Introduction

The United States Patent and Trademark Office (USPTO) *Office Action Research Dataset for Patents* contains detailed information derived from Office actions issued by patent examiners to applicants during the patent examination process. An "Office action" is a written notification to the applicant of the examiner's decision on patentability. It generally discloses the reasons for any rejections, objections, or requirements and includes relevant information or references that the applicant may find useful for responding to the examiner and deciding whether to continue prosecuting the application.

Office actions, as well as incoming and other outgoing documents related to examination, are accessible as image files for granted patent and publicly available patent applications via the USPTO Public Patent Application Information Retrieval (PAIR) system.[1] However, Public PAIR does not currently allow for bulk downloads and only provides files in image formats whereas, for Office actions, the text versions stored internally at USPTO are more compatible with standard text analytic tools. Moreover, because these documents contain structured and unstructured text, sophisticated text mining and analytic methods are necessary to systematically identify key data elements, such as the grounds for a rejection, the claims affected, and the prior art cited. This information is particularly useful for the growing body of empirical literature surrounding the patent examination process, examiner heterogeneity, and application and litigation outcomes (Cockburn et al. 2003; Lichtman 2004; Lemley and Sampat 2012; Mann 2014; Carley et al. 2015; Frakes and Wasserman 2015). To our knowledge, only a few recent papers utilize data from a comprehensive sample of Office actions collected through computational- and resource-intensive methods. Frakes and Wasserman (2017) leverage data on rejections collected from Public PAIR via the National Center for Supercomputing Applications at the University of Illinois. Kuhn et al. (2017) and Thompson and Kuhn (2016) use extensive cloud computing capability and optical character recognition to convert millions of Public PAIR image files to text. The relative inaccessibility of Office actions and the considerable effort required to obtain meaningful data therefrom has deterred broader use by researchers and largely prevented scholars and policymakers from fully exploiting the valuable information stored in these documents.

We aim to rectify this situation by using machine learning techniques to systematically extract information from Office actions and construct a relational database of key data elements. We are making these data readily available to the research community and other stakeholders as the USPTO *Office Action Research Dataset for Patents* (hereafter "*Dataset*"). This initial release of the *Dataset* consists of three data files derived from 4.4 million Office actions mailed from 2008 through mid-July 2017[2] for 2.2 million unique patent applications.[3] Rejections for obviousness are by far the most prevalent, occurring on 79 percent of Office actions in the *Dataset*. We observe rejections for lack of novelty in roughly 42 percent of actions and rejections related to the written description of the invention or clarity of the claims in just over one-third. Relatively few Office actions in the *Dataset*, about 11 percent, contain a rejection related to patent subject matter eligibility, statutory double patenting, utility, or inventorship.[4]

[1] Office actions are also available as image files from the USPTO Global Dossier, *see* https://globaldossier.uspto.gov.
[2] The time for filing a response to an Office action begins from the mail date whether the action is conveyed via electronic or paper delivery. The applicant is given three months to respond to the Office Action with a possibility of extension for added fees. The *Dataset* does include a small number of Office actions with mailing dates in 2001. However, because these Office actions were issued to patent applications filed in or after 2008, we suspect the 2001 mailing date is the result of human or encoding error.
[3] The *Dataset* includes published patent applications as well as non-published applications made publicly available upon publication of a child application. Published and publicly available applications were identified based on those available via PAIR Bulk Data, *see* https://pairbulkdata.uspto.gov/.
[4] The preceding paragraph includes various patent-related terms of art, including "obviousness", "lack of novelty", "patent subject matter eligibility", etc. We discuss these terms broadly in Section II and provide general definitions in Section IV.

The *Dataset* also provides new information on patents and published patent applications cited as "prior art" in the patent examination process. Economic and legal scholars often use patent citations as an indicator of patent value (e.g., Harhoff et al. 2002; Hall et al. 2005; Sampat and Ziedonis 2004), patent quality (e.g., Lanjouw and Schankerman, 2004), and knowledge diffusion (e.g. Jaffe et al. 1993; Jaffe et al. 2000; Jaffe and Trajtenberg 2002). More recent studies, however, call into question the effectiveness and validity of patent citation metrics. One criticism is that patent examiners, rather than inventors or their agents, account for a large share of citations on patent documents (Sampat 2010; Alcacer and Gittelman 2006; Alacer et al. 2009). Examiner added citations are not likely to reflect the knowledge available to or used by inventors and this undermines the interpretation that patent citations represent knowledge flows. Additionally, there are strategic motivations for patent agents or attorneys to search for and cite prior art, raising concerns regarding underreporting (Kesan 2002; Lemley and Tangri 2003; Sampat 2010) as well as overreporting (Cotropia et al. 2013). On the other side of the debate, assessments of patent citations using inventor surveys (Jaffe et al. 2000) or other metrics (Nelson, 2009; van Zeebroeck 2011) generally find that citations are a useful, albeit noisy, indicator of knowledge flows and patent value.

In the *Dataset*, we link prior art citations found in the text of Office actions to prior art cited by applicants and examiners on official USPTO forms. This allows *Dataset* users to identify the specific prior art used by the examiner as the basis for a rejection in an Office action. It also allows users to more precisely identify which patents were cited by the applicant and the examiner.[5] We see this as a significant contribution to the debate on the usefulness of patent citations in economic and legal research.

The *Dataset* serves as a proof of concept to solving the challenges of access to public Office action data. Our intention is to make regular updates and enhancements to the *Dataset* to enable researchers and policymakers to derive valuable insights from the wealth of information captured in Office actions. This effort is made possible by the USPTO Digital Services & Big Data (DSBD) portfolio in collaboration with the USPTO Office of the Chief Economist (OCE). The DSBD's mission is to improve public access and usability of USPTO data and investigate and standardize the agency's big data infrastructure to deliver advanced analytic capacity. The DSBD collaborated with OCE on this project to capitalize on the latter office's experience and ongoing efforts to make "research-ready" datasets available to economic and legal scholars and, thereby, foster research on the role of intellectual property in the economy.

The remainder of this paper is structured as follows. Section II provides a basic overview of the examination process as background for *Dataset* users. Section III details the methods used for generating the *Dataset* from the structured and unstructured text of the Office actions. In Section IV, we review the three main data files that comprise the *Dataset* and define the variables therein. Section V considers the coverage and other limitations of the *Dataset*. Section VI concludes.

II. Patent Examination Process Background

It is useful for *Dataset* users to have a basic understanding of how patents typically proceed through examination. In this section, we give a concise synopsis. Marco et al. (2017) provides a more thorough primer on the patent examination process and the examiner performance appraisal system. We also note additional references throughout this document to direct data users to more authoritative and detailed sources of information.

[5] Prior art patent citation data captured on the front page of a U.S. patent grant or U.S. pre-grant publication does identify whether a reference is cited by the examiner, the applicant, or a third party. However, if a prior art reference is cited by both the examiner and the applicant, the front page of the patent grant or pre-grant publication will only indicate that it is cited by the examiner.

In general, the patent examination process initiates when an applicant files for a patent with the USPTO. Upon receipt, the application goes through pre-examination review to ensure that the application is complete, all necessary forms are filed, and all relevant fees are paid. A complete application includes a written description of the invention (called a "specification"), at least one claim, and any necessary drawings.[6] As part of this pre-examination review, the claims of the application are classified and forwarded to the relevant USPTO technology center for examination. Within the technology center, the application is assigned to a patent examiner in one of the group art units.[7]

The examiner evaluates the claims in the application for compliance with applicable statutes and regulations.[8] She checks to make certain that the claims are directed to patent-eligible subject matter, that the written description is adequate to describe and enable the claimed invention, and that the claims clearly define the invention. She also conducts a prior art search to determine whether the claimed invention is novel and nonobvious. She looks for previous patent documents[9] or non-patent literature to determine whether the invention is anticipated by a single reference, or rendered obvious either by a single reference or by a combination of references.[10] Based on this examination, the examiner may either allow all claims or issue an Office action indicating a Non-Final Rejection that rejects or objects to one or more of the claims.[11]

A typical Non-Final Rejection Office action identifies the specific claims and the statutory or non-statutory grounds on which those claims are objected to and/or rejected. Generally, statutory rejections are based on non-compliance with applicable patent law statutes in United States Code Title 35 (35 U.S.C.).[12] A non-statutory basis for rejection or objection is grounded in judicial doctrine or patent rules found in Title 37 Code of Federal Regulations.[13] A single Office action may include multiple grounds for objection and/or rejection applying to different or intersecting sets of claims. For certain statutory rejections, the examiner will cite in the Office action the previous patent documents and/or non-patent literature references to support the rejection. The Office action may also identify specific claims that would be allowed should the applicant overcome the raised objections or claims that are allowable without any objection. Figure 1 contains extracts of Office actions, highlighting key elements of the text, including the action taken on the claims.

USPTO recommends that examiners utilize Office action templates with standardized headings and custom form paragraphs to render documents consistent, easy to read, and legally proper. Standardized headings and form paragraphs provide legal terms and definitions relevant to the objections and/or rejections being raised (see Figure 1).

There are multiple templates available to examiners, and they can design their own templates for subsequent use. Thus, the structure of and text included in Office actions can vary considerably between examiners. Additionally, because examiners are not required to use headings and form paragraphs, some

[6] *See* Manual of Patent Examining Procedure (MPEP) 601.01. A filing date is assigned when the application is complete.
[7] Technology centers are comprised of work groups which are further comprised of group art units. See the definition for the *art_unit* variable for more details.
[8] Prior to examining the claims, the examiner may issue a restriction if multiple inventions appear in the claims. The applicant would then be required to choose claims drawn to a single invention (see MPEP 803). If the applicant wishes to pursue patent protection on the additional inventions that are not chosen, one or more divisional applications may be filed.
[9] Patent documents include both U.S. and foreign issued patents and pre-grant publications.
[10] *See* MPEP 2103 for a detailed overview of the patent examination process.
[11] If the examiner decides to allow all claims at this stage, the communication sent to the applicant is referred to as a first action on the merits allowance. It is also possible for the examiner to issue an office action (called an Ex parte Quayle) indicating that although the subject matter of the examined claims are allowable, certain formal requirements still remain and must be addressed. First action allowance and Ex parte Quayle actions are not included in the *Dataset*.
[12] *See* https://www.uspto.gov/web/offices/pac/mpep/consolidated_laws.pdf.
[13] *See* https://www.uspto.gov/web/offices/pac/mpep/consolidated_rules.pdf.

Office actions consist more of prose or free-form text with no or minimal headings. Such Office actions generally contain relevant legal terms and definitions embedded in the text. See Figure 2 for an example of a double patenting rejection with no headings or legal form paragraphs.

Upon receiving a Non-Final Rejection, the applicant is generally given three months to respond but may take up to three additional months in exchange for added fees. The applicant typically submits a response with some combination of arguments and amendments to the claims to clarify them or to narrow their scope to avoid encompassing the prior art.[14] After the examiner receives the applicant's response, she re-evaluates the claims to determine whether the rejections or objections have been overcome. If no issues remain, the applicant is informed that the claims are allowable.[15] Otherwise, the examiner may find the applicant's arguments to be insufficient to overcome the rejections or objections, or that the applicant's amendments raise further issues that preclude allowance of the claims. The examiner then issues an Office action indicating a Final Rejection, which generally follows the same format and legal elements as a Non-Final Rejection. Although the Final Rejection closes the first round of the examination process, the applicant may continue to seek patent protection through various mechanisms.[16]

III. Data Generation Process

To construct the *Dataset*, we first retrieve Office actions for patent applications in the 12, 13, 14, and 15 series. The series, or the first two digits of the application number, gives a rough indication of the order in which applications were received by the USPTO. Generally, series 12 through 15 cover applications with filing dates in the 2008 to 2017 period.[17] Utility patent applications comprise the vast majority of applications in the 12 through 15 series, but these series also include applications for plant patents as well as reissues. Applications for design patents are excluded.

For applications in the 12 through 15 series, we attempt to locate and process, from internal USPTO servers, all Office actions identified as a Non-Final Rejection or a Final Rejection.[18] We successfully process these text documents for the vast majority of applications in the 12 through 15 series. However, in a number of cases, document quality issues interfere with processing, resulting in less than complete coverage (see Section V on Limitations). We also remove any applications not published or made publicly available.[19] The resulting set consists of 4,384,532 Office actions issued from examiners to the applicants of 2,188,039 unique patent applications.

When an examiner raises an objection and/or rejection to claim(s) in an Office action, she will typically follow a three step process consisting of: (1) entering an appropriate heading for the action, (2) inserting an appropriate set of form paragraphs to specify the legal grounds for the action, and (3) applying a

[14] The applicant may also file information disclosure statements, which are used to comply with the applicant's duty to disclose any information relevant to patentability. This information typically includes potential prior art, particularly when revealed to the applicant during the examination of a related foreign or domestic application. The applicant may also ask for a telephone or in-person interview with the examiner.
[15] The Notice of Allowability indicates which claims are allowed. Notices of Allowability are not included in the *Dataset*.
[16] The applicant may file an "after final" response to the examiner including arguments and/or amendments to the claims. The examiner may then either allow the application, or alternatively respond in an "Advisory Action" and address each of the applicant's arguments, or explain that the amendments are not entered because they require further search and/or consideration. Advisory Actions are not included in the *Dataset*. Alternatively, the applicant may continue to seek patent protection before the examiner by filing a Request for Continued Examination (RCE). The applicant may also file an appeal to the examiner's rejections with the USPTO's Patent Trial and Appeal Board (PTAB) arguing that the PTAB should reverse the examiner's rejections. Lastly, the applicant may pursue a separate invention that was disclosed in their original specification by filing a new continuation application, which is issued a new application number but is entitled to the benefit of the filing date of the original application.
[17] Because patent application serial numbers are assigned chronologically to patent applications filed at USPTO, application serial numbers and filing dates will generally correspond. However, there are exceptions. *See* https://www.uspto.gov/web/offices/ac/ido/oeip/taf/filingyr.htm.
[18] We distinguish office actions indicating a Non-Final Rejection or a Final Rejection based on Image File Wrapper (IFW) document codes "CTNR" and "CTFR," respectively. *See* https://www.uspto.gov/sites/default/files/patents/process/status/top_40_eOA_doc_codes.xls. See the definition for the *document_cd* variable for more details.
[19] Published and publicly available applications were identified based on those available via PAIR Bulk Data, see https://pairbulkdata.uspto.gov/.

