Using analytics to generate glossaries in patent applications

Computer-driven glossaries would lead to higher-quality patents and less litigation. The technology is available to put them in place

By **Manny Schecter** and **Alison Mortinger**

Three years ago we published an article entitled, "A case for adopting controlling dictionaries in the USPTO" (*Intellectual Asset Management*, issue 39, January/February 2010, pages 51 to 55). The article explained the way in which ambiguities in claims were resolved and proposed the use of a controlling dictionary (or a hierarchy of dictionaries or treatises) established by the US Patent and Trademark Office (USPTO). The article accounted for different controlling dictionaries by technology and allowed patent applicants to override such dictionaries by providing their own definitions or referring to other dictionaries. Ensuring that definitions of claim terms are available would significantly enhance the clarity and predictability of claim meaning, thereby reducing disputes over claim interpretation. In addition, the application of controlling dictionaries would be part of the patent file history, and therefore intrinsic evidence that would be considered more reliable should a claim interpretation dispute arise during litigation.

Common criticisms of our proposal were the difficulty in choosing the most appropriate dictionaries and the hierarchy among them (if using more than one dictionary), and the burden on patent applicants to access potentially multiple sources in order to verify their satisfaction with the definitions that they provide (and to provide overriding definitions if they were not satisfied). Our comeback has been that any dictionary would be better than none, provided that the benefits outweighed the downside. Nevertheless, the criticisms will inevitably be alleviated by the evolution of technology. Increasingly sophisticated computerised data analytics should and will eventually be used to define claim terms and dramatically reduce the verification burden on patent applicants.

It is well known that computers are capable of storing and searching large databases more quickly than humans. Structured queries find structured data at electronic speeds. Where computers have traditionally not fared so well is in understanding an unstructured query in natural human language. That is because natural language is highly nuanced. Consider the questions, "What breeds bark the most?" and "What species have the thickest bark?" In the first context the term 'bark' refers to the bark of a dog, while in the other the same term refers to the bark of a tree. Computers are now capable of analysing the context, determining the intended meaning and identifying the correct responses to queries such as these. More sophisticated analysis would be required for, "What is the best way to prevent harm to trees caused by dogs urinating on their bark?" because the term 'bark' is used in the context of both dogs and trees.

The famed success of the IBM Watson computer system on the television show *Jeopardy!* demonstrated the cutting-edge capability of computerised data analytics to understand natural language. IBM Watson successfully deciphered sophisticated clues intentionally designed to stump the most talented humans and which would have confounded ordinary computers. The same technology can also be applied to the patent examination process. Computerised data analytics should be used to examine a patent application and all related material to create a glossary of the most substantive claim terms and their definitions.

Empowering examiners

Glossaries created by applicants at the time of filing ensure that claims will be attributed to their intended scope; however, most applicants do not include glossaries, most likely because ambiguities can often be exploited at the time of assertion. The use of analytics-driven glossaries created by the USPTO would allow examiners to:

- Appreciate the scope of claims fully.
- Shape initial prior art searches or extend existing searches for additional prior art, depending on the timing.
- Consider properly the claims in view of the prior art.

USPTO-created glossaries would be presented to patent applicants before or within the first office action on the merits. Applicants would then have the opportunity to correct improperly defined terms so that once a patent issues, the meaning of claim terms is unambiguous and on the record for the life of the patent.

There would be several inputs to the process of generating a glossary. The primary input to an analytics tool built for this purpose would be the patent application, including any materials incorporated by reference. Secondary inputs would include material relevant to the application, such as related applications (including priority or divisional applications), applicant-supplied or default dictionaries, treatises or USPTO class definitions (the use of default dictionaries was described in our earlier article). Tertiary inputs include applicant-supplied prior art and publicly available information, such as the Internet.

The next phase in the process would be the creation of a proposed glossary of claim terms based on those inputs, as shown in Figure 1.

Figure 1. **Creation of a proposed glossary of claim terms**

Step 1	• Identify common words used (eg, 'a', 'an', 'the', 'comprising'). • Identify in marked-up claims (first occurrence only) and discard from the analysis.
Step 2	• Perform a combinatorial search on claim terms to find which appear as phrases in literature from the specified technical field. • Identify those terms as phrases for subsequent steps.
Step 3	• (Lex*) Check for applicant-defined glossary in specification. • If match, identify source in mark-up and copy to proposed glossary.
Step 4	• (Lex*) Check for term definitions from context within specification. • If discernible, identify source and copy to proposed glossary.
Step 5	• (Dictionary, applicant) Check for applicant-specified dictionaries (with ranking). • If dictionaries cited, attempt match according to ranking. • If match, identify source in mark-up and copy to proposed glossary.
Step 6	• (Dictionary, tech centre) Check for common terms for USPTO tech centre. • Compare terms to a tech centre-specified dictionary, if there is one. • If match, identify source in mark-up and copy to proposed glossary.
Step 7	• (Class definition) Check for terms in USPTO class definitions. • If match, identify source in mark-up and copy to proposed glossary.
Step 8	• (Common meaning) Consult tertiary sources for definition. • If match, identify source in mark-up and copy to proposed glossary.

