The IDM Free Online Platform for Dictionary Publishers

Vincent Lannoy[1]
IDM

Printed dictionaries have built a genuine identity over the years. Lexicographers work for renowned publishers according to specific rules and processes; distribution channels are well-organized and efficient at delivering to educational or public markets. The emergence of new actors, exclusively focused on the Web, is a major upheaval as they deliver large corpora to a worldwide audience. Those Pure Players are now dominating the online dictionary market not only in terms of audience but also by establishing their own brands, independent of existing print brands.

These new actors bring their own vision of what an online dictionary should be. This presents a great opportunity for the industry to rethink the way dictionaries are written and published, inspired by the distinctive strengths of the Internet as a medium which call for clarity of the information, easiness of the service, and above all, intrinsic value of linguistic, i.e. lexicographic data.

Our experience, built through day-to-day management of several major free online dictionary websites, demonstrates the strong draw of dictionary content. Since dictionary websites encompass a very broad spectrum of the language and make it available for free on the Internet, users discover online dictionaries by very diverse means. Their distinct paths to a dictionary reflect their different interests in the content, and also their different expectations for the content delivered.

Making dictionary data amenable to favourable placement in search engines, for searches made in many languages, requires close involvement of lexicographers. These lexicographers must adapt to a process of creating entries for dynamic display on screen in addition to static display in print; understanding the impact of Search Engine Optimization (SEO) on entry structure; integrating a rich network of hyperlinks and making use of non-textual media to enrich their lexical content. Lexicographers are in the spotlight of the digital paradigm!

Quality of the content and publishers' care over data play a key role in building user loyalty and depth of visit on the Website. On average, in a language learning context, we observe that visits last between 5 and 7 pages, providing the publisher with the opportunity to be in contact with its users for several pages. The question is to do what? For the moment, most of the dictionary websites are dead ends: a user enters for one or several definitions and leaves though his needs or interests can be much deeper. He may require course books, vocabulary lists, exercises for learners, novels, reference content, etc. Affiliation models help propose not only the publisher's own content but complementary contents, products or services coming from partners. We are currently successfully experiencing with a partner the efficiency of an up-sales model based on dictionary free entries. Dictionary content is not only an efficient attraction point but plays also the role of a user qualification filter for targeted up-sales. Dictionary is an intermediary between a query and a targeted product.

Let's detail the opportunities offered by the online dictionary market in three areas:

- *Search Engine Optimization (SEO): why dictionary content is a marvellous resource to answer a wide range of queries in search tools such as Google, Bing, Yandex or Baidu,*
- *Reaching local markets worldwide with bilingual content,*
- *User Generated Content: an unmissable resource.*

1. Why Should you Care about SEO?[2]

A search in Google.com for [dictionary] yields nearly two hundred thousand results among which, more than the first hundred results – it is difficult to assess accurately afterwards – provide free dictionary content[3]. Now, if one amends this query – [english dictionary], [online dictionary] – the result lists change, promoting websites in a different order. The algorithmic decisions of a search engine greatly affect the traffic of dictionary sites. So dictionaries have to do with SEO and SEO has to do with lexicography.

[1] With the contribution of Orion Montoya, Ingénierie Diffusion Multimédia (IDM).

[2] The language used in this section for the case studies is mainly English as a convenient means to share examples.

[3] http://www.google.com/search?hl=en&q=dictionary.

1.1. A Dictionary is Applicable to a Broad Range of Interests

Let’s take the example of a monolingual English dictionary of approximately 40,000 entries. In a month, users reach the Website by querying for more than 150,000 keywords in search engines, varying from very generic – [dictionary] for example – to very specific phrases. A 250 free pages Premium website we consider by comparison is accessed monthly by 2,500 keywords.

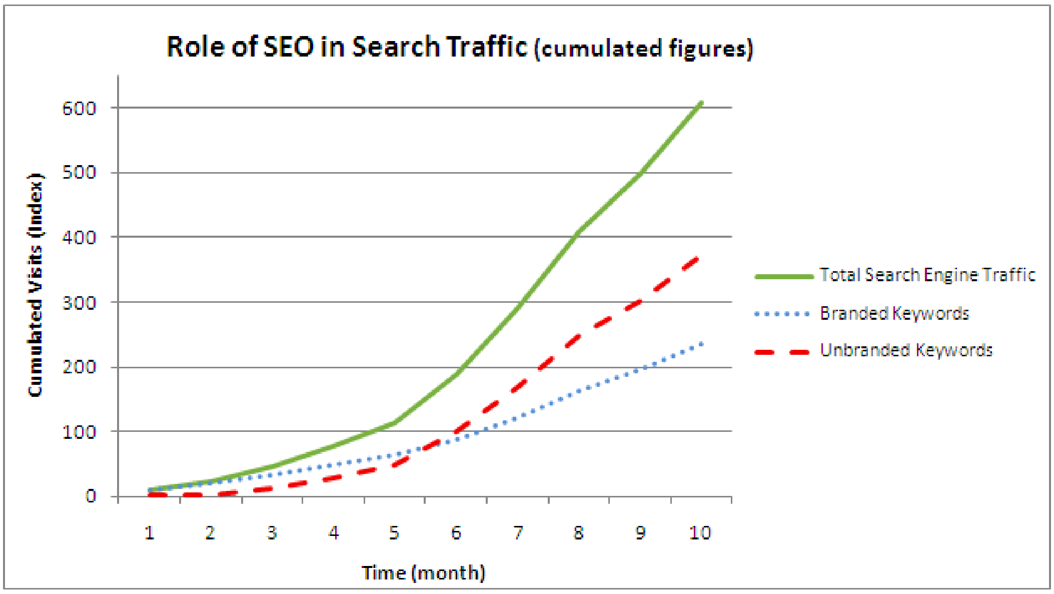


Figure 1. Interest for entries vs. interest for brand[4]

The figure 1 above illustrates the search engine traffic to a brand new dictionary website over time, broken down between branded queries (dictionary’s or publisher’s name) and unbranded queries ([*word*], [definition of *word*], [is *word* plural?], etc.).

