2021 年 3 月 15 日

データベース収録確認ワード（その1）

アジア特許情報研究会：伊藤徹男

1．はじめに

2009 年頃に中国特許情報を英語で検索できる中国特許庁の SIPO English(SIPO En)や CNPAT、C-Pat、SooPAT など無料中国英語データベースが、ワールドワイドな英語データベースでも freepatents online、Patent Lens などが次々と現れてきました。まだ商用英語データベースに中国、韓国の登録特許や実用新案は収録されておらず、台湾特許に至っては公開特許情報も満足に収録されている商用データベースもない時代でした。

「韓国特許調査では公開になる前に登録となる「公開前登録特許」の公開公報は発行されませんので登録特許情報が収録されていないと検索漏れを生じますよ」と指摘させていただいたのは 2011 年の JPO 主催韓国特許情報セミナーでした。韓国特許庁関係者も「韓国では当然の事実なのに日本で周知されてはいないのですか？」と。

DWPI や CAPlus など出力料金が従量制の高額な商用英語データベース（現在では定額制となっていますが）などは気軽に使えませんでしたので、書誌、抄録情報のみの（クレームは収録していない）無料データベースを予備検索的に使っていました。

最近では Espacenet や WIPO の PATENTSCOPE もアジアの情報なども収録されて英語だけでなく原語検索もできるようになり、PatSnap（商用）や台湾特許庁の Global Patent Search System(GPSS)なども英語と原語でハイブリッド検索できるデータベースとして出現しました。

そのような中、化学系の研究会で昔お世話になった方から「そういえば昔、データベースの収録確認ワードを紹介していましたよね。今でも活用できるのでしょうか」という問い合わせがありました。PC 内にほこりをかぶっていたデータを見つけてお送りしましたが、「今でも活用できるのだろうか」と疑問に思ったと同時に最近はまっている GPSS の収録状況についても確認したいと思い、ほこりを払って使ってみることにしました。

DWPI や CAPlus、Orbit.com などの商用データベースや PATENTSCOPE には抄録やクレームの収録確認コードが用意されていますが、多くの無料データベースにはそのような確認コードはありません。

2009 年に検証した「データベース収録確認ワード」が新たに出願したデータベース、特に英語と原語を共に収録してハイブリッド検索可能となっている台湾特許庁の Global Patent Search System(GPSS)、PATENTSCOPE および、まだわずかですが原語を収録し始めた Espacenet での英語情報と原語情報の収録状況を再検証し、ハイブリッド検索に活用できるか調べてみました。

2．データ収録確認用ワード（2009 年検証）

2009 年当時に検証した収録確認ワードと収録率データの一部を表1に示しました。抄録で示された英語情報を見ながら出現率の高そうな用語をトライ＆エラー的に確認したものです。中国特許庁英語データベース SIPO En と C-PAT から原始的で非効率な方法で求めたものです。（現在、C-PAT は存在しません）

表1．英語抄録収録確認ワードと収録率(2009 年 CN 公開特許：280,990 件)

	invention	method	provide	using	high	system	according	included
SIPO En	218045	131469	117841	106587	59426	55748	48246	48185
	77.6%	46.8%	41.9%	37.9%	21.1%	19.8%	17.2%	17.1%
C-PAT	218036	134839	116246	111764	59652	65914	48332	49840
	77.6%	48.0%	41.4%	39.8%	21.2%	23.5%	17.2%	17.7%
	improved	products	time	operate	surface	direct	position	other
SIPO En	44513	42273	41771	41325	40238	34098	35870	34612
	15.8%	15.0%	14.9%	14.7%	14.3%	12.1%	12.8%	12.3%
C-PAT	44476	48224	42225	42312	43787	39320	37126	40000
	15.8%	17.2%	15.0%	15.1%	15.6%	14.0%	13.2%	14.2%
	types	composition	value	light	element	change	source	quality
SIPO En	22409	22354	21064	18618	16970	15669	15562	14876
	8.0%	8.0%	7.5%	6.6%	6.0%	5.6%	5.5%	5.3%
C-PAT	24402	31784	21262	22293	23908	16492	17607	14939
	8.7%	11.3%	7.6%	7.9%	8.5%	5.9%	6.3%	5.3%
	present	data	layer	power	current	dry	result	film
SIPO En	31260	25863	25817	25656	14898	14755	12765	11370
	11.1%	9.2%	9.2%	9.1%	5.3%	5.3%	4.5%	4.0%
C-PAT	33990	36809	26108	27246	15206	15646	12887	11389
	12.1%	13.1%	9.3%	9.7%	5.4%	5.6%	4.6%	4.1%
	front	seal	way	phase	the	and	to	is
SIPO En	10390	10368	9927	9305	279404	275493	251779	232833
	3.7%	3.7%	3.5%	3.3%	99.4%	98.0%	89.6%	82.9%
C-PAT	10330	10499	10261	11977	280056	276058	280237	232536
	3.7%	3.7%	3.7%	4.3%	99.7%	98.2%	99.7%	82.8%

もちろん、アルファベット 1 文字や「i%」「in%」「inv%」など確認ワードの用語数が少ないものを並べて検索した方が収録率は高くなりますが Espacenet など他のデータベースとの収録率を比較するためには 3 文字以上が必要であり、検索フィールドへの入力文字数も10語まで、という制限もあることから以下の用語を候補として挙げました。

「com% + con% + pro% + inv% + met%」 ⇒ 中国特許庁英語 DB SIPO En 収録率(99.1%)

その当時は、前方一致検索ができないデータベースの場合には、同様に公報頻出用語をトライ＆エラー的に抽出し、以下の用語を使うこととしました。

「invention + method + provide + using + high + system + included + according + improved + products + time + operate + direct + surface + position + other」

⇒ SIPO En 収録率(98.6%)

中国語（簡体字）収録について確認した検索用語と検証結果を表2に示しました。中国特許庁のSIPO CN（中国語データベース）ではSIPO Enのように要約までしか収録していませんので、クレームや全文の収録が100%であるCNIPR（現在の中国版）を使ってクレームや全文用語についてみたものです。

表2. 簡体字収録確認ワードと収録率(2007年CN公開特許)

