24 Nov 2020

Measuring diversity in invention and patenting is easier said than done

Stakeholders across the IP community are keen to increase diversity in their ranks but, as Suzanne Harrison and Erik Oliver explain, even the most sophisticated companies face a challenge in building a full picture of precisely who is participating in the patenting process

Diversity has been a hot topic for several years. Whether it is racial, gender or economic diversity, there has been plenty of coverage of how the lack of it is affecting individuals, businesses and the economy. While increasing diversity is an important and worthwhile topic overall, this article focuses specifically on increasing gender diversity in intellectual property and specifically in patenting.

Members of the Gathering have been grappling how to achieve meaningful change in the gender diversity representation of group members' IP portfolios. By way of background, the Gathering is an IP best practices group comprising companies that meet to collectively explore, identify and create best practices around IP and other intangible assets. Current members include AT&T, Comcast, Dropbox, Facebook, Google, Imidomics, Intel, Juul, LinkedIn, Rambus, Red Hat, Richardson Oliver Law Group LLP, Seagate, Uber, US Navy, and View .

Naively, the group thought it would be relatively easy to get their portfolio diversity data, identify the problem(s), solve some issues and ultimately achieve change. This turned out to be very wrong. Very early on, the Gathering companies got stuck on the most foundational question: how to get accurate gender data for their own patent portfolios so that they could see how many female staff members were actively engaged in the patenting process.

Even within their own organisations, getting the data presented a number of hurdles to the members.

- First, was determining whether human resources (HR) had control of the data. If they did not, then who did? If yes, trying to work with HR to get the data and understand the internal rules and policies around its usage was often difficult. For example, gender data was only available for current employees. Previous employee gender data was not available so other than surfing LinkedIn or other sites, it was extremely difficult and time consuming to get that information. Thus a full portfolio gender view (over long time periods) was not possible.
- Second, if HR had the data, but was not willing to share it, then companies needed to resort to manually getting the data, for example through corporate directory pictures or sites like LinkedIn. It quickly became apparent that the group needed an easier, faster and cheaper solution while maintaining accuracy.

This challenge served as the launching point for examining other approaches to obtaining gender data for patent portfolios.

Setting up the experiment

The group quickly settled on the necessity of an algorithmic approach to determining gender from patent data and a test to determine which algorithms would be useful for this purpose. The test included companies sharing their actual data and comparing it to the algorithms to see which came closest to those numbers, and whether the algorithm was sufficiently accurate for corporate use.

For this test, we chose the following algorithms:

- **USPTO Patents View data** – this provides gender for specific issued patents only.
  - This is a freely available set of data files created by the USPTO and hosted by a third party. The data files have had commercial algorithms applied to assign genders to inventors by the USPTO's Office of the Chief Economist
  - Also, we tried an approximation of gender using the USPTO Patents View data. We extended the specific mappings provided in the Patents View data to try to map names to genders generically. Note, this will be imprecise since the Patents View data took into account many factors including country of origin whereas our approximation did not.
- **WIPO** – this article includes a data file to map first names and a country to a gender.
- **Social Security Baby Names Index** - we chose for this experiment to limit the data to babies born in the US between 1960 and 2002.

A dozen Gathering members provided us a data set of nearly 23,000 US and international patents covering over 65,000 total inventors, we tried all four of the above algorithms (we used two variants of USPTO Patents View).

In the course of this test, we gained a number of interesting insights:

- No one algorithm was perfect
- Anglicised Chinese and Indian names represented the bulk of the non-matches

- Undetermined inventors were nearly entirely male
- Manually checking that your most prolific inventors are correctly gendered is important.

Additionally we realised that there were some implicit requirements for an algorithm that the members had:

- It needed to be transparent, eg people wanted to be able to see how different inputs worked and results were calculated
- Users needed to be able to track the entire life cycle from patent disclosure through to grant
- The algorithm needed to be free
- It needed to be easy to use without a data scientist on your team
- The algorithm needed to be accurate, eg correlate well with “true North”/actual numbers

Congress, the USPTO and diversity

Experimentation

To help determine which algorithm to use, we built a simple Excel spreadsheet as the test bed for the group. We asked the 12 participating companies to provide us *at least* 100 US issued patents *and at least* 100 international patents. With the requirement that the company must know the gender of the inventors on the patents - this would constitute the “True North”. The companies had to provide their lists in the form shown in figure 1. Basically the format is one row per inventor-per patent. So a patent like US9123456B2 with eight inventors requires eight rows.

Figure 1. Data Input Sample

Patent or Pub	First Name	Country
US9123456B2	Takashi	JP
US9123456B2	Shotaro	JP
US9123456B2	Shinji	JP
US9123456B2	Taikou	JP
US9123456B2	Takamasa	JP
US9123456B2	Masanori	JP
US9123456B2	Akio	JP
US8512345B2	Adam	US
US10000000B2	Joseph	US
...	...	...

