

From Lexicography to Terminology: a Cline, not a Dichotomy

Thierry Fontenelle

Translation Centre for the Bodies of the European Union, Luxembourg

thierry.fontenelle@cdt.europa.eu

Abstract

In a paper presented at the Euralex 2012 conference, ten Hacken (2012) discusses the OED's problematic claim to be the "definitive record of the English language". He argues that what distinguishes the OED from other dictionaries is the information it provides about English words and the range of problems this information can be used to solve. Dictionaries are not descriptions of a language, he claims, but tools with which users of the dictionary solve problems of a particular type. The nature of the dictionary therefore determines which types of problems it can solve.

In this paper, I would like to extend the parallel made by ten Hacken between general dictionaries, learners' dictionaries and historical dictionaries such as the OED to what is traditionally perceived as a dichotomy, namely the distinction between dictionaries and terminological databases. Instead of viewing term bases as a totally distinct type of linguistic product, I would like to argue that they should rather be seen as a specific kind of tool which provides information that specific users will use in order to solve specific linguistic problems, usually related to translation. In the course of their careers, translators will indeed need to make use of a whole range of dictionaries, starting from learners' dictionaries when they learn foreign languages, to monolingual dictionaries and bilingual dictionaries to learn translation techniques, to term bases as soon as they start translating highly specialized and technical texts. We will focus on terminology databases such as IATE, the European Union's interinstitutional term base, which is the natural tool to which they turn to obtain information about technical terms in the medical, legal, environmental, chemical fields, to cite only a few domains covered by this resource. With 8.7 million terms covering the 24 official languages of the European Union, including 1.4 million English terms and half a million abbreviations, this database is a highly popular tool in the translation world (44 million queries in 2013).

In addition to general term bases such as IATE, we will also discuss other specialized EU terminological databases such as ECHA-Term, a term base compiled for the European Chemicals Agency (ECHA) to help industry comply with the legal requirements of the REACH Directive and of the regulation on the classification, labeling and packaging of chemicals. We will show how user requirements have been taken into account to meet the needs of the users of the database, who resort to ECHA-Term to obtain reliable, coherent and up-to-date multilingual terminology in the chemicals field, a *sine qua non* for clear specialized communication. The description of these databases will make it clear that the distinction between 'traditional' lexicography and terminology is more a cline than a dichotomy,

insofar as the types of linguistic information included in the respective products created by both disciplines all correspond to the specific needs of their users.

Keywords: term banks; terminological database; translation; European Union; LSP dictionaries; IATE; ECHA-term

1 Introduction

In a paper presented at the Euralex 2012 conference, ten Hacken (2012) discusses the Oxford English Dictionary's problematic claim to be the "definitive record of the English language". He points out that the OED is often regarded as authoritative and that one of the aspects of authority is the comprehensive lexical coverage of the dictionary. Yet, he argues, even if lexicographers such as Simpson (2000:1) call the OED "the principal dictionary of record for the English language", there is in fact no empirical entity corresponding to "the English language" for which the OED could be taken as a description (ten Hacken 2012: 838). Simpson himself is aware of the impossibility of providing a comprehensive coverage in a dictionary. It is therefore a myth to assume that a dictionary should contain every word. A dictionary can therefore only be a partial record and it is unrealistic to assume that a dictionary can provide a full record of a language.

Ten Hacken argues that what distinguishes the OED from other dictionaries is the information it provides about English words and the range of problems this information can be used to solve. Dictionaries are not descriptions of a language, he claims, but tools with which users of the dictionary solve problems of a particular type. The nature of the dictionary therefore determines which types of problems it can solve.

It is interesting to note that the controversy around the comprehensive nature of a dictionary such as the OED usually involves a comparison with other general-purpose monolingual dictionaries, including learners' dictionaries. The inclusion of usage notes is also used as a criterion to distinguish descriptive and prescriptive dictionaries. Surprisingly, no mention is usually ever made of a different type of dictionary, namely terminological databases, which should also be seen as specific types of dictionaries designed to solve specific types of linguistic problems. Why is that nobody ever questions the comprehensiveness of a terminological database? Why would anybody expect the OED to include "all possible" words of the English language, while recognizing that the list of acronyms and abbreviations in a language is potentially infinite and that even a huge term base such as IATE, the interactive terminology database of the European Union described below, can only provide a partial record of specialized terminology, even with 1.4 million English terms and half a million abbreviations and acronyms?

2 From Learner's Dictionaries to Mono- and Bilingual Dictionaries to Terminological Bases

I would like to extend the parallel made by Ten Hacken between general dictionaries, learners' dictionaries and historical dictionaries such as the OED to what is traditionally perceived as a dichotomy, namely the distinction between dictionaries and terminological databases. Instead of viewing term bases as a totally distinct type of linguistic product, I would like to argue that they should rather be seen as a specific kind of tool which provides information that specific users will use in order to solve specific linguistic problems, usually related to translation.

If, as proposed by Ten Hacken (2012, 2009), dictionaries are tools with which users solve problems of a particular type, the various types of dictionaries available on the market actually correspond to the range of problems users are faced with at different moments of their career. A teenager or a university student who learns a foreign language will most probably require a small bilingual dictionary at the beginning of the learning process because a learner's dictionary can only be used by someone who is not a total beginner in this foreign language. Once knowledge of the foreign language reaches a certain level, the student will be encouraged to make use of a learners' dictionary, which will provide useful information in a decoding and an encoding perspective, thanks to its simplified definitions, its system of grammar codes, its illustrative examples, etc. General-purpose monolingual dictionaries target a different kind of public, made up of advanced native speakers or of non-native speakers who have an in-depth knowledge of the language of the dictionary. Historical dictionaries such as the OED, with their focus on etymology and the evolution of words, are yet for other users who expect the dictionary to provide descriptive records of the development and use of words over time.