** Steps 3 and 4 allow an applicant to be, as is known in the patent world, "his or her own lexicographer" and as such defer first to an applicant-supplied definition, either in a glossary or in the text of the description*

For example, using parts of a patent application as shown in Figures 2 and 3, we can show how the tool would work to generate a proposed glossary for the examiner and the applicant in Figure 4.

Step by step

In Step 1 of the process outlined in Figure 1 the analytics tool would parse the terms of the claim shown and eliminate common words (eg, 'a', 'comprising', 'located', 'on', 'top', 'surface', shown in green) by means of comparison to a master list. This master list can be specific to the desired degree of granularity (eg, tech centre, class or subclass), and can be automatically generated by analysis of a number of issued patents, along with input from examiners.

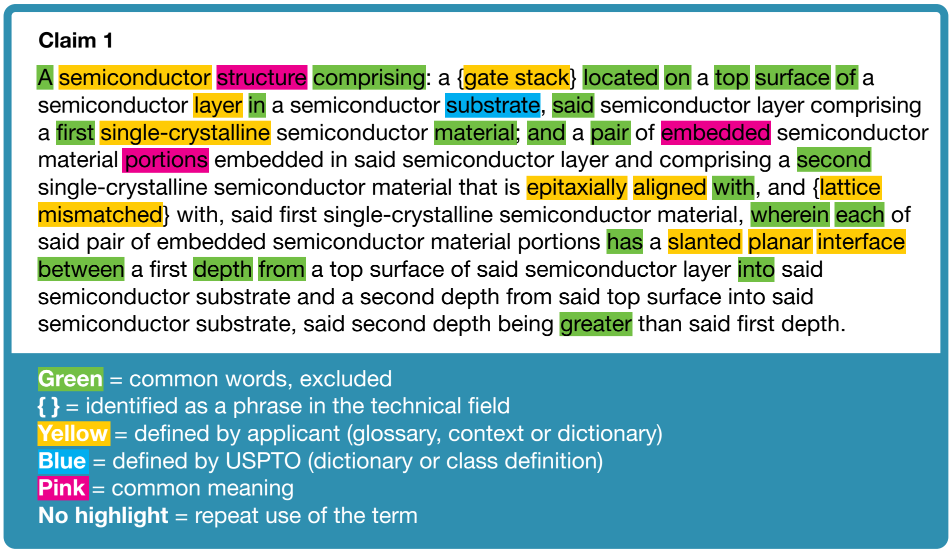
Once the common words have been discarded, the remaining words are significant claim terms. Each of the following steps is performed for each significant claim term.

In Step 2, a search of the semiconductor field results in 'gate stack' and 'lattice mismatched' being identified as phrases. These phrases will be treated as single terms in subsequent steps.

Continuing with Step 3, deference is given to an applicant-supplied glossary. Computationally, this step is relatively simple. There is no glossary in this case.

Step 4 is more complex in that sophisticated analytics will be required to determine the meaning of a claim term from the context of the specification (which includes text from items incorporated by reference). Here, for the term 'semiconductor', the analytics tool finds the first statement that appears to be a definition in the text as highlighted. In order to perform this step successfully, the tool can be trained on a technical domain-specific body of annotated patents to recognise not only ordinary definitional clues such as 'ie', but also often-used phrases in patent applications such as 'may be selected from' or 'may comprise'. With existing natural