Search engines reveal the nature of a users’ interest in the information: they might be looking for a definition, a pronunciation, a linguistic particularity, etc., and they express their need in their own way into the engines. There is no standard way to search: two different people are likely to query a search engine differently to get to the same entry. In the benchmark we use here, we can state that on average, each dictionary page is accessed by three different search queries monthly. This is a key difference from print where the user is driven to the content only by alphabetical order. The way users search for content must be taken into account when writing and shaping the dictionary, and this process has nothing to do with the print paradigm. Figure 2 below illustrates a SEO phenomenon: users access the dictionary through a very wide range of queries: 60% of the incoming visits come from 99% of the keywords. This is the familiar ‘Long Tail’ effect. In the figure below only the 2,000 keywords bringing the most traffic per month are represented.

[4] The figures are cumulated to smooth seasonal variations.

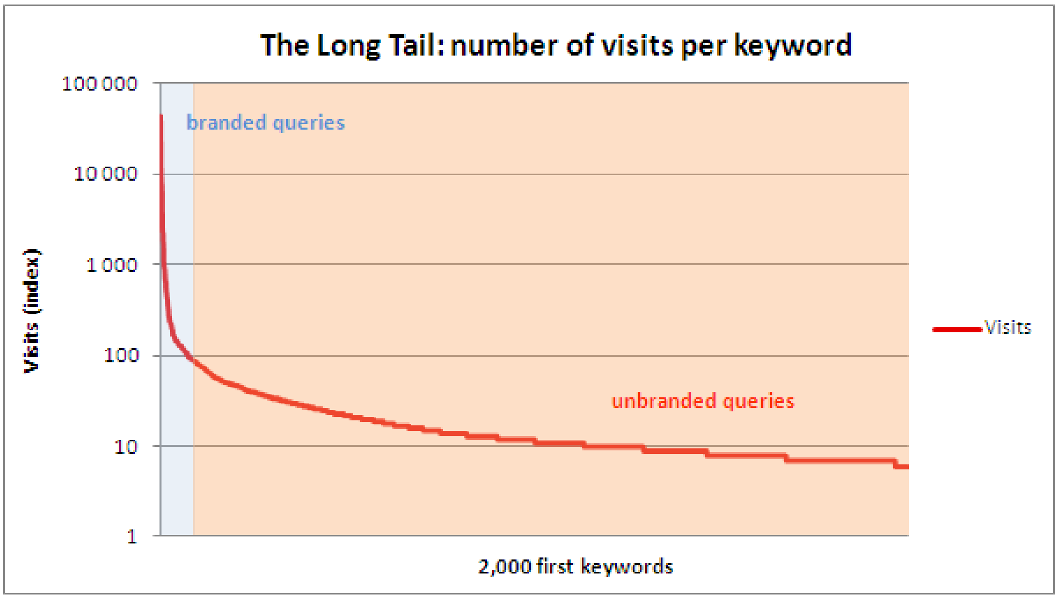


Figure 2. Dictionary content is open to many different search queries

Stating it differently, most of the traffic is coming from words which, individually, account for a very little share of the usage. This advocates for a huge number of entries in the product in order to be applicable to the widest variety of searches. An online dictionary is not a physical object that must stay at a manageable size. It is therefore the lexicographers' responsibility to design a product where users find their way easily between frequent and infrequent words, between different topical sets, etc. The consistency of the product is driven first by the data, not by the output medium.

1.2. Interest in Phrases: a Job for Lexicographers, an Opportunity for Publishers

In the previous section, we have demonstrated the role of unbranded queries to drive traffic to a free website. But what are those queries, in particular as applied to a dictionary?

Some of the unbranded queries very logically contain the standard keywords *dictionary* or *English*. However, early players on the free dictionary market have a stronghold at the top of the search engines result lists for such standard queries.

But searches for English phrases like [flushed with a howling success], [see to the children's breakfast] or [humanitarian grounds definition] for instance, yield very different result lists, leaving room for websites with a smaller audience but with carefully edited and search engine-optimized data. In the free dictionary website benchmark index we use in this presentation, 90% of the visits generated by unbranded queries are phrase-centric instead of being worded around generic dictionary keywords. However, it is important to underline that we cannot identify demand for single word definitions (like [flower]), since our benchmark index is too largely outscored by pure players.

The same applies to examples – users do want more examples than the limited number available in print – to synonyms and antonyms (users specifically searching for synonym of *word*); to audio pronunciations, especially for learners; and to collocations.