A parallel may be drawn with the tools used by translators. At the beginning of their career, students in translation will primarily make use of general-coverage bilingual dictionaries which will provide them with information about collocations, idioms, sense distinctions, etc. The role of translations in bilingual dictionaries is to provide target-language equivalents of the source-language headword (see also Fontenelle forthcoming). At a later stage, however, seasoned translators will tend to consult their bilingual dictionaries less and less, and will turn more frequently to terminological databases (a.k.a. term banks), which will enable them to translate highly-specialized texts and to make communication possible between specialists, or between specialists and the general public. Such tools have become a *sine qua non* in our multilingual world where access to technical information across multiple domains is a must.

3 Terminology and Term Banks

Understanding terminology, i.e. the specialized vocabulary which is used in a specific domain, is a key element in communication. This poses a number of challenges in the case of translation in a multi-

lingual context, since knowing the exact meaning of a technical term is necessary to understand a text, but also to reproduce the text as faithfully as possible in another language. It therefore no surprise that the various translation services of the European Union have traditionally dedicated significant resources to the compilation of terminological information in the official languages of the EU. In order to describe the vocabulary of special subject fields, terminologists create term banks, which are compilations of the collections of words associated with a given domain. A terminology database will then be seen as a repository of descriptions of concepts, which are seen as mental constructs which are distinct from the terms they correspond to in a given language (see also Fontenelle and Rummel, forthcoming). The traditional approach assumes that terms can be organised into networks of concepts to structure a given domain. This is indeed well-suited for normalization, but, as is pointed out by Jacquemin and Bourrigault (2003), there seems to be a flaw in this reasoning, because this approach is not really suitable for computational term analysis. A terminologist indeed bases his or her work upon the analysis of textual data (corpora) and the term base is actually the result of this analysis, and not the result of some introspection whereby abstract conceptual maps would be derived.

Even if terminological databases are traditionally seen as distinct from dictionaries, it cannot be denied that terminological entries have a lot in common with dictionary entries. Of course, at the macrostructural level, it is clear that some entries found in a traditional dictionary will not be found in a term base: some parts of speech will be absent from term banks. Prepositions or adverbs for instance will most probably not be found in term banks. Even verbs are traditionally underrepresented in such resources because the vast majority of terms are noun phrases. This is not enough to consider that a term base is not a dictionary, however. After all, rare scientific words will also be excluded from learners' dictionaries. At the microstructural level, definitions will be as essential in term bases as in a traditional monolingual dictionary and the NLP community has always been interested in how candidate terms could be extracted from corpora, together with possible definitions (see Person 1998, who proposed an analysis of the defining mechanisms signalling the presence of a term in a corpus, using linguistic patterns such as 'X is known as Y' or "X is called Y" to link a definiens and a definiendum).

4 IATE: The European Union Terminological Database

4.1 An Interinstitutional Database

IATE stands for 'Inter-Active Terminology for Europe' and is the term base of the language services of the European Union. This concept-oriented, large-scale multilingual database covers all fields of activity of the European Union. IATE was initially launched in 1999 by the Translation Centre for the Bodies of the European Union, located in Luxembourg. Today, the Translation Centre manages the technical aspects of the project on behalf of the project partners: the European Commission, the Eu-

ropean Parliament, the Council, the Court of Auditors, the Court of Justice of the European Union, the European Investment Bank, the European Central Bank, the Economic and Social Committee, the Committee of the Regions and the Translation Centre. Before the launch of the project and its opening to the public in 2007, nearly every institution had its own term base (Fontenelle and Mergen 1998; Reichling 1998), while, today, IATE can be seen as the shared terminology database of all existing terminology collections of the translation services of all EU institutions and bodies. It can be consulted free of charge at <http://iate.europa.eu>.

IATE can be searched for specific terms or abbreviations in a given source language and for its equivalent(s) in any of the 23 other languages (IATE contains mainly terminology in the 24 official languages of the EU, as well as some content in non-official languages). After the accession of Croatia on 1 July 2013, the 24 official languages are: Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish, Swedish.

The database is constantly updated by the terminologists and translators of the various participating institutions. In 2013, around 97 000 new terms were added and over 150 000 existing terms were modified. The contents of the database as of 1 January 2014 can be broken down as in Table 1.

Language	Number of terms	Language	Number of terms
en - English	1402006	ga - Irish	57879
fr - French	1339496	lt - Lithuanian	53802
de - German	1034088	hu - Hungarian	49858
it - Italian	701735	et - Estonian	41489
nl - Dutch	691801	sl - Slovenian	41337
es - Spanish	617124	cs - Czech	38043
da - Danish	603481	sk - Slovak	37219
pt - Portuguese	533623	mt - Maltese	35732
el - Greek	523086	ro - Romanian	35451
fi - Finnish	329491	bg - Bulgarian	34420
sv - Swedish	314220	lv - Latvian	28617
la - Latin	64159	hr - Croatian	8863
pl - Polish	59576		

Table 1: IATE: Number of terms (1/1/2014).

The table reflects the history of the European Union and its successive enlargements, of course. Altogether, the IATE database contained 8,705,334 terms on 1/1/2014. The languages with the most terms were the European Union's most often used working languages, viz. English, with more than 1.4 million terms in first place, followed by French (1.3 million terms) and then German (1 million terms). It is interesting to note that Latin is also fairly well-represented, with 64159 terms: taxonomies of animal

and vegetal species are crucial for translators who deal with the translation of texts related to the Common Fisheries Policy or the Common Agricultural Policy, as well as texts written by the European Environment Agency, the Community Plant Variety Office or the European Food Safety Agency.

