A recursive process for mapping and clustering technology literatures: case study in solid-state lighting

Kevin W. Boyack* and Jeffrey Y. Tsao

Sandia National Laboratories,
PO Box 5800, Albuquerque, NM 87185, USA
E-mail: kboyack@mapofscience.com
E-mail: jytsao@sandia.gov
*Corresponding author

Ann Miksovic and Mark Huey

Strategic Perspectives, Inc.,
Boulder, CO, USA
E-mail: ann@perspectivesweb.com
E-mail: mark@perspectivesweb.com

Abstract: We present a method for partitioning article and patent literatures for the purpose of identifying sub-domains within those literatures at multiple levels. The method is based on bibliographic coupling, and is used recursively to identify fine-grained clusters of articles or patents, along with higher level aggregations of those clusters. This method is applied in an analysis of the literatures of solid-state lighting, using a comprehensive dataset of 35,851 English-language articles and 12,420 US patents published or issued during the years 1977–2004. An analysis of trends by nation and continent is reported for the entire knowledge domain as well as at the highest level of clustering. The fine-grained clusters are used to identify the hottest recent topics in both the scientific (articles) and technical (patents) literatures.

Keywords: recursive; mapping; clustering; literature; solid-state lighting; SSL; bibliographic coupling; patents; scientific articles; trends; hot topics.

Reference to this paper should be made as follows: Boyack, K.W., Tsao, J.Y., Miksovic, A. and Huey, M. (2009) ‘A recursive process for mapping and clustering technology literatures: case study in solid-state lighting’, *Int. J. Technology Transfer and Commercialisation*, Vol. 8, No. 1, pp.51–87.

Biographical notes: Kevin W. Boyack spent 17 years at Sandia National Laboratories where he worked in various areas including combustion (experimental and modelling), transport processes, and socio-economic war gaming. His past several years were spent in developing and establishing the accuracy of techniques to map science and technology literatures for strategic purposes. He is currently Chief Technical Officer at SciTech Strategies, Inc. (www.mapofscience.com), focusing on science and technology mapping with associated metrics and indicators, network analysis, and the integration and analysis of multiple data types.

Jeffrey Y. Tsao is a Principal Member of Technical Staff at Sandia National Laboratories, where he studies the integrated science, technology and economics of solid-state lighting and other areas, and, more recently, network models of the 'science of science'. His previous positions include: VP of R&D at E2O Communications, Manager of the Chemical Process Science and Semiconductor Materials Departments at Sandia National Laboratories, and Visiting Lecturer at the Institute of Materials Research and Engineering in Singapore. He has co-authored over 150 publications, is author of the monograph *Materials Fundamentals of Molecular Beam Epitaxy*, and is a Fellow of the American Physical Society.

Ann Miksovic is the Director of Research for Strategic Perspectives, Inc., a management consulting firm focused on the market intelligence needs of R&D organisations, where she has directed hundreds of studies for government and commercial clients. In her 25+ year career as a Researcher, she has specialised in research project design and management, data analysis and technical writing. In recent years, she has focused primarily on projects for DOE national laboratories, where she developed and completed technology benchmarking studies; technical and instrumentation area summaries; federal agency funding studies; and business, science and technology tracking reports.

Mark Huey is Principal with Strategic Perspectives, Inc., a management consulting firm focused on the market intelligence needs of R&D organisations, where he works with project teams in areas ranging from materials science to biosecurity to informatics. He assists clients with technical and market intelligence, competitive positioning, external interactions and project strategy. He has also led several benchmarking studies on R&D portfolios.

1 Introduction

Solid-state lighting (SSL) is an emerging technology for producing visible white light from semiconductors for illumination purposes. It has tremendous potential, both to reduce energy consumption through improved efficiency in the conversion of energy to light, as well as to enhance human productivity and the human visual experience through real-time tailoring of the colour properties, brightness and spatial positioning of light.

SSL is also a global technology in virtually every sense. It will benefit consumers and nations throughout the world. It is now, and likely will continue to be, manufactured by companies throughout the world. And it rests on a foundation of knowledge that has been, and continues to be, developed by scientists and technologists worldwide.

There have been a number of recent reviews of various aspects of SSL: its benefits (EERE, 2003; Tsao, 2003; Schubert and Kim, 2005), national initiatives (Strategic Perspectives, 2006) and the science (BES, 2006) and technology (EERE, 2006; OIDA, 2002a; OIDA, 2002b) itself. In this report, we analyse the foundational knowledge domain associated with SSL, with special attention paid to international trends in the contributions to that knowledge domain, and to particular sub-domains that are evolving rapidly.

The remainder of this report is organised as follows. In Section 2, we discuss the starting point of our analysis: the creation of a relatively comprehensive dataset consisting of articles and patents considered foundational to SSL. In Section 3, we discuss some of the overall features of this dataset, including size, growth rate and key contributions. In Section 4, we partition the dataset according to contributions by nation and continent and discuss international trends to these contributions. In Section 5, we describe an iterative process for mapping and clustering the article and patent datasets into a number of emergent knowledge sub-domains on the basis of bibliographic coupling metrics. In Sections 6 and 8, we discuss international trends in national and continental contributions to the emergent article and patent knowledge sub-domains. In Sections 7 and 9, we discuss the hottest article and patent knowledge sub-domains.

We note that the analysis described in this report is somewhat similar to a recent analysis of international patent trends in nanotechnology (Huang et al., 2004), but differs in several important ways. First, the descriptive area of our analysis is much narrower, and hence is more amenable to self-consistency checks by domain experts. Second, by including both the article and patent literature, we can assess trends in underlying science as well as in technology.[1] Third, we do not focus or provide statistics on individual institutions (e.g., companies or universities). Fourth, we use the data itself to provide emergent categorisation schema rather than relying on preset schema such as the US patent classification system.

2 Creating the dataset

The starting point for this study was the creation of a dataset consisting of articles and patents considered foundational to SSL. En route to the creation of this dataset, a number of choices were made:

1. the primary database(s) of which the dataset would be a subset
2. the boundaries of the knowledge domain we would like the dataset to represent
3. the strategy for constructing the dataset from the primary database(s).

2.1 *Primary databases*

For primary databases, we used two:[2] Thomson Scientific's Science Citation Index for journal articles, and the US Patent and Trademark Office's database for US patents. Both of these databases provide bibliographic (backward referencing) information from which similarity metrics could be deduced and used, as discussed in Section 5, to quantitatively define sub-domains within the larger SSL knowledge domain. Both databases also provide citation (forward referencing) information from which impact metrics could be deduced and, as discussed in Sections 7 and 9, used to quantitatively identify important emerging sub-domains within the larger SSL knowledge domain.

We note, however, that a drawback to using these two databases is the first's bias toward English language journals and the second's bias toward US patents. Hence, it must be kept in mind that our analyses, throughout this report, will tend to over-represent the strength of English-speaking nations (particularly the USA), and under-represent the strength of non-English-speaking nations (particularly China). We call China out specifically because very few of its technical journals are English language [only 25 of the 1,411 source journals in the 2001 China Scientific and Technical Papers and Citations database (Wu et al., 2004)] and very few of its inventions are patented in the USA [in

2001, only 621 USPTO patents were of Chinese origin (USPTO, 2004), while 99,278 SIPO patents were of Chinese origin (SIPO, 2004)]. There are a number of possible reasons for this (e.g., patent filing costs, home-market-targeting of businesses, weak intellectual property protection), but the net effect is that China's strength is especially under-represented in our analyses.

2.2 Knowledge-domain boundaries

For the knowledge domain that we would like the dataset to represent, we chose what might be loosely called 'electroluminescent materials and phenomena' or materials and phenomena through which electricity is converted into light. No distinction was made with respect to wavelength, to reflect the fact that similar concepts underlie the design and fabrication of devices that emit light at different wavelengths. Hence, though SSL is concerned explicitly with visible wavelengths, we consider that the knowledge domain foundational to SSL includes non-visible wavelengths. We do, however, make a distinction with respect to the process by which light is induced to emit from a material or structure. We include processes involving conversion of electricity into light, but exclude processes involving conversion of other forms of energy into light, such as photoluminescence, chemiluminescence, bioluminescence or sonoluminescence. The one instance in which this may exclude too much is in the area of phosphors, which are used in some SSL applications to down-convert light of a higher energy (shorter wavelength) into light of a lower energy (longer wavelength).

2.3 Search strategy

To maximise self-consistency in our analyses of the article and patent literature, for our strategy for constructing the dataset we sought a Boolean search string compatible (within minor formatting changes) with both the Thomson Scientific Science Citation Index and the US Patent and Trademark Office's database. The search string was developed iteratively to optimise around maximising inclusion of articles and patents lying inside, and around maximising exclusion of articles and patents that lie outside, the knowledge domain discussed above. In the language of information storage and retrieval, the search string was developed to simultaneously maximise recall and precision (Korfhage, 1997).

To do this, we relied on human technical judgment. To maximise recall, the search string was tested against a 'gold standard' set of articles and patents known to be from scientists[3] or institutions[4] at the forefront of SSL research and development and verified by our own judgment to be in the target knowledge domain. To maximise precision, we scanned the results of the search string to assess the percentage of articles and patents that by our judgment were indeed within the target knowledge domain.

The final search string is as follows:

$$\begin{aligned} S = S_1 < \text{or} > (S_{2a} < \text{or} > S_{2b} < \text{or} > S_{2c}) < \text{or} > (S_{3a} < \text{or} > S_{3b}) \\ < \text{or} > (S_{4a} < \text{or} > S_{4b} < \text{or} > S_{4c}) < \text{or} > S_5, \end{aligned}$$

where

S_1 = semic* < and > lum* < and > (copolym* < or > polym*
< or > organic < or > "II-VI" < or > "III-V" < or > "III-nitride"
< or > "gallium nitride" < or > "GaN")

S_{2a} = ("light" < order >< near / 1 > emi*) < and > ("active layer"
< or > "active region" < or > "clad layer" < or > "cladding layer"
< or > "well layer" < or > epit* < or > hetero* < or > "pn junction"
< or > "II-VI" < or > "III-V" < or > "III-nitride" < or > "gallium nitride"
< or > "GaN")

S_{2b} = (("light" < order >< near / 1 > emi*) < order >< near / 2 > layer*)
< not > ("plasma" < or > "noble gas" < or > flouresce*)

S_{2c} = ("light" < order >< near / 1 > emi*) < near / 2 > (copolym*
< or> polym* < or > organic < or > diode* < or > semi*)

S_{3a} = electrolum* < or > (electro < order >< near / 1 > (lum* < or > phos*))

S_{3b} = "EL" < order >< near / 1 > (dev* < or > display* < or > "element"
< or > "elements" < or > lamp* < or > panel* < or > phosphor*)

S_{4a} = "LEDs" < or > OLED* < or > ("a LED" < not > ("led to" < or > "led from"))

S_{4b} = ("an" < or > "HB" < or > "white" < or > "UV" < or > "blue"
< or > "green" < or > "amber" < or > "red") < order >< near / 1 > "LED"

S_{4c} = "LED" < order >< near / 1 > (array* < or > dev* < or > display*
< or > element*)

S_5 = "semiconductor light source" <or> (solid state light*)

We estimate that this search string has a recall and precision of roughly 70%–90%, which, though far from perfect, is within a reasonable range for such an exercise (Lundberg et al., 2006). However, we would inject a note of caution – because of the role of human technical judgment in our process and because of the wide range of lexical phrases associated with this knowledge domain, this search string no doubt contains considerable room for improvement.

Finally, records were further segmented into two knowledge sub-domains, one corresponding to inorganic electroluminescent materials and phenomena (the basis of inorganic light-emitting diodes or LEDs) and one corresponding to organic electroluminescent materials and phenomena (the basis of organic light-emitting diodes or OLEDs). For this segmentation, we used the sub-query:

O = oled* < or > polym* < or > monom* < or > "ligand" < or > hydroxy*
< or > macromol* < or > ("organic" < not > "metal-organic" < not >
"metalorganic")

Records in the dataset matching this sub-query (S <and> O) were categorised in the organic sub-domain; those not matching (S <not> O) were categorised in the inorganic sub-domain. For brevity, we will sometimes refer to the inorganic and organic sub-domains as the LED and OLED datasets, respectively, though we really mean the inorganic and organic electroluminescent materials and phenomena knowledge sub-domains, respectively.

3 Overall features of the dataset

The dataset consists of a total of 48,271 records. 35,851 records were articles, of which 27,972 were in the inorganic sub-domain and 7,879 in the organic sub-domain. 12,420 records were US patents, of which 10,500 were in the inorganic sub-domain and 1,920 were in the organic sub-domain.

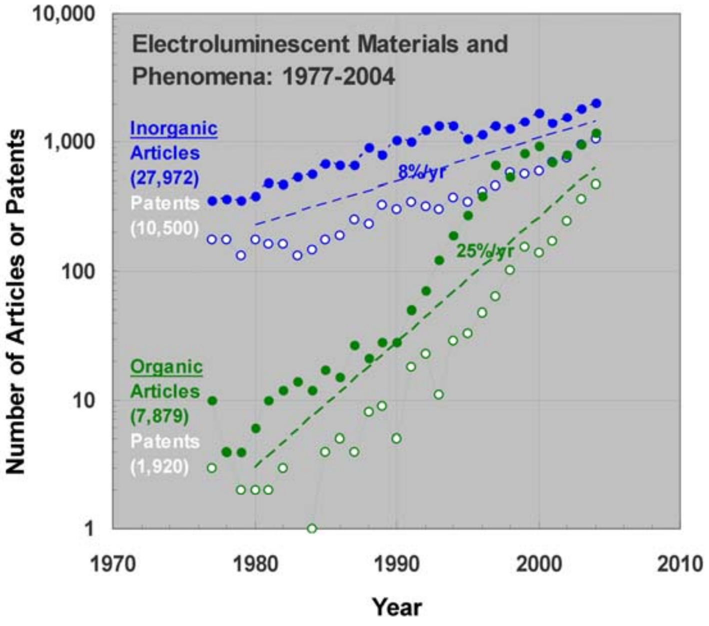
A yearly time series of the numbers of articles and patents in the two (inorganic and organic) knowledge sub-domains is shown in Figure 1. The inorganic knowledge sub-domain has historically been much larger than the organic knowledge sub-domain, due to important early applications in optical communications, storage and signalling. The organic knowledge sub-domain has more recently been growing much more rapidly (~25%/yr compared to ~8%/yr), however, due to emerging applications in flat-panel displays. If current trends continue, its combined rate of article and patent production will exceed that of the inorganic knowledge sub-domain within the coming decade.