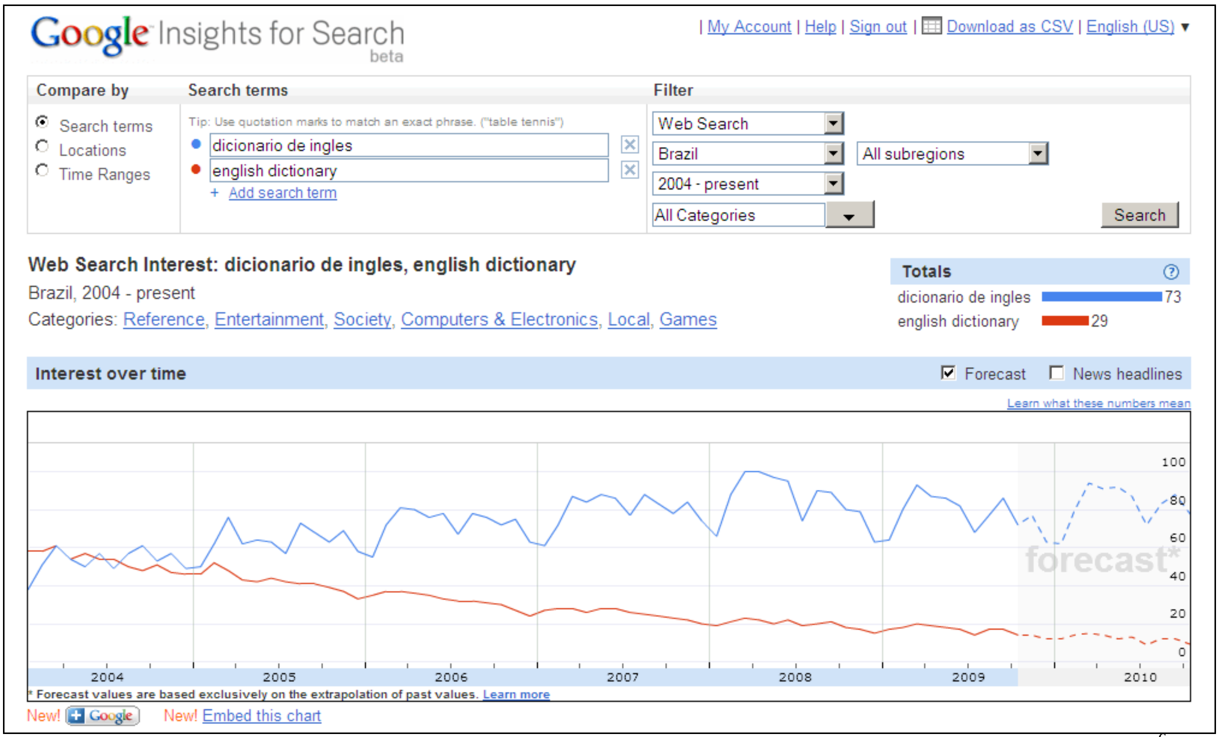
A lot is still to be done in those areas to both properly promote this targeted content toward search engines and also to provide users with the appropriate lexicographic guidelines to make good use of such content once they find it. This work requires specific linguistic, editorial and publishing skills which lexicographers definitely possess. But it also requires that lexicographers attain a deep understanding of how websites work and how users search, navigate, and experience the Web.

2. Local Languages and Local Markets

To learn or understand a foreign language, people need good bilingual content in which to root their own translation skills. The ability to bridge the gap between two different languages is of major interest and urgent need. As it often is, the Internet is both a creator of, and a solution to this need. Let’s ponder to what extent.

2.1. Native vs. English: Diversity, not Supremacy[5]

Internet users search more and more in their native language. This statement may sound either obvious or debatable, but it can be taken as a fact based on statistical data.



Compare by: ● Search terms ○ Locations ○ Time Ranges

Tip: Use quotation marks to match an exact phrase. ("table tennis")

● dicionario de ingles ☒
● english dictionary ☒

Filter: Web Search | Brazil | All subregions | 2004 - present | All Categories

Web Search Interest: dicionario de ingles, english dictionary
Brazil, 2004 - present
Categories: Reference, Entertainment, Society, Computers & Electronics, Local, Games

Totals	
dicionario de ingles	73
english dictionary	29

* Forecast values are based exclusively on the extrapolation of past values. Learn more

Figure 3. Comparison of volume of searches in Brazil for [english dictionary] and [dicionario de ingles][6]

Figure 3 compares *(upper line)* the number of times [dicionario de ingles] has been searched in Google from Brazil with *(lower line)* the number of searches for [english dictionary], also in Brazil, from 2004 to 2009. One thing appears clearly: Brazilian Internet users are searching more and more for the same information by typing it in Portuguese rather than in English.