The public version of IATE received 44 million queries in 2013 and the number of queries in the internal version of IATE (accessible only to EU staff) amounted to 14,206,137 (vs. 11,178,323 queries in 2012). Queries in the public version of IATE came from all over the world, from France to Somalia, Argentina to South Korea. The country where the most queries originated from in 2013 was Italy, followed by France, Spain, Germany, Belgium, Greece, Portugal, the United Kingdom, the Netherlands and Switzerland.

Terms in IATE have a fairly standard data structure. One of the challenges which the creators of the database faced was the mapping rules between the data structures of the existing databases and the new format of this interinstitutional database. A concept-oriented approach was adopted to express the various aspects of concepts via a series of three interrelated levels:

- (1) a language-independent top level, containing information pertaining to the whole concept. Information about domains is a case in point (i.e. the field of knowledge in which the concept is used). Other types of information can also be stored at that level, including pictures or images;
- (2) An intermediate 'language' level for definitions, explanations and comments, which can be stored for each of the languages of the terminological record;
- (3) The 'term' sub-level, at which several terms can be distinguished to store synonyms of a given concept or abbreviations. Reliability codes, term references and the date when a record was last edited will typically appear at that level.

For instance, *BSE*, *bovine spongiform encephalopathy* and *mad cow disease* are three distinct terms that are synonyms and refer to the same concept. The definitions would be coded at level 2 (the three terms will have the same definition in a given language), while the domain (Animal health) is at level 1, the conceptual level, as is illustrated in Figure 1 and in Figure 2 below.

The screenshot shows the IATE search interface. At the top left is the IATE logo with the text 'InterActive Terminology for Europe'. On the top right, there is a language dropdown menu set to 'English (en)' and links for 'Search Screen' and 'Help'. Below this is a search input field containing 'BSE' and a 'Search' button. Underneath the search bar, it indicates 'en > de fr it nl (domain: Any domain, type of search: All)'. The main results area shows 'Result 1- 10 of 32 for BSE'. The results are organized into a table with columns for language codes, terms, quality indicators (stars), and icons. The first result is under the domain 'Animal health [COM]' and lists 'mad cow disease'. Subsequent results are grouped by language: EN (bovine spongiform encephalopathy, BSE), DE (Rinderwahnsinn, spongiforme Rinderenzephalopathie, BSE), FR (maladie de la vache folle, maladie des vaches folles, encéphalite spongieuse bovine, encéphalite spongiforme bovine, encéphalopathie spongiforme bovine, ESB), IT (follia bovina, encefalopatia spongiforme bovina, BSE), and NL (bovine spongiforme encefalopathie, BSE). Each entry includes a quality indicator of five stars and a plus sign, and some have additional icons like a magnifying glass or a document icon.

Language	Term	Quality	Icon
Animal health [COM]			
	mad cow disease	★★★★+@	
EN	bovine spongiform encephalopathy	★★★★+@	📖
	BSE	★★★★+@	
	bovine spongiforme Enzephalopathie	★★★★+@	🔍
DE	Rinderwahnsinn	★★★★+@	📖
	spongiforme Rinderenzephalopathie	★★★★+@	
	BSE	★★★★+@	
	maladie de la vache folle	★★★★+@	
	maladie des vaches folles	★★★★+@	🔍
FR	encéphalite spongieuse bovine	★★★★+@	📖
	encéphalite spongiforme bovine	★★★★+@	🔍
	encéphalopathie spongiforme bovine	★★★★+@	🔍
	ESB	★★★★+@	
	follia bovina	★★★★+@	
IT	encefalopatia spongiforme bovina	★★★★+@	📖
	BSE	★★★★+@	
	gekkekoeienziekte	★★★★+@	
NL	bovine spongiforme encefalopathie	★★★★+@	📖
	BSE	★★★★+@	

Figure 1: Query on BSE in IATE.

Domain	Animal health
en	
Definition	progressive, fatal, neurologic disease of adult domestic cattle that resembles <i>scrapie</i> (IATE:1257587) of sheep and goats
Definition Ref.	The Merck Veterinary Manual > Nervous System > Bovine Spongiform Encephalopathy > Introduction, http://www.merckvetmanual.co... [10.1.2012]
Note	<p>It was first diagnosed in the UK in 1986. Epidemiological studies conducted in the UK suggest that the source of BSE was cattle feed prepared from bovine tissues, such as brain and spinal cord, that was contaminated by the BSE agent. Speculation as to the cause of the appearance of the agent causing the disease has ranged from spontaneous occurrence in cattle, the carcasses of which then entered the cattle food chain, to entry into the cattle food chain from the carcasses of sheep with scrapie.</p> <p>For further information please refer to: WHO > Programmes and projects > Media centre > Fact sheets > Bovine spongiform encephalopathy, http://www.who.int/mediacent... [10.1.2012]</p>
Term	mad cow disease
Reliability	3 (Reliable)
Term Ref.	Centers for Disease Control and Prevention > CDC A-Z Index > BSE (Bovine Spongiform Encephalopathy, or Mad Cow Disease), http://www.cdc.gov/ncidod/dv... [10.1.2012]
Date	10/01/2012
Term	bovine spongiform encephalopathy
Reliability	3 (Reliable)
Term Ref.	The Merck Veterinary Manual > Nervous System > Bovine Spongiform Encephalopathy > Introduction, http://www.merckvetmanual.co... [10.1.2012]
Date	10/01/2012
Abbreviation	BSE
Reliability	3 (Reliable)
Term Ref.	The Merck Veterinary Manual > Nervous System > Bovine Spongiform Encephalopathy > Introduction, http://www.merckvetmanual.co... [10.1.2012]
Date	10/01/2012

Figure 2: set of synonyms for bovine spongiform encephalopathy.