We note for reference that growth rates for the journal article and patent literatures, over all knowledge domains, are 2.7% and 4.7%, respectively, for the time period of this study.[5] Thus, the inorganic knowledge sub-domain has been growing slightly faster, while the organic knowledge sub-domain has been growing much faster, than the historical average.

To give a visual feel for the very large size of the dataset, we show, in Figures 2 and 3, temporal scatter plots of the two knowledge sub-domains. Each data point represents a record in the dataset, plotted according to its publication or issue date and the number of times it has been cited (by articles published through 2004 and by patents issued through 2005). The coloured points represent articles, the white points represent patents. The citation scale starts at 1, so articles or patents that have not been cited are not plotted. Note that we are using a log scale for the citation scale, so the uppermost data points have orders of magnitude more citations than the lowermost data points. Also note that citations do not begin to accrue until after an article or patent has been published or issued, leading to the apparent trail-off in citations for articles and patents published or issued in the most recent five years.

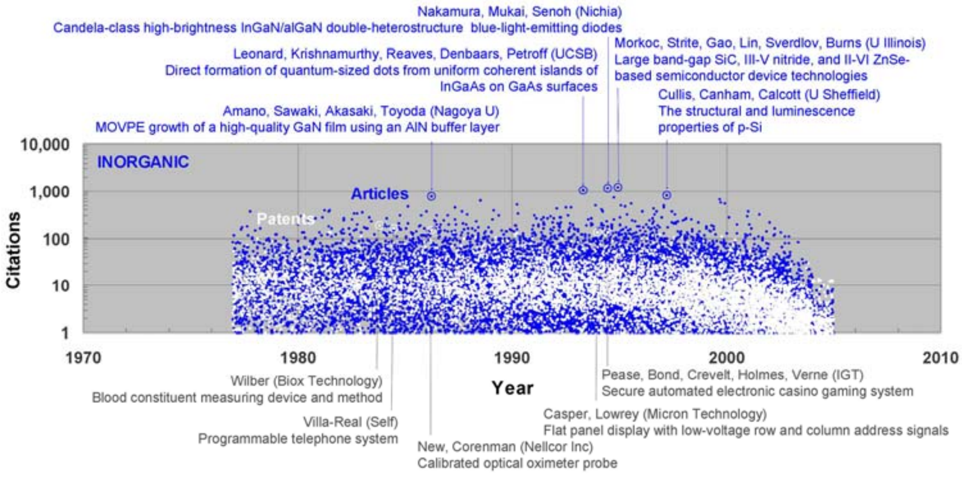
We have also listed the five most highly cited articles and patents in each of the sub-domains. These extremely influential articles have been cited on the order of 1,000–5,000 times and these extremely influential patents have been cited on the order of 100–500 times.

Figure 1 Numbers of articles and patents in the broad knowledge domain of electroluminescent materials and phenomena, as a function of year (see online version for colours)



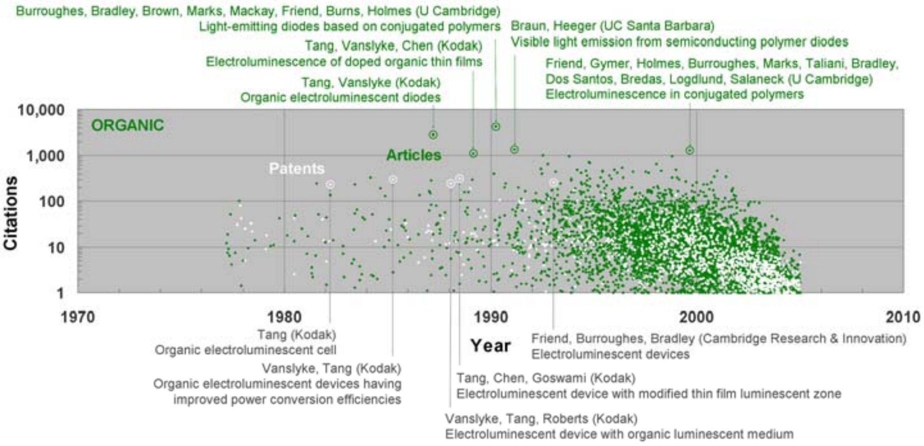
Notes: Blue points represent the inorganic, green points represent the organic, knowledge sub-domains. Filled points represent articles, white points represent patents.

Figure 2 Scatter plots of articles and patents in the inorganic (LED) knowledge sub-domain, arranged according to number of times cited and year published or issued (see online version for colours)



Notes: Blue points represent articles, white points represent patents. The most highly cited five articles and five patents are circled and labelled.

Figure 3 Scatter plots of articles and patents in the organic (OLED) knowledge sub-domain, arranged according to number of times cited and year published or issued (see online version for colours)



Notes: Green points represent articles, white points represent patents. The most highly cited five articles and five patents are circled and labelled.

- *Inorganic articles.* In the inorganic knowledge sub-domain, two of the most highly-cited articles involve GaN materials and devices: the 1986 article by Amano et al. (1986) on a method for growing GaN on lattice-mismatched sapphire and the 1994 article by Nakamura et al. (1994) on high-brightness blue LEDs. Both of these breakthroughs were unexpected, and together they were a crucial foundation for high-brightness visible and white light emitters. A third article, the review of wide-gap semiconductors by Morkoc et al. (1994) is an indication of the importance of the GaN materials system for SSL. The fourth and fifth papers described the synthesis and properties of two kinds of nanostructures whose properties are of interest to electroluminescent devices: the article by Leonard et al. (1993) on self-assembled growth of quantum dots and the article by Cullis et al. (1997) reviewing the structural and luminescence properties of nano-porous silicon.

- *Inorganic patents.* For the inorganic knowledge sub-domain, the five most highly cited patents were for various applications of light-emitting devices (rather than for the devices themselves). Two of these were for uses in healthcare and medicine: Wilber's (1983) patent and New and Corenman's (1986) patent, in which the spectral transmittance of LEDs is used as a measure of blood constituents, including oxygen content. Two others were for uses as information indicators for consumer products: Villa-Real's (1984) patent for programmable telephones and Pease et al. (1994) patent for gaming systems. A final highly cited patent was Casper and Lowrey's (1993) patent on flat-panel (including electroluminescent) displays.

- *Organic articles.* For the organic knowledge sub-domain, two of the most highly-cited articles involved the discovery at Eastman Kodak Company of organic electroluminescence in small molecules: the article by Tang and VanSlyke (1987), and the article by Tang et al. (1989). The other three most highly-cited articles involved the subsequent discovery at the University of Cambridge and at the University of California Santa Barbara of polymer-based organic electroluminescence: the article by Burroughes et al. (1990), the review article by the same group (Friend et al., 1999) and the article by Braun and Heeger (1991).

- *Organic patents.* For the organic knowledge sub-domain, the five most highly cited patents, in contrast to the case for the inorganic knowledge sub-domain, were for electroluminescent devices themselves, as opposed to applications of these devices. Four of these (Tang, 1982; VanSlyke and Tang, 1985; Tang et al., 1988; VanSlyke et al., 1988) were from the group at Eastman Kodak Company, and were for electroluminescent devices based on small molecules. The fifth (Friend et al., 1993) was from the University of Cambridge group, for an electroluminescent device based on polymers.

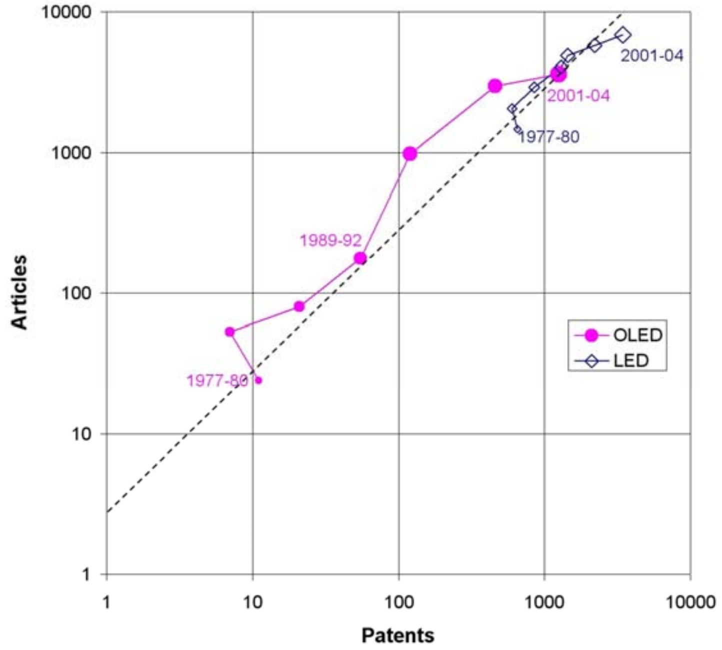
Note that the two datasets we are juxtaposing, articles and patents, are both indicators of the strength of the knowledge base foundational to SSL. However, they are different indicators.

On the one hand, articles tend to have greater scientific content, with a longer-term and farther-reaching impact on technologies and applications. Articles also have a more streamlined review process, and hence can respond more quickly to the latest technical breakthroughs. And articles often contain scientific insights that are themselves not patentable, though they may strengthen areas of patent activity. On the other hand, patents tend to have greater technological content, with a shorter-term and more immediate potential impact on applications. In cases where the commercial value is high, patents may even substitute for articles as the primary vehicle for codifying new knowledge.

Hence, in this study, we analyse both literatures. To do this in a manner that places them both on equal footing, we show the LED and OLED datasets as trajectories in the ‘article-patent’ space shown in Figure 4. Each data point represents simultaneously the number of articles and patents in a sequence of four-year time periods. The larger data points are for the more recent periods and points from successive periods are connected to show changes over the seven data points that span the period 1977 to 2004.

The dashed line indicates the ratio between numbers of articles to numbers of patents (2.9), for the entire dataset. This ratio is substantially lower than the ratio between the total numbers of articles in the Science Citation Index to the total number of US patents in the USPTO database (6.7 in 1980, decreasing to 4.5 in recent years). This is an indication that the knowledge domain we are considering here has relatively more relevance to commercial application than the average. The slight right-leaning curvature in the more recent years of both trajectories indicates a somewhat increasing relative emphasis on patents. This is not unique to the knowledge domain considered here, but is characteristic of the primary databases themselves. However, note that although relative growth is higher for patents than for articles, in terms of absolute numbers, articles are still growing at a slightly higher rate.

Figure 4 Numbers of articles and patents in sequential four-year time periods (1977–1980, 1981–1984, …, 1997–2000, 2001–2004) for the inorganic (LED) and organic (OLED) knowledge sub-domains (see online version for colours)



Finally, note that the organic knowledge sub-domain is growing at a much higher percentage rate than the inorganic knowledge sub-domain, as evidenced by the greater spacing between points in Figure 4. The implication is that, if these relative growth rates continue, the OLED dataset will catch up with the LED dataset within the coming decade in both patents (horizontal distances between points) and in articles (vertical distances between points).

4 International trends

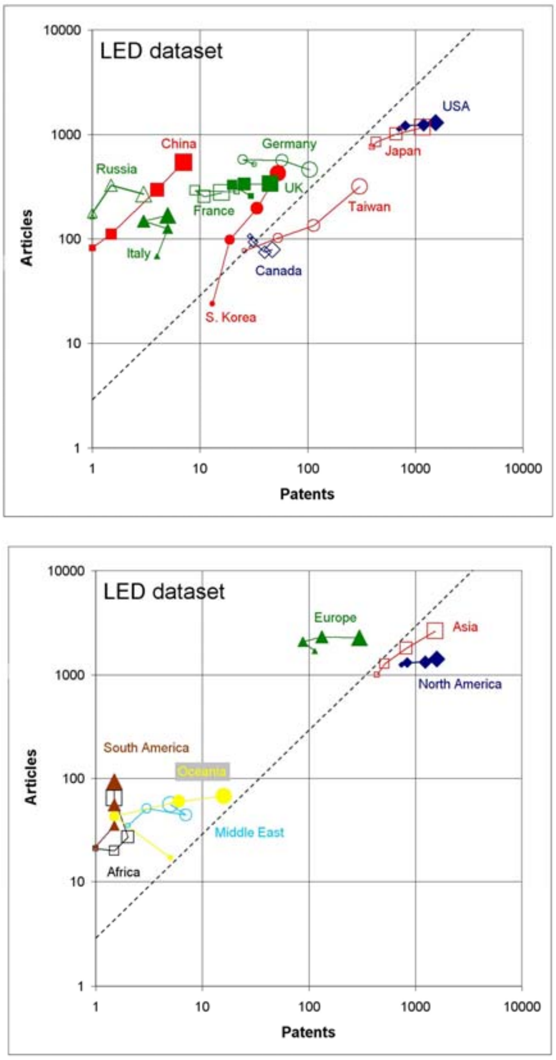
Thus far, we have associated each article and patent in the dataset with either the inorganic or organic knowledge sub-domains, so that trends in these separate sub-domains can be separately quantified and compared. We can also associate each article and patent in the dataset with one or more national origins, determined by the origins of its associated authors, inventors or institutions. There are various methods of making this association; here, we use the simplest method, in which an article or patent is assigned a unique national origin according to first author or first inventor.