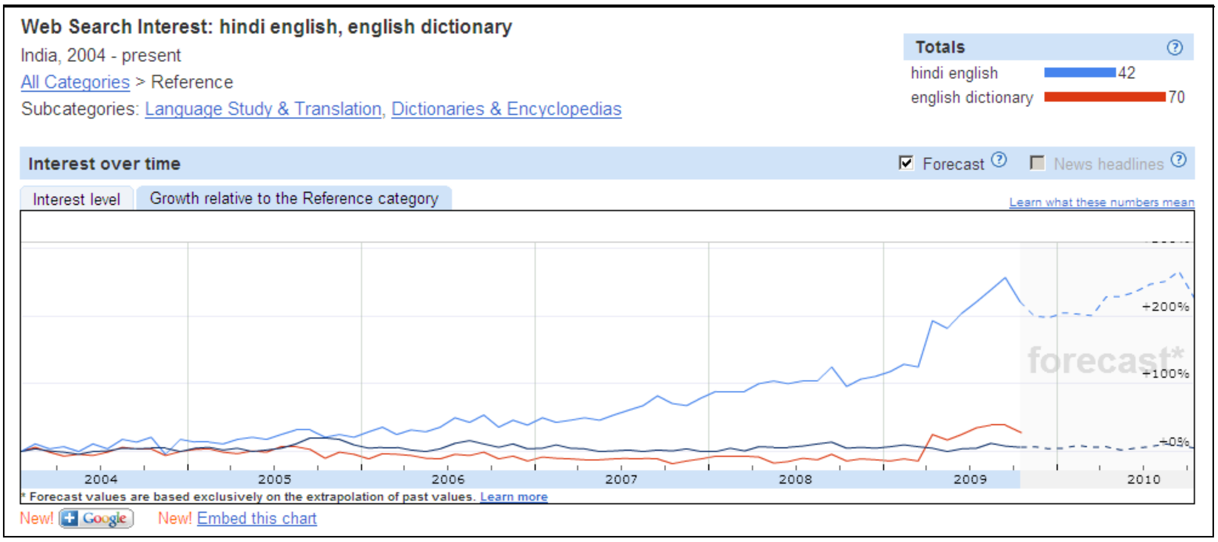
We take here the example of the Brazilian market but the same statement can be made, with differences only in volume but not in trends, about other markets such as Japan, Italy, France or Spain. It is the case in China too but more balanced; and interestingly not in Germany. These facts invite us to consider each market on its own – which is exactly what publishers have done for years with printed editions.

[5] The language used in this section for the case study is mainly English as a convenient means to share examples.

[6] Use of Google Insights for Search: http://www.google.com/insights/search/.

2.2. To Grow, Provide Bilinguals!

Users are searching for bilingual content and this demand is increasing quickly.



Web Search Interest: hindi english, english dictionary
India, 2004 - present
All Categories > Reference
Subcategories: Language Study & Translation, Dictionaries & Encyclopedias

Totals	
hindi english	42
english dictionary	70

Interest over time — ☑ Forecast □ News headlines

Interest level | Growth relative to the Reference category

* Forecast values are based exclusively on the extrapolation of past values. Learn more

New! + Google New! Embed this chart

Figure 4. Searches growth rate in India: [hindi english] vs. [english dictionary] Google queries

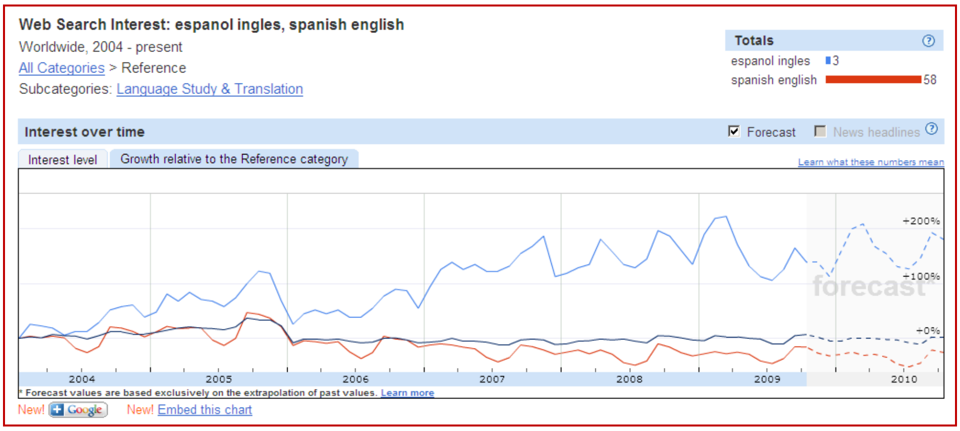
In Figure 4, contrary to Figure 3, the line chart does not graph volumes of searches but growth rate of the search queries, relative to the growth rate of a given Google category of queries. In this case, we consider the Google category *Reference*, which contains all the reference works. The dark middle line represents the average growth rate of searches within the Reference category. The upper line stands for the [hindi english] query when the lower line stands for the [english dictionary] one. Those two queries must be read against the Reference line.

The conclusion we can draw from this graph is that demand for Hindi-English bilingual content is growing fast, while apparent demand for monolingual English content is flat. As in the previous section, this statement applies to many other markets than India.

2.3. Conclusion: A Local Approach

According to the previous statements, we can assume that users search the Web for bilingual content in their native language.

Using the same growth rate comparison feature of Google Insights for Search, we can graph the following:



Web Search Interest: espanol ingles, spanish english
Worldwide, 2004 - present
All Categories > Reference
Subcategories: Language Study & Translation

Totals	
espanol ingles	3
spanish english	58

Interest over time — ☑ Forecast □ News headlines

Interest level | Growth relative to the Reference category

* Forecast values are based exclusively on the extrapolation of past values. Learn more

New! + Google New! Embed this chart

Figure 5. Searches growth rate worldwide: [espanol ingles] vs. [spanish english] Google queries

According to the graph, in rough volumes (visible in the ‘Totals’ panel, top right corner of the screenshot above), demand for Spanish-English bilingual content, queried in English, is still much higher than queried in Spanish (index of 58 versus index of 3 on the whole period 2004-2009). However, if we consider growth rates instead of volumes, demand for bilingual content queried in Spanish is growing very fast (upper line of the chart above).