"standard" sentence structure to express the action taken on the claim(s), as illustrated in Figure 1. We take advantage of this process to extract key data elements from the text of each Office action.

We first perform text segmentation, dividing the text of each document into the three units which correspond to each step: (1) heading, (2) form paragraph, and (3) action sentence. We then develop two methods for processing the text in each segment in order to construct the *Dataset*. Figure 3 illustrates the overall process flow, which we discuss in detail below.

a. Classifying Headings and Form Paragraphs

The first method classifies headings and form paragraphs into types of action taken based on similarity to a pre-labeled set of standardized headings and form paragraphs derived from examiner tools and manuals. The pre-labeled set includes those headings and form paragraphs for which the "type" of action taken is unambiguous. The type of action includes rejection by statute section, non-statutory double patenting rejection, objection, and allowance, and the action type is further subcategorized into "subtypes" indicating the relevant statute paragraph or keyword. Action type and subtype categories appear in Table 1 and Table 2, respectively. We discuss these tables in more detail in Section IIIc.

We construct our pre-labeled set from the headings and form paragraphs stored within the Office Action Correspondence Subsystem (OACS) as well as those appearing on a sample of Office actions. Generally, examiners draft Office actions within OACS, which provides tools for inserting standard headings and form paragraphs consistent with the Manual of Patent Examining Procedure (MPEP). The headings and form paragraphs stored in OACS are largely uniform and labeled based on the type of action taken and statutory basis for rejection. However, because some examiners do not use standardized headings from OACS, we also extract headings from a sample of roughly 90,000 Office actions. We tabulate the most frequently occurring non-standardized headings from this sample and manually identify the type of action taken. This set of non-standardized headings supplements the OACS headings. The final pre-labeled set includes 76 unique headings and 37 unique form paragraphs.[20]

For the 4.4 million Office actions, we extract the headings and form paragraphs and match them to the pre-labeled set using textual similarity. For matching extracted headings, we use two text similarity measures: (1) Jaro-Winkler distance and (2) a variation of the Jaccard Index. Jaro-Winkler distance measures the edit distance between two text strings, giving higher weight to strings that match from the beginning. We calculate a modified version of the Jaccard Index as the intersection of two text strings divided by the minimum length of those two strings. This variation of the Jaccard Index is intended to yield scores that quantify similarity of extracted headings to the standardized headings (See Appendix A for additional details). If both measures exceed a predefined threshold of 0.85, we classify the extracted heading as the one that it is most similar to in the pre-labeled headings set. We establish this threshold through manual validation of sample data.

For the extracted form paragraphs, we only use the first 600 characters of the form paragraphs for matching to the set of pre-labeled form paragraphs.[21] We again use the Jaro-Winkler distance measure, but we also compute a cosine similarity score. For the latter, each form paragraph is characterized as a term frequency vector. We calculate the cosine similarity between two form paragraphs as the cosine distance between their frequency vectors (See Appendix B for additional details). We compute a

[20] Note that the form paragraphs set consists of the legal definitions of rejections under the statute and, therefore, is a subset of all the form paragraphs defined in the MPEP. For a full list of form paragraphs currently in the MPEP, *see* https://www.uspto.gov/web/offices/pac/mpep/FPs.html.

[21] We determine that matching on the first 600 characters of form paragraph text is appropriate through iterative experimentation and manual validation of results using sample data.

combined similarity score as the weighted sum of the Jaro-Winkler distance and the cosine distance. If this combined similarity score exceeds a predefined threshold of 0.80, we classify the extracted form paragraph as the one that it is most similar to in the pre-labeled form paragraph set. Again, we establish this threshold through manual validation of sample data.

Applying text matching, we classify headings and form paragraphs for roughly 84 percent of the 4.4 million Office actions in the *Dataset*. The remaining documents comprise Office actions for which the examiner included no or minimal headings and/or form paragraphs (See Figure 2 for an example). They also include Office actions for which matching headings and/or form paragraphs to our pre-labeled set yields similarity scores below predefined thresholds. For such documents, we rely on a second, more sophisticated method, which we discuss in detail below, for extracting information from the text indicating the action taken on the claims.

b. Extracting Data from Action Sentences

Our second method employs Natural Language Processing (NLP), domain knowledge, and heuristic logic to extract relevant data from the sentence(s) expressing the action taken on the claim(s) in each Office action. Typically, an examiner uses a single sentence to express the action taken. These single sentences generally follows a consistent structure, exemplified in Figure 4. The claims to be addressed are the subject (e.g., "claims 1-2"), followed by a verb phrase reflecting the action taken (e.g., "are rejected"), a prepositional phrase stating the legal grounds (e.g., "under pre-AIA 35 U.S.C. 102(b)")[22], and another prepositional phrase declaring prior art references (e.g., "as being anticipated by Schmitt et. al. (e.g., US 2010/0080426 A1)").[23]

To identify these different elements, we apply a sentence parser from the Stanford CoreNLP package[24] and generate a constituency-based parsing tree that represents the syntactic structure of the sentence according to English language grammar rules. Figure 5 depicts the parsing tree generated from our prior example with each constituent of the sentence labeled with a part-of-speech tag (e.g., NP: noun phrase, VP: verb phrase, PP: prepositional phrase, etc.).[25] We leverage this parsing tree to identify the different constituents and map them to four main data elements of interest, namely:

(1) Claim(s) in question – claims 1-2
(2) Action taken on claim(s) – rejected
(3) Legal ground for the action – 35 U.S.C. 102(b) or 102(b) in short
(4) Prior art(s) cited in the action – Schmitt et al. (US 2010/0080426 A1)

Thus, by analyzing the sentence structure and applying NLP algorithms with grammar rules to label terms, we extract these four main data elements, which we then encode in the *Dataset*.

Since Office actions are freely formatted text documents written by thousands of different examiners, there is considerable variation in the syntactical structure of action sentences. We adopt additional NLP tools to accommodate this variation. To illustrate this, Figure 6 shows a sample action sentence that does not conform to the typical standard shown in Figure 4. In this action sentence, the prior art cited ("Schmitt et al.") is only mentioned by name, rather than by full citation, because it has been previously mentioned

[22] Pre-AIA indicates the statute prior to being amended per the Leahy-Smith America Invents Act (AIA). *See* the definition for the *action_subtype* variable in Section IVb for more information on the impact of AIA.
[23] The interface within OACS provides examiners guidance on the structure and content of the standard sentence based on the inserted form paragraph.
[24] *See* https://stanfordnlp.github.io/CoreNLP/index.html.
[25] We apply part-of-speech tags used in the Penn Treebank Project, *see* https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

in the document. To handle this, we adopt a Named Entity Recognition algorithm and some heuristic logic (e.g., a noun before et al. is very likely to be a person's name) to identify all the "person" names appearing in the document.[26] When a person name and citation string is first referenced as prior art in the document, we store it in a canonical form. If the same person name is subsequently referenced, we retrieve the stored canonical form using a simple Entity Resolution process that applies name matching based on Jaro-Winkler distance. We then include only the patent or publication number of the full citation in the *Dataset*.

c. Encoding Action Types and Subtypes

Combining results from the headings/form paragraph analysis with the NPL sentence structure analysis, we identify the type and subtype of the action(s) taken in each of the 4.4 million Office actions. The action types from both methods agree in 95 percent of cases. Such a high rate of agreement between the two methods provides some validation of each approach and supports the overall quality of the data in the *Dataset*. For the remaining 5 percent of cases, there is at least one type label mismatch between methods, which we explicitly indicate in the *Dataset*.[27]

Table 1 shows the frequency of Office actions in the *Dataset* by action type and document code. A single Office action typically includes multiple action types. The mean number of action types per Office Action in the *Dataset* is 2.1.[28] Thus, the action type categories in Table 1 are not mutually exclusive and the percentage figures represent the proportion of documents (Non-Final Rejections, Final Rejections, and combined) that include at least one of the designated action types. Office actions with a rejection for obviousness under 35 U.S.C. 103 are by far the most prevalent in the *Dataset*. Roughly 77 percent of Non-Final Rejections and 84 percent of Final Rejections include a 103 rejection. Nearly half the Non-Final Rejections in the *Dataset* contain a rejection for lack of novelty under 35 U.S.C. 102, but only 32 percent of Final Rejections include such a rejection. Similarly, 38 percent of Non-Final Rejections contain a rejection related to the written description or clarity of claims under 35 U.S.C. 112, compared to only 29 percent of Final Rejections. Relatively few Office actions in the *Dataset* contain a 35 U.S.C. 101 rejection. About 13 percent of Non-Final Rejections and 7 percent of Final Rejections include a 101 rejection.

We are able to derive consistent action subtype labels for rejections raised under 35 U.S.C. 102, 35 U.S.C. 103, and 35 U.S.C. 112 based on the relevant section paragraphs.[29] However, because there are no section paragraphs *per se* under 35 U.S.C. 101, we develop action subtype labels based on keywords observed in form paragraphs and action sentences. Because the case law related to 35 U.S.C. 101 continues to evolve, the USPTO periodically issues new guidance and revised training material to assist examiners.[30] Over time, these resources can provide varying recommendations regarding the use of specific form paragraphs or how to apply explanations corresponding to court decisions. This complicates the process for identifying consistent action subtype labels for 101 rejections.

---

[26] We specify heuristic logic based on standard English language and writings. We then test the validity of rules by iteratively experimenting and manually validating results for sample data.

[27] The *rejection_fp_mismatch* field in the *Dataset* indicates when there is a mismatch between the action type identified by heading/form paragraph analysis and that identified via NPL sentence structure analysis. Specifically, the variable indicates when a form paragraph included in the Office action does not match a rejection raised in the action sentence. *See* the variable definition in Section IVa.

[28] This number reflects all action types, including allowed claims and objections. The mean number of rejections (101, 102, 103, 112, double patenting) per document is 1.8.

[29] To ensure subtype labels are consistent over time, we map section paragraphs in post-America Invents Act (AIA) to their pre-AIA equivalent for rejections under 112. See the definition for the *action_subtype* variable for more details.

[30] For additional information, including a discussion of historical developments, related to Patent Subject Matter Eligibility, see https://www.uspto.gov/sites/default/files/documents/101-Report_FINAL.pdf.

To account for this, we first create a basic 35 U.S.C 101 ontology using the categorizations found in the MPEP: double patenting, subject matter eligibility, utility, and improper naming of inventor.[31] These categories, particularly subject matter eligibility, are generally too broad for most research purposes. To develop more precise action subtype categories, we review training materials[32], memoranda to the examination corps[33], and recommended form paragraphs[34] spanning the 2008 to early 2016 period. We derive an initial set of key search terms and phrases from these materials. We query the text of each 101 rejection for these keywords and phrases to identify an initial categorization. We then further refine categories into 12 action subtype labels based on significant Supreme Court decisions and/or topics (see Appendix C for a mapping of search terms and phrases to action subtype labels). If multiple subtypes apply to a 101 rejection, we prioritize Supreme Court decisions (in reverse temporal order) followed by keywords. However, to account for instances where multiple Supreme Court decisions apply, we generate a set of fields to indicate if each of the four decisions were mentioned at all in the action (see Section IVb for more details on these fields).