Having assigned an article or patent a unique national origin, we then use United Nations conventions (UN, 2006) to assign articles or patents a unique macro-geographical (continental) origin: Africa, Latin America and the Caribbean, Northern America, Asia, Europe and Oceania. For many purposes, other roll-ups based on geo-political or socio-cultural considerations would also be interesting (e.g., by affinity to the major language families of the world, such as English, Spanish and

Chinese). However, the United Nations convention based on geography does provide a measure of normalisation in that Northern America, Asia and Europe are reasonably close in terms of some measures of economic activity.

Using these associations, we can now chart trajectories in 'article-patent' space, just as in Figure 4, but of the contributions to the LED and OLED datasets from various nations and continents. To minimise the overlap in the trajectories, we have truncated them and show only the most recent four time periods.

Figure 5 Numbers of articles and patents in sequential four-year time periods by nation (top) and continent (bottom) for the inorganic knowledge sub-domain (see online version for colours)



Notes: Larger symbols in each series represent more recent four-year time periods. Asian nations are coloured red, North American nations dark blue and European nations green.

The trajectories for the LED dataset are shown in Figure 5: the top eleven nations on the top and the seven continents on the bottom. The diagonal dashed lines represent the average ratio of articles to patents for the LED dataset. A data point that lies along the dashed line represents a ratio of articles to patents equal to the LED dataset average. A data point that lies below (or above) the dashed line represents a ratio of articles to patents lower (or higher) than the LED dataset average.

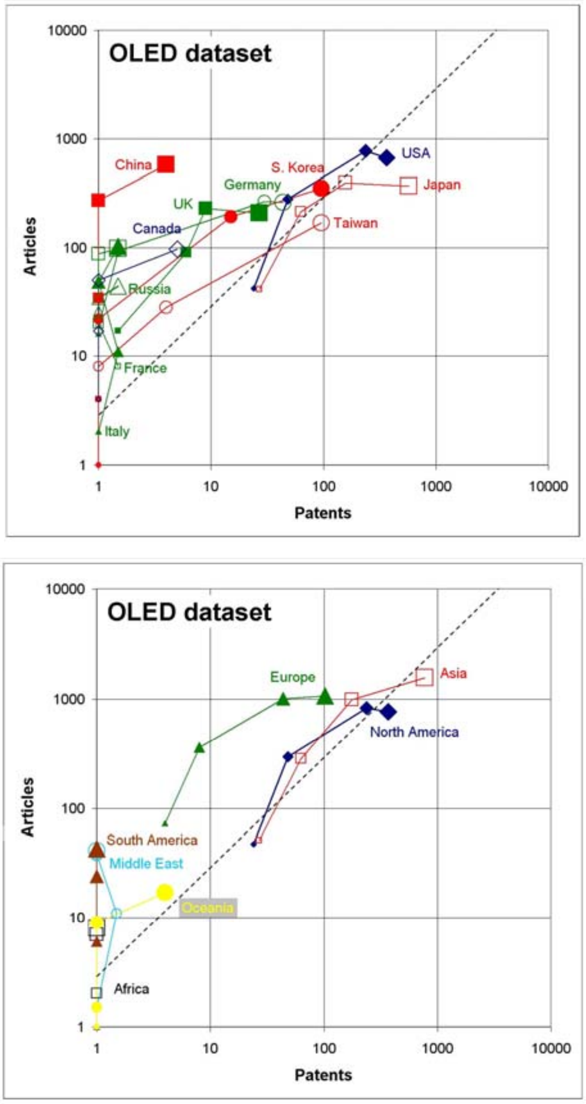
- *LED dataset: nation trajectories.* From the top panel of Figure 5, one can see that the USA and Japan are the two dominant nations, with the USA leading slightly in the most recent four-year time period. One can also see that with its faster trajectory (distance between data points), Japan can be expected to approach or surpass the USA in the next four-year time period. From the rightward curvature of their trajectories, both show an increasing recent emphasis on patents over articles. Interestingly, almost all of the other nations (except Taiwan and Canada) show an opposite emphasis on articles over patents.
- *LED dataset: continent trajectories.* From the bottom panel of Figure 5, one can see that Asia is the dominant continent. It has significantly more articles than North America with a comparable number of patents and has significantly more patents than Europe with a comparable number of articles. Moreover, its trajectory (distance between data points) is such that its dominance is likely to continue and grow. From the rightward curvature of the North America and Europe trajectories, both show an increasing emphasis on patents over articles. However, in absolute numbers, Europe still has a significant emphasis on articles over patents, while North America has a significant emphasis on patents over articles.

The trajectories for the OLED dataset are shown in Figure 6: the top eleven nations on the top and the seven continents on the bottom. The diagonal dashed lines represent the average ratio of articles to patents for the OLED dataset. A data point that lies along the dashed line represents a ratio of articles to patents equal to the OLED dataset average. A data point that lies below (or above) the dashed line represents a ratio of articles to patents lower (or higher) than the OLED dataset average. Note that, because of the recent ramp-up in the field, all countries are starting from a trivial base in terms of patents.

- *OLED dataset: nation trajectories.* From the top panel of Figure 6, one can see that, just as for the LED dataset, the USA and Japan are the two dominant nations. The USA leads slightly in articles, but Japan leads slightly in patents. Also, their relative trajectories are such that Japan's patent dominance is likely to continue, while the USA's article dominance is likely to wane. Just as with the LED dataset, the rightward curvature of their trajectories indicates an increasing recent emphasis on patents over articles, offset by a greater emphasis on articles over patents in almost all of the other nations (except Taiwan).
- *OLED dataset: continent trajectories.* From the bottom panel of Figure 6, one can see that, even more so than for the LED dataset, Asia is the dominant continent. It has significantly more articles and patents than either North America or Europe. Moreover, its trajectory (distance between data points) is such that its dominance is likely to continue and grow. And, just as for the LED dataset, from the rightward curvature of the North America and Europe trajectories, both show an increasing emphasis on patents over articles. However, in absolute numbers, Europe still has a

significant emphasis on articles over patents, while North America has a significant emphasis on patents over articles.

Figure 6 Numbers of articles and patents in sequential four-year time periods by nation (top) and continent (bottom) for the organic knowledge sub-domain (see online version for colours)



Notes: Larger symbols in each series represent more recent four-year time periods. Asian nations are coloured red, North American nations dark blue and European nations green.

To summarise, the USA is the dominant nation in both the LED and OLED datasets, but Japan is a very close second and is on a faster trajectory. While Canada is a modest contributor within North America, Taiwan, South Korea and China are significant contributors within Asia. As a result, Asia is the dominant continent in both the LED and OLED datasets, and is also on a faster trajectory.

In fact, we should remind ourselves at this point that, for the reasons mentioned in Section 2, our dataset over-represents English-language and US contributions to the overall knowledge domain. Hence, Asia is even stronger than appears. For example, from Figures 5 and 6 one sees that China has a significantly greater emphasis on articles than on patents in this dataset. Insofar as strong science is a leading indicator for strong technology, one can anticipate that China is in a good position to generate more US patents in the future. However, China's domestic market is growing very quickly, and before too long may be the largest market in the world for SSL. From that perspective, Chinese companies and inventors have less incentive to apply for US patents for their intellectual property – it may be sufficient for them to be protected within China. Also, China lacks a history of a strong intellectual property protection and companies therefore have less incentive to patent locally or overseas.

Table 1A Nations with the most articles during the 2000–2004 time period

Rank	*Country*	*# articles 2000–2004*	*Avg # cites 2000–2004*	*# articles all years (rank)*
1	USA	2460	8.65	8747 (1)
2	Japan	1974	4.50	6050 (2)
3	China	1362	2.96	1970 (5)
4	Germany	957	6.93	3418 (3)
5	South Korea	937	3.23	1311 (7)
6	UK	712	6.24	2275 (4)
7	Taiwan	541	4.67	866 (9)
8	France	468	4.15	1781 (6)
9	Italy	331	3.65	744 (10)
10	Russia/USSR	314	2.12	1082 (8)
11	Canada	220	5.40	632 (11)
12	India	188	2.06	530 (12)
13	Poland	179	2.62	507 (14)
14	Singapore	173	3.69	268 (21)
15	Sweden	158	5.20	523 (13)
16	Netherlands	155	5.63	411 (15)
17	Spain	144	3.00	326 (17)
18	Brazil	144	3.33	270 (20)
19	Switzerland	136	6.79	379 (16)
20	Belgium	114	9.93	292 (18)

Note: This is a five-year time period, rather than the four-year time periods used for the trajectories in Figures 4 through 6.

Table 1B Nations with the most patents during the 2000–2004 time period

Rank	*Country*	*# patents 2000–2004*	*avg # cites 2000–2004*	*# patents all years (rank)*
1	USA	2280	3.22	6428 (1)
2	Japan	1982	2.38	3950 (2)
3	Taiwan	437	1.60	601 (3)
4	Germany	174	2.38	379 (4)
5	South Korea	161	0.92	232 (6)
6	Netherlands	81	1.22	89 (9)
7	UK	79	1.44	242 (5)
8	Canada	59	1.80	197 (7)
9	Sweden	20	2.30	64 (10)
10	Switzerland	19	0.95	62 (11)
11	France	18	0.89	107 (8)
12	Australia	15	2.73	31 (12)
13	Belgium	15	4.07	26 (15)
14	Hong Kong	13	2.08	25 (16)
15	China	12	2.17	16 (18)
16	Finland	9	5.67	29 (13)
17	Israel	7	2.57	19 (17)
18	Italy	6	0.67	29 (14)
19	Austria	6	4.83	10 (20)
20	Ireland	6	2.00	9 (21)

Note: This is a five-year time period, rather than the four-year time periods used for the trajectories in Figures 4 through 6.

Similar conclusions can be drawn from Table 1, which shows a more detailed view of the article and patent contributions of the various nations to the overall dataset.

The US and Japan dominate the dataset for both articles and patents. But because of their relatively stronger emphasis on patents than articles, they ‘completely’ dominate in patents, with 78.5% of all patents in the most recent five years, and only somewhat dominate in articles, with 35% of all articles in the same time period. With the exception of Taiwan (which ranks 3rd in patents but 7th in articles), most other countries have the reverse emphasis of articles over patents. We have already mentioned China, which ranks 3rd in articles but 15th in patents. India is of interest in that it ranks 12th in articles, but does not have a single US patent in our dataset. We do not explore the potential reasons for these differences here, but merely note that they exist.

Also, the trajectories are such that dominance of Asia will only become greater in the future. For instance, China ranks 3rd in scientific papers during the most recent five years, but only 5th over the full time period. Thus, China has stepped up its investment in the science foundation of this knowledge domain as compared to the rest of the world. In contrast, the UK shows the opposite trend, ranking 4th over the full time period, but only 6th during the most recent five years. A separate analysis (not shown) indicates that Asia became stronger than Europe in the early 1990s, and became stronger than North

America in the early 2000's. North America is still stronger than Europe, though, with no sign of an impending cross-over.

5 Mapping and clustering the dataset

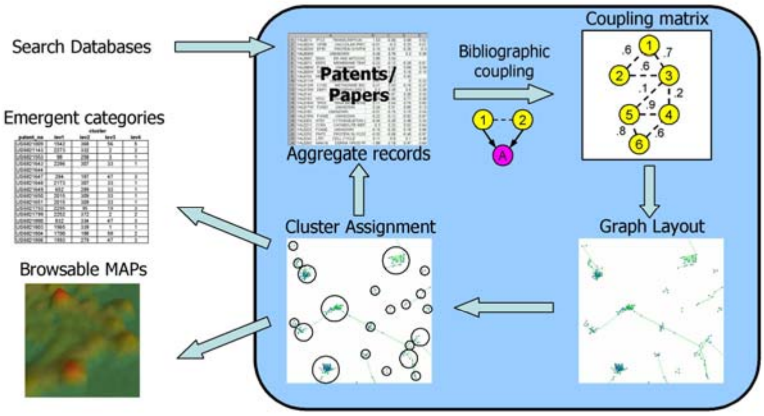
Thus far, we have associated each article and patent in the dataset with either the inorganic or organic knowledge sub-domain and with a unique national and continent origin. In terms of the technical disciplines that are associated with the knowledge domain, these are still very coarse associations. It would be useful to drill deeper into the dataset, and cluster articles and patents into knowledge 'micro'-domains that could then be separately quantified and analysed.

There are a number of methods for doing such clustering. The majority of surveys of particular fields cluster articles or patents into pre-existing categories. Journals and journal types are a handy categorisation for articles and patent classes find similar usage for patents. However, many journals, especially large or multidisciplinary journals, cover a variety of topics that can bridge multiple categories and conversely several journals can cover similar topics. Likewise, the US patent classification system is a handy categorisation schema for patents. However, the system is inconsistent and time-varying: some classifications are broad and others narrow, classifications change definitions over time, patents are reassigned at varying intervals and most patents are multiply classified. In fact, a cursory review indicates that such inconsistencies and time variations were present in our SSL dataset.

Given the ambiguities associated with such pre-existing categories, we chose instead to use recently developed bibliographic techniques to generate maps of the dataset from which clusters emerge automatically from the properties of the dataset. These clusters represent a 'natural' way of understanding the organisation and ongoing evolution of the knowledge domain. They also provide a means for understanding at one finer level of detail relative to national strengths in the various knowledge micro-domains.

The general five-step process we used to cluster both articles and patents is illustrated in Figure 7.