The conclusion that can be drawn is that Internet strategy needs to be designed market-by-market and to take a local approach by a) providing bilingual content and b) translating the interfaces to maximize SEO efficiency. This is not exactly news: some parties are already entering local markets by these means. But it is reassuring to be able to confirm the validity of their choices. The combinations of markets and languages leave room for many players.

3. User-Generated Content: a Vital Liberality

Let’s make the case roughly before going into the details. Dictionary publishers are in the same situation as encyclopedia publishers several years ago: they simply cannot ignore the groundswell of user-generated content. Openness and users’ involvement are not an option, they are an obligation.

3.1. Why is User-Generated Content Indispensable to Lexicography?

Obviously, language is a living material that cannot fully comply with our publishing schedules: a learner may search for a word not considered to be a ‘learner’ word by a publisher; another may search for a highly technical/informal/business/etc. word and not find it in the family dictionary. Such users can now search the Web for those words, and their web searches are subject to the SEO principles presented above -- which, in turn, privilege dictionaries with a very wide range of entries. So lexicographers have to learn from dictionary users to keep their lexicographical databases comprehensive. Being receptive to users’ submissions not only allows for input from fields where lexicographers cannot be experts (the deep variety of scientific and technical terminology, professional jargons, slang words) but also allows them to stay in touch with actual usage of the language. In a sense, contributors become language panellists.

3.2. How do you Deal with User-Generated Content?

Practicality and quality are key concerns. Of course, we cannot expect all of our users to be capable of defining a word or structuring an entry. Editing a dictionary entry is not like writing an encyclopedia article. We should expect a close collaboration between users – adding new entries, new senses, examples, pronunciations, etc. – and lexicographers transforming these ‘raw lexicographic materials’ into homogeneous and reliable dictionary content.

Such an editorial process requires specific tools to publish updates on a scale of hours. They consist of light and simple tools to gather data from users, responsive back-office tools to review and edit data, streamlined tools to publish updates easily and very regularly. This constitutes a microcosm of the existing publisher’s editing process, not a revolution.

A starting point could be the unanswered queries. In our measurement, word searches made in a monolingual advanced learners English dictionary originated for print, approximately 12% of searches end up at either ‘no result’ or ‘did you mean...?’ spellchecker pages. This figure is high and indicates a significant level of disappointment for users: the dictionary may not be able to answer their questions, even though the figure measures only searches made directly on the Website, not Google (etc.) searches that sent them to websites other than their preferred dictionary because either the word was missing or it was not ranked high enough in the search results.

3.3. Differentiation from Print Products

If entries and senses are submitted by users, it helps making the online dictionary radically different from the printed dictionaries. Lexicographers and publishers can then write and market entry sets differently according to the targeted product. It helps solving the issue of duplicating the content for print and online.

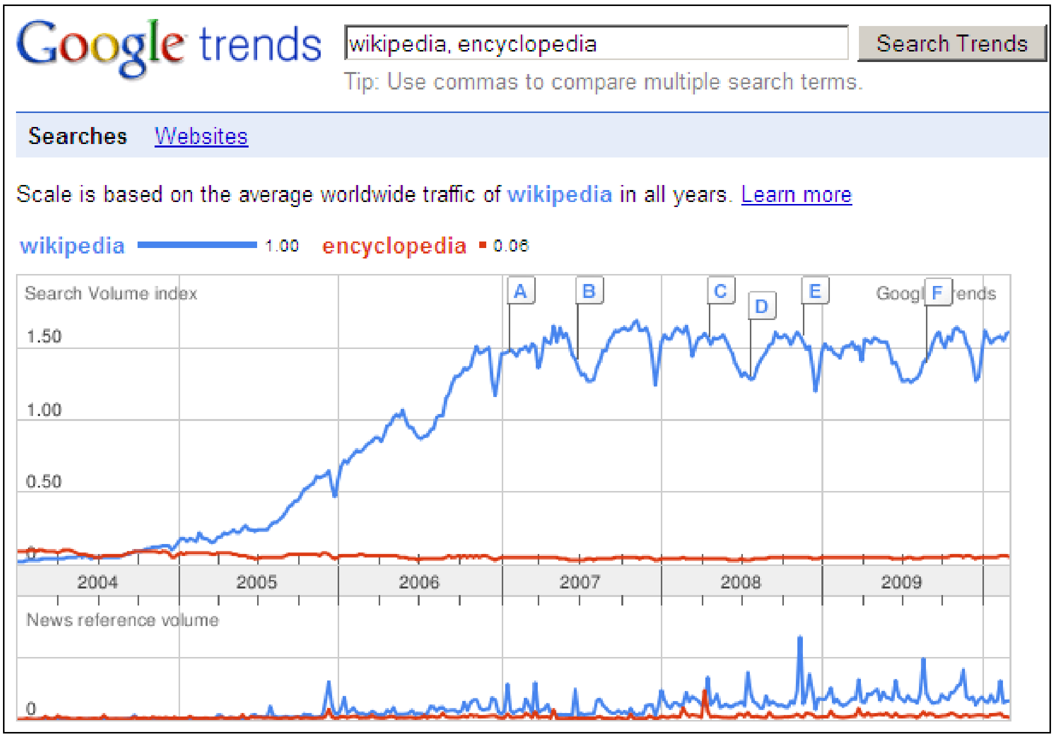
User-generated content is a democratic modification of existing editorial processes: the acceptance of an ongoing, open critical review of the dictionary. As such, it is a vital source of inspiration and improvement. Without a doubt, user-generated content is worth considering.