Figure 2. **Mark-up of a claim with sources of definitions**

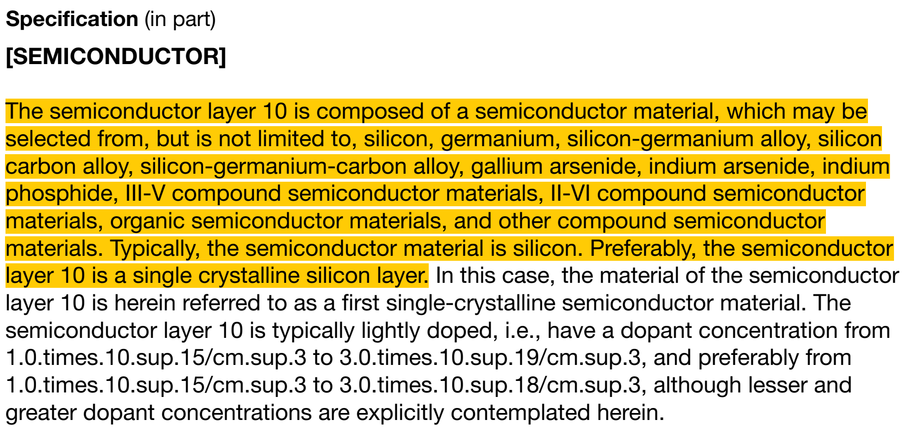


Claim 1

A semiconductor structure comprising: a {gate stack} located on a top surface of a semiconductor layer in a semiconductor substrate, said semiconductor layer comprising a first single-crystalline semiconductor material; and a pair of embedded semiconductor material portions embedded in said semiconductor layer and comprising a second single-crystalline semiconductor material that is epitaxially aligned with, and {lattice mismatched} with, said first single-crystalline semiconductor material, wherein each of said pair of embedded semiconductor material portions has a slanted planar interface between a first depth from a top surface of said semiconductor layer into said semiconductor substrate and a second depth from said top surface into said semiconductor substrate, said second depth being greater than said first depth.

Green = common words, excluded
{ } = identified as a phrase in the technical field
Yellow = defined by applicant (glossary, context or dictionary)
Blue = defined by USPTO (dictionary or class definition)
Pink = common meaning
No highlight = repeat use of the term

Figure 3. **Checking for a term definition from context**



Specification (in part)

[SEMICONDUCTOR]

The semiconductor layer 10 is composed of a semiconductor material, which may be selected from, but is not limited to, silicon, germanium, silicon-germanium alloy, silicon carbon alloy, silicon-germanium-carbon alloy, gallium arsenide, indium arsenide, indium phosphide, III-V compound semiconductor materials, II-VI compound semiconductor materials, organic semiconductor materials, and other compound semiconductor materials. Typically, the semiconductor material is silicon. Preferably, the semiconductor layer 10 is a single crystalline silicon layer. In this case, the material of the semiconductor layer 10 is herein referred to as a first single-crystalline semiconductor material. The semiconductor layer 10 is typically lightly doped, i.e., have a dopant concentration from 1.0.times.10.sup.15/cm.sup.3 to 3.0.times.10.sup.19/cm.sup.3, and preferably from 1.0.times.10.sup.15/cm.sup.3 to 3.0.times.10.sup.18/cm.sup.3, although lesser and greater dopant concentrations are explicitly contemplated herein.

language processing systems, such as IBM Watson, achieving 100% accuracy will not be possible, so the tool will also show a confidence level in the definition. If so desired, the tool can be tailored to leave out any definition with a confidence level below a certain threshold. Once the tool has determined the contextual definition, it is copied to the proposed glossary.

Steps 3 and 4 allow an applicant to be his or her own lexicographer, as is permitted under US law, and to use any term or even create a new term, provided that "any special meaning assigned to a term is clearly set forth in the specification" (see Memorandum to Technology Centre Directors and Patent Examining Corps from John Love, deputy commissioner for patent examination policy, "Indefiniteness rejections under 35 USC 112, second paragraph").

In Step 5, if the applicant has supplied his or her own dictionary, the tool attempts to locate a definition of the claim terms. In this example, the applicant has not supplied one.

Steps 6 and 7 attempt to find a definition for each significant claim term from USPTO sources: either a default dictionary for the tech centre or definitions for the class that has been assigned on receipt of the application. Here, there is no tech centre dictionary, but fortunately there is a match for 'semiconductor' in the text of the class definition at the URL shown.

Finally, in Step 8 the tool attempts to determine the common meaning of any terms used. Here, one was found for 'semiconductor' using dictionary.com, but many potential sources exist. For example, the Public Patent Foundation provides glossaries of claim terms in different technical fields, assembled from US case law.

Determining definitions

The colour-coded claim in Figure 2 shows how the tool determined the definitions of the terms. Where more than one meaning is found, the tool will choose the first identified source from the steps of the process shown in Figure 1. Here, multiple meanings were found for 'semiconductor', but the term is coded yellow to show that an applicant-supplied definition was used in the proposed glossary. Blue indicates a USPTO source and pink indicates a common meaning. No highlighting is used for the second and subsequent uses of a claim term.

Once generated, the proposed glossary report is presented to both the examiner and the applicant. The proposed glossary will likely show at least one definition, because if nothing else, a common meaning is probable, considering the vastness of the Internet. If there is more than one definition, the examiner will have more than one source to check that the terms are being given their ordinary meaning, provided that they are not inconsistent with the specification required by 35 USC 112, Paragraph 2. If the examiner determines that functional claim language is being used, more weight can be given to the definition determined from the specification under 35 USC 112, Paragraph 6. The colour-coded mark-up of the claim will be a clear indicator of which source is being used.