	一种	设备	提取	内容	发明	方法	使用	技术
AB	162721	22267	6111	3798	148955	94674	41643	24033
	78.1%	10.7%	2.9%	1.8%	71.5%	45.4%	20.0%	11.5%
CL	187822	26147	10604	7084	3840	121561	41629	6028
	90.1%	12.5%	5.1%	3.4%	1.8%	58.3%	20.0%	2.9%
FULL	201186	80416	28616	194509	207808	162226	172819	205285
	96.6%	38.6%	13.7%	93.4%	99.7%	77.9%	82.9%	98.5%
	具有	说明	通过	用户	根据	结果	用于	进行
AB	74413	1142	62401	11850	25700	6789	68541	52074
	35.7%	0.5%	30.0%	5.7%	12.3%	3.3%	32.9%	25.0%
CL	86984	1000	96319	15112	128869	13479	82761	84625
	41.7%	0.5%	46.2%	7.3%	61.9%	6.5%	39.7%	40.6%
FULL	178386	182869	184966	42043	147075	96746	151405	179437
	85.6%	87.8%	88.8%	20.2%	70.6%	46.4%	72.7%	86.1%

収録率が 70%以上のワード部分にはマークしましたが、抄録（要約）収録率が高い用語が必ずしもクレーム収録も高いとは言えませんでした。中国語の場合には1文字でも検索が可能なのでやはり頻出すると思われる用語で各フィールドから確認したのが以下です（表3）。

表3. 中国語収録確認ワードと収録率(2007年CN公開特許:208,345件)

	TI		AB		CL		FULL	
种	40,644	20%	168,123	81%	191,127	92%	206,427	99%
一	41,697	20%	184,995	89%	202,963	97%	207,496	100%
二	4,170	2%	41,757	20%	92,452	44%	156,058	75%
1	911	0%	40,311	19%	207,786	100%	204,541	98%
发	8,698	4%	161,049	77%	67,389	32%	29,117	14%
的	113,202	54%	206,940	99%	207,433	100%	208,326	100%
用	46,066	22%	157,737	76%	164,681	79%	197,441	95%
和	33,480	16%	140,020	67%	162,848	78%	200,280	96%
以	8,219	4%	126,039	60%	155,027	74%	184,773	89%
中	14,884	7%	132,952	64%	182,859	88%	203,983	98%
处	7,294	4%	40,691	20%	73,905	35%	96,917	47%
上	2,764	1%	99,320	48%	147,297	71%	194,679	93%
下	1,649	1%	65,921	32%	132,047	63%	185,060	89%
含	4,736	2%	39,744	19%	70,602	34%	73,684	35%
于	22,394	11%	138,297	66%	192,926	93%	197,463	95%
在	6,302	3%	147,861	71%	198,540	95%	206,721	99%
具	11,024	5%	86,301	41%	101,715	49%	16,299	8%
法	97,914	47%	99,020	48%	126,324	61%	57,710	28%
量	5,658	3%	59,518	29%	86,391	41%	99,181	48%
式	12,134	6%	41,987	20%	73,401	35%	91,109	44%
性	12,467	6%	75,976	36%	78,421	38%	167,519	80%
剂	12,202	6%	36,253	17%	50,212	24%	60,506	29%

その結果、簡体字データベース(CNIPR)の要約、クレーム、全文の収録確認ワードとしては以下の用語で問題ないことがわかりました。

2007 年 CN 公開特許:208,345 件

AB　(种 or 一 or 发 or 的 or 用 or 在)208,332 件(≒100%)

CL　(种 or 一 or 1 or 的 or 中 or 于 or 在)　208,325 件(≒100%)

FULL (种 or 一 or 1 or 的 or 用 or 和 or 中 or 上 or 于 or 在)　208,331(≒100%)

台湾特許の繁体字での収録確認ワードも同様に検証した結果を表4に示しました。

表4. 繁体字収録確認ワードと収録率(2007 年 TW 公開/公告特許)

公開	、	。	種	發	一	1(半)	的
AB	21934	46926	37319	30764	42795	7052	32903
	47%	100%	79%	65%	91%	15%	70%
CL	31967	46728	46459	12918	46726	46755	38995
	68%	99%	99%	27%	99%	100%	83%
FULL	45884	46753	46122	46464	46495	44120	45954
	98%	100%	98%	99%	99%	94%	98%
公告	、	。	種	發	一	1(半)	的
AB	9954	22202	17869	13554	20403	3149	15548
	45%	100%	80%	61%	92%	14%	70%
CL	15292	22198	22148	18886	22218	22180	18255
	69%	100%	100%	85%	100%	100%	82%
FULL	568	22214	578	583	584	541	579
	3%	100%	3%	3%	3%	2%	3%
公開	用	以	具	其	之	本	含
AB	29931	34314	22462	30433	41788	30171	17595
	64%	73%	48%	65%	89%	64%	37%
CL	36320	43168	32897	46543	46023	5488	30757
	77%	92%	70%	99%	98%	12%	65%
FULL	46454	46487	45339	46417	46456	46484	37369
	99%	99%	96%	99%	99%	99%	80%
公告	用	以	具	其	之	本	含
AB	14194	16729	11143	14302	19920	13375	7683
	64%	75%	50%	64%	90%	60%	35%
CL	18591	20743	16747	22139	21994	18652	14121
	84%	93%	75%	100%	99%	84%	64%
FULL	584	584	569	583	583	584	17265
	3%	3%	3%	3%	3%	3%	78%

2007 年公開特許／公告特許:46986 件／22218 件(TWPAT)

AB　一 or 種 or 發 or 的 or 用 or 以 or 含 or 具 or 其 or 之 or 本 or 。

⇒ 公開 46981 件(≒100%)／公告 22218 件(=100%)

CL　一 or 1 or 種 or 的 or 用 or 以 or 下 or 含 or 具 or 其 or 之 or 。

⇒ 公開 46759 件(≒100%)／公告 22218 件(=100%)

FULL　一 or 1 or 種 or 發 or 的 or 用 or 以 or 下 or 具 or 其 or 之 or 本 or 。

⇒ 公開 46766 件(≒100%)／公告 22214 件(≒100%)

句読点も収録確認ワードとして使えることもわかりました。

3. データベース収録確認ワードの再検証

10年以上前に検証したデータベース収録確認ワードが今でも有用なのか、を確認することが本稿の目的です。無料、商用を問わずデータベースの多彩な機能に目を奪われる前に「収録内容を調査前に把握しておくことが最も重要」として調査に当たってきました。最近では PATENTSCOPE や Espacenet に CN, TW, KR などの東アジアだけでなく ASEAN 各国からも英語情報と原語情報が収録されてハイブリッド検索できるようになってきましたが、英語や原語情報の収録状況は確認しないまま過ごしてきました。(単に発行日からの出願推移については ASEAN 各国も含め、検索 Tips「【別表】東アジアおよび ASEAN6か国の公開特許収録状況」で紹介しています [1)]。