Table 2 presents the frequency of claim-level rejections (i.e., document-rejection pairs) in the *Dataset* by action type and subtype, including 101 subtype categories. Rejections for obviousness under 35 U.S.C. 103(a) are the most common at the rejection level, comprising about 30 percent of all rejections in the *Dataset*. Roughly 12 percent of rejections are for use or publication more than one year before filing under pre-AIA 35 U.S.C. 102(b). Another 12 percent of rejections are for claims that fail to distinctly claim the invention under 35 U.S.C. 112(b). Otherwise, rejections are largely dispersed across action subtypes. Rejections under 35 U.S.C. 101 are predominantly for subject matter not eligible for patenting under the statute ("non-statutory"), particularly data (27 percent of 101 rejections), laws of nature (10 percent), and processes performed mentally, verbally or without a machine (5 percent). Roughly, 12 percent of 101 rejections indicate a judicial exception under the decision in *Alice Corp. v. CLS Bank International*.[35]

Deriving consistent action subtype labels for 101 rejections spanning the *Dataset* was a challenge given the evolving case law and shifting use of form paragraphs to particular topics. Our effort should be viewed only as an initial attempt intended to increase the utility of the *Dataset* for researchers. Action subtype labels should not be considered an authoritative classification by the USPTO. We expect this initial effort will stimulate further discussion and efforts to classify rejections under 35 U.S.C. 101.

d. Prior Art Citations

To more accurately identify examiner- versus applicant-cited prior art[36], we supplement the prior art citations extracted from the Office action text with data from two additional sources. First, we obtain U.S. patent numbers and U.S. pre-grant publication numbers cited by the patent examiner during prosecution from the Form PTO-892 "Notices of References Cited".[37] Second, we extract U.S. patent numbers and U.S. pre-grant publication numbers cited by the applicant from the Information Disclosure Statement (IDS) Form PTO-1449.[38] These two forms are stored in MS-Word format on internal USPTO servers. For

[31] *See* MPEP 706.03(a).
[32] https://www.uspto.gov/patent/laws-and-regulations/examination-policy/subject-matter-eligibility
[33] https://www.uspto.gov/patent/laws-and-regulations/examination-policy/memoranda-examining-corps
[34] https://www.uspto.gov/web/offices/pac/mpep/form-paragraph-book.pdf
[35] *See* https://www.supremecourt.gov/opinions/13pdf/13-298_7lh8.pdf.
[36] *See* note 5.
[37] MPEP 707.05
[38] The applicant is required to submit an Information Disclosure Statement (IDS) listing all patents, publications, applications, or other information known to the applicant to be material to patentability of the claims in the application. References to U.S. patents and U.S. patent published applications are to be listed separately from the citations of other documents and were historically captured on Form PTO-1449 (currently the Form PTO/SB/08A and 08B). *See* 37 C.F.R. 1.98(a)(1).

each application included in the *Dataset*, we convert the Form PTO-892 and Form PTO-1449 into XML format. We then parse the XML to generate a relational database table. Note that both Form PTO-892 and Form PTO-1449 may contain citations to foreign patent documents and non-patent literature, which are not included in the *Dataset*.[39] The *Dataset* does include a small number of foreign patent document numbers and non-patent literature cited as prior art in Office actions. These include some foreign patent document numbers recorded in standard formats (e.g. "WO 2012/095284 A1"), which enable us to properly identify and parse numbers. Note, however, that our second method typically does not correctly identify and parse citations that consist of non-patent literature. Thus, for a small number of 102 and 103 rejections in which the examiner referenced only non-patent literature in the Office action, we do not record the prior art in the *Dataset*.

The prior art cited in the Office action typically correspond to the citations listed on the Form PTO-892 and/or Form PTO-1449.[40] However, for various reasons, examiners may cite in the action prior art that is not recorded on either of those forms. Thus, in the *Dataset*, we indicate whether the citation was referenced in a specific Office action as well as whether it was listed on the Form PTO-892 and/or Form PTO-1449.

Note that applicants can submit multiple Forms PTO-1449 during examination. An applicant may submit an additional Form PTO-1449 in the later stages of prosecution, for example, to disclose prior art revealed to the applicant during the examination of a related foreign or domestic application. Consequently, prior art citation data from Forms PTO-1449 in the *Dataset* may not be fully observed for patent applications still pending with the Office as of July 2017.

IV. Dataset Files and Variables

This release of the *Dataset* consists of three data files derived from 4.4 million Office actions issued for 2.2 million applications with filing dates predominately in the 2008 to 2017 period. Figure 7 displays the organizational structure of the *Dataset*. We describe each file and its variables in more detail in the following subsections.

The first data file is called **office_actions** and includes basic information regarding the Office action and a set of indicators for the type of action(s) taken. There are 4.4 million observations in this data file, with each observation representing a unique Office action (as identified by the *ifw_number* field). See Table 3 for a list and brief description of all variables in the **office_actions** file.

Each Office action document may include multiple actions taken against claims. The second file is called **rejections** and indicates the type (*action_type*) and subtype (*action_subtype*) of each action taken on claims in the Office action. There are 10.1 million observations in **rejections**, with each observation representing a unique document-action pair (as identified by the document *ifw_number* and *action_type/action_subtype* combination). Table 4 includes a list and brief description of the variables in the **rejections** file.

The third file is called **citations** and includes the U.S. patent grants and published U.S. patent applications cited in the prosecution of each application. The **citations** data file is derived from the citations referenced on the Form PTO-892, on the Form PTO-1449, and in the text of the Office actions. There are 58.9 million unique application-citation pairs in the **citations** data file, of which roughly 21 percent represent prior art cited in an Office action issued to the applicant. For citations not referenced in an

[39] Because there is considerable variation in the formats of citation to foreign patent documents and non-patent literature, we were unable parse them in a systematic way for this initial release.
[40] MPEP 1302.12

action, the unique observation in the **citations** file is the application-citation pair (as identified by *app_id* and *citation_pat_pgpub_id* fields). For those that are referenced in an action, the **citations** file contains additional fields to enable users to link the application-citation pair to the document-action pair (via the document *ifw_number* and *action_type*/*action_subtype* combination) in the **rejections** file. See Table 5 for a list and brief description of all variables included in the **citations** file.

a. Variables in office_actions

Table 3 provides a brief description of the following variables in the **office_actions** file.

Application Number
Each application received by the USPTO is given a unique application number (*app_id*). The number is eight digits long and used to keep track of the application while it is being processed and examined. The application number is comprised of two parts. For all applications that were not filed under the Patent Cooperation Treaty (PCT), the first two digits indicate the application's series number. For the most part, the series number gives a rough indication of the order in which applications were received by the USPTO. This release includes Office actions issued for applications in series 12, 13, 14, and 15.

Image File Wrapper (IFW) Number
Each Office action issued by the USPTO is given an Image File Wrapper (IFW) identifier. The *ifw_number* identifier is a unique, alpha-numeric code that serves as the primary key for linking observations across the three data files.

Document Code
Each Office action has a document code (*document_cd*) identifying the type of document issued by the examiner to the applicant. This release only includes Non-Final Rejections and Final Rejections indicated by the following document code values, respectively:

- CTNF – Non-Final Rejection – An Office action issued by the examiner to the applicant rejecting one or more claims that does not close-out prosecution. A Non-Final Rejection can also include objections to claims and/or other requirements.

- CTFR – Final Rejection – A second or any subsequent Office action issued by the examiner to the applicant rejecting one or more claims that is made final indicating that the examiner intends to close prosecution. A Final Rejection may include grounds for objections, rejections, and/or other requirements. Upon receiving a Final Rejection, the applicant no longer has the right to amend the application unless the amendment merely cancels claims or complies with a formal requirement made earlier.

Mail Date
The *mail_dt* variable is the date the Office action was mailed-out from USPTO.

Examiner Art Unit
The *art_unit* variable indicates the group art unit to which the examiner issuing the Office action belongs. Group art units are designated as four digit numbers. The first two digits indicate the technology center

(TC) to which the group art unit is assigned. The designations for the TCs have changed over the years, but currently there are eight TCs for examining regular utility applications.[41]

- 1600 – Biotechnology
- 1700 – Chemical and Materials Engineering
- 2100 – Computer Architecture, Software, and Information Security
- 2400 – Computer Networks, Multiplex Communication, Video Distribution and Security
- 2600 – Communications
- 2800 – Semiconductors, Electrical and Optical Systems and Components
- 3600 – Transportation, Construction, Electronic Commerce, Agriculture, National Security and License & Review
- 3700 – Mechanical Engineering, Manufacturing, Products

Classification Codes
When the USPTO processes new patent applications, they assign the application into one primary US Patent Classification (USPC) technology class and subclass. Classification of new applications is used in (1) the assignment of the applications to the most relevant examiner group art units and (2) the searches for relevant prior art during patent examination. The USPTO began transitioning to the Cooperative Patent Classification system in 2013 to enable prior art searching, but continues to use the USPC system for routing applications to examiners. Since the data files capture examiner-issued Office actions, the USPC class is more relevant.[42] Each USPC class and subclass is identified by a code, provided in *uspc_class* and *uspc_subclass* fields, respectively.

Heading Structure Missing
The USPTO recommends examiners utilize an Office action template with standardized headings to make the product consistent and easy to read. The *header_missing* field is an indicator that identifies when the Office action does not include standard headings or contains no headings.

Form Paragraph Missing
The USPTO recommends that examiners insert standard form paragraphs in Office actions for each type of rejection raised. Standard form paragraphs contain relevant legal definitions and provide for a streamlined and consistent format. The *fp_missing* field is an indicator that identifies when the Office action does not contain the form paragraph(s) for the rejection(s) raised.

Rejection and Form Paragraph Mismatch
The *rejection_fp_mismatch* field is an indicator that identifies when the form paragraph(s) included in the Office action do not match the rejection(s) raised in the action sentence(s).

[41] For more details regarding the current group art units and the technology centers to which they belong, please refer to http://www.uspto.gov/patent/contact-patents/patent-technology-centers-management. See Marco et al. (2014) for a description of how the older group art units map into the current TCs.
[42] *Dataset* users can retrieve the current and at issue CPC, as well as the at issue International Patent Classification, for applications that result in a patent grant via the USPTO PatentsView web-tool www.Patentsview.org. *See* the PatentsView data download page at http://www.patentsview.org/download/; data query builder at http://www.patentsview.org/query; or application programming interface at http://www.patentsview.org/api/doc.html.

Closing Paragraph Missing
At the end of the Office action, the examiner is to provide specific contact information, such as a phone number and alternative contact. The *closing_missing* field is an indicator that identifies if such information is missing from the Office action.

101 Rejection – Subject Matter Eligibility, Statutory Double Patenting, Utility, etc.
The *rejection_101* field indicates whether a 35 U.S.C. 101 rejection is raised in the Office action. Generally, an examiner rejects a claim under 35 U.S.C. 101 if any one of the following four requirements is not met: (1) whoever invents or discovers an eligible invention obtains only one patent therefor (i.e., there is no statutory double patenting); (2) the inventor is the applicant for applications filed prior to September 16, 2012 and each inventor is identified in an application filed on or after that date; (3) the claimed invention falls within one of four patent eligible categories of invention, i.e., process, machine, manufacture, or composition of matter, as these categories have been interpreted by the courts[43]; and (4) the claimed invention is useful or has utility that is specific, substantial, and credible.[44]

If a 101 rejection appears, it will be further categorized into different subtypes based on the form paragraphs and specific triggering keyword(s) in the rejection text. These subtypes appear in the **rejections** data file.

102 Rejection – Lack of Novelty
The *rejection_102* field indicates whether a 35 U.S.C. 102 rejection is raised in the Office action. In general, an examiner rejects a claim as not novel under 35 U.S.C. 102 if she finds it is anticipated, i.e., expressly or inherently described, by a single prior art reference.[45] If such a rejection appears, it will be further specified into different subtypes based on the identified statute paragraph, such as 102(a), 102(b), etc. These subtypes appear in the **rejections** data file.

103 Rejection – Obviousness
The *rejection_103* field indicates whether a 35 U.S.C. 103 rejection is raised in the Office action. Generally, an examiner rejects a claim as obvious under 35 U.S.C. 103 if she determines the claimed invention would have been obvious to a person having ordinary skill in the field to which the invention pertains. She must provide objective evidence to support her rejection.[46] If such a rejection appears, it will be further specified in the **rejections** data file into different subtypes based on the identified statute paragraph, principally 103(a).

112 Rejection – Written Description, Indefinite Claims, etc.
The *rejection_112* field indicates whether a 35 U.S.C. 112 rejection is raised in the Office action. In general, an examiner rejects a claim under U.S.C. 112 if she determines it does not meet requirements regarding the adequacy of the disclosure of the invention.[47] If such a rejection appears, it will be further specified into different subtypes based on the identified statute paragraph, such as 112(a), 112(b), etc. These subtypes appear in the **rejections** data file.

---

[43] Additionally, to determine that a claimed invention is a judicial exception, the examiner is to perform a two-part subject matter eligibility test. For detailed guidance on how patent examiners should evaluate claims for patent subject matter eligibility under 35 USC 101, *see* https://www.uspto.gov/patent/laws-and-regulations/examination-policy/subject-matter-eligibility.
[44] *See* MPEP 2104 through MPEP 2106 for a detailed overview of patent subject matter under 35 USC 101.
[45] *See* MPEP 2131 through MPEP 2138 for a detailed overview of compliance with 35 USC 102 and MPEP 2150 through MPEP 2153 for a detailed discussion of changes to 35 USC 102 as amended by the AIA.
[46] *See* MPEP 2141 through MPEP 2146 for a detailed overview of compliance with 35 USC 103 and MPEP 2150 through MPEP 2152 and MPEP 2158 for a detailed discussion of changes to 35 USC 103 as amended by the AIA.
[47] *See* MPEP 2161 through MPEP 2186 for a detailed overview of compliance with 35 USC 112.

Double Patenting Rejection
The *rejection_dp* field indicates whether a non-statutory double patenting rejection is raised in the Office action. Generally, non-statutory double patenting occurs when similar, but not identical, scope is claimed by a common inventor and/or assignee.[48]

Objection
The *objection* field indicates whether an objection is raised in the Office action. Objections are generally raised for minor informalities or violation with patent rules, such as when claims are not properly grouped together or figure elements are not properly referenced in the specification.[49]

Allowable Claims
The *allowed_claims* field indicates when the Office action includes text specifying that one or more claims are allowable if the objection can be overcome. The field can also indicate when certain claims are allowed without any objection.

Table 1 shows the frequency of Office actions with each action type – rejection by statute, objection, and allowance – by document type in the **office_actions** file.

Greater than One Citation in 102 Rejection
The *cite102_gt1* field indicates when more than one reference is cited as the basis to reject certain claim(s) under 35 U.S.C. 102 in the Office action. Note that examiners may include definitional or evidentiary references to support a rejection based on a single citation. Examiners may also use additional references to provide clarity regarding the publication date of a citation or document evidence of prior use or sale.

Greater than Three Citations in 103 Rejection
The *cite103_gt3* field indicates when more than three references are cited as the basis to reject certain claim(s) under 35 U.S.C. 103 in the Office action.