The proposed glossary can be generated either before or at the time of the first office action. The former is recommended, so that the examiner can factor the definition into the search strategy. If this takes place at the time of the first office action, any prior art found by the examiner can be used as another tertiary input to the definitions. The examiner and the applicant will have an opportunity to review and correct any inaccuracies in the proposed glossary before it becomes final and part of the record, and ultimately published as part of the issued patent. This review will ensure that the applicant is satisfied with all definitions, including any that have been created by the applicant choosing to be his or her own lexicographer.

The glossary should become final early in the examination process so that examination can proceed based on a common understanding of the claim terms. Suitable disincentives can be provided to ensure that applicants cannot avoid a definition for every term – for example, a deadline can be set for response, beyond which the applicant's consent to the proposed glossary terms is presumed. The final glossary must be similar to the proposed glossary, but will only have the finally approved (by the examiner and the applicant) definition for each claim term. The burden on the applicant, although not zero, is minimised, because a concise document with all relevant information is provided and the applicant can simply indicate acceptance or choose an alternate definition from among those presented. Rarely, an applicant will need to refer to the specification or other source for a definition that was not found, and those results can be fed back into the tool to provide increased reliability for future applications.

Achievable process

The proposed process is certainly achievable, but there will be technical hurdles – most notably the determination of a claim term meaning based on context from the specification and any materials incorporated by reference. We are confident that these hurdles can be overcome during the implementation phase and, once complete, the tool could be made public so that it can optionally be used on applications before submission to the USPTO. Incentives could be provided for applicants to do so (eg, lower fees or accelerated examination), since all claim terms would have a clear definition and thus the claims as a whole would be much more likely to satisfy 35 USC 112, Paragraph 2.

The creation of analytics-driven glossaries as described above would occur at electronic speeds and result in patents with significantly less ambiguity. The scope of claims would be clearer and less disputable, reducing litigation – including the need for Markman hearings – to resolve disputes and easing the ability of the public to design around. The advantages would outweigh the slight additional burden on applicants. There would likely continue to be issues with the doctrine of equivalents; however, the metes and bounds of the invention would be more apparent to everyone – patentees, licensees and other inventors, as well as the courts. We predict that computerised data analytics will be used as described as soon as an application specific to patent examination can be completed, and as soon as USPTO resources allow for implementation.

Figure 4. **Proposed glossary report**

COMMON WORDS EXCLUDED, In order of appearance
A, comprising, located, one, top, surface, of, in, first, and, pair, second, with, wherein, each, has, between, depth, from, into, greater, than

DEFINITION INPUTS
GLOSSARY: No
DICTIONARY CITED BY APPLICANT: No
TECH CENTRE DICTIONARY: None
CLASS DEFINITION:
http://www.uspto.gov/web/patents/classification/uspc257/defs257.htm

CLAIM TERMS, In order of appearance
Semiconductor
LEX/APPLICANT DEFINED
Source: Context
Confidence level: 70%
Meaning: selected from, but is not limited to, silicon, germanium, silicon-germanium alloy, silicon carbon alloy, silicon-germanium-carbon alloy, gallium arsenide, indium arsenide, indium phosphide, III-V compound semiconductor materials, II-VI compound semiconductor materials, organic semiconductor materials, and other compound semiconductor materials. Typically, the semiconductor material is silicon. Preferably, the semiconductor layer 10 is a single crystalline silicon layer.
DICTIONARY, APPLICANT PROVIDED: None
DICTIONARY, TECH CENTRE: None
CLASS DEFINITION
Source: http://www.uspto.gov/web/patents/classification/uspc257/defs257.htm
Meaning: A material whose electrical resistivity is between that of insulators and conductors. The resistivity is commonly changed by light, heat, electric, or magnetic fields incident on the material. Current flow is achieved by transfer of positive holes as well as by movement of electrons.
COMMON MEANING:
Source: dictionary.com
Meaning: substance, as silicon or germanium, with electrical conductivity intermediate between that of an insulator and a conductor.

Action plan

- Computerised data analytics should be used to generate claim term glossaries from a hierarchy of sources.
- The first consulted source will be the specification, preserving the ability for an applicant to be his or her own lexicographer.
- The applicant will have the opportunity to review and correct the glossary.
- A glossary on the record for every patent will reduce ambiguity, make claim scope clearer for everyone and reduce litigation.

Manny W Schecter is associate general counsel, IP law, and chief patent counsel with IBM Corporation.
Alison D Mortinger is counsel, IP law strategy and policy at the company.

The authors would like to thank **Scott Spangler**, IBM Almaden Research Lab, for his assistance on the subject of data mining, and **Anthony Levas**, IBM Watson Research Lab, for his assistance on the subject of natural language processing.

The opinions expressed in this article are those of the authors, and not necessarily those of the IBM Corporation.