The system offers today the following features (see also Fontenelle and Rummel in press):

- One common database for all institutions and agencies containing all legacy data;
- Basic and advanced search features (including stemming and base character conversion);
- On-line access in read and write mode, i.e. the possibility for users to carry out modifications, to add entries directly to the central database and hence to allow their colleagues to benefit from this work in real time;
- A validation workflow that ensures that all newly added or modified terminology is reviewed;
- Role-based user management;
- Auditing features that keep track of all changes made to the terminology in the database;
- Features for the export and import of data;
- Statistics on the content of the database and user activity;
- A basic messaging system as communication mechanism between the actors in the terminology workflow.

In 2014, a new functionality will also be offered to allow users to download or copy the contents of the IATE database which is not protected by third-party copyrights, for research or for commercial purposes.

4.2 Language and Terminology: A Dynamic Organism

Languages are dynamic organisms. New terms are created every day (nobody was talking about 3D printers five years ago and the translation of selfie is a hot topic in many linguistic communities in 2014). It is therefore not surprising to see that IATE is not just concerned with terms proper, but also with abbreviations and acronyms, which are condensed versions of often long and complex terms. Table 2 shows the contents of the IATE database per term type.

Abbrev	519161
Formula	698
Phrase	142784
Short Form	20579
Term	8022112

Table 2: Number of terms per type in IATE.

With more than half a million abbreviations, one can immediately see that the question of the ‘completeness’ of a dictionary discussed in the context of the OED above is a myth, indeed, and a quick glance at the list of abbreviations included in such a database will convince anybody that there cannot be such a thing as a “definitive record” of any language.

DROIT [CdT]		Entrée entière	
EN	the Administrative Board shall adopt rules of procedure	★★★★+	+@
DE	der Verwaltungsrat gibt sich eine Geschäftsordnung	★★★★+	+@
ES	el Consejo de Administración adoptará su reglamento interno	★★★★+	+@
FR	le Conseil d'administration arrête son règlement intérieur	★★★★+	+@
IT	il Consiglio di amministrazione adotta il proprio regolamento interno	★★★★+	+@

DROIT [CdT]		Entrée entière	
EN	an Opposition Division shall consist of three members	★★★★+	+@
DE	eine Widerspruchsabteilung setzt sich aus drei Mitgliedern zusammen	★★★★+	+@
ES	una División de Oposición constará de tres miembros	★★★★+	+@
FR	une division d'opposition se compose de trois membres	★★★★+	+@
IT	una divisione di opposizione è formata da tre membri	★★★★+	+@

DROIT [CdT]		Entrée entière	
EN	the Secretariat for the Administrative Board shall be provided by the Office	★★★★+	+@
DE	die Sekretariatsgeschäfte des Verwaltungsrates werden vom Amt wahrgenommen	★★★★+	+@
ES	la Oficina se encargará de la secretaría del Consejo de Administración	★★★★+	+@
FR	le secrétariat du Conseil d'administration est assuré par l'Office	★★★★+	+@
IT	l'Ufficio svolge la funzione di segretariato del Consiglio di amministrazione	★★★★+	+@

DROIT [CdT]		Entrée entière	
EN	the Administrative Board shall take its decisions by a simple majority	★★★★+	+@
DE	der Verwaltungsrat faßt seine Beschlüsse mit der einfachen Mehrheit	★★★★+	+@
ES	el Consejo de Administración tomará sus acuerdos por mayoría simple	★★★★+	+@
FR	le Conseil d'administration prend ses décisions à la majorité simple	★★★★+	+@
IT	il Consiglio di amministrazione prende le sue decisioni a maggioranza semplice	★★★★+	+@

Figure 3: semi-fixed phrases including shall in IATE.

The inclusion of “formulae” and “phrases” in the term base is also a sign that terminologists are more and more concerned with phraseological units and various types of formulaic expressions. The need to standardize language (especially in legal texts, but also in technical fields such as aviation and aeronautics) encourages people to make use of ready-made phraseological patterns which sometimes go way beyond the traditional notion of terms viewed as a noun phrase. In this context, the new European financial supervision authorities which organize stress tests at the European level in coordination with the European Central Bank regularly issue Guidelines which must be translated in all the official EU languages. The translators working in the financial field to provide the various linguistic versions of these Guidelines have received strict recommendations concerning the translation of modal auxiliaries like “shall” and “should” (e.g. should needs to be translated by *devoir* in French in the specific context of Guidelines for financial supervision). These questions are debated by the terminologists of the various language teams of the European Union and IATE now includes a fair amount of

“fixed phrases” which show how these modal auxiliaries should be translated (e.g. “shall” in English most frequently translates as a present tense in French legal texts, when in normal, non-legal language, the default is usually a future tense in French). Figure 3 above illustrates some of these “semi-fixed phrases” included in IATE (see “shall be provided by the Office” à “est assuré par l’Office”). It is clear that there has been an evolution over the last few years with respect to the traditional distinction between lexical items included in dictionaries and terms included in terminology databases and what used to be a clear-cut distinction now increasingly appears as a cline between lexicography and terminology.