2, 3年前には台湾特許庁データベース TWPAT とは別に Global Patent Search System (GPSS) も現れ、TW 以外のワールドワイドな情報が収録されています。TWPAT でも発明の名称に英語情報が入るようになってからは、英語情報を元に台湾の繁体字を検索用言語として抽出してきましたが、同じ台湾情報でも GPSS の方が何となく英語情報収録が多いように感じながら、「ではどのぐらいの収録率」なのかについては確認しないまま使っています。

昔の収録確認ワードを使って英語および原語収録を確認してもいいのですが、GPSS ではアルファベット1文字からでも検索できたりするので改めて各データベースの収録確認ワードについて検証してみました。

1)英語収録確認ワード

10年前の英語収録確認ワードを用いて GPSS で2020年公開特許の収録を確認してみると、以下のように 97%となりました。この程度で満足すべきかを検証してみたいと思います。

ID=2020 and (com* or con* or pro* or inv* or met*)@ab　45308 件/46569 件(97%)

アルファベット1文字から検索できる、ということであれば A～Z までを全部並べて検索すればよいのですが、ワード文字数が少ないと英訳文が収録されていなくても公報中の略称、単位などのアルファベットなども拾ってしまいます。

(a* or b* or c* or d* or e* or f* or g* or h* or i* or j* or k* or l* or m* or n* or o* or p* or q* or r* or s* or t* or u* or v* or w* or x* or y* or z*)

英語か原語いずれかが収録されていればよい、とすればそれでもいいのですが、英訳情報がどの程度収録されているのかを確認する場合には不充分となります。

また、GPSS で「ID=2020 not (a* or b* or ・・・ or y* or z*)@ti」と、「発明の名称」中に英語情報を含まないものとして not 演算しても「Semiconductor device」「POLISHING APPARATUSES」など英訳が付与された「発明の名称」が 19036 件も抽出されますので、アルファベット1文字検索では不具合があるものと思われます。これは TWPAT でも同様で、発明の名称や要約中に英訳が付与されたものも抽出されます。

そこで、公報原語中の「PC, PE, ABS, Xn, R(1～4), a), A1 ～A16, ℃」などの略称、単位のアルファベットも避けつつ英語収録を確認できるワードとしてアルファベット3文字について検証しました。アルファベット2文字でも略称や単位のアルファベットとの重複は避けられず、また、データベースによっては「Abstracts」「Claims」などのフィールド名があるのみで中身がない、というものもありますので「AB」「CL」なども避けて、GPSS で台湾公開特許中の英語情報2020年発行分から 1000 件以上収録する3文字ワードを抽出してみました。その結果が表6です。

（英語情報がほぼ100％収録されている日本版 CNIPR が利用できれば中国公開特許からの検証でもいいのですが、無料の GPSS で検証しました。）

表6. アルファベット3文字ワードの要約中出現率（TW2020 年公開特許：46829 件）

	収録数	収録率		収録数	収録率		収録数	収録率		収録数	収録率
acc*	1841	4%	dis*	21280	45%	loa*	1065	2%	ret*	1566	3%
aci*	1738	4%	eac*	8537	18%	loc*	4256	9%	sea*	2023	4%
act*	3011	6%	ele*	11027	24%	met*	18805	40%	sec*	16046	34%
ada*	1003	2%	ena*	1338	3%	mor*	5389	12%	sem*	3263	7%
add*	3322	7%	enc*	1637	3%	pho*	2136	5%	sep*	1808	4%
adj*	4095	9%	end*	4727	10%	pla*	7050	15%	ser*	2566	5%
aro*	1356	3%	exa*	1320	3%	plu*	8949	19%	str*	9015	19%
arr*	5036	11%	ext*	5836	12%	pol*	4449	10%	sub*	9889	21%
ass*	4536	10%	fil*	5672	12%	por*	6860	15%	sup*	6171	13%
can*	10271	22%	fir*	17129	37%	pos*	6563	14%	sur*	10064	21%
com*	24810	53%	hav*	9720	21%	pre*	19693	42%	ther*	11773	25%
con*	29782	64%	inc*	27504	59%	pro*	30985	66%	tra*	10128	22%
cos*	1003	2%	inv*	14477	31%	rea*	4510	10%	tre*	3031	6%
def*	3566	8%	lay*	7534	16%	rec*	9405	20%	use*	10212	22%
det*	7776	17%	liq*	2069	4%	rep*	3261	7%			

そこで、ここで得られた59件のワード集合（以下、英語ワード集合と略）を英語収録ワードとして

以下のようにして検索し、GPSS と TWPAT の台湾特許、および GPSS における中国や韓国公開特許について発明の名称、要約、クレームの出現率を求めて表7に示しました。

ID=2020 and (acc* or aci* or act* or ada* or add* or adj* or aro* or arr* or acc* or ass* or can* or com* or con* or cos* or def* or det* or dis* or eac* or ele* or ena* or enc* or end* or exa* or ext* or fib* or fil* or fir* or hav* or inc* or inv* or loa* or loc* or met* or mor* or pho* or pla* or plu* or pol* or por* or pre* or pro* or rea* or rec* or rep* or ret* or sea* or sec* or sem* or sep* or ser* or str* or sub* or sup* or sur* or ther* or tra* or tre* or use*)@ab

アルファベット3文字の全候補すべてを並べて抽出した方が確実なようですが、データベースへの入力制限で（入力できない、文字数が一定数以上では検索時間が異常に長くなったり、エラーとなる）、2020年 TW 公開特許1000件以上出現するワードに限定して検索しました。ちなみに500件～1000件未満のアルファベット3文字ワード185件を加えても収録率への影響は1％未満です。

アルファベット3文字ワードの収録率を検証する中で「are be do for have he is me of on the that this to」などは、発明の名称、要約、クレームのいずれからも GPSS では検索できません（検索件数0件）。検索には利用できないストップワードのようです。

表7. 英語ワード集合による収録率確認

	GPSS			TWPAT		
	TI	AB	CL	TI	AB	CL
TW	39613	46510	4153	17232	29614	2856
	85%	99%	9%	37%	63%	6%
CN	942884	1180165	50938			
	62%	78%	3%			
KR	94462	73934	0			
	63%	50%	0%			