One Citation in 103 Rejection
The *cite103_eq1* field indicates when only one reference is cited as the basis to reject certain claim(s) under 35 U.S.C. 103 in the Office action. Note that examiners may rely on established knowledge or "Examiner's official notice" to support a 103 rejection based on a single citation.

Max Citations in 103 Rejection
The *cite103_max* field indicates the largest number of references cited as the basis to reject certain claim(s) under 35 U.S.C. 103 in the Office action. For example, if claim 1 is rejected under 103 based on two citations, claim 3 is rejected under 103 based on four citations, and no other 103 rejections are raised, then this field will contain the value 4 as it is the highest number of references cited for any of the claims. Note that, in certain cases, the algorithm for generating the *cite103_max* field is less precise. Specifically, when an examiner cites non-patent literature with author names recorded individually or lists chemical compound codes, the algorithm tends to overcount the number of references, resulting in an overestimated value in the *cite103_max* field. This is most evident in technology areas that primarily rely on non-patent literature. We plan to address this issue in a subsequent release through a more robust detection algorithm.

[48] Statutory double patenting falls under 35 U.S.C. 101, *see* the definition for the *rejection_101* variable. *See* MPEP 804 for a comprehensive discussion of double patenting.
[49] 37 C.F.R. 1.75(g).

Signature Type
The *signature_type* field indicates the signature types of the Office action. We generate this field based on the examiner title(s) extracted from the signature block of the document. Generally, the title is inserted in OACS by default when an examiner signs the Office action. If more than one examiner worked on the action, each examiner's name and title are listed separately. The *signature_type* field contains one of the following values:

- 0 – Examiner
- 1 – Primary Examiner (PE)
- 2 – Examiner + PE
- 3 – Examiner + Supervisory Patent Examiner (SPE)
- 4 – Examiner + PE/SPE + Director

b. Variables in rejections

Table 4 provides a brief description of the following variables in the **rejections** file.

Action Type
For each grounds for rejection raised in the Office action, the *action_type* field indicates the relevant section of 35 U.S.C. This field also contains relevant categories where specific sections of the statute do not apply, including non-statutory double patenting, objections, and allowed claims.

Action Subtype
For each grounds for rejection raised in the Office action, the *action_subtype* field indicates the paragraph letter within the relevant section of 35 U.S.C. This field will also contain relevant categories where specific sections or paragraphs of the statute do not apply, principally 101 rejection subtypes. Table 2 shows the frequency of Office action document-action pairs by action type (*action_type*) and subtype (*action_subtype*) and provides a brief description of each action subtype.

Note that, because the Leahy-Smith America Invents Act (AIA) amended certain sections of the statute, action subtypes for 102 rejections may have different meaning before and after AIA implementation. AIA amended 35 U.S.C. 102 to establish the first-inventor-to-file system. As a result, two pre-AIA section paragraphs about novelty, 102(a) and 102(b), correspond with AIA subparagraph "102(a)(1)". The other pre-AIA section paragraph regarding novelty, 102(e), maps to subparagraph "102(a)(2)". There are no corresponding provisions in the AIA for the remaining paragraphs of the pre-AIA 102 statute: 102(c), 102(d), 102(f), and 102(g). The descriptions for these 102 paragraph subtypes identify them as "pre-AIA" in Table 2.[50]

AIA also amended 35 U.S.C. 112, adding labels to subparagraphs not previously labeled. To provide consistent labels in the *Dataset*, we map pre-AIA section paragraphs to their post-AIA labeled equivalents for action subtypes under 112.[51]

Claims Rejected
The *claim_numbers* field lists the application claims in question for each action (*action_type*/*action_subtype* combination) raised.

[50] For an overview of the impact of AIA on 35 U.S.C. 102, *see* https://www.uspto.gov/sites/default/files/aia_implementation/fitf_comprehensive_training_prior_art_under_aia.pdf.
[51] Specifically, we directly map the first through sixth pre-AIA paragraph to its post-AIA alphabetical equivalent a) through f).

Alice Decision
The *alice_in* field indicates if the form paragraph and/or text in the action references the Supreme Court decision in *Alice Corp. v. CLS Bank International*.[52]

Bilski Decision
The *bilski_in* field indicates if the form paragraph and/or text in the action references the Supreme Court decision in *Bilski v. Kappos*.[53]

Mayo Decision
The *mayo_in* field indicates if the form paragraph and/or text in the action references the Supreme Court decision in *Mayo v. Prometheus*.[54]

Myriad Decision
The *myriad_in* field indicates if the form paragraph and/or text in the action references the Supreme Court decision in *Association for Molecular Pathology v. Myriad Genetics, Inc.*[55]

c. Variables in citations

Table 5 provides for a brief description of the following variables in the **citations** file.

U.S. Patent or Pre-grant Publication Citation
The *citation_pat_pgpub_id* field contains references cited in prosecution of the application. The field predominantly includes those citations that consist of a prior U.S. patent grant or U.S. pre-grant publication. For a small subset of observations in the **citations** data file, this field includes the raw text extracted from an Office action. This raw text may include citations to foreign patent documents as well as non-patent literature.

Parsed
The *parsed* field contains the U.S. patent grant number or U.S. pre-grant publication number parsed from the *citation_pat_pgpub_id* field. This field may also contain the foreign patent document number if the *citation_pat_pgpub_id* field contains such a reference.

Form 892
The *form892* field indicates whether the prior U.S. patent grant or U.S. pre-grant publication citation was retrieved from the Form PTO-892.

Form 1449
The *form1449* field indicates whether the prior U.S. patent grant or U.S. pre-grant publication citation was retrieved from the Form PTO-1449. Note that, since applicants can submit multiple Forms PTO-1449 during examination, prior art citation data from this source may be incomplete for patent applications still pending with the Office as of June 2017.

Citation Referenced in Office Action
The *citation_in_oa* field indicates whether the citation was referenced in an Office action. When this field is positive (value of 1), the *ifw_number*, *action_type*, and *action_subtype* fields will identify the relevant Office action and rejection grounds for which the citation is referenced.

---

[52] *See* https://www.supremecourt.gov/opinions/13pdf/13-298_7lh8.pdf.
[53] *See* https://www.supremecourt.gov/opinions/09pdf/08-964.pdf.
[54] *See* https://www.supremecourt.gov/opinions/11pdf/10-1150.pdf.
[55] *See* https://www.supremecourt.gov/opinions/12pdf/12-398_1b7d.pdf.

V. Limitations

Researchers and other *Dataset* users should be conscious of the limitations of these data. As with most data derived from textual documents, typographical and other input errors are evident, though infrequent. For example, there are less than 200 Office actions in the *Dataset* with mailing dates prior to the filing date of application for which the action was issued. The accuracy of mailing dates is important because the impact of an Office action depends on the stage of prosecution in which it is issued by an examiner. The *Dataset* alone can only indicate whether the action is a Non-Final or Final Rejection and the sequence in which these documents were issued to the applicant based on mailing dates. The *Dataset* does not provide information regarding when the application was filed, when the application was placed on an examiner's docket, what amendments or other actions may have occurred prior to, after, or between Office actions, or the current status of the application (e.g., granted, abandoned, or still pending).

Users will need to merge the *Dataset* with other data sources to retrieve such key information. The Office of the Chief Economist's Patent Examination Research Dataset (or "PatEx") contains detailed information on publicly available patent applications filed with the USPTO. PatEx is sourced from Public PAIR and contains "research-ready" flat files of data on application characteristics, including filing and publication dates, continuation history, and prosecution events (Graham et al. 2015).[56] Some of the Public PAIR data elements are also available for more targeted search and download via the USPTO Patent Examination Data user interface and application programming interface (API).[57] Additionally, researchers interested in mapping rejected claims to their claim text can leverage the Patent Claims Research Dataset (Marco et al. 2016).[58] Combining the *Dataset* with these and other sources provides opportunities for augmenting emerging research on the patent examination process as well as advancing original inquiry in other areas.

Merging data, however, exposes an additional limitation of the *Dataset* related to systematic gaps in data coverage. To explore some of these limitations, we match the *Dataset* to internal USPTO administrative data on patent applications to indicate the extent of these coverage issues.[59] The matched sample is limited to publicly available applications for utility and plant patents as well as reissues in the 12, 13, 14, and 15 series. We also omit applications for which an examiner has yet to issue a first action on the merits of the application.[60]

For this matched sample, Figures 8 shows the number of applications with Office action data coverage by filing year. Generally, the *Dataset* contains some Office action data for about 250,000 to 300,000 published applications per filing year. This number declines for more recent filings because new applications are still awaiting prosecution and examiners have not yet issued an Office action. Still, even for older cohorts where fewer applications are likely to remain pending, the *Dataset* does not provide complete coverage for all published applications. Applications that are allowed by the examiner without any rejections or objections could help explain this coverage gap. Such "first action allowances" would not be captured in the *Dataset* since no Non-Final or Final Rejection would have been issued.

---

[56] PatEx is available for download at https://www.uspto.gov/learning-and-resources/electronic-data-products/patent-examination-research-dataset-public-pair.
[57] The Patent Examination Data interface and API currently provide for searching and downloading data from the bibliographic (tab) and transaction history (tab) at https://ped.uspto.gov/peds/.
[58] The Patent Claims Research Dataset is available for download at https://www.uspto.gov/learning-and-resources/electronic-data-products/patent-claims-research-dataset.
[59] For this exercise, we use USPTO administrative data that is largely the internal version of the PatEx dataset and, thus, includes patent applications that have yet to publish or be made publicly available. Note, however, that we exclude such non-publicly available applications from Figures 8, 9, and 10.
[60] *See* https://www.uspto.gov/web/offices/pac/mpep/s2660.html.

Figure 9 depicts published applications with Office action data by filing year, as a percentage of total published applications and as a percentage of published applications that were not allowed on first action. For most of the older filing year cohorts, the *Dataset* contains some Office action information for more than 80 percent of published applications, but more than 90 percent of published applications when excluding those allowed on first action. The 2010 filing year cohort is an evident exception. There is Office action data coverage for 64 percent of published applications filed in 2010, 72 percent of published applications that were not allowed on the first action.

To better isolate coverage gaps, in Figure 10, we plot application-Office action document pairs by application filing date and Office action mailing date. White areas reflect dates for which there is no Office action coverage in the *Dataset*. For the 2010 filing year cohort, Figure 10 shows the *Dataset* coverage is most limited for applications filed in the last quarter of that year. These Office action documents are missing due to data quality issues that interfered with processing these documents. We are currently pursuing quality and pre-processing fixes to include these documents in the next release of the *Dataset*.

Figure 10 also indicates that there can be a considerable lag between the application filing date and the mailing date of the Office action. *Dataset* users should be mindful of the potential implications of such lags for their research, particularly when applicants continue to seek patent protection after an initial Final Rejection.[61]

For all Office actions in the *Dataset*, the mailing date lags the application filing date by an average of 2.2 years. Clearly, this lag would tend to be shorter for the first Non-Final Rejection issued to an applicant relative to subsequent rejections. To better illustrate this, Figure 11 shows a box plot of the distribution of application-Office action document pairs by time from application filing to document mailing. It includes a separate box plot for the first, second, third, fourth, and fifth or subsequent Office action observed in the *Dataset* based on earliest mailing date. The number of Office actions per unique patent application is fairly skewed. Just under half (48 percent) of the unique applications in the *Dataset* have only one Office action. Another 28 percent have only two, 11 percent only three, and 7 percent four. The remaining 6 percent of unique applications have 5 to 22 Office actions in the *Dataset*.

As expected, Figure 11 shows the time between filing and action mailing increases with each additional action. The median lag from filing to first action mailing is 1.7 years, 2.2 years for the second, 2.8 years for third, and 3.3 years for the fourth. This suggests fairly consistent intervals between the mailing of first and second action, second and third action, etc. The median values and distributions are largely the same if we consider only applications that have been disposed of (via patent grant or application abandonment) as well as if we control for the total number of Office actions issued for a particular application. Users should be cautious of this as well as the coverage issues discussed previously.

VI. Conclusions

Policy makers and scholars are interested in understanding and improving the patent examination process. This interest is warranted since patents are an incentive mechanism that fosters innovation and helps to sustain economic growth and competitiveness. This paper describes the methods used to construct the *UPSTO Office Action Research Dataset for Patents* as well as its structure and content. This *Dataset* offers policy analysts and researchers new opportunities to explore and understand patent prosecution. It provides important information taken from Non-Final and Final rejections by patent examiners covering the 2008-2017 period and supplements this information with data on prior art citations by applicants and

[61] *See* note 16.

examiners. Office action data are particularly relevant to the growing body of empirical work on the patent examination process, examiner heterogeneity, patent quality, and application and litigation outcomes. Likewise, readily available data on actions, particularly the prior art used as the basis for a rejection, will greatly augment the longer established literature on innovation, knowledge diffusion, and technology change. Our intention is to make regular updates and enhancements to these data to ensure researchers and policymakers are gaining valuable insights from the wealth of information captured in Office actions.

VII. References

Alcacer, J., Gittelman, M., 2006. Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations. The Review of Economics and Statistics, 88 (4), 774-779.

Alcacer, J., Gittelman, M., Sampat, B., 2009. Applicant and Examiner Citations in US Patents: An Overview and Analysis. Research Policy, 38 (2), 415-427.

Carley, M., Hedge, D., Macro, A., 2015. What is the Probability of Receiving a US Patent? The Yale Journal of Law & Technology, 17, 203-223.

Cockburn, I., Korum S., Stern S., 2003. Are All Patent Examiners Equal? Examiners, Patent Characteristics, and Litigation Outcomes. In Cohen, W.M., Merrill, S.A. (Eds.), Patents in Knowledge-Based Economy. National Academies Press, Washington, DC.