4.3 Metadata for User Preferences

Very much like a traditional dictionary which makes use of usage notes and a variety of labels aimed at capturing levels of formality (formal, informal, slang, taboo...), a terminological database such as IATE makes extensive use of metalinguistic labels which guide translators in their daily work. Official terminology in any field may indeed change rapidly and terms which are commonly used today may become deprecated tomorrow. It is therefore important to capture the preferences expressed by the main “consumers” of a translation (official organizations and administrations, public bodies and authorities, scientific communities, etc). Such preferences may result from simple stylistic or even sometimes arbitrary preferences and conventions. They may also reflect historical evolutions and a need to avoid geo-political problems. IATE will therefore make use of metalinguistic labels such as “preferred”, “obsolete” or “deprecated” to provide information about changes in the pragmatic use that is made of these terms.

The concept for the disease known as A(H1N1) is a case in point. IATE indicates that several “synonymous” terms can be used to refer to that disease, which was first observed in Mexico in 2009, including the term Mexican flu as well as swine flu. The term Mexican flu was used at the very beginning of what later became a worldwide epidemic, but international organisations such as the European Centre for Disease Prevention and Control (ECDC), one of the Translation Centre’s clients, expressed a strong preference in favour of the term A(H1N1)v, rather than Mexican flu or swine flu. Such preferences are captured through the use of labels like Preferred (appearing in green on the IATE web site) or Deprecated (in red on the IATE web site), as illustrated in Figure 4.

Medical science [COM]		Full entry
	influenza A(H1N1)v virus (Preferred)	★★★★ +@
	novel flu virus	★★★★
EN	novel influenza virus	★★★★
	novel influenza virus A(H1N1)	★★★★ +@
	Mexican influenza virus (Deprecated)	★★★★
	swine influenza virus A(H1N1) (Deprecated)	★★★★

Figure 4: Preferred terms vs Deprecated terms.

The usage note included at the conceptual level of the terminological record reads as follows:

ECDC prefers to use the term influenza A(H1N1)v (where v indicates variant), which has been chosen by WHO's Global Influenza Surveillance Network and helps distinguish the virus from seasonal influenza A(H1N1) viruses and A(H1N1) swine influenza viruses. A name for the disease caused by the virus has yet to be determined by WHO but the term 'swine flu' is inaccurate for what is now a human influenza. REF: European Centre for Infectious Disease Control ECDC Interim Risk Assessment > Health Topics > Documents > Human cases of influenza A(H1N1)v, <http://www.ecdc.europa.eu/en...> (27.8.2009)

Metalinguistic labels such as Deprecated/Obsolete or Preferred are also complemented by the systematic use of reliability codes, which are captured in a different field in the database. Less reliable terms may indeed be included in the database in order to offer as much information as possible to the translators, but a low reliability code provides an indication that the term is based on information coming from less trusted sources, as is the case for the distinction between the deprecated term *dual fuel vehicle* (reliability =2) and the preferred term *bi-fuel vehicle* (reliability =3), as in Figure 5 below.

Domain ENERGY, Means of transport

en	
Definition	vehicle with two separate fuel storage systems that can run part-time on two different fuels and is designed to run on only one fuel at a time
Definition Ref.	Regulation (EC) No 692/2008 implementing and amending Regulation (EC) No 715/2007 on type-approval of motor vehicles with respect to emissions from light passenger and commercial vehicles (Euro 5 and Euro 6) and on access to vehicle repair and maintenance information 32008R0692/EN
Note	On internal combustion engines one fuel is petrol (gasoline) or diesel, and the other is an alternate fuel such as natural gas (CNG), LPG, or hydrogen. The two fuels are stored in separate tanks and the engine runs on one fuel at a time, unlike <i>flexible-fuel vehicles</i> (IATE:210005), that store the two different fuels mixed together in the same tank, and the resulting blend is burned in the combustion chamber. REF:Wikipedia > Bi-fuel vehicle, http://en.wikipedia.org/wiki... (28.8.2009)
Term	bi-fuel vehicle (Preferred)
Reliability	3 (Reliable)
Term Ref.	Regulation (EC) No 692/2008 implementing and amending Regulation (EC) No 715/2007 on type-approval of motor vehicles with respect to emissions from light passenger and commercial vehicles (Euro 5 and Euro 6) and on access to vehicle repair and maintenance information 32008R0692/EN
Date	06/12/2013
Term	dual fuel vehicle (Deprecated)
Reliability	2 (Minimum reliability)
Term Ref.	Wikipedia > Bi-fuel vehicle, http://en.wikipedia.org/wiki... [6.12.2013]
Term Note	'Dual fuel vehicle' may refer either to a 'bi-fuel vehicle' (as defined here) or to a 'flex fuel vehicle' (see IATE:210005).
Date	06/12/2013

Source: COM

IATE ID: 2242494

Figure 5: Preference labels and reliability codes.

A label such as Obsolete may be used to mark terms which are no longer in official use. Official denominations for countries or cities may indeed change, as was the case for Bombay, which was changed to Mumbai. Depending upon the context, translators may need to preserve the former name (in historical documents, for instance), which is why the two terms need to coexist in the database, with labels distinguishing them.

5 ECHA-term

IATE is a “generalist” database covering many subject fields. It is one of the key resources used by translators, who contribute to its enrichment alongside the terminologists who manage the database. The Translation Centre for the Bodies of the European Union also created much more specialized databases, such as the ECHA-term database, which was developed for the European Chemicals Agency (ECHA), located in Helsinki, Finland. ECHA was created in 2007 to implement the European Union’s

chemical legislation, and more particularly the REACH Directive which was adopted to improve the protection of human health and the environment from the risks that can be posed by chemicals, while enhancing the competitiveness of the EU chemicals industry. REACH also promotes alternative methods for the hazard assessment of substances in order to reduce the number of tests on animals. With the REACH regulation, companies are responsible for providing information on the hazards, the risks and the safe use of the chemical substances they manufacture, import and transport throughout the European Union. The Classification, Labelling and Packaging (CLP) Regulation introduced a globally harmonised system for classifying and labelling chemicals in the EU, thereby ensuring that the hazards presented by these chemicals are clearly communicated to workers and consumers.