台湾特許庁データベース TWPAT および GPSS のいずれもクレームには原則として英訳は付与されていませんから、英語ワード集合をここまで並べても10％未満のノイズが入ることは避けられないようです。その程度の手法でしかないとあきらめています。データベースへの英訳付与の概略を知る程度だと思った方がよさそうです。

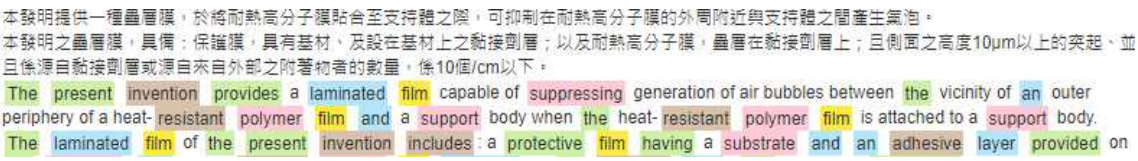
その前提で見てみると、従来からの台湾特許庁データベース TWPAT に比べ、GPSS の発明の名称や要約への英語付与率は高いことがわかります。また、GPSS では「韓国特許の全文は収録していない」との記述がありますが、公告特許も含め、現時点では韓国特許（実案を含む）のクレーム

も収録されていないことが確認できました。

表7の情報を裏付けるデータとして、図1に同一案件でも GPSS の要約には英語が付与されているが、TWPAT には英語の付与がないケースを示しました。

図1. GPSS と TWPAT 要約中の英語付与の違い

GPSS(AN:109128576) ・・要約に英語付与



TWPAT(AN:109128576) ・・要約に英語付与なし

本發明提供一種疊層膜，於將耐熱高分子膜貼合至支持體之際，可抑制
在耐熱高分子膜的外周附近與支持體之間產生氣泡。
本發明之疊層膜，具備：保護膜，具有基材、及設在基材上之黏接劑
層；以及耐熱高分子膜，疊層在黏接劑層上；且側面之高度10μm以上的

また、「英語ワード集合での検索」では以下のように要約やクレーム、全文中に物質名や化合物名のみが英訳されていたり(図2)、アルファベット略号が含まれるものもあります(図3)。

図2. 要約全体に英訳はないが特定の用語のみ英訳が付与されているもの

本發明乃是一種益生菌茶葉，其含有下述重量份數比的組分：益生菌及茶葉混合物，其中該益生菌為 Bacillus coagulans BC208，該益生菌之菌數約大於等於104 CFU／g至小於等於10000 CFU／g之間，並分別包裝於隔絕袋中保存。其中，該茶葉為不發酵茶、半發酵茶及全發酵茶至少一種。該茶葉為紅茶、綠茶、青茶、烏龍茶及普洱茶至少一種。該隔絕袋為PE塑膠袋、鋁箔袋或電鍍鋁箔袋。

図3. 化合物の示性式など

[式中，R a8 為可具有鹵素原子的烷基、氫原子或鹵素原子；Z a1 為單鍵或*-(CH2)h3- CO -L 54 -，h3為1～4的整數；L 51 、L 52 、L 53 及L 54 分別獨立地為-O-或-S-；s1為1～3的整數；s1'為0～3的整數；R 1 為氫原子或甲基；A 1 為單鍵或*- CO -O-；R 2 為鹵素原子、羥基、鹵代烷基或烷基；mi為1～3的整數；ni為0～4的整數，其中，mi+ni≦5。]

図4. アルファベット1文字検索では全文中の記号等を拾う。

的前端部31 a，是位於帶狀體30的前端部。即，前片31的前端部31 a，是位於後側竿體4的前端部。前片31的後端部31 b，是位於帶狀體30的後端部。即，前片31的後端部31 b，是位於後側竿體4的後端部。但是，前片31的後端部31 b是位於比後側竿體4的後端部更前側的位置也可以。後片32的前端部32 a，是位於後側竿體4的中間部附近。後片32的後端部32 b，是位於帶狀體30的後端部。即，後片32的後端部

2)データベースの英訳収録状況

東アジア各国特許庁から発行される特許情報は各国言語ですが、それらが英訳されて EPO に送られたものは DOCDB として Espacenet や商用英語データベースの基となり、WIPO に送られたものは PATENTSCOPE として収録されています。GPSS に収録されている情報も含めて上記で検証したアルファベット3文字英語ワード集合を用いて各データベースの英訳情報を見てみました。

a)中国特許英語情報

日本版 CNIPR の英語情報は要約、クレームとも問題なく100%収録していますが、GPSS では2008年発行以前の公開情報の要約英語収録は極めて悪いことが確認できました。表8の GPSS クレームの4%未満の英語収録は、既に紹介したように物質名や化合物名の部分的な英訳によるものでクレームそのものが英訳されている訳ではありません。

表8. 中国特許情報の英訳

	CNIPR					GPSS				
PD	CN公開	AB	収録率	CL	収録率	CN公開	AB	収録率	CL	収録率
2000	38,296	38,282	100.0%	38,281	100.0%	38,296	17,191	45%	508	1%
2001	50,364	50,348	100.0%	50,349	100.0%	50,365	20,671	41%	1,620	3%
2002	58,984	58,971	100.0%	58,966	100.0%	58,984	27,050	46%	2,002	3%
2003	77,472	77,444	100.0%	77,449	100.0%	77,472	37,819	49%	2,583	3%
2004	93,944	93,919	100.0%	93,921	100.0%	93,944	44,617	47%	3,218	3%
2005	155,446	155,419	100.0%	155,415	100.0%	155,447	74,636	48%	4,992	3%
2006	172,424	172,403	100.0%	172,376	100.0%	172,428	108,147	63%	5,265	3%
2007	208,345	208,318	100.0%	208,279	100.0%	208,348	162,379	78%	6,476	3%
2008	241,182	241,170	100.0%	241,124	100.0%	241,182	177,129	73%	7,172	3%
2009	281,006	280,995	100.0%	280,945	100.0%	281,007	280,797	100%	8,189	3%
2010	315,836	315,823	100.0%	315,746	100.0%	315,840	315,724	100%	9,403	3%
2011	368,434	368,427	100.0%	368,330	100.0%	368,434	344,161	93%	11,232	3%
2012	543,296	543,284	100.0%	543,124	100.0%	543,297	543,154	100%	15,099	3%
2013	632,585	632,567	100.0%	632,412	100.0%	632,585	632,426	100%	17,196	3%
2014	777,335	777,313	100.0%	777,121	100.0%	777,336	777,215	100%	20,809	3%
2015	955,341	955,320	100.0%	955,061	100.0%	955,341	955,205	100%	24,310	3%
2016	1,045,796	1,045,759	100.0%	1,045,490	100.0%	1,045,741	1,045,498	100%	25,825	2%
2017	1,270,655	1,270,646	100.0%	1,270,172	100.0%	1,270,363	1,270,174	100%	34,801	3%
2018	1,575,621	1,575,603	100.0%	1,575,237	100.0%	1,575,438	1,575,232	100%	39,185	2%
2019	1,532,682	1,532,676	100.0%	1,532,065	100.0%	1,532,539	1,531,697	100%	47,874	3%
2020	1,517,110	1,517,107	100.0%	1,514,168	99.8%	1,517,011	1,180,165	78%	50,938	3%