4. Free Dictionaries and their Business Model

Free does not mean there is no business model to support free online dictionary activity. There is no point here in assessing various models in detail, but we wish to give some indications about the existing models and how they impact the products designed by lexicographers and publishers.

4.1. Brand Building

By attracting a huge and loyal audience, free dictionary websites establish their brand, meaning that more and more users are familiar with the brand name. Figure 6 below shows the progressive shape of the Wikipedia's online reputation. The upper line represents volume of searches in Google for [wikipedia] over time, i.e. the brand name itself.



Tip: Use commas to compare multiple search terms.

Scale is based on the average worldwide traffic of wikipedia in all years. Learn more

Figure 6. The Wikipedia Brand: Building Reputation Google searches for [wikipedia] vs. [encyclopedia])

Establishing the brand name could be a goal by itself for a publisher in the perspective of selling other digital references, of course, but also and it is a key point, to strengthen the image of print products. Brand is a strong capital on the educational market for example. Correlation between a publisher's online traffic and its print sales should be studied[7].

4.2. Data and Service Licensing

A quick analysis of the top free dictionary websites reveals that most players publish data that they licensed from other publishers. It is a common model on the Web: for one single copyright owner, there is a wide population of distributors. Content licensing is a promising field of development for publishers, to license copyrights to other websites as their core content (dictionary websites), as a complement to their own data (online newspapers, blogs) or as a required service (corporations' Intranet).

Let's give a single example of dictionary online licensing. On The New York Times online edition[8], double-clicking on a word within an article opens a pop-up window displaying the word's definition. The dictionary content is provided by a free online dictionary website. We may guess that this publisher cannot only provide The New York Times but also many other partners.

The issue for publishers is to have easy means to select, extract and control the content to be licensed. We are positioning ourselves as providers of such tools.

4.3. Online Advertising

Advertising is a subject of its own which deserves a full paper. We are not discussing here whether publishers should enter or not the advertising model. We present its main characteristics.

Basically, advertising revenue works according to three main components: targeted audience, traffic and revenue per page (or eCPM, standing for *Cost per Mille*).

- *Targeted audience*, i.e. what are the users' demographics. The most obvious segmentation criterion is geographical. The ad revenue world can be split in four categories: the USA, the UK, Western Europe and Japan, Asia and South America. From our experience, the revenue per page is ten times higher in the USA than in China for instance. Advertisers value a website's capacity to target an identifiable and valuable audience. Brand plays a key role in this association,

- *Traffic*, i.e. how big the website's audience is. Ad revenue is basically proportional to the traffic: the more traffic you have, the more revenue you make and the more appealing you are to advertisers,

- *eCPM*, i.e. how much ads are worth on your website's pages. It is where lexicography interacts with advertising by dealing with the affinity between the dictionary content and adverts, what printed newspapers and magazines have been experiencing for years.

[7] See the study *The Short-Term Influence of Free Digital Versions of Books on Print Sales* in references.

[8] http://www.nytimes.com/.

Advertising is still a developing market and ways to accommodate dictionary entries, reference work and advertising are still to be created.

4.4. Funnelling Toward Subscription: Affiliation and the Freemium Model

As presented in the previous sections, online dictionaries today are considered ‘dead ends’, that is to say that they use SEO to attract a massive audience which is to be retained as long as possible on the dictionary’s free pages. In most cases, users are not converted into subscribers to a complementary service.

Dictionary content is powerful to attract visitors with different profiles and interests. But, how do you create a value proposition to convert a free visitor into a potential subscriber? This up-selling model can be envisaged in two levels:

- *Affiliation*. Free pages filter and qualify users according to rules agreed between parties (geographical, list of pages/entries, etc.) and then promote the affiliate’s product(s). There is a commissioning agreement between the publisher and the affiliate. Affiliation can be complementary to advertising,

- *Freemium*. Free pages play the same qualifying role to make targeted users subscribe to the publisher’s online Premium products. Examples on the Internet are Last.fm, Flickr.com, LinkedIn.com or Skype. This implies that the publisher has digital online products to sell. The link between the dictionary pages holding the Premium incentive and the Premium trial and e-payment pages is called a ‘funnel’ and must be designed with care to optimize conversion rate.

- Tools to be used to target users and display the offer accordingly (affiliated or Premium) are comparable to the ones used for online advertising.

5. IDM Involvement in Online Dictionaries

IDM offers a range of tools and services, from editing to publishing, for publishers wishing to distribute their content online: through the Web, on mobile devices, or licensing to third parties.

We highlight hereafter four main steps in the publishing process.

5.1. Prepare the Dictionary Data

As presented in the previous sections, it is critical to make the dictionary data suitable for digital media. It means that data available for print must be expanded (addition of meta-data for SEO), merged or split depending on what the publisher wants to achieve (for example, how to make a learners and an advanced learners dictionary work on the same website? What about monolinguals and bilinguals interlinking?), reviewed for an online publishing (check cross-references for example) and segmented (versions for desktop screen display and mobile devices are not the same).

This data preparation work must be handled within streamlined processes to make operations easy and clear to follow, within dedicated back-office tools to provide the publisher with the means to control what is delivered.

5.2. Index the Data for Search

Once data are prepared for digital publishing, it is indispensable to index them to provide end-users with a powerful built-in search engine. At that stage, publishers have the opportunity to value specific aspects of their lexicographic work, for example by indexing phrases on their own, giving particular visibility to identified linguistic elements. The publisher then proofs the data and the search feature to check that they are corresponding to their quality standards.