The European Chemicals Agency therefore has very important multilingual communication tasks (the Translation Centre produced over 25,000 pages of translations for ECHA in 2013, mainly leaflets, technical guidance, web content, IT manuals, administrative documents and news items translated into all the official languages of the European Union). In 2009, the ECHA-term project was launched with the objective to provide ECHA and its stakeholders with a reliable, coherent, and up-to-date source of terminology to harmonise the use of terminology in the REACH and CLP context, to enhance clear communication and ultimately to reduce costs for the stakeholders. The more general aim was to help industry comply with the legal requirements, to support the national authorities in their work and to improve the quality of the translated material. The Translation Centre collaborated closely with ECHA to create a terminological database, which included the compilation of a multilingual database of over 1200 terms related to REACH, CLP and the biocides regulation, as well as “substances of very high concern” (over 50 terms) in 23 languages. The project also included the development of a platform for the dissemination of the contents.

5.1 Compilation of Contents for ECHA-term

The general process for the creation of the terminological contents of the database can be described as follows:

- Definition of a relevant corpus in the source language (usually English);
- Semi-automatic extraction of concepts and completion with definition, reference, context, note etc. by terminologists;
- Validation of the monolingual glossary by 2 or 3 translators to ensure that the data is relevant (are any key concepts missing)
- Formal revision by English terminologist;
- Validation of the monolingual glossary by ECHA’s experts;
- Multilingual phase: target equivalents and relevant information are completed by the Translation Centre’s terminologists;
- Ideally target language equivalents are validated by experts (ECHA);

- Import of data in ECHA-term;
- Maintenance of data following user feedback

5.3 Main Features

The database, which can be consulted free of charge at <http://echa.cdt.europa.eu>, offers a range of search options. Users can search by:

- Terms
- EC numbers
- CAS numbers (unique numeric identifier designating one substance in the Chemical Abstracts Service)
- GHS codes (Globally Harmonised System)
- Hazard
- Precautionary statements

In addition to the search criteria above, an alphabetical list of terms can also be browsed. A word cloud also appears to the left of news items on the home page, displaying the most-frequently searched terms, as illustrated in Figure 6. A lemmatization tool is also included to provide a match with respect to the contents of the database even if the query is inflected (as in a query on the plural form in substances of very high concern).

ECHA-term Multilingual Chemical Terminology by ECHA

Search criteria
EN - English substances of very high concern Define
Translate to FR - French

Show alphabetical list

11/02/2014
More terms in Croatian on ECHA-term: the biocides terminology is now online
The key terms related to biocides can now be found on ECHA-term in Croatian. In total, the database contains 810 entries in Croatian (including REACH and CLP terminology and the substances of very high concern).
More terms related to biocides will be available in 23 EU languages later in March.
→ News archive

article authorisation candidate list clp compliance check dnel downstream user endpoint exposure scenario guidance hazard information requirement inquiry intermediate mixture notification pbt reach registrant registration registration dossier restriction risk management measure safety data sheet sief stakeholder substance svhc testing proposal use

Figure 6: ECHA-term home page.

The user can choose to either define the monolingual term or to translate it into one of the 23 supported languages. The presentation of term entries is very similar to the layout offered by IATE, as one can see in Figure 7 below.

The screenshot shows the ECHA-term website interface. At the top, there is a search bar with 'EN - anglais' selected and 'SVHC' entered. Below the search bar, there are buttons for 'Définir', 'Traduire en', and a dropdown menu for 'FR - français'. A link 'Afficher la liste alphabétique' is also visible. The search results show 'Résultats 1-1 sur 1 pour SVHC'. The first result is 'Chemical legislation, REACH' with an 'Entrée entière' link. Below this, there are two rows of results: one for 'EN' (English) and one for 'FR' (French). The English row shows 'substance of very high concern' with a star rating and a document icon. The French row shows 'substance extrêmement préoccupante' with a star rating and a document icon. A definition box is open, providing a detailed explanation of SVHC in the context of REACH, CMR, PBT, and vPvB, and listing criteria for substances of very high concern.

Figure 7: EN->FR translation of SVHC (“substance of very high concern” - “substance extrêmement préoccupante” in French).

Figure 8 below illustrates the inclusion in ECHA-term of precautionary statements. Such statements are phrases that describe the recommended measures to minimise or prevent adverse effects resulting from exposure to a hazardous substance or mixture due to its use or disposal. It is easy to understand why standardisation is crucial in this context, since the manufacturers and the chemical industry need to make use of “pre-fabricated language” which is in a way similar to the use of controlled language and vocabulary in the aviation and aeronautic domain. The labelling and packaging of dangerous goods need to be clear and unambiguous, which leaves little room for creativity and stylistic variations. It is therefore not surprising to find fairly lengthy phrases and linguistic material in this terminology database which goes beyond the traditional notion of a term typically viewed as a noun phrase. Consider the following examples of precautionary statements whose translations are provided into all the EU official languages:

- Do not spray on an open flame or other ignition source. [see Figure 8]
- Avoid contact during pregnancy/while nursing.
- Keep cool. Protect from sunlight.
- Rinse cautiously with water for several minutes.

- Do not pierce or burn, even after use.
- Keep away from any possible contact with water, because of violent reaction and possible flash fire.