Cotropia, C., Lemley, M.A., Sampat, B., 2013. Do Applicant Patent Citations Matter? Research Policy, 42 (2013), 844–54.

Frakes, M.D., Wasserman, M.F., 2015. Does the U.S. Patent and Trademark Office Grant Too Many Bad Patents?: Evidence from a Quasi-Experiment. Stanford Law Review, 67, 613.

Frakes, M.D., Wasserman, M.F., 2017. Is the Time Allocated to Review Patent Applications Inducing Examiners to Grant Invalid Patents?: Evidence from Micro-Level Application Data. The Review of Economics and Statistics, 99 (3), 550-563.

Graham, S.J.H., Marco, A.C., Miller, R., 2015. The USPTO Patent Examination Research Dataset: A Window on the Process of Patent Examination (November 2015). USPTO Economic Working Paper 2015-4; Georgia Tech Scheller College of Business Research Paper No. 2016-055. Available at SSRN: https://ssrn.com/abstract=2848549.

Hall, B.H., Jaffe, A., Trajtenberg, M., 2005. Market Value and Patent Citations. Rand Journal of Economics, 36 (1), 613-76.

Harhoff, D., Scherer, F., Vopel, K., 2002. Citations, Family Size, Opposition and Value of Patent Rights. Research Policy, 32 (8), 1343-63.

Jaffe, A., Trajtenberg. M., 2002. Patents, Citations, Innovations: A Window on the Knowledge Economy. Cambridge, MA: MIT Press.

Jaffe, A.B., Trajtenberg, M., Fogarty, M.S., 2000. Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors. American Economic Review, 90 (2), 215-218.

Jaffe, A.B., Trajtenberg, M., Henderson, R., 1993. Geographic Localization of Knowledge Spillovers as Evidence by Patent Citations. The Quarterly Journal of Economics, 108 (3), 577–598.

Kesan, J.P., 2002. Carrots and Sticks to Create a Better Patent System. Berkeley Technology Law Journal, 17 (2), 763-797.

Kuhn, J. M.,Younge, K. A., Marco, A. C., 2017. Patent Citations and Empirical Analysis (August 7, 2017). Available at SSRN: https://ssrn.com/abstract=2714954.

Lanjouw, J.O., Schankerman, M., 2004. Patent Quality and Research Productivity: Measuring Innovation with Multiple Indicators. The Economic Journal, 114 (495), 441-465.

Lemley, M.A., Sampat, B., 2012. Examiner Characteristics and Patent Office Outcomes. The Review of Economics and Statistics, 94 (3), 817–827.

Lemley, M.A., Tangri, R.K., 2003. Ending Patent Law's Willfulness Game. Berkeley Technology Law Journal, 18 (4), 1085-1125.

Lichtman, D., 2004. Rethinking Prosecution History Estoppel. University of Chicago Law Review, 71, 151–182.

Marco, A.C., Miller, R., Kesan, J.P., 2014. Perspectives on the Growth in Chinese Patent Applications to the USPTO (February 1, 2014). USPTO Economic Working Paper 2014-1; University of Illinois College of Law Legal Studies Research Paper No. 17-27. Available at SSRN: https://ssrn.com/abstract=2849622.

Marco, A.C., Sarnoff, J.D., deGrazia, C., 2016. Patent Claims and Patent Scope (October 2016). USPTO Economic Working Paper 2016-04. Available at Available at SSRN: https://ssrn.com/abstract=2844964.

Marco, A.C., Toole, A.A., Miller, R., Frumkin, J., 2017. USPTO Patent Prosecution and Examiner Performance Appraisal (June 1, 2017). USPTO Economic Working Paper No. 2017-08. Available at SSRN: https://ssrn.com/abstract=2995674.

Mann, R., 2014. The Idiosyncrasy of Patent Examiners: Effects of Experience and Attrition. Texas Law Review, 92, 2149–2176.

Nelson, A.J., 2009. Measuring Knowledge Spillovers: What Patents, Licenses and Publications Reveal about Innovation Diffusion. Research Policy, 38 (6), 994-1005.

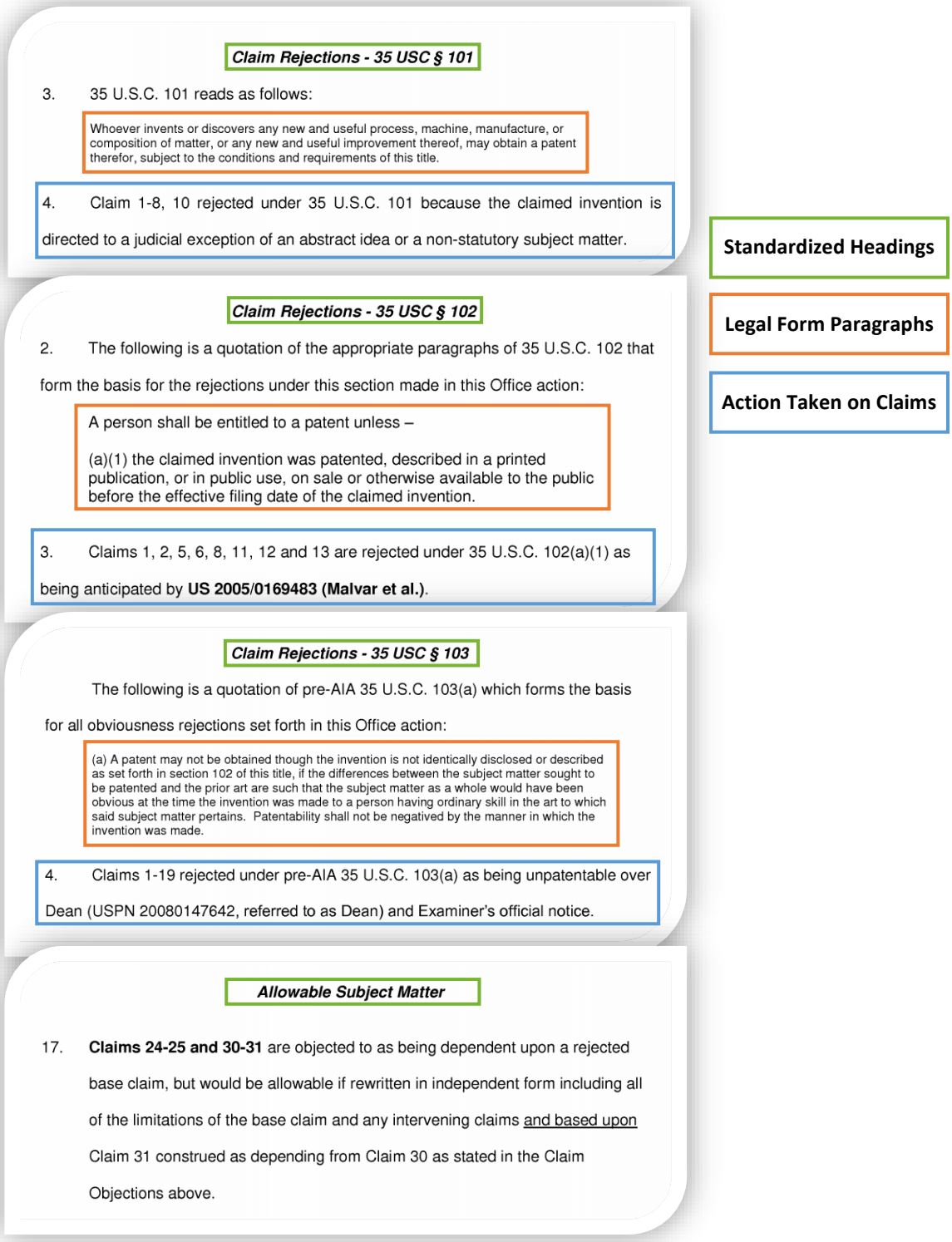
Sampat, B., 2010. When Do Applicants Search for Prior Art? Journal of Law and Economics, 53 (2), 399–416.

Sampat, B.N., Ziedonis, A., 2004): Patent Citations and the Economic Value of Patents. In: Moed, H.F., Glänzel, W., Schmoch, U. (Eds.), Handbook of Quantitative Science and Technology Research.

van Zeebroeck, N., 2011. The Puzzle of Patent Value Indicators. Economics of Innovation and New Technology, 20 (1), 33-62.

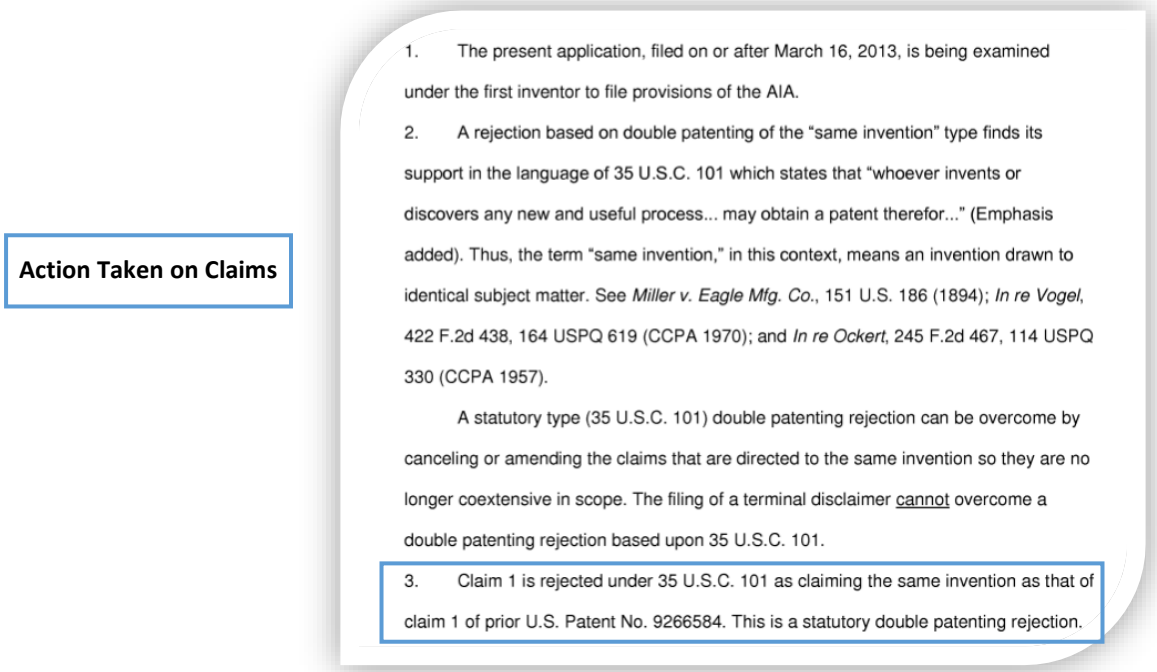
Younge, K. A., Kuhn, J. M., 2016. Patent-to-Patent Similarity: A Vector Space Model (July 30, 2016). Available at SSRN: https://ssrn.com/abstract=2709238.

Figure 1. Office actions with standard structural elements



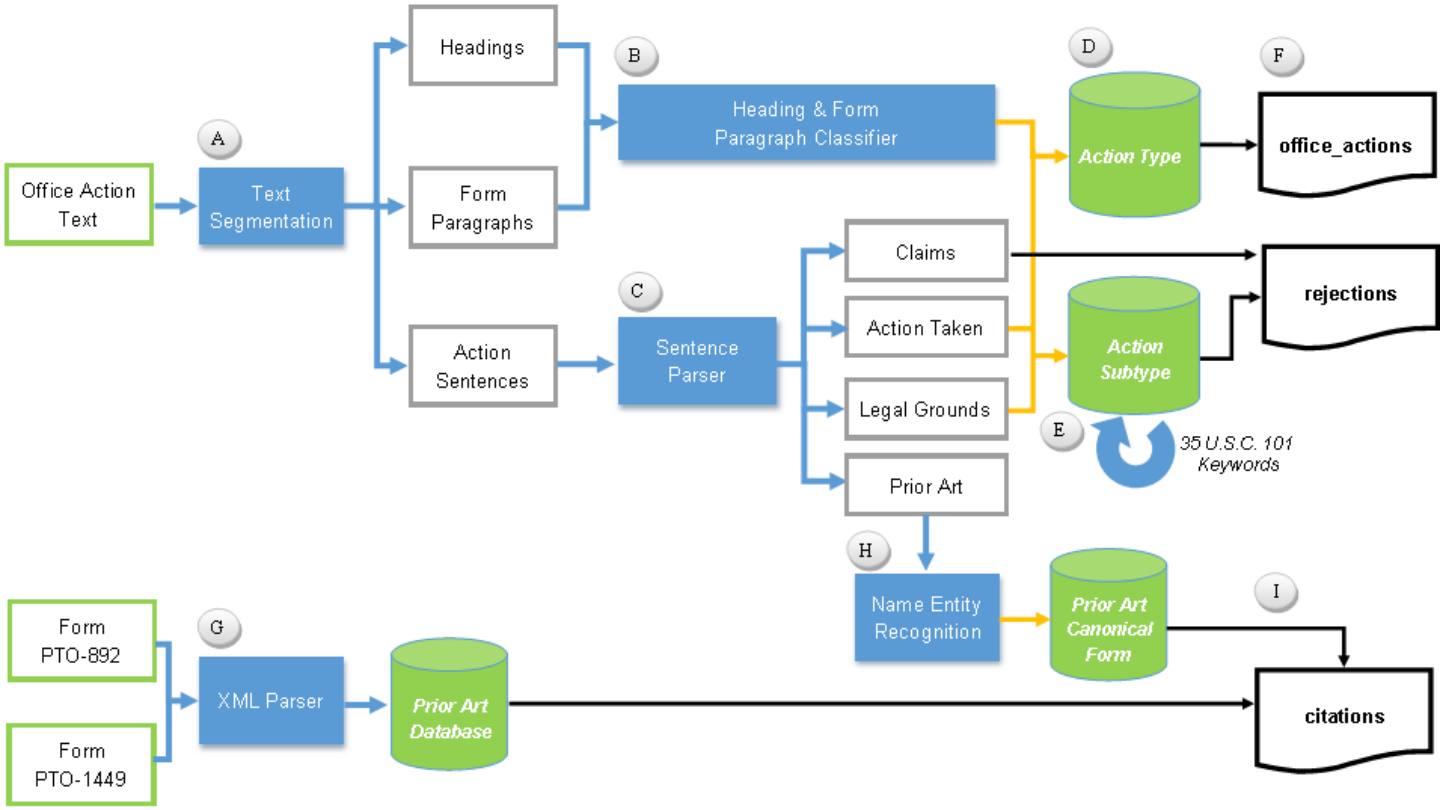
Notes: Sample text extracted from Office actions included in the *Dataset*. Standardize headings, legal form paragraphs, and sentences indicating the action taken by the examiner on the designated claims (or "action sentence") identified.