Figure 7 Schematic of the article and patent mapping and clustering process (see online version for colours)



1. First, bibliographic coupling counts (N_{ij}) were calculated for each pair of records. Bibliographic coupling occurs when two articles or patents cite common references. So-called cosine coefficients were then calculated from the coupling counts for each pair of records as $\mathrm{COS} = N_{ij} / sqrt(N_i \cdot N_j)$, where N_{ij} is the number of common references in records i and j, and N_i and N_j are the number of references in records i and j, respectively.
2. A graph is constructed in which each article or patent is a node, and the cosine coefficients are edge weights between the nodes. The graph is then 'laid out' into a two-dimensional map by assigning x, y positions to each node in such a way as to minimise the distance between pairs of nodes with high cosine coefficients. The mapping algorithm is the VxOrd algorithm (Davidson et al., 2001) contained in the VxInsight® tool (Boyack et al., 2002) for knowledge domain visualisation.
3. In the resulting map, articles or patents that are near each other are then clustered. There are a number of methods for doing such clustering. Here, we use a modified average-link clustering algorithm that assigns each node to a cluster based on edges and distances between nodes (Klavans and Boyack, 2006).
4. At this point we have a level-1 clustering, which constitutes a first-order emergent categorisation based on how authors and inventors (and examiners) perceive an article or patent with regard to other articles or patents. However, a typical level-1 cluster contains on average ten records, so there are still too many clusters to enable a high-level view of the dataset. To overcome this, we iteratively cluster the clusters. This involves aggregating together all records in a cluster, by assigning all references in a given record to its cluster. Then, we go through steps 1 through 3 again, using the level-1 clusters as the nodes. After calculating cosine coefficients for pairs of clusters, laying out the graph, and clustering the clusters, we then have a level-2 clustering.
5. Even higher-level clusterings can now be created by aggregating references and iterating steps 1 through 3, with the number of clusters being reduced by roughly an order of magnitude per iteration. In practice, we stopped after aggregating into roughly 50 level-3 clusters, and did a final manual aggregation into level-4 clusters, using our own technical understanding of the knowledge domain. The resulting level-4 clusters we call 'superclusters', and they represent a high-level emergent categorisation scheme for the knowledge sub-domains associated with the dataset.

The resulting numbers of clusters and records remaining in the maps at the various levels of aggregation are given in Table 2. Note that at each level of aggregation some records are 'lost' (e.g., 3,114 articles were lost in the aggregation leading to level-1 clusters). These articles could not be aggregated into any cluster of articles, usually because they have no common backward references with other articles in the dataset. At each level of aggregation, fewer articles are lost. Also note that fractionally more patents are lost than articles. This is because the average article has 25 references while the average patent has 8 references. Thus patents have less information available with which to calculate bibliographic coupling. This leads to a dropping of a higher fraction of patents at each level of aggregation, and also leads to more ambiguity for patent than for article clustering.

Table 2 Numbers of clusters at each level of aggregation for the article and patent maps

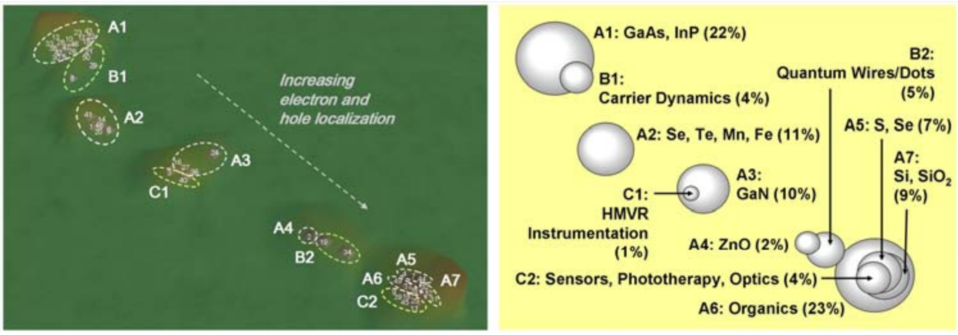
	# clusters	*# articles*	*# clusters*	*# patents*
Initial		35,890		12,621
Level-1	4,428	32,776	2,336	10,921
Level-2	491	32,561	388	10,512
Level-3	50	32,553	62	10,505
Level-4	11	32,553	10	10,505

6 Article superclusters

Two maps of the resulting superclusters of articles in the dataset are shown in Figure 8.

The left map is a topographical level-3 map taken directly from VxInsight, in which height represents density of articles. These articles have been aggregated into 50 level-3 clusters, and these level-3 clusters have in turn been aggregated manually into the 11 level-4 dashed-oval-outlined superclusters labelled A1-A7, B1-B2 and C1-C2. Note that the lateral extent of a supercluster on this map is not an indication of its size (number of articles) but an indication of the bibliographic-coupling 'diversity' of its aggregated level-3 clusters.

Figure 8 Level-3 and level-4 cluster maps of all articles in the dataset (see online version for colours)



Notes: (Left) VxInsight topographic map showing the level-3 clusters and their manual aggregation (dashed white lines) into level-4 'superclusters'.
(Right) a stylised map in which each supercluster's area is proportional to the number of articles it contains.

The right map is a stylised level-4 map, in which the superclusters' centre positions have been preserved, but their areas have been scaled to reflect the number of articles they contain. Though this stylised map no longer has information regarding bibliographic-coupling diversity within the superclusters, it retains information regarding bibliographic-coupling diversity between the superclusters. Also, because the stylised map is inherently two-dimensional (despite the three-dimensional sphere-like shading of the clusters), it is easier to interpret and will be used exclusively in the remainder of this article.

The superclusters fall into three groups: the first distinguished by materials system, the second by structure or phenomenon and the third by applications.

A *Materials-oriented superclusters.* Most (84%) of the articles aggregated into the seven materials-oriented superclusters labelled A1–A7. The largest single supercluster, A6 (23%), is that associated with organic materials. There are also two relatively large superclusters associated with the III-V compound semiconductors: A1 (22%) is associated with GaAs-based and InP-based materials and A3 (10%) is associated with GaN-based materials. Then there are three clusters associated with the II-VI compound semiconductors: A2 (11%) is associated with the Te and Se compounds, A4 (2%) is associated with ZnO materials and A5 (7%) is associated with the S compounds. Finally, a medium-sized cluster, A7 (9%), is associated with Si/SiO_2 photonics.

B *Structure- or phenomenon-oriented superclusters.* A much smaller (9%) fraction of the articles aggregated into the two structure – or phenomenon-oriented superclusters labelled B1–B2. B1 (4%) is associated with carrier and exciton dynamics and recombination and B2 (5%) is associated with low-dimensional structures, such as quantum wires and dots.

C *Applications-oriented superclusters.* A still smaller (5%) fraction of the articles aggregated into the two applications-oriented superclusters labelled C1–C2. C1 (1%) is associated with instrumentation for human motor and visual response experiments and C2 (4%) is associated with sensors, photo-therapies and optics.

Assigning a single high-level interpretation to the tens of millions of bibliographic coupling comparisons that were made to compute these two maps is of course non-trivial. However, a reasonable interpretation of the positioning of the materials-oriented superclusters appears to be based on the degree to which electrons and holes are spatially localised in the various materials or phenomena of interest. The materials associated with the extreme-upper-left superclusters are characterised by electrons and holes with very long mean-free paths, such as in the purest and crystalographically most ordered semiconductors in supercluster A1 (GaAs and InP). The materials associated with the extreme lower-right superclusters are characterised by electrons and holes that are localised in space, such as the molecule-scale electronic excitations associated with supercluster A6 (organics) or the localised impurities (e.g., rare earths) in the materials associated with superclusters A5 (S, Se) or A7 (Si, SiO_2). In between these extremes, from upper left to lower right, are the less-pure or crystalographically less ordered semiconductors associated with superclusters A2 (Se, Te, Mn, Fe) and A3 (GaN) and the often-nanostructured semiconductors associated with supercluster A4 (ZnO).

The positioning of the structure- or phenomenon-based superclusters is consistent with this interpretation. The phenomena associated with supercluster B1 (carrier dynamics) are characteristic of electrons and holes in motion over sometimes long distances and in large part made use of the materials associated with supercluster A1 (GaAs, InP). The phenomena associated with supercluster B2 (quantum wires/dots) are characteristic of electrons and holes that are confined in space and for the most part made use of nanostructured GaAs and InP materials.

The positioning of the applications-oriented superclusters is also consistent with this interpretation. The applications associated with supercluster C1 (instrumentation for human motor and visual response experiments) make extensive use of visible-wavelength light-emitting devices based on materials associated with supercluster A3 (GaN). The applications associated with supercluster C2 (sensors, photo-therapy, optics) are closely associated with the study of materials in supercluster A6 (organics) and related bio-organic materials.

We are now in a position to analyse the relative contributions various nations have made not only to the overall knowledge domain, but to the knowledge sub-domains that the various superclusters represent. This analysis is summarised in Tables 3(a), 3(b) and 3(c).

- Each column in Tables 3(a) and 3(b) lists, for each supercluster, the percentages of articles contributed by nine representative nations and the 'rest of the world'. The first eight of the nine nations chosen were the top eight article-producing nations from Table 1. For the ninth nation, we chose to retain Canada (#11 in Table 1) to give North America a more balanced presence in this analysis. Table 3(a) lists these percentages for the entire 28 years (1977–2004) and Table 3(b) lists them for the most recent five years (2000–2004). The rightmost columns in both Tables list the percentages of articles contributed by the various nations, summed over all superclusters. The vertical sum of percentages within each column is 100%.
- Table 3(c) lists, again for each supercluster, the ratio between the two percentages in Tables 3(b) and 3(a). This ratio is an indication of the trend in time of a nation's contribution to the articles in the various superclusters. Ratios greater than, equal to or less than, unity indicate that a nation's relative contribution to the articles in a supercluster is growing, staying the same, or shrinking.

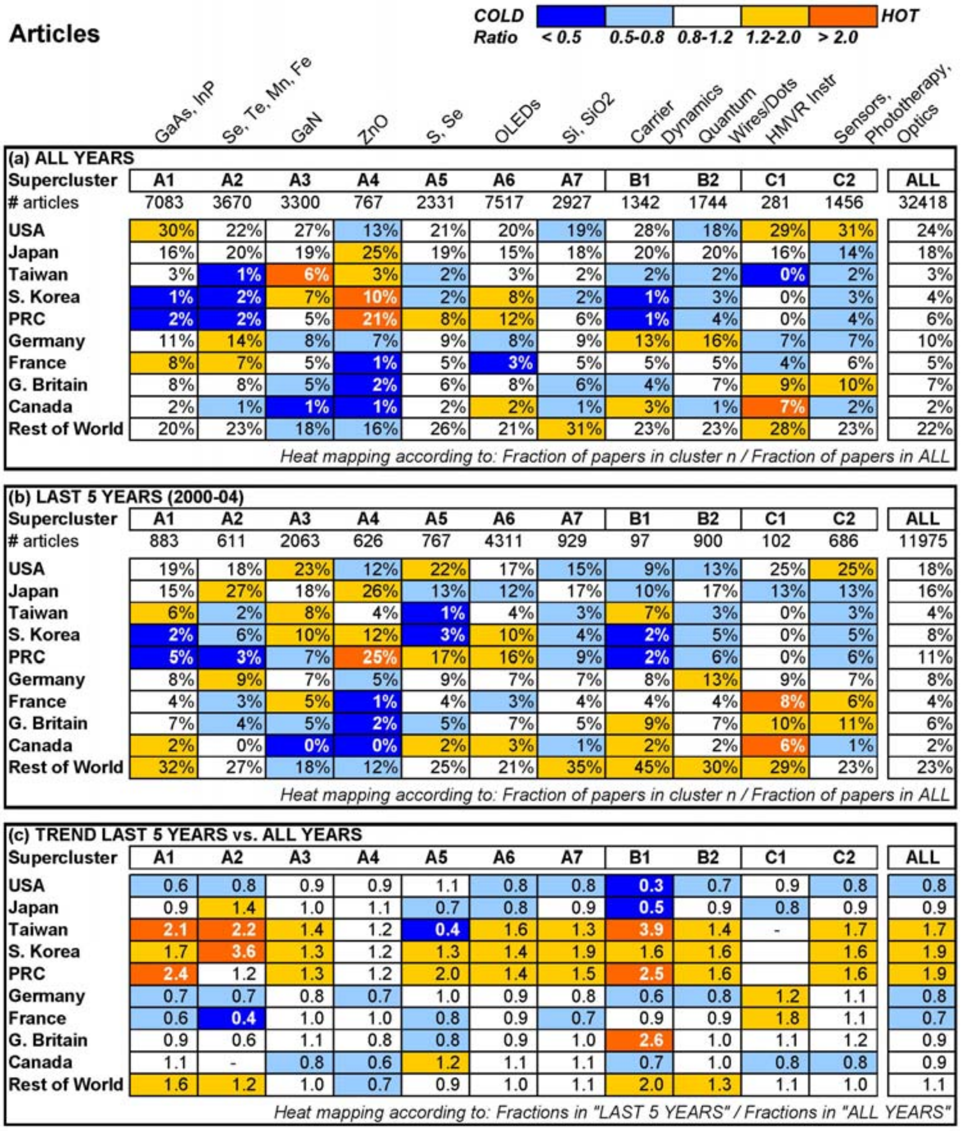
From Tables 3(a) and 3(b), one can see that the USA and Japan are the two dominant nations. The middle columns indicate that they are the only two nations that have contributed more than 9% of the articles in each of the superclusters. The rightmost columns indicate that together they contributed more than one-third of all the articles (summed over all superclusters) in both the full 28-year dataset and in the most recent five years of the dataset.

From the rightmost column of Table 3(c), however, one can also see that this overall dominance is shifting. The US's trend ratio is 0.8, indicating that its percentage contribution, summed over all superclusters, is shrinking (from 24% to 18%). Japan's trend ratio is 0.9, indicating that its percentage contribution is also shrinking (from 18% to 16%), though less quickly. In contrast, Taiwan's trend ratio is 1.7 and South Korea's and China's trend ratios are 1.9, indicating that their percentage contributions are growing rapidly (Taiwan's from 3% to 4%, South Korea's from 4% to 8% and China's from 6% to 11%).