The database also includes hazard statements such as:

- Contains gas under pressure, may explode if heated
- In contact with water releases flammable gases which may ignite spontaneously.
- Harmful to aquatic life with long lasting effects.
- Toxic by eye contact.

The examples above make it abundantly clear that the traditional distinction between lexicography and terminology is more and more blurred. The phraseological patterns displayed in the precautionary statements above sometimes correspond to entire sentences (in the imperative form). In some cases, they correspond to what would be considered a collocation in a traditional dictionary. *Keep cool* is a case in point: absent any context, the phrase is highly ambiguous ('stay calm and relaxed', 'gardez votre calme' in French, is a possible meaning). In the chemical legislation covered by our specialized database, the advice is not ambiguous and indicates that a product should not be exposed to high temperatures ('tenir au frais' in French). One can imagine the possible disastrous consequences of mistranslations of such labels if the appropriate terminology is not respected.

5.3 Pictograms as Terms

Terms included in terminology databases are most often noun phrases, although, as we have seen above, verbal collocations such as *keep cool*, complex imperative sentences, other types of prefabricated phraseological patterns and even modal auxiliaries may be granted term status in such databases. A more recent trend appears to be the inclusion of items which are traditionally seen as non-translatable material, such as images or diagrams. ECHA-term has innovated in this context, with the inclusion of pictograms used in the chemical CLP legislation. Figure 8 illustrates the information displayed for a pictogram corresponding to corrosion, which refers to a type of physical or health hazard. Other pictograms such as skull and crossbones, exclamation marks, gas cylinders or exploding bombs will allow users to quickly find out the meaning of such graphical representations, a crucial element for the industry, but also for firemen and civilian protection specialists in emergency situations. Such CLP pictograms can be used when creating safety data sheets and training material in national languages in the various Member States of the European Union.

[Back to search screen](#)

Alphabetical search

EN - English Precautionary statement Do not spray on an open flame or other ignition source.

[Feedback](#)

Other languages: [BG](#) [CS](#) [DA](#) [DE](#) [EL](#) [ES](#) [ET](#) [FI](#) [FR](#) [HR](#) [HU](#) [IT](#) [LT](#) [LV](#) [MT](#) [NL](#) [PL](#) [PT](#) [RO](#) [SK](#) [SL](#) [SV](#) [All](#)

Domain	Chemical legislation, CLP
Code	P211
FR	
Precautionary statement	Ne pas vaporiser sur une flamme nue ou sur toute autre source d'ignition.
Reliability	★★★★
Term Ref.	Annexe IV, Règlement (CE) n o 1272/2008 du Parlement européen et du Conseil du 16 décembre 2008 relatif à la classification, à l'étiquetage et à l'emballage des substances et des mélanges, modifiant et abrogeant les directives 67/548/CEE et 1999/45/CE et modifiant le règlement (CE) n o 1907/2006, 32008R1272/FR
Term Note	Conseils de prudence — Prévention
EN	
Precautionary statement	Do not spray on an open flame or other ignition source.
Reliability	★★★★
Term Ref.	Annex IV, Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006, 32008R1272/EN
Term Note	Precautionary statements — Prevention

Figure 8: Precautionary statements.

Alphabetical search

EN - English Pictogram corrosion

[Feedback](#)

Other languages: [BG](#) [CS](#) [DA](#) [DE](#) [EL](#) [ES](#) [ET](#) [FI](#) [FR](#) [HU](#) [IT](#) [LT](#) [LV](#) [MT](#) [NL](#) [PL](#) [PT](#) [RO](#) [SK](#) [SL](#) [SV](#) [All](#)

Domain	Chemical legislation, CLP
Code	GHS05
Pictogram	
Related	acrylamide , ammonium bichromate , ammonium chromate , ammonium dichromate , chromic acid [H2Cr2O7] diammonium salt , diammonium dichromate , dichromic acid diammonium salt
EN	
Pictogram	corrosion
Reliability	★★★★
Term Ref.	Annex V, Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008, 32008R1272/EN
Term Note	Physical hazard; health hazard

Figure 9: pictograms in ECHA-term (corrosion).

5.4 A User Perspective

In 2012, the European Chemicals Agency conducted a survey in order to better understand to what extent the ECHA-term users were satisfied with the glossary and who these users were. With around 3200 visitors per month and about 300 search queries per day, ECHA-term is obviously a much more 'confidential' database than IATE. Yet, contrary to original expectations, only 22% of the users who responded to the survey are translators. The majority (56%) of the respondents indicated that they actually work in the chemical industry, the remainder working for EU Member States, for international organisations and for NGOs such as environmental protection agencies. The vast majority of the users indicated that the database helps them understand the REACH and CLP Regulations and stressed that the terms were relevant to them. The multilingual nature of the database is considered a key feature for the users, who mainly visit the ECHA-term web site in order to look up the translation of specialized terminology. A feedback mechanism also allows users to suggest new terms (around 100 terms are added every year) and to provide comments about existing entries. Users can also download the entire database.

Such results explain why this project has been considered a success. They show that a small, but well-maintained glossary can make a difference for the Agency's stakeholders, who use it on a daily basis. It contributes to the use of a unified and consistent terminology in all the translations related to the regulations in the chemical field, a *sine qua non* in multilingual communication.