Figure 2. Office action without headings or form paragraphs



Notes: Sample text without standardize headings or legal form paragraphs extracted from Office action included in the *Dataset*. Sentence indicating the action taken by the examiner on the designated claims (or "action sentence") identified.

Figure 3. USPTO *Office Action Research Dataset for Patents* data generation process



A Text segmentation divides text of each document into three units: Headings; Form Paragraphs; and Action Sentences

B Heading and form paragraph classifier assigns an action type and sub-type based on text matching to a pre-labeled set of standardized headings and form paragraphs from examiner tools and manuals

C Sentence parser is applied to Action Sentences to generate constituency-based parsing trees. From trees, four elements are extracted: Claims in question; Action Taken on claims; Legal Grounds for action; and Prior Art cited

D Action type is assigned based on results from classifier (B) and action taken/legal grounds parsed from Action Sentences (C)

E Action subtype is assigned based on results from classifier (B) and legal grounds parsed from Action Sentences (C). Action subtype for rejections per 35 U.S.C. 101 assigned based on keyword matching

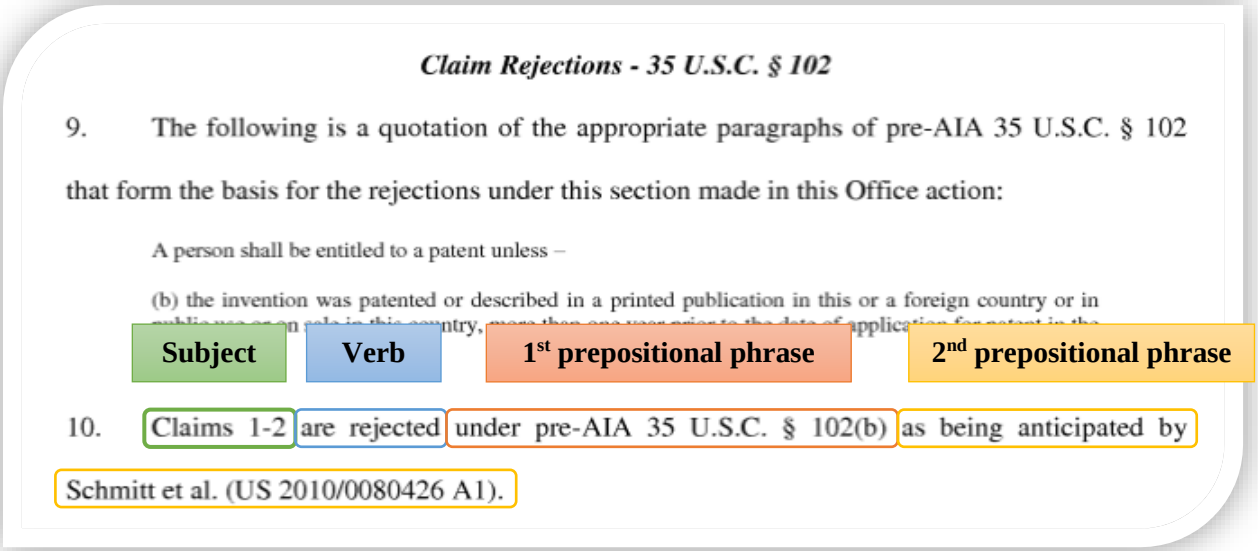
F Action types (D) are encoded and stored in **office_actions** data file. Claims in question and action subtypes (E) are encoded and stored in **rejections** data file

G XML parser is applied to prior art citations listed on Forms PTO-892 and PTO-1449 to generate prior art database of U.S. patent grants and pre-grant publications

H Name Entity Recognition are applied to identify, and store in canonical form, prior art cited by name in the Action Sentences (C)

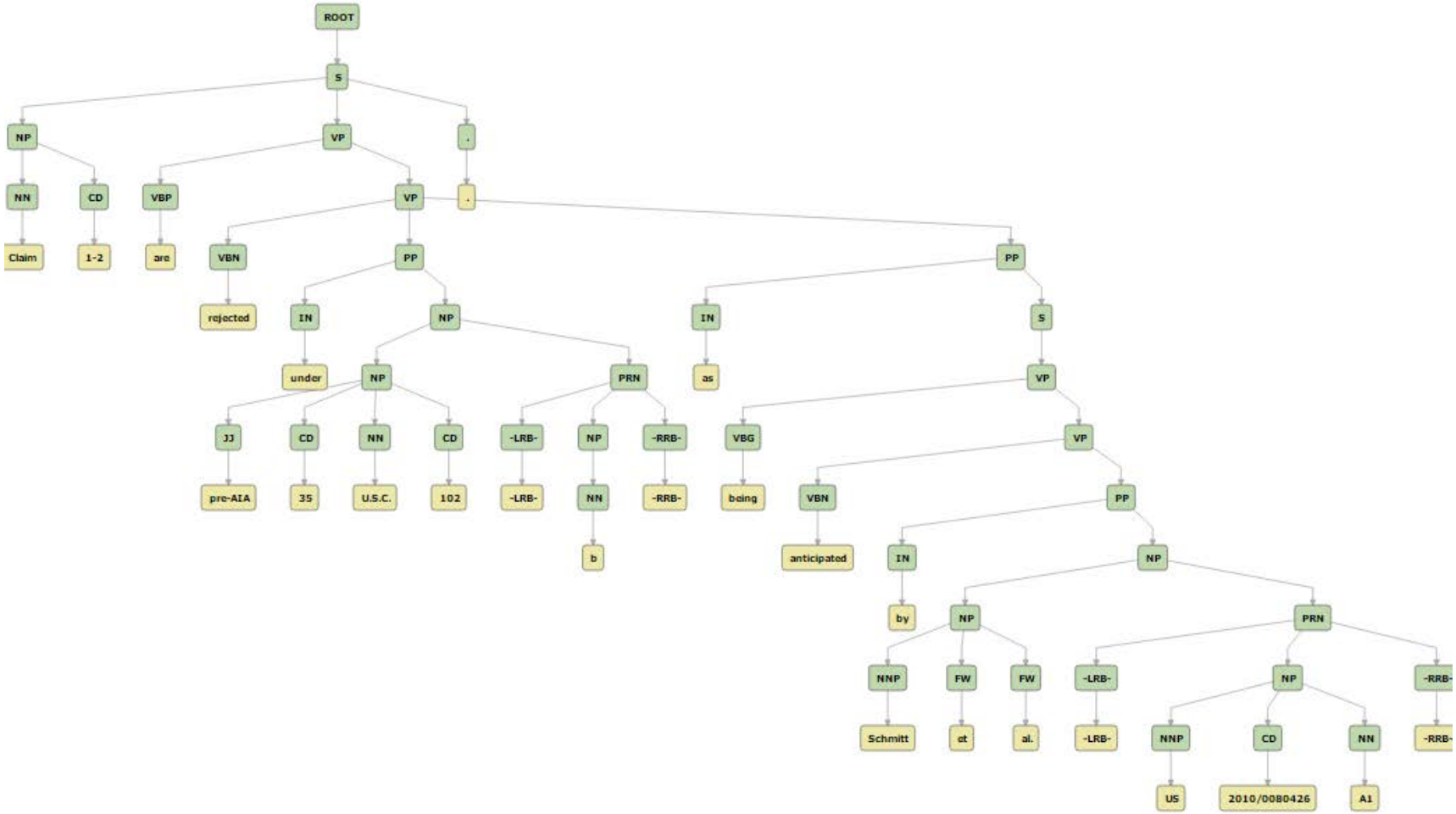
I Prior art citations from Forms PTO-892 and PTO-1449 (G) and Action Sentences (H) are matched and stored in **citations** data file

Figure 4. Standard action sentence syntax structure



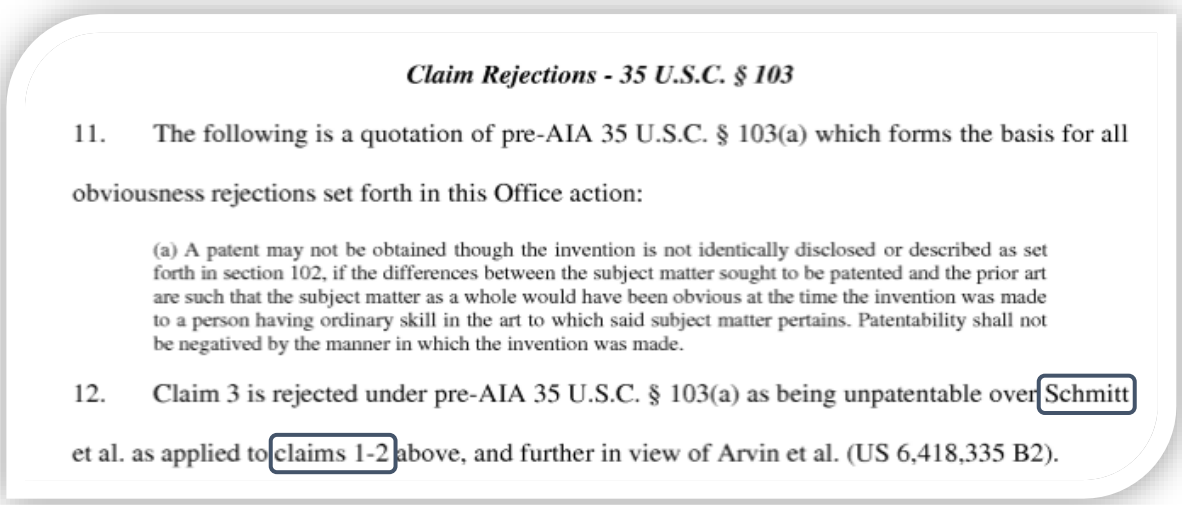
Notes: Sample text from Office action included in the *Dataset*. Sentence indicating the action taken by the examiner on the designated claims (or "action sentence") identified. Action sentences generally follow a consistent structure in which claims to be addressed are the subject (e.g., "claims 1-2"), followed by a verb phrase reflecting the action taken (e.g., "are rejected"), a first prepositional phrase stating the legal grounds (e.g., "under pre-AIA 35 U.S.C. 102(b)"), and a second prepositional phrase declaring prior art references (e.g., "as being anticipated by Schmitt et. al. (e.g., US 2010/0080426 A1)").

Figure 5. Parsing tree of action sentence labeled with part-of-speech tags



Notes: Parsing tree generated from the action sentence in Figure 4. Each constituent of the sentence is labeled with a part-of-speech tag (e.g., NP: noun phrase, VP: verb phrase, PP: prepositional phrase, etc.)

Figure 6. Non-standard action sentence syntax structure



Claim Rejections - 35 U.S.C. § 103

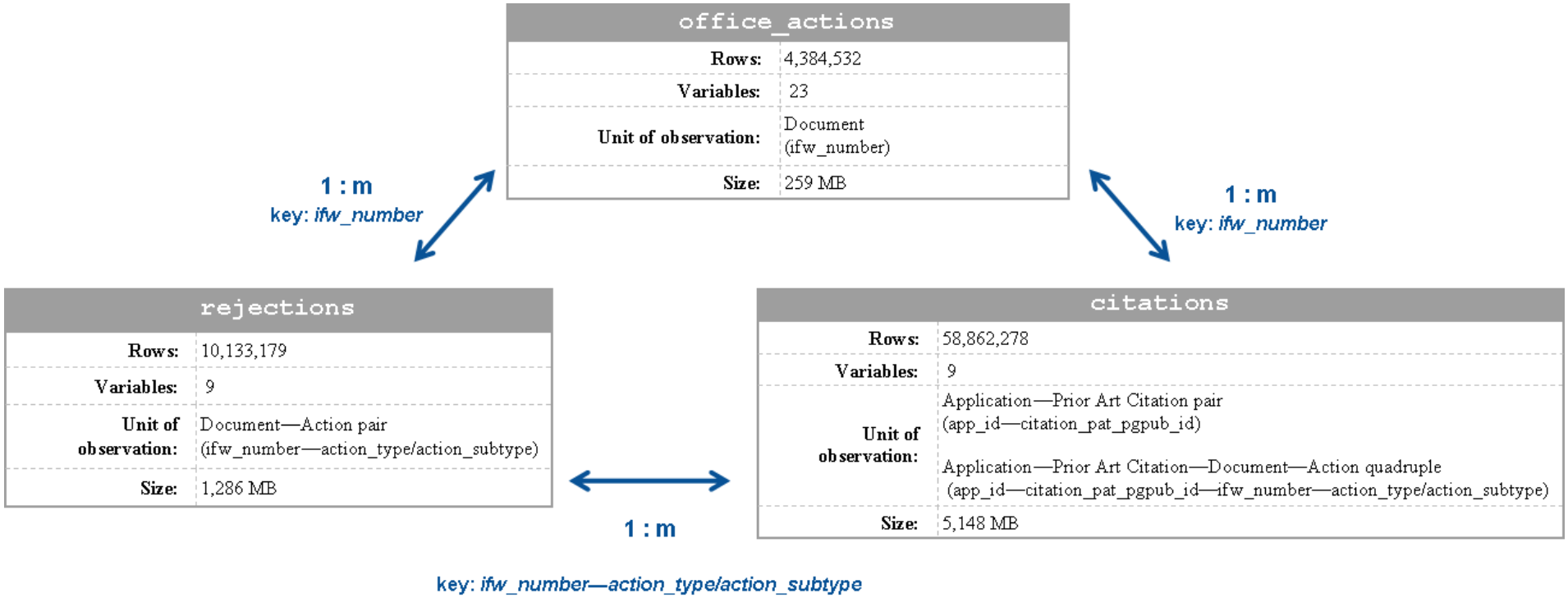
11. The following is a quotation of pre-AIA 35 U.S.C. § 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

12. Claim 3 is rejected under pre-AIA 35 U.S.C. § 103(a) as being unpatentable over Schmitt et al. as applied to claims 1-2 above, and further in view of Arvin et al. (US 6,418,335 B2).