b)台湾特許英語情報

表9から、従来から存在する台湾特許庁 TWPAT の要約英訳が予想外に低く、GPSS の収録が極めて高いことがわかりました。台湾特許庁データベースにおけるクレームではいずれもクレーム

では物質名、化合物名からの収録数ですが、図5に示すように例示物質名の英訳を網羅的に英訳してくれているので用語辞書として関連の用語を集める場合には効率的で助かっています。

いずれにしても物質名や化合物名の英訳を期待しなければクレーム中から英語検索しようと思わない方がいいでしょう。

表9. 台湾特許情報の英訳

		TWPAT				GPSS			
PD	TW公開	AB	収録率	CL	収録率	AB	収録率	CL	収録率
2003	8,194	151	2%	372	5%	8,071	98%	598	7%
2004	28,927	739	3%	2,108	7%	28,245	98%	3,052	11%
2005	41,439	1,314	3%	3,445	8%	39,183	95%	4,969	12%
2006	44,783	1,166	3%	3,678	8%	42,385	95%	5,232	12%
2007	46,986	1,047	2%	3,811	8%	44,997	96%	5,359	11%
2008	50,141	1,098	2%	4,228	8%	49,664	99%	5,992	12%
2009	52,618	989	2%	3,903	7%	52,160	99%	5,530	11%
2010	44,962	805	2%	3,398	8%	44,488	99%	4,823	11%
2011	46,157	692	1%	3,123	7%	45,568	99%	4,521	10%
2012	51,592	741	1%	3,063	6%	50,849	99%	4,462	9%
2013	52,126	655	1%	3,064	6%	51,414	99%	4,528	9%
2014	48,719	599	1%	2,899	6%	48,011	99%	4,202	9%
2015	47,367	32,370	68%	2,828	6%	46,585	98%	4,168	9%
2016	44,356	29,265	66%	2,902	7%	43,743	99%	4,148	9%
2017	43,677	28,197	65%	2,705	6%	43,209	99%	4,007	9%
2018	44,071	27,268	62%	2,704	6%	43,771	99%	3,936	9%
2019	47,989	30,068	63%	2,912	6%	47,697	99%	4,237	9%
2020	46,837	29,614	63%	2,856	6%	46,510	99%	4,153	9%

図5. クレーム中の英語用語

6.如申請專利範圍第5項所述之組成物，其中組分B選自由以下項所組成之除草劑群組： B1　氯草酮、吡草酮、苄草唑、異惡唑草酮、吡柔蘇氟酮、苯吡唑草酮、吡唑特、硝磺草酮、磺草酮、苯并雙環酮、特呋三酮、環磺酮、氟吡草酮、2-[2-(3,4-二甲氧基苯基)-6-甲基-3-側氧基-嗒咿-4-羰基]-環己烷-1,3-二酮和具有式 (II) 之化合物；B2　氨胺吡啶酸（aminopyralid）、二氯吡啶酸（clopyralid）、毒莠定（picloram）、綠草定（triclopyr）、氟草煙（fluroxypyr）、二氯喹啉酸（quinclorac）、氯甲喹啉酸（quinmerac）、2,4-D、2-4-DB、氯甲醯草胺（clomeprop）、mecroprop、滴丙酸（dichlorprop）、MCPA、MCPB、草滅平（chloramben）、麥草畏（dicamba）、TBA、氯氟吡啶酯（florpyrauxifen-benzyl）、氟氯吡啶酯（halauxifen- methyl）和草除靈； B3　禾草滅（alloxydim）、丁苯草酮（butroxydim）、烯草酮（clethodim）、噻草酮（cycloxydim）、環苯草酮（profoxydim）、稀禾定（sethoxydim）、哌喃草酮（tepraloxydim）、肟草酮（tralkoxydim）、唑啉草酯（pinoxaden）、炔草酯（clodinafop- propargyl）、氰氟草酯（cyhalofop-butyl）、禾草靈（diclofop- methyl）、精惡唑禾草靈（fenoxaprop-P-ethyl）、精惡氟禾草靈（fluazifop-P-butyl）、精吡氟氯禾靈（haloxyfop-P- methyl）、惡唑醯草胺（metamifop）、喹草酯（propaquizafop）、精噻禾靈-甲基（quizalofop-P- methyl）、噻禾糠酯（quizalofop-P-tefuryl）；以及 B4　草乃敵（diphenamid）、萘丙胺（naproanilide）、敵草胺（napropamide）、氟噻草胺（flufenacet）、苯噻醯草胺（mefenacet）、乙草胺（acetochlor）、甲草胺（alachlor）、丁草胺（butachlor）、二甲草胺（dimethachlor）、二甲酚草胺（dimethenamid）、吡唑草胺（metazachlor）、S-異丙甲草胺（S- metolachlor）、烯草胺（pethoxamid）、丙草胺（pretilachlor）、撲草胺（propachlor）、異丙草胺（propisochlor）、甲氧噻草胺（thenylchlor）、三唑醯草胺（ipfencarbazone）、四唑醯草胺（fentrazamide），以及莎稗磷（anilofos）、唑草胺（cafenstrole）、吡咯磺隆（pyroxasulfone）

c)PATENTSCOPE および Espacenet における東アジア各国の英語要約収録状況

台湾特許は PATENTSCOPE に収録されていないので中国と韓国特許に限定して Espacenet と共に英語収録状況を確認しました。英語ワード集合で要約、クレーム、全文の収録も確認しました。