The way a publisher indexes data is specific and provides added value to end-users. That's why we promote indexing and searching for licensing purposes as well: a third-party website may access the data through the search engine – technically speaking through a Webservice or an API – putting the publisher in a position not only to distribute its content but also to market a full service that they control, from the data to the searching facility.

5.3. Provide Website Building Tools

We feel that data editing, processing for publishing, indexing and broadcasting through dedicated webservice and API is our expertise. That is to say that we position our service by providing publishers with means to build dictionary websites according to the state-of-the-art web standards, without necessarily encompassing the delivery of the end-users website itself. Practically speaking, IDM provides publishers with a SDK (Software Development Kit) allowing them to choose any provider to build the end-users website. We believe this scheme provides a lot of flexibility and allows publishers to diversify their approach of the market.

5.3.1. Data, Features and SEO Bundled for Layout: the Role of the Template Engine

The Free Online Platform SDK embeds the use of a template engine.

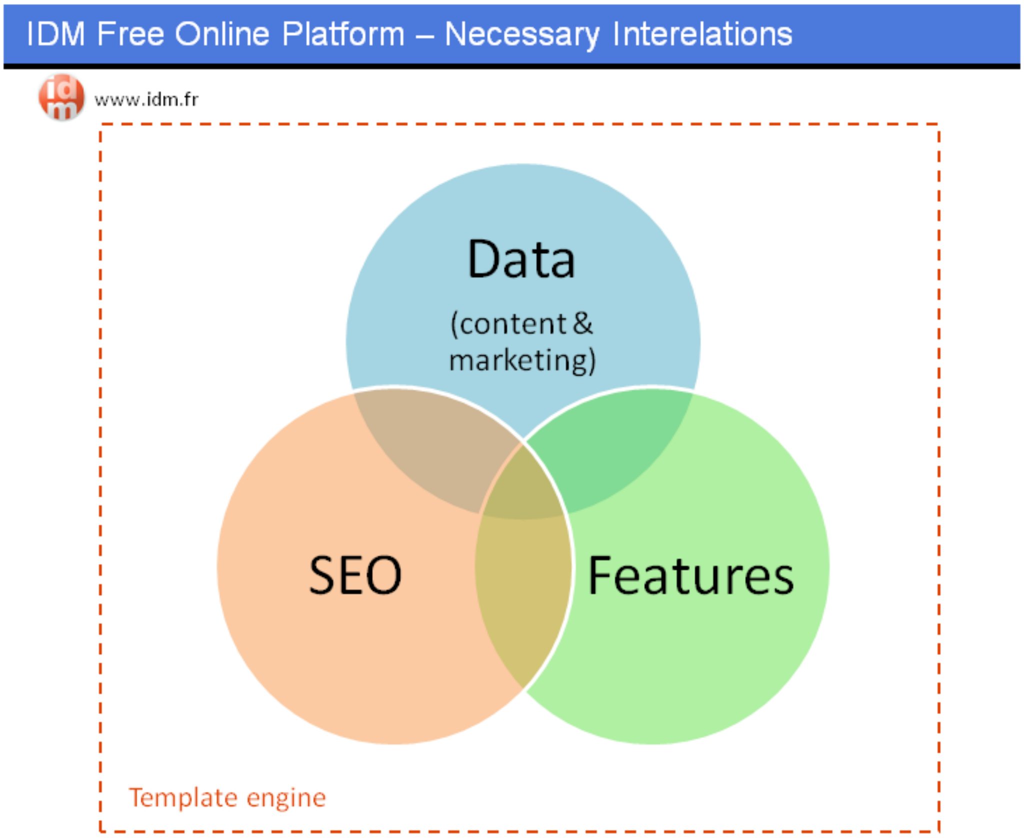


Figure 7. Combination of the Key Components within the Template Engine

The template engine allows the website developer to efficiently control the display of the pages, according to the data, SEO and features constraints. Moreover, the template engine allows the publisher to amend the display without calling for the developer's help. This architecture results in both tight control over the content and great flexibility in its management.

5.3.2. Compliance with Mobile Devices

The market is stretched between wide desktop screens on one side and tiny screen resolutions with mobile devices on the other side. Traffic metrics show clearly that more and more users access the free dictionary websites via smartphones over the Internet. It is necessary to address that demand.

The IDM Free Online Platform's control over the data combined with the use of the template engine offers the possibility to setup an alternate website, mirroring the main one, to comply with smaller screen resolutions. A user browsing the Website from a smartphone can be detected and automatically routed to the smartphone-compliant website. This second website should be simplified in terms of features and content to cope with smartphones' screen resolution and technical strains.