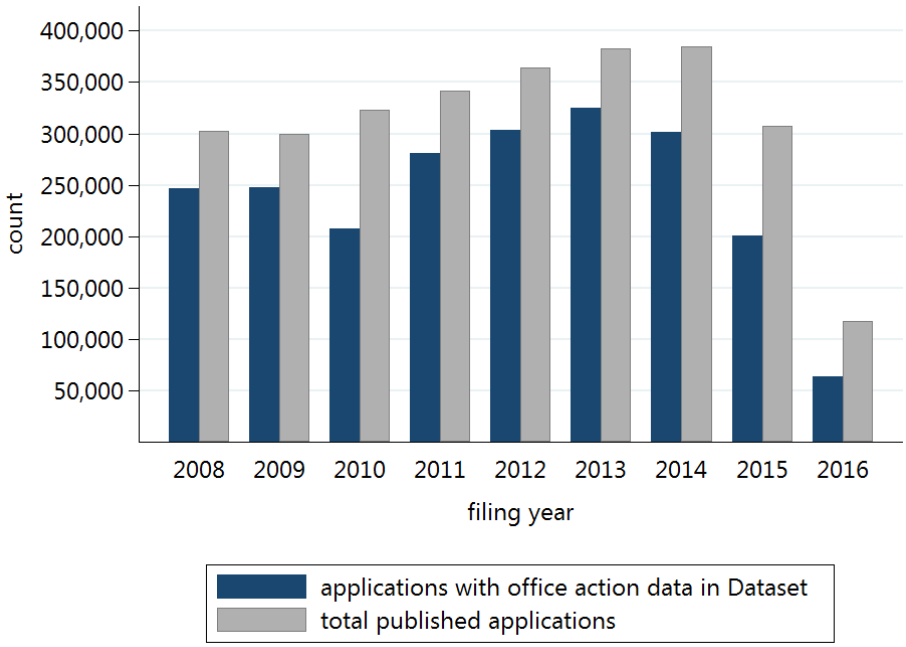
Notes: Sample text from an Office action included in the *Dataset*, in which the action sentence does not follow the standard structure depicted in Figure 4. The prior art cited ("Schmitt et al.") is only mentioned by name, rather than by full citation, because it has been previously mentioned in the document.

Figure 7. USPTO *Office Action Research Dataset for Patents* structure



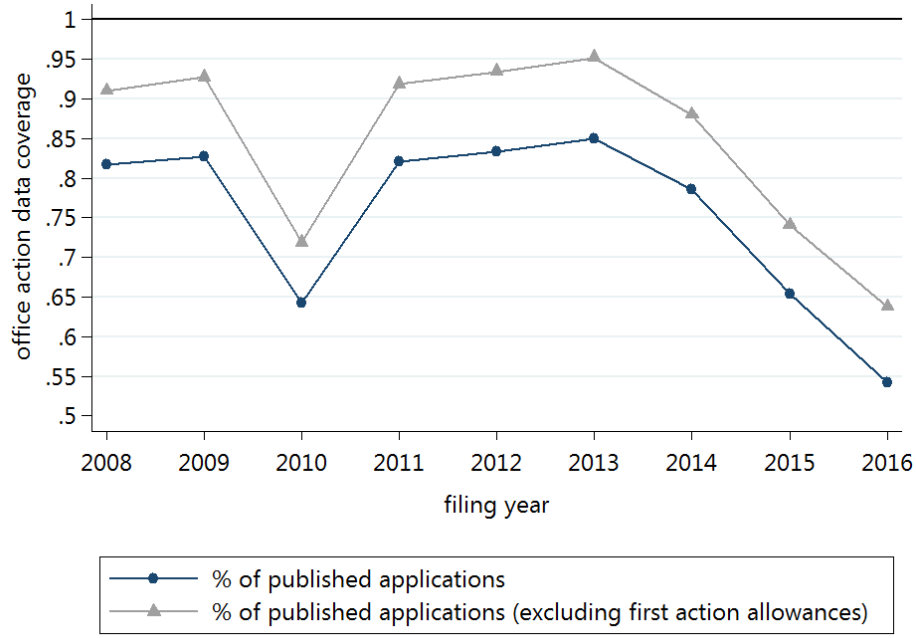
Notes: Figure depicts the organizational structure of the *Office Action Research Dataset for Patents*. The *Dataset* includes three data files that can be merged with the identified key variables. The **office_actions** file contains Office action document level data. Each Office action is identified by a unique *ifw_number* which serves as the key for performing a one to many join of the **office_actions** table with the other two data files. The **rejections** file contains data at the Office action document-action pair level. A unique pair is identified by the *ifw_number* and action type (*action_type*) and subtype (*action_subtype*) combination. The **citations** file is derived from citations referenced on the Form PTO-892, Form PTO-1449, and text of Office actions. For citations not referenced in an Office action, a unique observation in the **citations** file is the application-citation pair (as identified by *app_id* and *citation_pat_pgpub_id* fields). For citations that are referenced in an action, the **citations** file contains fields to enable linking the application-citation pair back to the document-action pair (via the document *ifw_number* and *action_type/action_subtype* combination) in the **rejections** file.

Figure 8. *Dataset* coverage by filing year cohort – application count



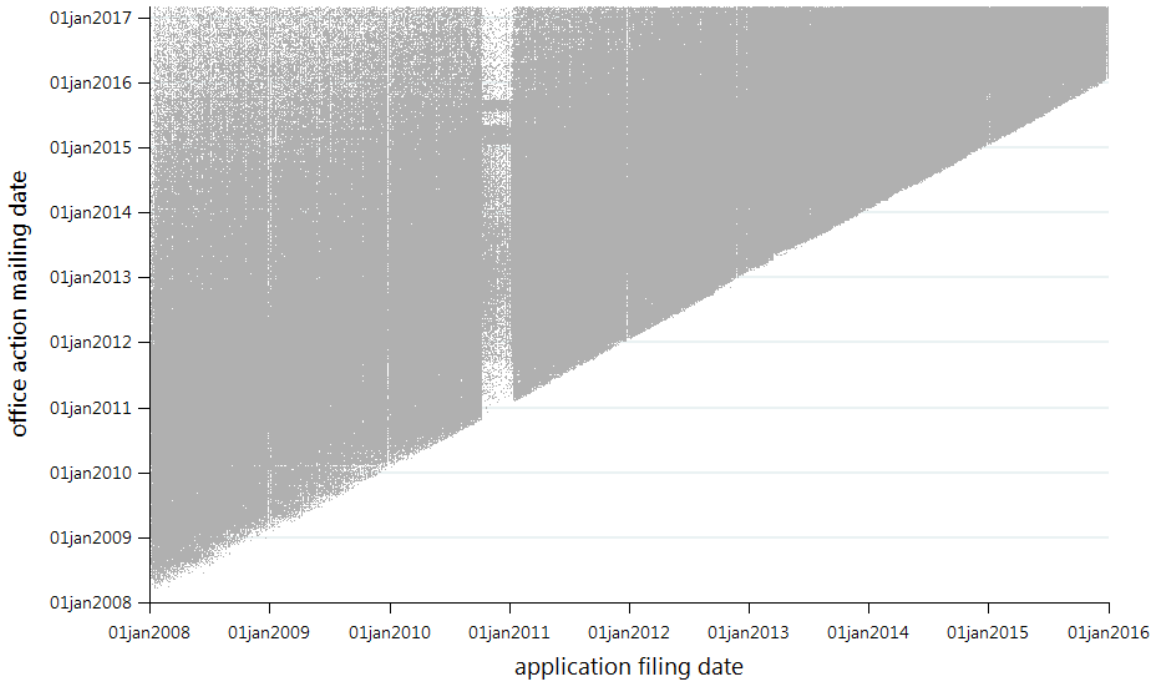
Notes: Figure plots the total number of published applications and the subset of published applications with at least one Office action in the *Dataset* by application filing year. A small number of published applications with filing dates prior to 2008 and at least one Office action in the *Dataset* are excluded from the Figure.

Figure 9. *Dataset* coverage by filing year cohort – percentage



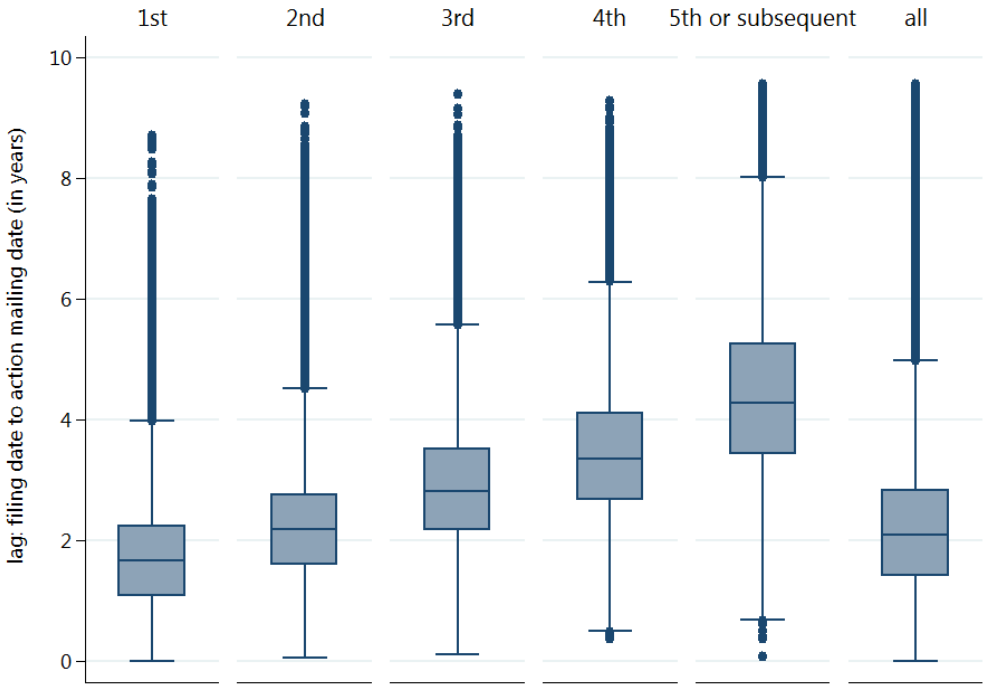
Notes: Figure plots published applications with at least one Office action in the *Dataset* by application filing year, as a percentage of total published applications and as a percentage of published applications that were not allowed on first action. A small number of published applications with filing dates prior to 2008 and at least one Office action in the *Dataset* are excluded from the Figure.

Figure 10. *Dataset* coverage by application filing date and Office action mailing date



Notes: Figure plots the application-Office action document pairs observed in the *Dataset* by the application filing date and Office action mailing date. White areas reflect dates for which there is no Office action coverage in the *Dataset*. Vertical white areas indicate coverage gaps by application filing date. Horizontal white areas indicate coverage gaps by office action mailing date. A small number of published applications with filing dates prior to 2008 and at least one Office action in the *Dataset* are excluded from the Figure.

Figure 11. Distribution of time from application filing to *Dataset* Office action mailing



Notes: Figure depicts a box plot of the distribution of application- Office action document pairs observed in the *Dataset* by the time from application filing date to Office action mailing date in years (lag). Figure includes separate plots for the first, second, third, fourth, and fifth or subsequent Office action observed in the *Dataset* based on earliest mailing date.