PATENTSCOPE の中国特許は書誌情報自体が2013年以降、収録も悪く（2019年調査と同様）、英語要約は2012年以降の収録が悪いことが確認できました（表10）。韓国特許の書誌収録が KIPRIS より多いことは不明ですが、要約は2011年以降、あまりよくありません（表11）。

PATENTSCOPE では完璧とは言えないまでも59の英語ワード集合で収録率を求めましたが、Espacenet では検索入力制限のため「com* or con* or dis* or inc* or pro*」の5ワードによるもので収録率も不充分なものとなっています。そこで次節に紹介する具体的なタームで収録比較しました。

表10. PATENTSCOPE、Espacenet の中国特許収録

	CNIPR	PATENTSCOPE			Espacenet		
PD	CN公開	CN公開	AB	収録率	CN公開	AB	収録率
2000	38,296	38,296	38,274	100%	38,016	14,599	38%
2001	50,364	50,365	50,349	100%	50,029	20,175	40%
2002	58,984	58,984	58,970	100%	58,536	26,600	45%
2003	77,472	77,472	77,443	100%	77,013	37,460	49%
2004	93,944	93,943	93,915	100%	92,861	45,418	49%
2005	155,446	155,445	155,412	100%	154,278	81,176	53%
2006	172,424	172,428	172,391	100%	170,477	118,426	69%
2007	208,345	208,348	208,274	100%	205,611	173,956	85%
2008	241,182	241,183	241,051	100%	238,692	204,233	86%
2009	281,006	281,005	280,727	100%	278,109	276,114	99%
2010	315,836	315,844	315,168	100%	313,713	311,777	99%
2011	368,434	368,452	366,007	99%	366,236	341,557	93%
2012	543,296	542,959	312,861	58%	540,355	537,429	99%
2013	632,585	594,187	190,658	32%	629,386	626,456	100%
2014	777,335	751,525	662,250	88%	773,919	770,606	100%
2015	955,341	845,386	575,392	68%	949,450	946,167	100%
2016	1,045,796	734,845	353,658	48%	1,036,897	1,033,427	100%
2017	1,270,655	807,292	763,778	95%	1,262,400	1,259,066	100%
2018	1,575,621	1,595,336	1,432,545	90%	1,565,145	1,561,405	100%
2019	1,532,682	814,759	804,262	99%	1,524,462	786,728	52%
2020	1,517,110	171,626	167,860	98%	1,508,908	1,144,899	76%

表11. PATENTSCOPE、Espacenet の韓国特許収録

	KIPRIS	PATENTSCOPE			Espacenet		
PD	KR公開	KR公開	AB	収録率	KR公開	AB	収録率
2000	77,495	97,283	95,799	98%	77,289	68,804	89%
2001	114,264	122,774	108,620	88%	113,688	73,034	64%
2002	97,487	103,539	102,141	99%	97,123	80,176	83%
2003	97,909	101,973	95,357	94%	97,486	76,037	78%
2004	110,582	115,168	114,647	100%	109,869	85,645	78%
2005	123,492	130,240	130,030	100%	122,657	43,968	36%
2006	135,946	157,370	157,272	100%	134,929	41,221	31%
2007	122,581	162,915	162,877	100%	121,280	88,079	73%
2008	122,581	146,537	146,514	100%	113,295	112,618	99%
2009	133,121	147,622	147,568	100%	131,953	119,926	91%
2010	139,151	152,886	140,035	92%	137,736	124,952	91%
2011	140,131	157,823	128,503	81%	138,819	93,214	67%
2012	139,463	159,842	128,954	81%	138,024	108,107	78%
2013	141,194	169,094	139,542	83%	139,480	110,914	80%
2014	148,110	185,559	120,431	65%	145,989	103,532	71%
2015	146,108	178,769	127,023	71%	144,121	106,672	74%
2016	150,599	177,638	131,085	74%	148,600	111,818	75%
2017	143,454	180,263	129,065	72%	141,395	75,377	53%
2018	138,541	175,567	138,001	79%	136,632	99,827	73%
2019	143,824	182,062	122,173	67%	141,737	66,310	47%
2020	146,008	191,134	93,509	49%	143,810	56,601	39%

また、PATENTSCOPE のクレームや全文中からの英語ワード集合での抽出では、各年代数1000 件の収録があり、収録率も数%を示していますが、やはり物質名、化合物名等の部分訳であり、クレームや全文の英訳は収録されていないことも確認しました（図9，図10）。

図6. クレームや全文などを収録していない場合の PATENTSCOPE 詳細情報画面 TOP

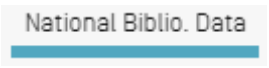


図7. クレーム、全文などを収録している場合の詳細情報画面 TOP

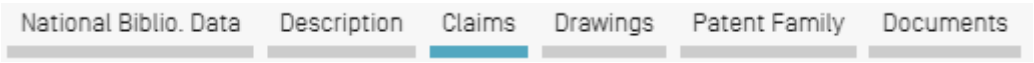


図8. 要約は書誌情報（National Biblio Data）中に表示される。

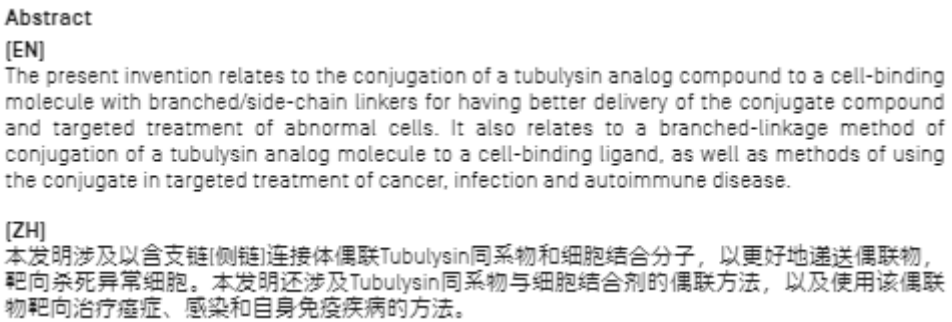


図9. クレームや全文中から英語ワード集合で抽出したものは物質名など部分訳を拾う（CN）

NH_2，$OS(O)_2OH$，$OS(O)_2OR_1$，$CH_2S(O)_2OR_1$，Ar，ArR$_1$，ArOH，$ArNH_2$，ArSH，$ArNHR_1$或$(Aa)_{q1}$；$(Aa)_{q1}$为含有相同或不同序列的天然或非天然氨基酸的肽；X_1和X_2独立地为O，CH_2，S，S(O)，NH,$N(R_1)$，$^+NH(R_1)$，$^+N(R_1)(R_2)$，C(O)，OC(O)，OC(O)O，OC(O)NH，NHC(O)NH；