This data was then analysed through the tool using all four algorithms (see figure 2) and combined with the True North provided by each company separately:

Figure 2. Results for a Company

Inventor Counts	True North		WIPO Data		Social Security Data		USPTO Algorithm Approximated		USPTO Patents View Results	
Item	Total	Percent	Total	Percent	Total	Percent	Total	Percent	Total	Percent
Number of Inventors	216	100%	216	100.00%	216	100.00%	216	100.00%	161	100.00%
Number of Female Inventors	15	7%	17	7.87%	26	12.04%	18	8.33%	6	2.78%
Number of Male Inventors	201	93%	150	69.44%	150	69.44%	176	81.48%	114	52.78%
Number of Undetermined Inventors	0	0%	49	22.69%	40	18.52%	22	10.19%	41	18.98%
Disclosure (or Patent) Counts										
Item	Total	Percent	Total	Percent	Total	Percent	Total	Percent	Total	Percent
Number of Unique Disclosures (or Patents)	101	100%	101	100.00%	101	100.00%	101	100.00%	69	100.00%
Number with at least one Female Inventor	14	14%	15	14.85%	23	22.77%	17	16.83%	6	5.94%
Number with at least one Male Inventor	94	92%	79	78.22%	82	81.19%	87	86.14%	66	65.35%
Number with solo Female Inventor	7	7%	5	4.95%	10	9.90%	8	7.92%	3	2.97%
Number with solo Male Inventor	37	36%	24	23.76%	28	27.72%	32	31.68%	18	17.82%
Weighted count of Female Disclosures	9.75	10%	8.83	8.75%	15.42	15.26%	11.42	11.30%	3.92	3.88%
Weighted count of Male Disclosures	91.25	89%	65.37	64.72%	66.85	66.19%	79.10	78.32%	50.97	50.46%

By combining this data for the 12 companies, we could see which algorithms were better correlated with True North (see figure 3).

In figure 3, the focus was on the percentage of all inventors who were female inventors. For this statistic, the WIPO algorithm was closer to True North across the companies than the USPTO Patents View data based on mean squared error.

Figure 3. Example comparison

Company Id	# of Pats Provided	# of Inventors	True North - % Female	WIPO - % Female	SS Baby - % Female	USPTO Approx - % Female	USPTO Patents View - % Female	Patents View - Mean Squared Error	WIPO - Mean Squared Error
1	19,000	54,000	19.00%	11.31%	9.89%	10.22%	10.55%	0.007140	0.005914
2	100	300	6.64%	7.17%	6.23%	6.23%	3.74%	0.000841	0.000028
3	100	200	7.00%	7.87%	12.04%	8.33%	2.78%	0.001781	0.000076
4	100	300	4.44%	3.55%	2.96%	2.96%	4.44%	0.000000	0.000079
5	700	2,800	13.47%	13.94%	10.31%	11.66%	7.59%	0.003457	0.000022
6	100	200	5.20%	8.09%	8.67%	6.36%	5.20%	0.000000	0.000835
7	1700	5,600	6.65%	6.89%	4.29%	7.62%	5.12%	0.000234	0.000006
8	100	250	9.24%	11.34%	10.50%	8.40%	4.62%	0.002134	0.000441
9	100	400	4.22%	16.89%	7.12%	17.15%	17.15%	0.016718	0.016053
10	50	200	2.55%	4.46%	1.27%	1.27%	1.27%	0.000164	0.000365
11	100	200	1.72%	2.15%	2.15%	2.15%	3.00%	0.000164	0.000018
12	100	300	10.78%	6.21%	7.84%	10.13%	9.80%	0.000096	0.002088
Totals	**22,250**	**64,750**	**7.58%**	**8.32%**			**6.27%**	**0.002728**	**0.002160**

Lower #s are better

Note: The number of patents and inventors have been rounded, only US issued patents records were used in this comparison to enable comparison to Patents View

Following the selection of the WIPO algorithm, we now come back to the original questions around bias in patenting. To download a free Excel tool to test this for yourself please click here.

Where do we go from here?

The current version of our algorithm only handles male/female and not non-binary genders. Future versions could better handle that. Similarly, we picked only three primary algorithms to test in this project. There is significant research on other name-to-gender approaches that could be incorporated in future versions. However, for the scope of this project and providing a quick-and-useful tool, the WIPO algorithm was found to be sufficient.

The Gathering is continuing its work on gender diversity in patenting and is moving forward on understanding and identifying bias within corporations, outside counsel and/or at the USPTO. In addition, the group is working on identifying internal metrics for quickly identifying bias and/or areas for gender diversity improvement. Additionally we hope to be able to create and circulate a few metrics for external benchmarking soon. This is too important a subject to not push forward on. We encourage other companies to share their work and learnings.

Erik Oliver

Chief operating officer | Richardson Oliver Insights

Suzanne Harrison

Co-founder and principal at Percipience | CEO of The Gathering

TAGS

Market Developments, Patents, North America, United States of America