Table 1: Frequency of Office actions in the *Dataset* by action type and document code

	CTNF: Non-Final Rejection		CTFR: Final Rejection		Non-Final & Final Rejections	
Action Type[1]	count	percent[2]	count	percent[3]	count	percent[4]
101 rejection - *Subject Matter Eligibility, Statutory Double Patenting, Utility, etc.*	374,865	12.7%	104,199	7.3%	479,064	10.9%
102 rejection - *Lack of Novelty*	1,386,671	46.8%	451,863	31.8%	1,838,534	41.9%
103 rejection - *Obviousness*	2,276,011	76.9%	1,195,329	84.0%	3,471,340	79.2%
112 rejection - *Written Description, Indefinite Claims, etc.*	1,118,364	37.8%	409,285	28.8%	1,527,649	34.8%
Double Patenting rejection (non-statutory)	355,048	12.0%	114,689	8.1%	469,737	10.7%
Objection	897,179	30.3%	280,583	19.7%	1,177,762	26.9%
Allowable claim	195,542	6.6%	118,119	8.3%	313,661	7.2%
Total	2,961,350		1,423,182		4,384,532	

1. Action types are not mutually exclusive

2. Proportion of Non-Final Rejection Office actions in the *Dataset* with designated Action Type

3. Proportion of Final Rejection Office actions in the *Dataset* with designated Action Type

4. Proportion of Non-Final and Final Rejection Office actions in the *Dataset* with designated Action Type

Note: A single office action document may include multiple action types. Accordingly, the type categories are not mutually exclusive and the percentage figures represent the proportion of non-final rejection, final rejection, and all documents in the *Dataset* that include the designated action type.

Notes: Table shows the frequency of Office actions (count of unique *ifw_number* observations) by action type and document code. A single Office action may include multiple action types. Accordingly, the type categories are not mutually exclusive and the percentage figures represent the proportion of Non-Final Rejections, Final Rejections, and combined documents in the *Dataset* that include the designated action type.

Table 2: Frequency of actions in the *Dataset* by type and subtype

Action Type[1]	Action Subtype[2]	count	percent of type subtotal	percent of total	Description
101 rejection *Subject Matter Eligibility, Statutory Double Patenting, Utility, etc.*	non-statutory-alice	56,709	12.4%	0.6%	Judicial exception under Alice decision
	non-statutory-bilski	31,143	6.8%	0.3%	Judicial exception under Bilski decision
	non-statutory-data	124,216	27.1%	1.2%	Non-statutory subject matter - data *per se*
	non-statutory-mayo	8,725	1.9%	0.1%	Judicial exception under Mayo decision
	non-statutory-mentally	24,188	5.3%	0.2%	Judicial exception - invention performed mentally, verbally or w/o machine
	non-statutory-myriad	3,599	0.8%	0.0%	Judicial exception under Myriad decision
	non-statutory-nature	48,094	10.5%	0.5%	Judicial exception - law of nature
	non-statutory-other	99,414	21.7%	1.0%	Non-statutory subject matter - all other
	non-statutory-software	23,114	5.0%	0.2%	Non-statutory subject matter - software
	statutory double patenting	31103	6.8%	0.3%	Invention claimed in multiple patents by the same inventor and/or assignee
	inventorship	52	0.0%	0.0%	Improper naming of inventor
	useful	8,075	1.8%	0.1%	Invention not useful or lacks specific, substantial, and credible utility
	subtotal	458,432	100.0%	4.5%	
102 rejection[3] *Lack of Novelty*	a	336,892	17.8%	3.3%	Taught, used, or known by others before invention
	b	1,233,221	65.2%	12.2%	Taught, used, or sold more than one year before applying (pre-AIA)
	c	54	0.0%	0.0%	Abandoned an invention (pre-AIA)
	d	89	0.0%	0.0%	Foreign patenting more than one year before applying (pre-AIA)
	e	298,987	15.8%	3.0%	Filed by others before invention (pre-AIA)
	f	880	0.0%	0.0%	Applicant is not the actual inventor (pre-AIA)
	g	83	0.0%	0.0%	Interference proceeding establishes that another invented first (pre-AIA)
	(blank)	21,759	1.2%	0.2%	
	subtotal	1,891,965	100.0%	18.7%	
103 rejection *Obviousness*	a	3,001,862	86.2%	29.6%	Obvious to person having ordinary skill in the art
	(blank)	479,613	13.8%	4.7%	
	subtotal	3,481,475	100.0%	34.4%	
112 rejection[4] *Written Description, Indefinite Claims, etc*	a	438,046	24.6%	4.3%	Written description of invention must enable its production and use
	b	1,245,752	70.1%	12.3%	Claims must particularly point out and distinctly claim the invention
	c	2	0.0%	0.0%	Claims may be independent, dependent, or multiple dependent
	d	71,121	4.0%	0.7%	Dependent claims include all limitations of another claim plus further limitations
	e	78	0.0%	0.0%	Multiple dependent claims must reference the other claims in the alternative
	f	1,209	0.1%	0.0%	Conditions for interpretation of the claim using limitations from the specification
	(blank)	21,845	1.2%		
	subtotal	1,778,053	100.0%	17.5%	
Double Patenting rejection		456,595		4.5%	Similar scope claimed by the same inventor and/or assignee
Objection		1,177,245		11.6%	Minor informalities or patent rule violations
Allowable claim		278,348		2.7%	Claims allowable if rejection and/or objection can be overcome
Cancellation		611,066		6.0%	Claim status identifed as cancelled in action text
Total		10,133,179		100.0%	

1. Action types derived from *action_type* field in **rejections** data file.

2. Action sub-types derived from *action_subtype* field in **rejections** data file.

3. Because the Leahy-Smith America Invents Act (AIA) amended certain sections paragraphs of the statute, action subtypes for 102 rejections may have different meaning before and after AIA implementation. AIA amended 35 U.S.C. 102 to establish the first-inventor-to-file system. As a result, two pre-AIA section paragraphs about novelty, 102(a) and 102(b), concord with AIA subparagraph "102(a)(1)". The other pre-AIA section paragraph regarding novelty, 102(e), concords with subparagraph "102(a)(2)". There are no corresponding provisions in the AIA for the remaining paragraphs of the pre-AIA 102 statute: 102(c), 102(d), 102(f), and 102(g). The descriptions for these 102 paragraph subtypes identify them as "pre-AIA"

4. The AIA amended certain section paragraphs under 35 U.S.C. 112. To provide consistent labels in the *Dataset*, we map pre-AIA paragraphs under 35 U.S.C. 112 to their post-AIA equivalents. Specifically, we directly map the first through sixth pre-AIA paragraph to its post-AIA alphabetical equivalent a) through f).

Notes: Table shows the frequency of Office action document-action pairs by action type and subtype combination. In the **rejections** data file each action is specific to the claims in questions. A single Office action may include multiple rejections within the same action type (e.g., there can be both a 112(b) rejection and a 112(d) rejection for the same or an overlapping set of claims) as well as multiple rejections within the same action type-subtype combination (e.g., one set of claims may be rejected under 102(b) based on certain prior art and a separate set of claims may be rejected under 102(b) based on different prior art). Consequently, rejection subtotals will not equate to Office action counts by action type in Table 1.

Table 3: List of variables included in office_actions

Variable Name	Description	Type	Format
app_id	Application number	str8	%-8s
ifw_number	Image File Wrapper (IFW) number of the Office action	str15	%-15s
document_cd	Office action document code	str4	%-4s
mail_dt	Date Office action mailed from USPTO	int	%td
art_unit	Examiner group art unit	str5	%-4s
uspc_class	Invention U.S. Classification	str4	%-4s
uspc_subclass	Invention U.S. Subclassification	str6	%-6s
header_missing	Equals 1 if Office action is missing standard heading(s) or has no heading(s), 0 otherwise	byte	%9.0g
fp_missing	Equals 1 if Office action has no form paragraphs, 0 otherwise	byte	%9.0g
rejection_fp_mismatch	Equals 1 if rejection(s) raised in action sentence(s) does not match form paragraph(s), 0 otherwise	byte	%9.0g
closing_missing	Equals 1 if Office action has no closing paragraph, 0 otherwise	byte	%9.0g
rejection_101	Equals 1 if 35 USC 101 rejection(s) raised, 0 otherwise	byte	%9.0g
rejection_102	Equals 1 if 35 USC 102 rejection(s) raised, 0 otherwise	byte	%9.0g
rejection_103	Equals 1 if 35 USC 103 rejection(s) raised, 0 otherwise	byte	%9.0g
rejection_112	Equals 1 if 35 USC 112 rejection(s) raised, 0 otherwise	byte	%9.0g
rejection_dp	Equals 1 if non-statutory double patenting rejection(s) raised, 0 otherwise	byte	%9.0g
objection	Equals 1 if objection(s) raised, 0 otherwise	byte	%9.0g
allowed_claims	Equals 1 if allowable claim(s) in Office action, 0 otherwise	byte	%9.0g
cite102_gt1	Equals 1 if more than 1 prior art citation referenced in 102 rejection, 0 otherwise	byte	%9.0g
cite103_gt3	Equals 1 if more than 3 prior art citations referenced in 103 rejection, 0 otherwise	byte	%9.0g
cite103_eq1	Equals 1 if only 1 prior art citation referenced in 103 rejection, 0 otherwise	byte	%9.0g
cite103_max	Max number of prior art citations referenced in 103 rejection	byte	%9.0g
signature_type	Signature Type (0-Examiner; 1-Primary Examiner (PE); 2-Examiner + PE; 3-Examiner + Supervisory Patent Examiner (SPE); 4-Examiner + PE/SPE + Director)	byte	%9.0g

Notes: Table lists the name and description of the variables included in the **office_actions** data file of the *Dataset.*

Table 4: List of variables included in rejections

Variable Name	Description	Type	Format
app_id	Application number	str8	%-8s
ifw_number	Image File Wrapper (IFW) number of the Office action	str15	%-15s
action_type	Type of action raised, indicated by section of 35 USC or category (double patenting, objections, or allowed claims)	str29	%-20s
action_subtype	Subtype of action raised, indicated by section paragraph of 35 USC or keyword	str22	%-20s
claim_numbers	Claims in the application for which the action is raised	strL	%-20s
alice_in	Equals 1 if Alice decision referenced in Office action text, 0 otherwise	byte	%9.0g
bilski_in	Equals 1 if Bilski decision referenced in Office action text, 0 otherwise	byte	%9.0g
mayo_in	Equals 1 if Mayo decision referenced in Office action text, 0 otherwise	byte	%9.0g
myriad_in	Equals 1 if Myriad decision referenced in Office action text, 0 otherwise	byte	%9.0g

Notes: Table lists the name and description of the variables included in the **rejections** data file of the *Dataset.*

Table 5: List of variables included in citations

Variable Name	Description	Type	Format
app_id	Application number	str8	%-8s
citation_pat_pgpub_id	Patent or pre-grant publication prior art cited	strL	%-20s
parsed	Patent or pre-grant publication number parsed from citation_pat_pgpub_id	str16	%-20s
form892	Equals 1 if citation listed on Form PTO-892, 0 otherwise	byte	%9.0g
form1449	Equals 1 if citation listed on Form PTO-1449, 0 otherwise	byte	%9.0g
citation_in_oa	Equals 1 if citation referenced in the action sentence of the Office action	byte	%9.0g
ifw_number	Image File Wrapper (IFW) number of the Office action	str15	%-20s
action_type	Type of action raised, indicated by section of 35 USC or category (double patenting, objections, or allowed claims)	str3	%-15s
action_subtype	Subtype of action raised, indicated by section paragraph of 35 USC or keyword	str1	%-11s

Notes: Table lists the name and description of the variables included in the **citations** data file of the *Dataset.*

Appendix A: Jaro-Winkler distance and Jaccard Index

The Jaro-Winkler distance between two text strings s_1 and s_2 is:

$$d_{w} = d_j + (lp(1 - d_j))$$

where:

$$d_j = \begin{cases} 0 & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & otherwise \end{cases}$$

$|s_i|$ is the length of s_i

m is the number of matching characters within half the length of the longest string or $\frac{\max(|s_1|,|s_2|)}{2} - 1$

t is the number of transpositions, i.e. number of characters out of sequential order divided by two

l is the length of the common prefix at the start of the string up to a maximum of 4 characters, and

p is a constant scaling factor for common prefixes or 0.1

The standard Jaccard Index is defined as:

$$J(A, B) = |A \cap B|/|A \cup B|$$

where:

A is a set of stemmered (a Porter stemmer) words of one string, and

B is a set of stemmered words of another string.

To compensate for the nature of headings (normally two or three words), a variation is defined as:

$$J(A, B) = |A \cap B|/|MIN(A, B)|$$

in which MIN function takes the smaller number in two sets.

Appendix B: Cosine Similarity

We calculate the cosine similarity between two form paragraphs as the cosine distance between their frequency vectors A and B as:

$$Cosine(A,B) = \frac{A \cdot B}{\|A\|_2 \|B\|_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

where:

A is a Term Frequency vector of a string, and

B is a Term Frequency vector of another string.

Appendix C: Mapping of key search terms and phrases to 101 rejection action subtype labels

Action Subtype label	Search terms and phrases
non-statutory-nature	Nature, natural, law of nature, natural process, human organism
non-statutory-alice	Alice
non-statutory-bilski	Bilski
non-statutory-data	Transitory, Software, Computer program, Computer Storage, Machine readable, Readable Media
non-statutory-mayo	Mayo
non-statutory-mentally	Directed merely to an abstract idea
non-statutory-myriad	Myriad
non-statutory-other	Being directed to non-statutory subject matter, the claimed invention is directed to non-statutory subject matter, be directed to an abstract idea, consideration of all, not directed to an abstract idea, based upon an analysis with respect to the claim, different than a judicial exception, directed to a judicial exception, significantly more than abstract idea
non-statutory-software	Transitory, Zletz, Nuijten
statutory double patenting	Independent or distinct restriction, Same invention as that of claim, more than one patent, same invention as of prior, same invention as of copending, Judicially created doctrine
inventorship	Incorrect inventorship, correct inventorship
useful	Applicant is intending to encompass, Asserted utility, Lacks patentable utility, inoperative