It is also interesting to compare a nation's percentage contribution to particular superclusters to its percentage contribution to the overall dataset. Ratios between these percentages that are greater than, equal to or less than, unity, indicate that a nation is placing relatively more, equal or less emphasis on the knowledge sub-domain associated with a particular supercluster than on the overall knowledge domain associated with all the superclusters. To enable these ratios to be visualised quickly, we have 'heat mapped' the colours of the various cells in Tables 3(a) and 3(b) so that ratios greater than unity are light orange (somewhat hot) and orange (very hot), while ratios less than unity are light blue (somewhat cold) and dark blue (very cold). For example, the upper left cell in

Table 3(a) shows that the USA contributed 30% of the articles in the A1 supercluster, while the upper right cell of that same row shows that the USA contributed only 24% of the articles in the entire dataset. Thus, the USA contributed 1.2 times more than expected to supercluster A1's articles and thus that cell is heat-mapped to light orange (somewhat hot).

Table 3 Absolute and percentage contributions of representative nations to the various article superclusters, (a) contributions to the entire 28-year dataset (b) contributions to the most recent five years of the dataset (c) the ratio between the two (see online version for colours)



Articles

COLD	🟦	🟦	⬜	🟨	🟧	HOT
Ratio	< 0.5	0.5-0.8	0.8-1.2	1.2-2.0	> 2.0	

(a) ALL YEARS

	GaAs, InP	Se, Te, Mn, Fe	GaN	ZnO	S, Se	OLEDs	Si, SiO2	Carrier Dynamics	Quantum Wires/Dots	HMVR Instr	Sensors, Phototherapy, Optics	
Supercluster	A1	A2	A3	A4	A5	A6	A7	B1	B2	C1	C2	ALL
# articles	7083	3670	3300	767	2331	7517	2927	1342	1744	281	1456	32418
USA	30%	22%	27%	13%	21%	20%	19%	28%	18%	29%	31%	24%
Japan	16%	20%	19%	25%	19%	15%	18%	20%	20%	16%	14%	18%
Taiwan	3%	1%	6%	3%	2%	3%	2%	2%	2%	0%	2%	3%
S. Korea	1%	2%	7%	10%	2%	8%	2%	1%	3%	0%	3%	4%
PRC	2%	2%	5%	21%	8%	12%	6%	1%	4%	0%	4%	6%
Germany	11%	14%	8%	7%	9%	8%	9%	13%	16%	7%	7%	10%
France	8%	7%	5%	1%	5%	3%	5%	5%	5%	4%	6%	5%
G. Britain	8%	8%	5%	2%	6%	8%	6%	4%	7%	9%	10%	7%
Canada	2%	1%	1%	1%	2%	2%	1%	3%	1%	7%	2%	2%
Rest of World	20%	23%	18%	16%	26%	21%	31%	23%	23%	28%	23%	22%

Heat mapping according to: Fraction of papers in cluster n / Fraction of papers in ALL

(b) LAST 5 YEARS (2000-04)

Supercluster	A1	A2	A3	A4	A5	A6	A7	B1	B2	C1	C2	ALL
# articles	883	611	2063	626	767	4311	929	97	900	102	686	11975
USA	19%	18%	23%	12%	22%	17%	15%	9%	13%	25%	25%	18%
Japan	15%	27%	18%	26%	13%	12%	17%	10%	17%	13%	13%	16%
Taiwan	6%	2%	8%	4%	1%	4%	3%	7%	3%	0%	3%	4%
S. Korea	2%	6%	10%	12%	3%	10%	4%	2%	5%	0%	5%	8%
PRC	5%	3%	7%	25%	17%	16%	9%	2%	6%	0%	6%	11%
Germany	8%	9%	7%	5%	9%	7%	7%	8%	13%	9%	7%	8%
France	4%	3%	5%	1%	4%	3%	4%	4%	4%	8%	6%	4%
G. Britain	7%	4%	5%	2%	5%	7%	5%	9%	7%	10%	11%	6%
Canada	2%	0%	0%	0%	2%	3%	1%	2%	2%	6%	1%	2%
Rest of World	32%	27%	18%	12%	25%	21%	35%	45%	30%	29%	23%	23%

Heat mapping according to: Fraction of papers in cluster n / Fraction of papers in ALL

(c) TREND LAST 5 YEARS vs. ALL YEARS

Supercluster	A1	A2	A3	A4	A5	A6	A7	B1	B2	C1	C2	ALL
USA	0.6	0.8	0.9	0.9	1.1	0.8	0.8	0.3	0.7	0.9	0.8	0.8
Japan	0.9	1.4	1.0	1.1	0.7	0.8	0.9	0.5	0.9	0.8	0.9	0.9
Taiwan	2.1	2.2	1.4	1.2	0.4	1.6	1.3	3.9	1.4	-	1.7	1.7
S. Korea	1.7	3.6	1.3	1.2	1.3	1.4	1.9	1.6	1.6		1.6	1.9
PRC	2.4	1.2	1.3	1.2	2.0	1.4	1.5	2.5	1.6		1.6	1.9
Germany	0.7	0.7	0.8	0.7	1.0	0.9	0.8	0.6	0.8	1.2	1.1	0.8
France	0.6	0.4	1.0	1.0	0.8	0.9	0.7	0.9	0.9	1.8	1.1	0.7
G. Britain	0.9	0.6	1.1	0.8	0.8	0.9	1.0	2.6	1.0	1.1	1.2	0.9
Canada	1.1	-	0.8	0.6	1.2	1.1	1.1	0.7	1.0	0.8	0.8	0.9
Rest of World	1.6	1.2	1.0	0.7	0.9	1.0	1.1	2.0	1.3	1.1	1.0	1.1

Heat mapping according to: Fractions in "LAST 5 YEARS" / Fractions in "ALL YEARS"

From the Tables 3(a) and 3(b) heat maps, one can see that the USA has had both a historical and a recent emphasis on supercluster C2 (sensors, phototherapy, optics). This is consistent with ongoing interest in the USA in the use of LEDs and lasers as sensors and sources in medical, environmental and communications applications.

From the Tables 3(a) and 3(b) heat maps, one can also see that the USA's historical emphasis on superclusters A1 (GaAs, InP) and C1 (HMVR instrumentation) has given way to a recent emphasis on superclusters A3 (GaN) and A5 (S, Se). A waning of interest in supercluster A1 is consistent with the maturing of our understanding of GaAs- and InP-based materials and their now widespread use in commercial applications ranging from cellular telephony to optical-fibre communications. A waning of interest in supercluster C1 is consistent with the maturing of light-emitting-diode technology for signalling (direct view) applications such as in human motor and visual response experiments. A waxing of interest in supercluster A3 is consistent with the emergence in the mid 1990's of GaN-based materials for SSL. A waxing of interest in supercluster A5 is consistent with growing interest in the early 2000's in nanotechnology, for which the S- and Se-based II-VI semiconductors are a common materials platform.

Note that the degree to which a nation emphasises particular superclusters at the expense of other superclusters depends mostly on its size. Nations that contribute significantly to the overall dataset (e.g., USA, Japan, Germany) contribute relatively evenly to the various superclusters. Nations that contribute less significantly to the overall dataset (Taiwan, South Korea, China, France, Canada) contribute relatively unevenly to the various superclusters. This may in part be due to the coarse graininess associated with smaller numbers of institutional contributors. It may also in part be due to the national competitive advantage that derives from building critical mass in a supercluster (Porter, 1990), coupled with the difficulty that small nations have in doing so in all superclusters.

For example, Taiwan's emphasis is supercluster A3 (GaN), consistent with its extensive commercial activity in GaN-based LED technology. Canada's and Great Britain's emphasis is supercluster C1 (human motor and visual response instrumentation), consistent with their strengths in animal and human psycho-physiology. South Korea's and China's emphasis is supercluster A4 (ZnO). In fact, this is the supercluster for which nations differ most in their emphases, the Asian nations having generally a much stronger emphasis than the North American and European nations. This may be a reflection of the relative newness of this supercluster: 82% (626 out of 767) of its articles were published in the last five years, a much higher percentage than the 37% (11,975 out of 32,418) of the articles in the entire dataset. The Asian nations, whose R&D investments are growing (percentage-wise) very rapidly, might be expected to target their investments in newer areas of science and technology where they may more easily find competitive advantage.

Finally, in addition to the smaller nations' greater ongoing selective emphasis on particular superclusters, they are also broadening their contributions in the other superclusters. This is illustrated by the hot (orange) cells in Table 3(c), which represent those superclusters for which a nation's percentage contribution has increased the most in recent years. These are, by and large, not the superclusters that the smaller nations are emphasising, but the superclusters that they are 'not' emphasising but that they are broadening their contributions to – because their contributions have been historically so small, even modest recent contributions represent large percentage increases.

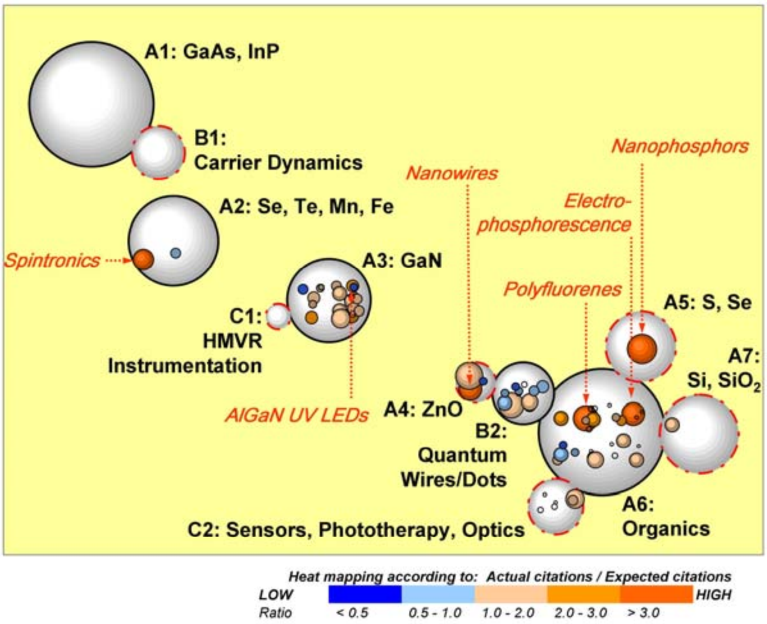
7 Article hot topics

Thus far, we have focused on the level-4 superclusters. These superclusters divide the knowledge domain finely enough to distinguish, as just discussed in Section 6, national contribution patterns. They do not divide the knowledge domain finely enough to be of use in analysing technical communities with shared paradigms, language and interests. In other words, few researchers would be likely to view themselves as ‘expert’ in any of the 11 level-4 superclusters, or even in any of the 50 level-3 clusters, broad as they are.

To analyse technical communities, we use the level-2 clusters, and for brevity we call these level-2 clusters ‘topics’. As indicated in Table 2, there were 491 topics, of which 72 (or 15%) can be considered young, in that their articles have a mean age of four years or less. These 72 young topics are shown in Figure 9 as coloured spheres on a stylised supercluster map. The superclusters on this map have been shifted manually from those shown on Figure 8, so that they do not overlap, and so that the topics can be placed unambiguously in the supercluster to which it was ultimately aggregated.

The areas of the topics have been scaled to reflect the number of articles they contain. The colouring (or heat mapping) of the topics have been chosen to reflect their ‘impact’. For an individual article, impact is the ratio between its number of citations and the average number of citations for all articles in the dataset published that same year. For a topic, impact is the average impact of all articles that aggregated into the topic. Topics that are both young and high impact are considered ‘hot’ and are coloured orange, while those with low impact are considered ‘cold’ and are coloured blue.

Figure 9 Distribution of young article topics (level-2 clusters) on a stylised article supercluster map (see online version for colours)



Notes: Areas of each topic are proportional to numbers of articles, colour of topics reflect impact. The six hottest topics are labelled in orange.

As illustrated in Figure 9, the young topics are not dispersed evenly amongst the superclusters. Superclusters A1 (GaAs, InP), B1 (carrier dynamics) and C1 (human motor and visual response instrumentation) have no young topics, and superclusters A2 (Se, Te, Mn, Fe) and A7 (Si, SiO_2) each have only one or two small young topics. But superclusters A3 (GaN), A4 (ZnO), A6 (organics) and B2 (quantum wires/dots) each have multiple young topics. Supercluster A6 (organics) has both the most young (26) and hot (17 with an impact greater than 1.0) topics.

The six hottest topics are specifically called out in Figure 9, and their details, including the seminal articles associated with them, are listed in Table 4. They are all relatively large and active areas of publishing activity:

- *Spintronics:* This area involves the manipulation of spin-polarised electrons. Such manipulation is mostly of interest for computation and memory applications, but may also someday be useful for applications involving the absorption and emission of polarised light.
- *AlGaN ultraviolet LEDs:* This area is aimed at LEDs that emit light in the ultraviolet. Such LEDs are mostly of interest for use as excitation sources for fluorescence spectroscopy and for water purification, but they may also someday be of use as excitation sources of white-light-emitting phosphors.
- *Nanowires:* This area takes advantage of modified electronic, optical and mechanical properties in very small diameter, near-one-dimensional nanowires. They are of interest for a wide variety of applications, including as light emitters in which nanoscale effects have suppressed strain-related structural defects.
- *Nanophosphors:* This area takes advantage of modified electronic, optical and mechanical properties in very small, near-zero-dimensional nanodots. Nanophosphors are of special interest when functionalised and attached to biomolecules, for use as light-absorbing and light-emitting 'tags'. They are also of interest for conversion of monochromatic coloured into multi-chromatic white light.
- *Polyfluorenes:* These are a promising class of 'wide-gap' organic material emitting efficiently in the blue and, with suitable modification, in the green and red. They are of interest for blue light emission in flat-panel colour displays and may also someday be of use in producing multi-chromatic white light.
- *Electrophosphorescence:* This area involves the harvesting of light emission from singlet and triplet states (rather than just singlet states), thereby enhancing significantly overall light-emission efficiency. Electrophosphorescence is of interest for flat-panel colour displays and ultimately also for white-light production.