Y_2为O，NH，NR_1，CH_2，S，NHNH，Ar；p_1，p_2和p_3独立地是0-100，但是不同时为0；q_1和q_2独立地是0-24；

R_1，R_2，R_3和R_3'独立地为H，C_1-C_8烷基；C_2-C_8杂烷基或杂环；C_3-C_8芳基、芳基烷基、环烷基、烷基环烷基、杂环烷基、杂烷基环烷基，碳环、或烷羰基；

図10. クレームや全文中から英語ワード集合で抽出したものは物質名など部分訳を拾う(KR)

청구항 16

제14항에 있어서,

상기 링커는 브랜칭 유닛(Branching unit, BR), 커넥션 유닛(connection unit), 또는 바인딩 유닛(Binding Unit)을 포함하고, 상기 커넥션 유닛은 약물과 브랜칭 유닛 또는 바인딩 유닛을 연결하고, 바인딩 유닛 또는 커넥션 유닛은 항체와 연결되는 리간드-약물 접합체.

政治的な関係から台湾特許情報は PATENTSCOPE には収録されていませんが、Espacenet には収録されています。そこで参考情報として台湾の Espacenet の要約収録状況を表12に示しました。収録ワードは中国、韓国と同様、「com* or con* or dis* or inc* or pro*」の5ワードです。

表12. Espacenet の台湾特許収録

	TWPAT	Espacenet		
PD	TW公開	TW公開	AB	収録率
2003	8,194	8,401	8,247	98%
2004	28,927	29,652	29,146	98%
2005	41,439	41,049	40,393	98%
2006	44,783	43,880	43,152	98%
2007	46,986	44,605	43,916	98%
2008	50,141	49,383	48,791	99%
2009	52,618	52,087	51,477	99%
2010	44,962	44,585	44,146	99%
2011	46,157	45,611	45,194	99%
2012	51,592	50,951	50,536	99%
2013	52,126	51,453	51,031	99%
2014	48,719	48,328	47,969	99%
2015	47,367	47,228	46,848	99%
2016	44,356	45,228	44,864	99%
2017	43,677	44,553	44,177	99%
2018	44,071	44,354	41,760	94%
2019	47,989	47,221	36,612	78%
2020	46,837	41,245	38,606	94%

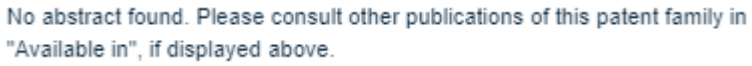
中国、韓国の Espacenet 収録状況からするとかなり収録率も高く、ほぼすべてのレコードに要約が収録されていそうです。次項の具体的検索タームの収録検証でも確認したいと思います。

公開日から求めたEspacenetの2003年、2004年の公開数が台湾特許庁TWPATより多くなっている理由については未検証です。

ちなみに、Espacenetでの以下の5個の英語ワード集合のNOT演算から得られる(要約が収録されていない)レコードを確認したところ、英訳要約が存在しなければほとんどで図11に示すようなアナウンスが表示されますので(極めてわずか)、ほぼ英訳要約は存在するものと思われます。

(pn=TWA AND pd=2020) NOT ab=(com* OR con* OR dis* OR inc* OR pro*)

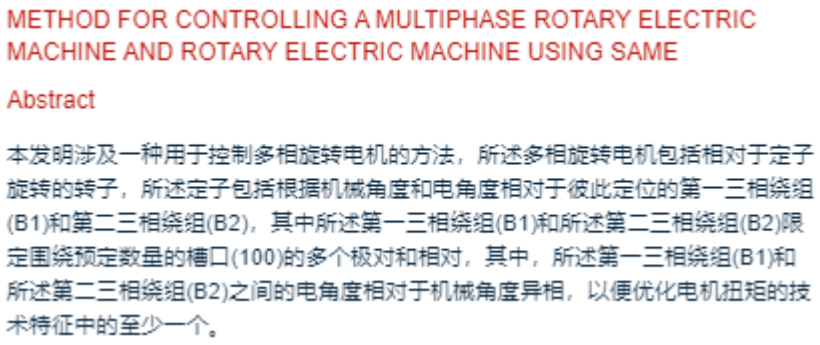
図11. 要約が未収録である、との表示



中国や韓国のEspacenet要約の英訳未収録についても同様に確認したところ、図12のように英語要約があるものは見つかっていません。案外、この5ワードでそこそこカバーできているのかもしれません。

(pn=CNA AND pd=2020) NOT ab=(com* OR con* OR dis* OR inc* OR pro*)

図12. 要約は収録されているが原語のみで英訳要約はない。



d) 検索タームによる東アジア各国の英語要約収録状況

収録検証用英語ワード集合に制限のあるEspacenetの英語要約収録状況を別の観点から確認するために出願人名、IPCなどの書誌情報と共に要約中の英語ワードの出現数を調べました。

Espacenetにおける中国、台湾、韓国の書誌収録については、既に「Espacenetで東アジアの特許調査(その1)」[2)]で紹介していますが、単に発行日からの検索であり、要約やクレームの収録については触れていません。

CNIPRとWIPS GLOBALの商用データベースとも比較してみました(表13)。

表13. 要約中の英語用語(laminate*)の確認

		2000年	2005年	2010年	2015年	2020年
CN	CNIPR	324	1018	1679	3645	5952
	GPSS	52	224	1681	3657	4546
	PS	324	1018	1678	2159	1088
	Espace	52	284	1674	3615	4571
	WIPS GL	324	1018	1679	3660	6087
TW	TWPAT		4	3	407	250
	GPSS		488	510	941	814
	Espace		525	516	938	747
	WIPS GL		501	517	959	829
KR	KIPRIS	19	37	45	27	21
	GPSS	561	1020	1748	1196	628
	PS	680	1055	1889	1190	802
	Espace	410	418	1680	1051	524
	WIPS GL	566	1032	1974	2017	2766
	KPA AB	706	1235	2070	1263	695