5.3.3. IDM Free Online Platform Software Architecture Overview

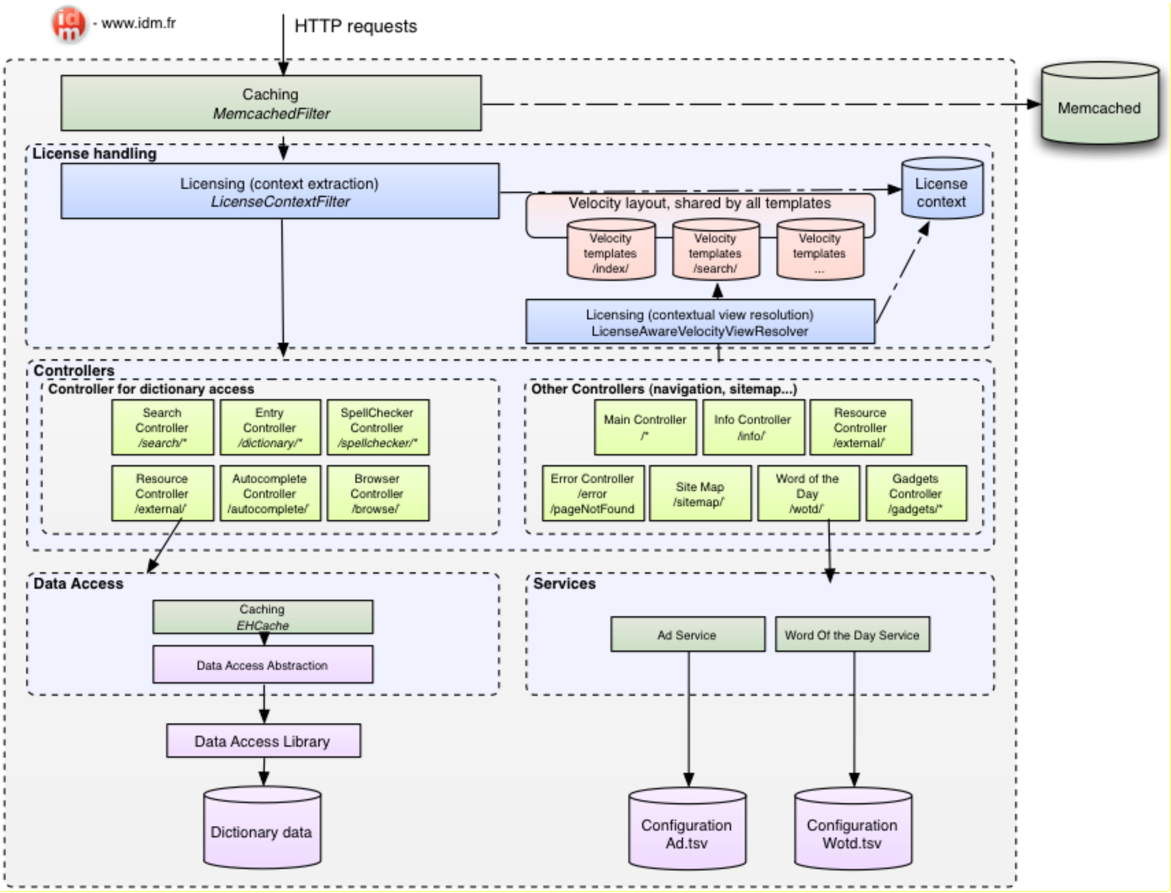

- www.idm.fr
HTTP requests
Caching
MemcachedFilter
Memcached
License handling
Licensing (context extraction)
LicenseContextFilter
License context
Velocity layout, shared by all templates
Velocity templates /index/
Velocity templates /search/
Velocity templates ...
Licensing (contextual view resolution)
LicenseAwareVelocityViewResolver
Controllers
Controller for dictionary access
Search Controller /search/*
Entry Controller /dictionary/*
SpellChecker Controller /spellchecker/*
Resource Controller /external/
Autocomplete Controller /autocomplete/
Browser Controller /browse/
Other Controllers (navigation, sitemap...)
Main Controller /*
Info Controller /info/
Resource Controller /external/
Error Controller /error /pageNotFound
Site Map /sitemap/
Word of the Day /wotd/
Gadgets Controller /gadgets/*
Data Access
Caching
EHCache
Data Access Abstraction
Data Access Library
Dictionary data
Services
Ad Service
Word Of the Day Service
Configuration Ad.tsv
Configuration Wotd.tsv

Figure 8. Free Online Platform Technical Components

5.4. Response Time is the Prime Virtue

Response time – the time a user has to wait before displaying a web page – is a crucial factor and must be addressed as such. Once engaged in free online business, there is no choice between speed and features: speed must always prevail.

This strong statement is built over experience, not only on the side of IDM but confirmed by the Internet giants such as Amazon, Yahoo, Google, Facebook, etc.

To handle response time properly, three factors are to be considered:

- Lightness of the pages. We provide 'GUI guidelines for dictionary websites' to advise graphic designers,

- Optimized hosting platform for high traffic. We can host end-users websites and webservices, if required by the publisher,

- Monitoring of performance.

Bibliography

Campy Cubillo, M.C. (2002). ‘General and specialised free online dictionaries’. In *Teaching English with Technology* 2/3. http://www.iatefl.org.pl/call/j_review9.htm; access date 14 February 2010.

Dean, J. (2005). *Lively legacy of ‘dull work’*. Bookseller. http://www.allbusiness.com/bookseller/20050415/4635798-1.html; access date 14 February 2010.

Hutchins, J. (2009). ‘Multiple Uses of Machine Translation and Computerised Translation Tools’. In *Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages – ISMTCL 2009*. http://www.hutchinsweb.me.uk/Besancon-2009.pdf; access date 14 February 2010.

Kilgarriff, A. (2010). ‘How to monetise a web presence (and hoover a moose). A report on the e-lexicography conference at Louvain-la-Neuve, Belgium, 22-24 October 2009’. Paper to be presented at EURALEX 2010.

Hilton III, J.; Wiley, D. (2010). *The Short-Term Influence of Free Digital Versions of Books on Print Sales*.
http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno= 3336451.0013.101.