Table 4 Details of the six hottest article topics from Figure 9

Topic	*Topic #*	*Size*	*Age (yrs)*	*Avg. cites*
Spintronics	256	212	2.8	12.1
AlGaN UV LEDs	260	46	1.6	4.9
Nanowires	374	246	2.6	11.4
Nanophosphors	389	407	4.0	21.1
Polyfluorenes	417	231	3.0	14.8
Electrophosphorescence	298	308	2.7	13.0

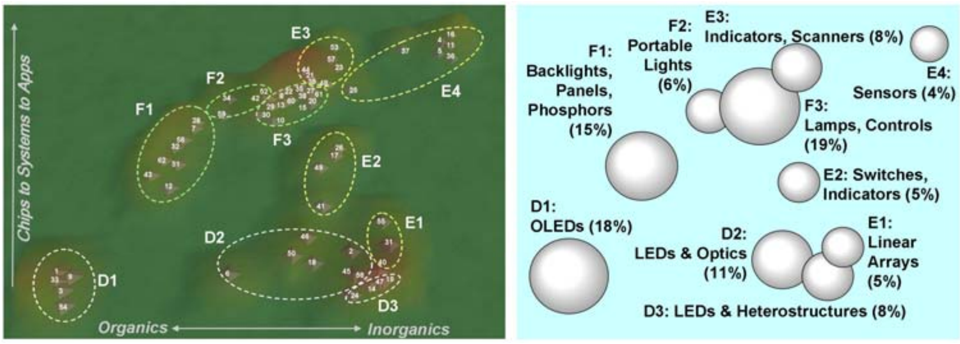
Topic	*Impact*	*Super-cluster*	*Seminal article*
Spintronics	2.9	A2	Ohno et al. (1999)
AlGaN UV LEDs	3.1	A3	Zhang et al. (2002)
Nanowires	3.0	A4	Duan et al. (2001)
Nanophosphors	3.2	A5	Chan and Nie (1998)
Polyfluorenes	3.1	A6	Grice et al. (1998)
Electrophosphorescence	3.2	A6	Baldo et al. (1999)

Note that none of these topics are hot specifically because of their perceived potential impact on SSL. They are hot because of a combination of internal ‘push’ (the perceived richness of the topic itself) and broad external ‘pull’ (the perceived richness of potential applications, of which SSL is one). ‘Nevertheless, the serendipitous and unpredictable nature of the interplay between science and technology suggests that SSL-relevant breakthroughs could emerge from any of these (as well as other) topics’. Given similar levels of targeted encouragement, whether an SSL-relevant breakthrough is more likely from a hot topic of broad relevance than from a colder topic of more SSL-exclusive relevance is a question of great interest to national innovation strategy, but beyond the scope of this report.

8 Patent superclusters

Two maps of the resulting superclusters of patents in the dataset are shown in Figure 10. The maps are similar to those in Figure 8 for articles, but with some differences. The left map is a topographical map taken from VxInsight, in which articles have been aggregated into 62 level-3 clusters, and these level-3 clusters have in turn been aggregated manually into the ten level-4 dashed-oval-outlined superclusters labelled D1-D3, E1-E4 and F1-F3. The right map is a stylised level-4 map, in which the superclusters’ centre positions have been preserved, but their areas have been scaled to reflect the number of patents they contain.

Figure 10 Level-3 and level-4 cluster maps of all patents in the dataset (see online version for colours)



Notes: (Left) VxInsight topographic map showing the level-3 clusters and their manual aggregation (dashed white lines) into level-4 superclusters.
(Right) a stylised map in which each supercluster's area is proportional to the number of patents it contains.

Examination of the full dataset shows that these patent superclusters are not nearly as 'clean' as the article superclusters, in that there is less mutual exclusivity between them. Roughly speaking, but only roughly, the superclusters fall into three groups: the first distinguished by materials and physics, the second by application in the low-power regime, and the third by application in the medium-to-high-power regime.

- *D materials- and physics-oriented superclusters.* A large percentage (37%) of the patents aggregated into the three materials- and physics-oriented superclusters labelled D1–D3. The first supercluster, D1 (18%), is that associated with OLEDs. The other two superclusters, D2 and D3 are associated primarily with inorganic light-emitting diodes: D3 (8%) is associated with the materials and heterostructure device aspects of inorganic light-emitting diodes and D2 (11%) is associated with the optics aspects of inorganic light-emitting diodes.
- *E low-power-applications-oriented superclusters.* A smaller (22%) fraction of the patents aggregated into the four low-power-applications-oriented superclusters labelled E1-E2. Supercluster E1 (5%) is associated with linear (1-D) arrays of LEDs. Supercluster E2 (5%) is associated with switches and indicators, and supercluster E3 (8%) is associated with indicators and scanners. Supercluster E4 (4%) is associated with sensors.
- *F medium-to-high-power-applications-oriented superclusters.* A somewhat larger (40%) fraction of the parents aggregated into the three medium-to-high-power-applications-oriented superclusters labelled F1–F3. Supercluster F1 (15%) is associated with backlights, panels and phosphors. Supercluster F2 (6%) is associated with portable lights. Supercluster F3 (19%) is associated with lamps and controls.

As for the articles, assigning a single high-level interpretation to the millions of bibliographic coupling comparisons that were made to compute these two maps is non-trivial. However, a reasonable interpretation of the positioning of the superclusters is the following.

In the vertical dimension, the lower row of superclusters, D1 (OLEDs), D2 (LEDs and optics), D3 (LEDs and heterostructures) and E1 (linear arrays), are based primarily on improvements in light-emitters at the materials and chip levels. The upper row of superclusters, E3 (indicators, scanners) and E4 (sensors), are based primarily on end-use applications of systems (e.g., liquid-level indicators, oximeters) containing light-emitters as an essential element. The superclusters in between are based primarily on improvements in light-emitting systems.

In the horizontal dimension, the lower left supercluster, D1 (OLEDs), is based primarily on organic light emitters, while the lower right supercluster, D3 (LEDs and heterostructures), is based primarily on inorganic light emitters. As a consequence, systems and applications that tend to use OLEDs, like the flat-panel displays contained in supercluster F1 (backlights, panels, phosphors), are positioned towards the left; while systems and applications that tend to use LEDs, like the oximeters contained in supercluster E4 (sensors), are positioned towards the right.

In other words (and very roughly), superclusters are dispersed vertically according to their emphasis on chips, systems or applications, while superclusters are dispersed horizontally according to their emphasis on organic or inorganic light emitters.

Interestingly, this high-level interpretation of the patent map is very different from that discussed in Section 6 of the article map. Two possible explanations seem particularly likely.

A first explanation might be that the intellectual content is in fact different for the article and patent datasets. For example, articles will have a greater science and technology content, while patents will have a greater technology and societal-use content. Hence, some areas of science would likely not be represented in the patent dataset, just as some areas of applications would likely not be represented in the article dataset.

A second explanation might be that even when articles and patents having similar intellectual content are present in both datasets, differences in citation practices might lead to differences in bibliographic coupling, and in turn lead to differences in record, cluster and supercluster positioning. For example, patents, whose intent is to establish priority (so as to exclude knowledge from being used by others), tend to cite patents that they are competitive with, but not patents that they are supported by. In contrast, articles, whose intent is partly pedagogical (to communicate knowledge in order to enable that knowledge to be used by others), tend not just to cite articles that they are competitive with, but also articles that they are supported by. The result is that articles that have similar intellectual content, and which cite a common set of supporting articles, will be strongly coupled bibliographically; while patents that have similar intellectual content, and which mostly cite each other, may not be as strongly coupled bibliographically. For this reason, it might be interesting in future work to explore differences between article maps based on bibliographic coupling and patent maps based on citation coupling.

We note here that in the early stages of our work we did attempt to map ‘simultaneously’ the article and patent datasets, using a lexical similarity metric based on common words used in article and patent abstracts. The words used in the articles and patents were sufficiently different, however, that the resulting map (not shown) consisted of one region populated mostly with articles, and another region populated mostly with patents. Additional analysis on word distributions showed that of the most widely used words (those occurring in at least 0.1% of the abstracts), only 14% were used relatively frequently (fractional ratios between 2:1 and 1:2) in both articles and patents. The remaining 86% appeared preferentially (fractional ratios greater than 2:1 or less than 1:2) in either articles or patents. These differences in language may simply reflect differences between the writing styles of researchers and patent attorneys, but they may also reflect real differences between the intellectual content of articles and patents.

We are now in a position to analyse the relative contributions various nations have made not only to the overall patent domain, but to the patent sub-domains that the various superclusters represent. This analysis is summarised in Tables 5a, 5b and 5c. The interpretations of Tables 5(a), 5(b) and 5(c) for patents are identical to those of the analogous Tables 3(a), 3(b) and 3(c) for articles. Each column in Tables 5(a) and 5(b) lists, for each supercluster, the percentages of patents contributed by nine representative nations and the 'rest of the world': Table 5(a) lists these percentages for the entire 28 years (1977–2004) and Table 5(b) lists them for the most recent five years (2000–2004). Table 5(c) lists, again for each supercluster, the ratio between the two percentages in Tables 5(b) and 5(a). This ratio is an indication of the trend in time of a nation's contribution to the articles in the various superclusters.

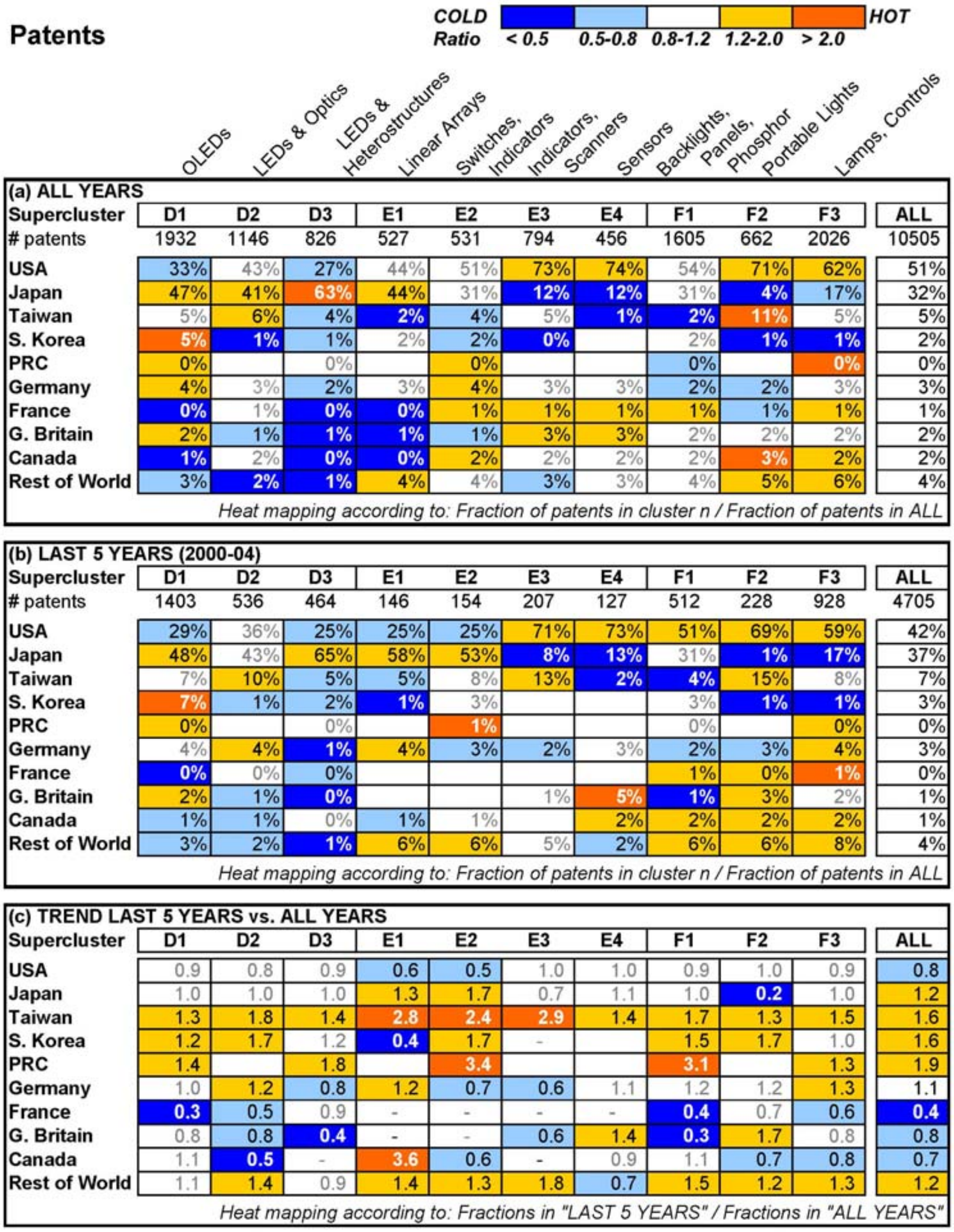
From Tables 5(a) and 5(b), one can see that the USA and Japan are the two dominant nations in patents, even more so than they were in articles. The rightmost columns indicate that together they contributed more than ¾ of all the patents (summed over all superclusters) in both the full 28-year dataset and in the most recent five years of the dataset.

From the rightmost column of Table 5(c), however, one can also see that this overall dominance is shifting. The USA's trend ratio is 0.8, indicating that its percentage contribution, summed over all superclusters, is shrinking (from 51% to 42%). Japan's trend ratio is 1.2, indicating that its percentage contribution is growing (from 32% to 37%). Indeed, the other Asian nations are also growing, even more rapidly than Japan: Taiwan's and South Korea's trend ratios are 1.6 and China's trend ratio is 1.9, indicating that their percentage contributions are growing rapidly (Taiwan's from 5% to 7% and South Korea's from 2% to 3%).

Just as in Table 3 for articles, in Table 5 for patents we have 'heat mapped' the colours of the various cells in Tables 3(a) and 3(b) to reflect a nation's percentage contribution to particular superclusters relative to its percentage contribution to the overall dataset.