Espacenet を含み、5種のデータベースで発行年ごとの公開特許の存在数です。各データベースの略称は次の通りです。

CNIPR：日本版 CNIPR（商用）

GPSS：台湾特許庁 Global Patent Search System

PS：WIPO PATENTSCOPE

Espace：EPO 新 Espacenet

WIPS GL：WIPS GLOBAL PATENT（商用）

TWPAT：台湾特許庁データベース

KIPRIS：韓国特許情報院 KIPRIS

KPA：韓国特許情報院英語データベース

マークした部分は他のデータベースに比べ異常な数値を示すものです。差分の検証はしていません（宿題です）。台湾の TWPAT の収録数が少ない点は、前にも触れましたように GPSS に比べ要約英訳が少ないためです。

KIPRIS の要約収録数も異常に低くなっていますが、発明の名称の英訳に比べ要約の英訳率は低いようです。2020 年発行の公開特許では「laminate」の発明の名称では 823 件の英訳が存在するのに要約は 21 件です。他の用語「“3D print”」でも発明の名称 136 件に対し、要約は 1 件となっています。

KIPRIS には英語要約までを収録した KPA というデータベースもありますが、本稿ではハイブリッド検索を対象に英訳収録率について検証しているため対象にしていません。英語情報のみから韓国特許を予備的に検索するには KPA は KIPRIS より有効です。

WIPS GLOBAL でも 2015 年、2020 年の韓国特許の要約数が突出している理由は未検証です。

要約中の英訳収録率とは関係ありませんが、用語以外での各データベースの収録状況を参考表として示しました。

参考表1. 出願人収録比較

		2000年	2005年	2010年	2015年	2020年
CN	CNIPR	31	1710	3047	4051	7069
	GPSS	30	1702	3019	4005	5859
	PS	33	1707	3713	1504	3028
	Espace	31	1726	3019	3992	6226
	WIPS GL	31	1711	3019	4006	7068
TW	TWPAT		664	192	415	1365
	GPSS		665	193	415	1365
	Espace		644	196	419	1338
	WIPS GL		665	193	415	1365
KR	KIPRIS	9500	11810	6442	7237	5799
	GPSS	7071	11781	6922	7359	5789
	PS	12973	13169	6919	6973	3573
	Espace	9541	11767	6835	5973	5718
	WIPS GL	9500	11810	6455	7237	5799

それぞれ各国における 2020 年の TOP 出願人です。

CN：HUAWEI TECH*

TW：TAIWAN SEMICONDUCTOR

KR：SAMSUNG ELECTRONICS

参考表2. IPC(B32B-027*)収録比較

		2000年	2005年	2010年	2015年	2020年
CN	CNIPR	169	461	868	4216	5177
	GPSS	267	755	772	4219	5177
	PS	126	411	880	3800	954
	Espace	263	780	853	4203	5146
	WIPS GL	244	687	868	4216	5177
TW	TWPAT		121	270	724	746
	GPSS		116	270	724	746
	Espace		276	279	723	715
	WIPS GL		300	278	725	745
KR	KIPRIS	240	403	730	915	1142
	GPSS	338	455	625	798	1140
	PS	222	260	650	921	1306
	Espace	338	449	609	815	1123
	WIPS GL	264	434	784	1018	1142

さらに、参考情報として WIPS GLOBAL で中国、台湾、韓国の公開特許収録率を Espacenet で使用した5ワード（要約①）とその他データベースで使用した59ワード（要約②）について検証した結果を参考表3として示しました。

参考表3. WIPS GLOBAL の英語ワード集合の違いによる収録率

	2000年	2005年	2010年	2015年	2020年
CN公開日	38296	155446	315836	955341	1517011
CN要約①	37110	151661	313856	951840	1514029
	97%	98%	99%	100%	100%
CN要約②	38290	155425	315825	955334	1517005
	100%	100%	100%	100%	100%
TW公開日		41438	44962	47367	46828
TW要約①		40480	44468	46983	46418
		98%	99%	99%	99%
TW要約②		41093	44916	47365	46824
		99%	100%	100%	100%
KR公開日	77504	123498	139151	146108	146030
KR要約①	77379	123017	138552	144302	143864
	100%	100%	100%	99%	99%
KR要約②	77504	123495	139146	146079	145992
	100%	100%	100%	100%	100%

要約①:(com* or con* or dis* or inc* or pro*).AB.

要約②:(acc* or aci* or act* or ada* or add* or adj* or aro* or arr* or acc* or ass* or can* or com* or con* or cos* or def* or det* or dis* or eac* or ele* or ena* or enc* or end* or exa* or ext* or fib* or fil* or fir* or hav* or inc* or inv* or loa* or loc* or met* or mor* or pho* or pla* or plu* or pol* or por* or pot* or pre* or pro* or rea* or rec* or rep* or ret* or sea* or sec* or sem* or sep* or ser* or str* or sub* or sup* or sur* or ther* or tra* or tre* or use*).AB.

次回は「データベース収録確認ワードの再検証」(その2)として東アジア各国原語での要約、クレーム、全文中からの収録数および英語＋原語ハイブリッド検索における収録について紹介します。

英語＋原語検索で互いに補完して収録率が上がればいい訳で、英語、原語の収録を別々に扱う必要もありませんが、新たに出現するデータベースで英語のみの収録、原語のみの収録を確認しておきたいときに活用していただければ幸いです。

データベースの収録確認ワードで問い合わせいただきました H 氏の協力で、上記検証中の日本版 CNIPR のデータをお寄せいただきました。ここに感謝申し上げます。

1)【別表】東アジアおよび ASEAN6か国の公開特許収録状況
https://sasiapi.org/2020/08/%e3%80%90%e5%88%a5%e8%a1%a8%e3%80%91%e6%9d%b1%e3%82%a2%e3%82%b8%e3%82%a2%e3%81%8a%e3%82%88%e3%81%b3asean%ef%bc%96%e3%81%8b%e5%9b%bd%e3%81%ae%e5%85%ac%e9%96%8b%e7%89%b9%e8%a8%b1%e5%8f%8e%e9%8c%b2/

2) Espacenet で東アジアの特許調査(その1)
https://sasiapi.org/2020/08/espacenet%e3%81%a7%e6%9d%b1%e3%82%a2%e3%82%b8%e3%82%a2%e3%81%ae%e7%89%b9%e8%a8%b1%e8%aa%bf%e6%9f%bb%ef%bc%88%e3%81%9d%e3%81%ae%ef%bc%91%ef%bc%89/

以上