6 Conclusion

I started this paper with a reference to ten Hacken's remark that "dictionaries are not descriptions of a language, but tools with which users of the dictionary solve problems of a particular type". The nature of the dictionary therefore determines which types of problems it can solve, he argued in his 2012 Euralex paper. The same is true of terminology databases, as we have seen. The use of such databases in today's globalized economy and multilingual world accounts for the nature of the linguistic information included in these electronic resources. Translations, definitions, acronyms, subject field labels, usage notes and examples are similar to what can be found in monolingual or bilingual dictionaries. Some other types of information are used somewhat differently or to a larger extent in terminology databases, however. References and reliability codes are crucial in term bases, although they are virtually absent in traditional monolingual dictionaries, even if historical dictionaries such as the OED do make use of reference information, an essential element when sketching the historical development of a given lexical item. Information about the author of a term entry is also important in term bases, given that terminology is frequently associated with standardized language used by specific communities of users.

Metalinguistic labels, which are not different from prescriptions, as is pointed out by ten Hacken (2012: 843), are found both in traditional dictionaries and in terminology databases. Labels such as *nonstandard*, *preferred* or *obsolete* reflect the lexicographer's or the terminologist's attempt to capture a judgment which will be exploited by the end user of the linguistic resource. This label will help the user decide whether it is pragmatically appropriate to use a given form (e.g. a translation), depending upon the context, the nature of the document produced, the client for whom the translation is made, etc.

The distinction between terminological items and lexical items is also more and more blurred. The nature of the linguistic items discussed by terminologists has undoubtedly evolved over the last 10-15 years. The inclusion in specialized electronic glossaries and term bases of items such as modal auxiliaries, complete sentences, collocational or phraseological patterns, images, diagrams and pictograms is driven by the needs of the target users and the requirements of modern multilingual communication. In this respect, the road from lexicography to terminology is more a continuum, a cline, rather than a hard-and-fast dichotomy.

Another issue which will also need to be addressed in the future is the level at which terminology should be managed (see also Fontenelle and Rummel, in press). Should terminology management be centralized or should it rather be done at the local level, down to the level of individual translators in big translation services? How then should the data be made available to its users? Clearly, web technologies have made it possible to disseminate terminological knowledge to millions of users (the publicly available version of IATE received 44 million queries in 2013). However, one of the major stumbling blocks in the dissemination process is that it is still up to the individual translator or user to 'suspect' that a term base such as IATE or ECHA-term is able to provide interesting and useful information about a given term. What is therefore needed is a mechanism which can alert a translator that a word or a sequence of words appearing in the source text she is dealing with corresponds to a term entry in a specialized database for which an equivalent exists in the target language. Such tools exist at the local level, but will need to be linked to huge databases like IATE, without forcing the translator to host a local copy of a 9-million-term database, which is not recommended for obvious performance reasons. A number of initiatives are currently under way to tackle this crucial issue. Another burning issue is also related to the use of term checkers which ensure that only recommended (read 'validated') terminology is used and that 'dispreferred', obsolete or deprecated terms are not used by the translator. Once again, such obstacles require some level of linguistic processing to match inflected forms in a text and the canonical forms recorded in the quality assurance mechanisms. Organization challenges are also at stake here, since it is crucial to determine who is doing what. Such challenges are different from the question revolving around the distinction between terminological and lexical items, but they are equally important from a user's perspective. Should translators themselves take care of the terminological work? Where should they capture the preferences expressed by the "clients"? Should it be done centrally or locally? How can we make sure these preferences are not one-off information, but can be recycled in future translations to avoid repeating the same mistakes?

These questions do not seem to have clear-cut answers: what is clear, however, is that the solutions can only be effective if they combine technological innovation, using the appropriate amount of linguistic processing, together with organizational changes to make the best use of what modern technology can offer to language workers.

7 References

- Fontenelle, Th. and Mergen, C. (1998): « Les interfaces terminologiques au Service de Traduction de la Commission européenne », *Terminologie et Traduction*, 1.1998, Commission des Communautés Européennes, Luxembourg, 210-221.
- Fontenelle, T. (in press) 'Bilingual Dictionaries'. In Durkin, P. (ed.) *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press.
- Fontenelle T., Rummel, D. (in press) 'Term banks'. In Hanks, P., De Schrijver, G.-M. (eds) *International Handbook of Modern Lexis and Lexicography*. Springer.
- Jacquemin, C. and Bourrigault, D. (2003) 'Term extraction and automatic indexing', in Mitkov, R. (ed.) *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press. 599-615.
- Pearson, J. (1998) *Terms in Context*. Studies in Corpus Linguistics. John Benjamins Publishing Co. Amsterdam.
- Reichling, A. (1998) 'Gestion centrale de la terminologie, EURODICAUTOM et ses outils satellites'. *Terminologie et Traduction*, 1.1998, Commission des Communautés Européennes, Luxembourg, 172-201.
- Rey, A. (1992) *La terminologie - Noms et notions*. 2ème édition. Collection « Que sais-je ? ». Presses Universitaires de France, Paris.
- Sager, J. (1990) *A Practical Course in Terminology Processing*. John Benjamins Publishing Company. Amsterdam.
- Simpson, J. (2000) 'Preface to the third edition of the OED'. Accessed at: <http://www.oed.com/public/oed3preface/preface-to-the-third-edition-of-the-oed>. [05/04/2014]
- Ten Hacken, P. (2012) 'In what sense is the OED the definitive record of the English language?', Fjeld, R., Torjusen, J. M. (eds) *Proceedings of the 15th EURALEX International Congress*, University of Oslo, 834-845.
- Ten Hacken, P. (2009) 'What is a Dictionary? A View from Chomskyan Linguistics'. *International Journal of Lexicography*. 22.4: 399-421.
- Wright, S. E. and Budin, G. (2001) *Handbook of terminology management (Volume 2): Application-oriented terminology management*. John Benjamins Publishing Company.