The US, for example, has had both a historical and recent emphasis on superclusters E3 (indicators, scanners) and E4 (sensors). Interestingly, this is consistent with the USA's historical and recent emphasis on article supercluster C2 (sensors, phototherapy, optics), as discussed in Section 6, and in ongoing interest in the use of LEDs and lasers as sensors and sources in medical, environmental and communications applications. Japan, in contrast, has had both a historical and recent emphasis on supercluster D3 (LEDs and heterostructures), consistent with its leadership role in the GaN-based materials breakthroughs discussed in Section 3. The implication is that, amongst these two dominant nations, Japan is emphasising improvements in light-emitters at the materials and chip levels, while the USA is emphasising improvements in light-emitting systems and their end-use applications.

Table 5 Absolute and percentage contributions of representative nations to the various parent superclusters, (a) contributions to the entire 28-year dataset (b) contributions to the most recent five years of the dataset (c) the ratio between the two (see online version for colours)



	OLEDs	LEDs & Optics	LEDs & Heterostructures	Linear Arrays	Switches, Indicators	Indicators, Scanners	Sensors	Backlights, Panels, Phosphor	Portable Lights	Lamps, Controls	
(a) ALL YEARS											
Supercluster	**D1**	**D2**	**D3**	**E1**	**E2**	**E3**	**E4**	**F1**	**F2**	**F3**	**ALL**
# patents	1932	1146	826	527	531	794	456	1605	662	2026	10505
USA	33%	43%	27%	44%	51%	73%	74%	54%	71%	62%	51%
Japan	47%	41%	63%	44%	31%	12%	12%	31%	4%	17%	32%
Taiwan	5%	6%	4%	2%	4%	5%	1%	2%	11%	5%	5%
S. Korea	5%	1%	1%	2%	2%	0%		2%	1%	1%	2%
PRC	0%		0%		0%			0%		0%	0%
Germany	4%	3%	2%	3%	4%	3%	3%	2%	2%	3%	3%
France	0%	1%	0%	0%	1%	1%	1%	1%	1%	1%	1%
G. Britain	2%	1%	1%	1%	1%	3%	3%	2%	2%	2%	2%
Canada	1%	2%	0%	0%	2%	2%	2%	2%	3%	2%	2%
Rest of World	3%	2%	1%	4%	4%	3%	3%	4%	5%	6%	4%

Heat mapping according to: Fraction of patents in cluster n / Fraction of patents in ALL

(b) LAST 5 YEARS (2000-04)											
Supercluster	**D1**	**D2**	**D3**	**E1**	**E2**	**E3**	**E4**	**F1**	**F2**	**F3**	**ALL**
# patents	1403	536	464	146	154	207	127	512	228	928	4705
USA	29%	36%	25%	25%	25%	71%	73%	51%	69%	59%	42%
Japan	48%	43%	65%	58%	53%	8%	13%	31%	1%	17%	37%
Taiwan	7%	10%	5%	5%	8%	13%	2%	4%	15%	8%	7%
S. Korea	7%	1%	2%	1%	3%			3%	1%	1%	3%
PRC	0%		0%		1%			0%		0%	0%
Germany	4%	4%	1%	4%	3%	2%	3%	2%	3%	4%	3%
France	0%	0%	0%					1%	0%	1%	0%
G. Britain	2%	1%	0%			1%	5%	1%	3%	2%	1%
Canada	1%	1%	0%	1%	1%		2%	2%	2%	2%	1%
Rest of World	3%	2%	1%	6%	6%	5%	2%	6%	6%	8%	4%

Heat mapping according to: Fraction of patents in cluster n / Fraction of patents in ALL

(c) TREND LAST 5 YEARS vs. ALL YEARS											
Supercluster	**D1**	**D2**	**D3**	**E1**	**E2**	**E3**	**E4**	**F1**	**F2**	**F3**	**ALL**
USA	0.9	0.8	0.9	0.6	0.5	1.0	1.0	0.9	1.0	0.9	0.8
Japan	1.0	1.0	1.0	1.3	1.7	0.7	1.1	1.0	0.2	1.0	1.2
Taiwan	1.3	1.8	1.4	2.8	2.4	2.9	1.4	1.7	1.3	1.5	1.6
S. Korea	1.2	1.7	1.2	0.4	1.7	-		1.5	1.7	1.0	1.6
PRC	1.4		1.8		3.4			3.1		1.3	1.9
Germany	1.0	1.2	0.8	1.2	0.7	0.6	1.1	1.2	1.2	1.3	1.1
France	0.3	0.5	0.9	-	-	-	-	0.4	0.7	0.6	0.4
G. Britain	0.8	0.8	0.4	-	-	0.6	1.4	0.3	1.7	0.8	0.8
Canada	1.1	0.5	-	3.6	0.6	-	0.9	1.1	0.7	0.8	0.7
Rest of World	1.1	1.4	0.9	1.4	1.3	1.8	0.7	1.5	1.2	1.3	1.2

Heat mapping according to: Fractions in "LAST 5 YEARS" / Fractions in "ALL YEARS"

Unlike for articles, where the smaller nations have a higher degree of emphasis on particular superclusters at the expense of others, for patents all nations have a similar degree of emphasis. Interestingly, Japan's emphasis at the materials and chip levels is shared by the other Asian nations (particularly Taiwan and South Korea), while the

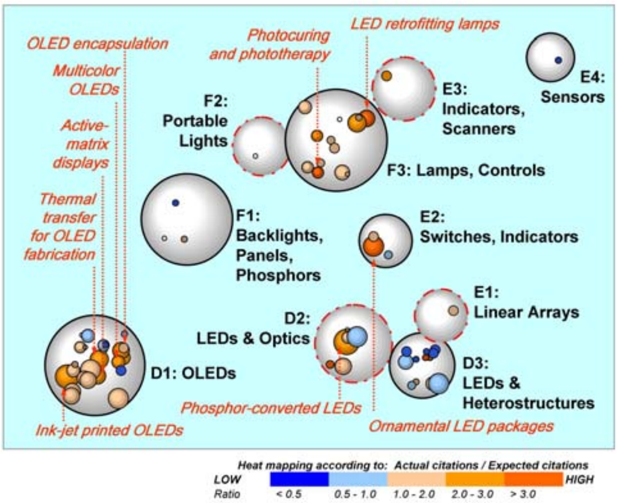
USA's emphasis on systems and end-use applications is shared by the North American and European nations. This is consistent with the growth of the semiconductor chip industries in Asia, and of the system integration and service industries in North America and Europe.

Finally, in addition to the Asian nations' ongoing selective emphasis on the materials and chips oriented superclusters, they are also, just as they are in the article superclusters, broadening their contributions in the other patent superclusters. This is illustrated by the hot (orange) cells in Table 5(c), which represent those superclusters for which a nation's percentage contribution has increased the most in recent years. These are, by and large, not the superclusters that the smaller nations are emphasising, but the superclusters that they are 'not' emphasising – but in which they are broadening their contributions. Because their historical contributions have been so small, even modest recent contributions represent large percentage increases.

9 Patent hot topics

As with articles, for patents we can also examine technical communities. These are the level-2 clusters, and again for brevity we call these level-2 clusters 'topics'. As indicated in Table 2, there were 388 topics, of which 65 (or 17%) can be considered young, in that their patents have a mean age 4 years or less. These 65 young topics are shown in Figure 11 as coloured circles on a stylised supercluster map. The superclusters on this map have been shifted manually from those shown on Figure 10, so that they do not overlap, and so that the topics can be placed unambiguously in the supercluster to which it was ultimately aggregated.

Figure 11 Distribution of young patent topics (level-2 clusters) on a stylised patent supercluster map (see online version for colours)



Notes: Areas of each topic are proportional to numbers of articles, colour of topics reflect impact. The nine hottest topics are labelled in orange.

Just as for the analogous map in Figure 9 for articles, the areas of the topics have been scaled to reflect the number of articles they contain. The colouring (or heat mapping) of the topics have been chosen to reflect their 'impact'. For an individual patent, impact is the ratio between its actual number of citations and the average number of citations for all patents in the entire US patent database issued the same year and having the same patent class (this is a slightly different definition of impact than that used for articles). For a topic, impact is the average impact of all patents that aggregated into the topic. Topics that are both young and high impact are considered 'hot' and are coloured orange, while those with low impact are considered 'cold' and are coloured blue.

Table 6 Details of the nine hottest patent topics from Figure 11

Topic	*Topic #*	*Size*	*Age (yrs)*	*Avg. cites*
OLED encapsulation	115	46	2.8	6.8
Active-matrix OLED displays	285	48	2.5	4.4
Multicolour OLEDs	362	49	4.0	14.6
Thermal transfer for OLED fabrication	339	67	2.1	2.1
Ink-jet printed OLEDs	258	100	2.5	3.2
Phosphor-converted LEDs	332	130	3.9	4.5
Ornamental LED packages	380	80	3.1	4.2
Photocuring and phototherapy	235	50	3.4	8.4
LED retrofitting lamps	294	70	3.3	5.2

Topic	*Impact*	*Super cluster*	*Seminal Patent*
OLED encapsulation	2.8	D1	Sheats et al. (2000)
Active-matrix OLED displays	2.2	D1	Ozawa (2001)
Multicolour OLEDs	2.6	D1	Forrest et al. (1998)
Thermal transfer for OLED fabrication	2.5	D1	Wolk et al. (2001)
Ink-jet printed OLEDs	2.8	D1	Sturm et al. (2000)
Phosphor-converted LEDs	2.1	D2	Lowery (1999)
Ornamental LED packages	3.0	E2	Kamada (2003)
Photocuring and phototherapy	3.7	F3	Kennedy and Keyser (1997)
LED retrofitting lamps	2.4	F3	Anderson (1996)

As illustrated in Figure 11, the young topics are not dispersed evenly amongst the superclusters. Superclusters E1 (linear arrays), E3 (indicators, scanners), E4 (sensors), F1 (backlights, panels, phosphors) and F2 (portable lights) have very few young topics. But superclusters D1 (OLEDs), D2 (LEDs and optics), D3 (LEDs and heterostructures), E2 (switches, indicators) and F3 (lamps, controls) each have multiple young topics. Supercluster D1 (OLEDs) has both the most young and hot topics. This is consistent with article supercluster A6 (organics) having also both the most young and hot topics. Interestingly, supercluster D3 (LEDs and heterostructures) has many young topics, but none of them are hot. The recent topics within medium-power lamps and controls (F3) all have higher than average impact.

The nine hottest topics are specifically called out in Figure 11, and their details, including the seminal articles associated with them, are listed in Table 6. They are all relatively large and active areas of patenting activity:

- *OLED encapsulation:* This area involves developing materials and designs for encapsulating OLEDs, with particular attention to preventing water or oxygen from reaching and causing degradation to OLED devices. Because OLEDs are extremely sensitive to water and oxygen, this is an area of great interest to any OLED application, including flat-panel displays and SSL.
- *Active-matrix OLED displays:* This area involves the development of flat-panel OLED displays in which each pixel is controlled by its own thin-film transistor. Once turned on, the pixel stays on throughout an entire frame refresh cycle, rather than decaying as in a passive-matrix display. This is an area of great interest to flat-panel displays, with possible impact on SSL in configurations where real-time tailoring of brightness and colour is important.
- *Multicolour OLEDs:* This area involves the development of OLEDs that emit more than one colour through integration of various organic active materials. This is an area of great interest for wide-colour-gamut flat-panel video displays, as well as for producing white light of variable colour temperature for SSL.
- *Thermal transfer for OLED fabrication:* This area involves thermal methods for transferring organic materials from one substrate to another, as a means of fabricating OLED devices. This is an area of interest for patterned devices in flat-panel displays, and possibly for patterned devices in SSL.
- *Ink-jet printed OLEDs:* This area involves ink-jet methods for depositing organic materials, as a means of fabricating OLED devices. This area is also of interest for patterned devices in flat-panel displays, and also possibly for patterned devices in SSL.
- *Phosphor-converted LEDs:* This area involves the development of phosphors and configurations for white light LEDs, in which a phosphor is used to convert monochromatic light into multi-chromatic white light. This method is the current dominant design for white light LEDs and is of direct interest to SSL.
- *Ornamental LED packages:* This area involves the development of LED packages for numerous signalling and display purposes. It is an active area, with possible interest for SSL.
- *Photocuring and phototherapy:* This area involves the use of LED light to induce chemical changes in non-biological and biological materials. Because the wavelength of the light can be targeted at particular chromophores and chemical reactions, the light can be used very efficiently. It is likely only of indirect interest to SSL.
- *LED retrofitting lamps:* This area involves the system aspects of packaging LEDs in such a way as to enable them to be plugged directly into systems which currently use traditional light sources. It is of direct interest to SSL, as one possible route to consumer acceptance.

Note that none of these topics, except perhaps for phosphor-converted LEDs and LED retrofitting lamps, are hot specifically because of their perceived potential impact on SSL. They are hot because of other nearer-term applications in displays, signalling and medical treatments. Nevertheless, just as for the article hot topics, the serendipitous and unpredictable nature of the interplay between technology and applications suggests that breakthroughs of importance to SSL could emerge from any of these (as well as other) topics.

10 Future directions

We have presented analyses of the literature of SSL, based on a comprehensive dataset of articles and US patents in the foundational knowledge domain of electroluminescent materials and phenomena. The analyses include: identification of knowledge sub-domains of historical and recent importance, and trends over time of the contributions of various nations and continents to the knowledge domain and its sub-domains.

However, there are some deficiencies in our analyses and we mention in closing two of these, along with possible approaches to circumventing them.

First, the dataset was created purely lexically: a search string acting on the lexical content of records in the primary databases. With such a purely lexical approach, it is difficult to simultaneously capture all knowledge sub-domains relevant to SSL, while avoiding capturing some knowledge sub-domains of marginal relevance. In our case, phosphors, because they involve photoluminescent (not electroluminescent) materials and phenomena, were poorly captured; while infrared (not visible) electroluminescence was perhaps captured too well.

To correct for this deficiency, an alternative approach to creating the dataset might be to make use of bibliometric distances. Starting from a carefully (and perhaps iteratively) chosen core set of articles and patents, one might successively add articles and patents that are bibliometrically ‘near’ the core set. Increasing the size of the core set would increase the precision, while increasing the allowed bibliometric distance away from the core set would increase the recall.

Second, because there is very little bibliographic coupling between the article and patent datasets, the article and patent datasets were analysed independently. The result was that our two sets of emergent knowledge sub-domains are independent of each other, even though clearly there are intellectual relationships between them. It would be interesting to be able to analyse a combined dataset, resulting in only one set of emergent knowledge sub-domains composed of both articles and patents.

A possible approach might be based on identifying ‘pairs’ of articles and patents with similar intellectual content based on coincidence of author sets, publication and issue dates, and word/phrase usage. These pairs would then be assigned artificially large edge weights, but otherwise the mapping of articles and patents to x, y positions based on bibliometric coupling would proceed as before. The greater the number of pairs the more robust the combined mapping might be expected to be. And by varying the strength of the artificially large edge weights, the coupling ‘between’ the article and patent datasets relative to the coupling ‘within’ the article and patent datasets would also be varied.

References

Amano, H., Sawaki, N. and Akasaki, I. (1986) 'Metalorganic vapor phase epitaxial growth of a high quality GaN film using an AlN buffer layer', *Appl. Phys. Lett.*, Vol. 48, pp.353–355.

Anderson, R.A. (1996) 'Light emitting diode lamp', US Patent Number 5,575,459, Uniglo Canada Inc.

Baldo, M.A., Lamansky, S., Burrows, P.E., Thompson, M.E., Forrest, S.R. (1999) 'Very high-efficiency green organic light-emitting devices based on electrophosphorescence', *Appl. Phys. Lett.*, Vol. 75, p.4.

BES (2006) 'Basic research needs for solid-state lighting', September, US Department of Energy, Office of Basic Energy Sciences, available at http://www.sc.doe.gov/bes/reports/files/SSL_rpt.pdf.

Boyack, K.W., Wylie, B.N. and Davidson, G.S. (2002) 'Domain visualization using VxInsight® for science and technology management', *J. of the American Society for Information Science and Technology*, Vol. 53, pp.764–774.

Braun, D. and Heeger, A.J. (1991) 'Visible light emission from semiconducting polymer diodes', *Appl. Phys. Lett.*, Vol. 58, pp.1982–1984.

Burroughes, J.H., Bradley, D.D.C., Brown, A.R., Marks, R.N., Mackay, K., Friend, R.H., Burns, P.L. and Holmes, A.B. (1990) 'Light-emitting diodes based on conjugated polymers', *Nature*, Vol. 347, pp.539–541.

Casper, S.L. and Lowrey, T.A. (1993) 'Flat panel display in which low-voltage row and column address signals control a much higher pixel activation voltage', US Patent Number 5,210,472, Micron Technology, Inc.

Chan, W.C.W. and Nie, S.M. (1998) 'Quantum dot bioconjugates for ultrasensitive nonisotopic detection', *Science*, Vol. 281, p.2016.

Cullis, A.G., Canham, L.T. and Calcott, P.D.J. (1997) 'The structural and luminescence properties of porous silicon', *J. Appl. Phys.*, Vol. 82, pp.909–965.

Davidson, G.S., Wylie, B.N. and Boyack, K.W. (2001) 'Cluster stability and the use of noise in interpretation of clustering', *Proceedings of IEEE Information Visualization 2001*, pp.23–30.

Duan, X.F., Huang, Y., Cui, Y., Wang, J.F., Lieber, C.M. (2001) 'Indium phosphide nanowires as building blocks for nanoscale electronic and optoelectronic devices', *Nature*, Vol. 409, p.66.

EERE (2003) 'Energy savings estimates of solid-state lighting in general illumination applications', November, US Department of Energy, Office of Energy Efficiency and Renewable Energy, Building Technologies Program, available at http://www.netl.doe.gov/ssl/PDFs/SSL%20Energy%20Savi_ntial%20Final.pdf.

EERE (2006) 'Solid-state lighting research and development portfolio: multi-year program plan FY'07-FY'12', March, US Department of Energy, Office of Energy Efficiency and Renewable Energy, Building Technologies Program, available at http://www.netl.doe.gov/ssl/PDFs/SSLMultiYearPlan.pdf.

Forrest, S.R., Thompson, M.E., Burrows, P.E., Sapochak, L.S. and McCarty, D.M. (1998) 'Multicolor organic light emitting devices', US Patent Number 5,707,745, The Trustees of Princeton University.

Friend, R.H., Burroughes, J.H. and Bradley, D.D. (1993) 'Electroluminescent devices', US Patent Number 5,247,190, Cambridge Research and Innovation Limited.

Friend, R.H., Gymer, R.W., Holmes, A.B., Burroughes, J.H., Marks, R.N., Taliani, C., Bradley, D.D.C., Dos Santos, D.A., Bredas, J.L., Logdlund, M. and Salaneck, W.R. (1999) 'Electroluminescence in conjugated polymers', *Nature*, Vol. 397, pp.121–128.

Grice, A.W., Bradley, D.D.C., Bernius, M.T., Inbasekaran, M., Wu, W.W. and Woo, E.P. (1998) 'High brightness and efficiency blue light-emitting polymer diodes', *Appl. Phys. Lett.*, Vol. 73, p.629.

Hicks, D. and Narin, F. (2001) 'Strategic research alliances and 360 degree bibliometric indicators', *Strategic Research Partnerships: Proceedings from an NSF Workshop*, November, Arlington, VA, available at http://www.nsf.gov/statistics/nsf01336/p1s6.htm.

Huang, Z, Chen, H., Chen, Z-K. and Roco, M.C. (2004) 'International nanotechnology development in 2003: country, institution and technology field analysis based on USPTO patent database', *J. Nanoparticle Research*, Vol. 6, pp.325–354.

Kamada, K. (2003) 'Light emitting diode (LED)', US Patent Number D482,337, Nichia Corporation.

Kennedy, J. and Kayser, R. (1997) 'Portable light emitting apparatus with a semiconductor emitter array', US Patent Number 5,634,711, Self.

Klavans, R. and Boyack, K.W. (2006) 'Quantitative evaluation of large maps of science', *Scientometrics*, Vol. 68, No. 3, pp.475–499.

Korfhage, R.R. (1997) *Information Storage and Retrieval*, Chapter 8, John Wiley and Sons, New York.

Leonard, D., Krishnamurthy, M., Reaves, C.M., Denbaars, S.P. and Petroff, P.M. (1993) 'Direct formation of quantum-sized dots from uniform coherent islands of InGaAs on GaAs surfaces', *Appl. Phys. Lett.*, Vol. 63, pp.3203–3205.

Lowery, C.H. (1999) 'Multiple encapsulation of phosphor-LED devices', US Patent Number 5,959,316, Hewlett-Packard Company.

Lundberg, J., Fransson, A., Brommels, M., Skar, J. and Lundkvist, I. (2006) 'Is it better or just the same? Article identification strategies impact bibliometric assessments', *Scientometrics*, Vol. 66, pp.183–197.

Morkoc, H., Strite, S., Gao, G.B., Lin, M.E., Sverdlov, B. and Burns, M. (1994) 'Large-band-gap SiC, III-V nitride and II-VI ZnSe-based semiconductor device technologies', *J. Appl. Phys.*, Vol. 76, pp.1363–1398.

Nakamura, S., Mukai, T. and Senoh, M. (1994) 'Candela-class high-brightness InGaN/AlGaN double-heterostructure blue-light-emitting diodes', *Appl. Phys. Lett.*, Vol. 64, pp.1687–1689.

New, W. Jr., Corenman, J.E. (1986) 'Calibrated optical oximeter probe', US Patent Number 4,621,643, Nellcor Incorporated.

Ohno, Y., Young, D.K., Beschoten, B., Matsukura, F., Ohno, H. and Awschalom, D.D. (1999) 'Electrical spin injection in a ferromagnetic semiconductor heterostructure', *Nature*, Vol. 402, p.790.

Optoelectronics Industry Development Association (OIDA) (2002a) 'Organic light emitting diodes (OLEDs) for general illumination update 2002', August, available at http://www.netl.doe.gov/ssl/workshop/Report%20OLED%20August%202002_1.pdf.

Optoelectronics Industry Development Association (OIDA) (2002b) 'Light emitting diodes (LEDs) for general illumination: an OIDA technology roadmap update 2002', November, available at http://www.netl.doe.gov/ssl/workshop/Report%20led%20November%202002a_1.pdf.

Ozawa, T. (2001) 'Display device with thin film transistor (TFT) and organic semiconductor film in a luminescent element', US Patent Number 6,194,837, Seiko Epson Corporation.

Pease, L.L., Bond, E.T., Crevelt, D.E. and Holmes, V.F. (1994) 'Secure automated electronic casino gaming system', US Patent Number 5,326,104, International Game Technology.

Porter, M. (1990) *The Competitive Advantage of Nations*, Free Press.

Schubert, E.G. and Kim, J.K. (2005) 'Solid-state light sources getting smart', *Science*, 27 May, Vol. 308, pp.1274–1278.

Sheats, J.R., Hueschen, M.R., Seaward, K.L., Roitman, D.B. and Briggs, G.A.D. (2000) 'Transparent, flexible permeability barrier for organic electroluminescent devices', US Patent Number 6,146,225, Agilent Technologies, Inc.

State Intellectual Property Office of Peoples Republic of China (SIPO) (2004) '2004 Annual Report', Chapter 14, available at http://www.sipo.gov.cn/sipo_English/ndbg/nb/ndbg2004/200509/t20050902_53480.htm.

Strategic Perspectives (2006) *US Competitiveness in Solid-state Lighting: Illuminating a Gathering Storm?*, April, Strategic Perspectives, Inc.

Sturm, J.C., Wu, C.C., Marcy, D. and Hebner, T.R. (2000) 'Fabrication of organic semiconductor devices using ink jet printing', US Patent Number 6,087,196, The Trustees of Princeton University.

Tang, C.W. (1982) 'Organic electroluminescent cell', US Patent Number 4,356,429, Eastman Kodak Company.

Tang, C.W. and VanSlyke, S.A. (1987) 'Organic electroluminescent diodes', *Appl. Phys. Lett.*, Vol. 51, pp.913–915.

Tang, C.W., Chen, C.H. and Goswami, R. (1988) 'Electroluminescent device with modified thin film luminescent zone', US Patent Number 4,769,292, Eastman Kodak Company.

Tang, C.W., VanSlyke, S.A. and Chen, C.H. (1989) 'Electroluminescence of doped organic thin films', *J. Appl. Phys.*, Vol. 65, pp.3610–3616.

Tsao, J.Y. (2003) 'Roadmap projects significant LED penetration of lighting market by 2010', *Laser Focus World*, May.

UN (2006) 'Composition of macro geographical (continental) regions, geographical sub-regions and selected economic and other groupings', revised 29 June, United Nations Statistics Division, available at http://unstats.un.org/unsd/methods/m49/m49regin.htm.

United States Patent and Trademark Office (USPTO) (2004) 'Patents by country, state and year – all patent types', December, available at http://www.uspto.gov/go/taf/cst_all.htm.

VanSlyke, S.A. and Tang, C.W. (1985) 'Organic electroluminescent devices having improved power conversion efficiencies', US Patent Number 4,539,507, Eastman Kodak Company.

VanSlyke, S.A., Tang, C.W. and Roberts, L.C. (1988) 'Electroluminescent device with organic luminescent medium', US Patent Number 4,720,432, Eastman Kodak Company.

Villa-Real, A.C. (1984) 'Programmable telephone system', US Patent Number 4,481,382, Self.

Wilber, S.A. (1983) 'Blood constituent measuring device and method', US Patent Number 4,407,290, Biox Technology, Inc.

Wolk, M.B., Baude, P.F., McCormick, F.B. and Hsu, Y. (2001) 'Thermal transfer element and process for forming organic electroluminescent devices', US Patent Number 6,194,119, 3M Innovative Properties Company.

Wu, Y., Pan, Y., Zhang, Y., Ma, Z., Pang, J., Guo, H., Xu, B. and Yang, Z. (2004) 'China Scientific and Technical Papers and Citations (CSTPC): history, impact and outlook', *Scientometrics*, Vol. 60, pp.385–397.

Zhang, J.P., Chitnis, A., Adivarahan, V., Wu, S., Mandavilli, V., Pachipulusu, R., Shatalov, M., Simin, G., Yang, J.W. and Khan, M.A. (2002) 'Milliwatt power deep ultraviolet light-emitting diodes over sapphire with emission at 278 nm', *Appl. Phys. Lett.*, Vol. 81, p.4910.

Notes

1 Still, our analyses of the two literatures should be considered independent. It would have been very interesting to explore an integrated analysis, but such analysis would have introduced considerable complexity, and indeed has thus far been only rarely attempted. See, e.g., Hicks and Narin (2001).

2 We used the time period 1977–2004 for articles: 1977 is the first full year for which we had access to complete bibliometric data from Thomson Scientific. For consistency, we used the same time period for US patents, even though complete bibliometric data is available for earlier years.

3 These included, for organic electroluminescent materials and phenomena, R.H. Friend, A.J. Heeger, C.W. Tang and S.R. Forrest, and for inorganic electroluminescent materials and phenomena, H. Amano, I. Akasaki, S. Nakamura, M.G. Craford, J.I. Pankove, R.L. Gunshor and S.P. Denbaars.

4 These included, for organic electroluminescent materials and phenomena, Cambridge Display, Universal Display, RiTdisplay, Samsung SDI, and for inorganic electroluminescent materials and phenomena, Lumileds, Nichia, Cree, Epistar, Osram Opto and United Epitaxy.

5 Deduced directly from Thomson Scientific's Science Citation Index for journal articles and from the US Patent and Trademark Office's database for US patents.