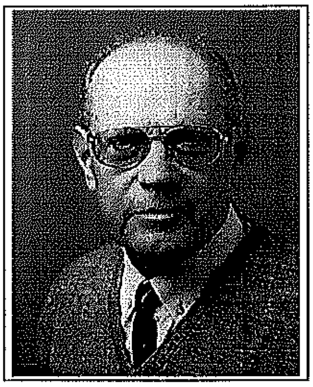
Database Merging and the Compatibility of Indexing Languages

Jacques Maniez
Formerly at the Institute of Technology, University of Dijon, France



Jacques Maniez taught many years on the IUT (Institut Universitaire de technologie) at the University of Dijon (France) in the knowledge organization area. His special emphases include linguistics and semiotics. Since he retired, he has been the French coordinator of ISKO.

Maniez, J. (1997). Database Merging and the Compatibility of Indexing Languages. *Knowledge Organization*, 24(4), 213-224. 10 refs.

ABSTRACT: Compatibility is the paradise lost of information scientists, the dream of a universal communication between information languages. Paradoxically the information languages increase the difficulties of cooperation between the different information databases. This noxious side-effect has become flagrant for the latest decade since the shared cataloguing and the telecharging facilities have increased the exchanges. After defining the notion of information compatibility, the author shows that it meets the same care of semantic coherence as the information languages themselves. Then, relying on the lessons of linguistics and automatic translating, he describes two types of viable solutions: the harmonization of several information languages (an uneasy and costly processing); and the automatic harmonization of the indexing formulas through prefabricated concordance tables, an easier solution which can however be hampered by structural discrepancies. Last he sketches a critical view of the concept of switching language.

1. A Few Prolegomena

1.1. The Need for a General Study of Compatibility Problems

The compatibility of information languages has been studied extensively, as shown by the bibliography issued by Ingetraut Dahlberg on the occasion of the Warsaw seminar "Compatibility and Integration of Order Systems" (Compatibility, 1996, p. 193-231*). (See the attached Appendix A for a report on this seminar.)* Dahlberg's bibliography records more than 500 documents published on that subject over the last 40 years. The 70s were the most productive decade, with 200 papers, 3 international conferences, and a directory from UNESCO. After a slight decline during the 80s, the study of compatibility problems is now again a central issue in the information field, due to the growing number of exchanges between different databases, chiefly via the WEB.

In a library or an information center, merging many external documents into the main base gives rise to discrepancies between index terms, which result in much noise and silence when the base is used. Therefore, we think it is time to define a general theory of compatibility in Information Science, for while most papers present local experiments, a small number of studies deal with general problems (Schmitz-Esser, Riesthuis, Dahlberg in: Compatibility, 1996) and the studies are often more descriptive than critical, for instance (Lancaster 86, p. 179-215).

Moreover, some new factors have modified compatibility issues over the last decade. These factors are improvements in automatic language processing and improvements in on-line searching. Computer-aided indexing is no longer perceived as an experimental process and can now be considered the natural complement to manual indexing. Therefore, we wish to propose abandoning the conventional opposition of 'INDEXING LANGUAGES vs. NATURAL LANGUAGES', to be replaced by the following two-level scheme:

INDEXING LANGUAGES

INFORMATION LANGUAGES
NATURAL LANGUAGES

And accordingly, we will extend our survey of compatibility problems to natural languages.

On the other hand, automatic information processing has become so fast that a computer can convert an indexing formula into equivalent statements almost immediately, thus allowing for the use of simpler technical solutions.

https://doi.org/10.5771/0943-7444-1997-4-213 https://www.inlibra.com/de/agb - Open Access -

1.2. How Can We Define "Information Language Compatibility"?

At first glance, the concept of "linguistic compatibility " is not clear (Riesthuis, in Compatibility, 1996, p. 24), being itself "incompatible" with the usual meaning of the term «compatibility», which is "The quality of things or beings which are capable of existing together in harmony" (Webster) or "...that can be used together" (Collins). Using this definition, it makes little sense to say that "two languages are compatible", as they are autonomous and do not have to be adapted.

However, the term has been allowed for in the framework of Information Science through decades of use in literature of the field and also through use by UNESCO, and its particular meaning(s) can be clarified when used in the environment of our field, where it is often associated with words such as 'integration', 'harmonization', reconciliation' or 'concordance'. These terms all imply the concept of convergence, which means that the compatibility of languages is not an intrinsic quality but rather a target to be reached, as in the phrases "the objectives of compatibility " or "problems of compatibility ".

Moreover, interlinguistic convergence can be reached either at the level of languages themselves or at the level of texts. We might call to mind the classical Saussurian dichotomy between 'language', i.e. a system of verbal signs and 'speech', i.e. the utterances produced through a language (Saussure, 1949). Accordingly in Information Science we can distinguish indexing languages (systems of signs designed for document description) and the utterances produced through indexing languages (for lack of a usual expression, we'll call these utterances 'indexing formulas').

The convergence of indexing languages is the most ambitious kind of interlinguistic compatibility. Here the term means the 'search for proximity or similarity', the highest degree of which is identity or total fusion. In this last case, the identity of utterances is logically derived from the identity of languages.

As for the convergence of indexing formulas, it can be reached by the classical device of translation. Here 'compatibility' is close to 'translatability' or 'convertibility'. This is the most commonly given definition of the term. For example, "compatibility means that for each term A of an information language P there is a term A' in an information language Q with the same meaning, so that we can convert A into A' without changes in meaning" (Riesthuis, in Compatibility, 1996, p.25.).

The distinction between the unification of languages and the unification of indexing formulas gives us a convenient division for this survey. But before making that distinction, it will be useful to examine the origins of compatibility problems in the context of the communication of information.

1.3. The Paradox of Information Languages Incompatibility

The search for the compatibility of information systems is not a minor problem, as it is related to their essential objective, which is to simplify the connection between those who need information sources about a theme and potentially relevant documents.

At the starting point of a search there is a 'subject', which is both a psychological and a linguistic entity. It in fact implies an interest in a special field of knowledge (psychological aspect) but which nonetheless cannot be communicated without some formulation (linguistic aspect). This process of formulation is itself a first kind of translation, from our internal or conceptual language into a so-called natural language, and as with any translation, it involves a bit of hesitation and some conceptual distortion.

On the other hand, there is a huge amount of information stored in documents. In order for each of them to be easily located by the user, it is necessary to label them with a short surrogate representing ones subject. Such is the task of indexers, who, after analyzing a document, reduce it to its essential field of interest, and first formulate it in their usual language. Let us stress that the process of indexing is the result of a double 'translation': a conceptual one (reducing to the essential) and a linguistic one (formulation), so that it implies a double risk of distortion from the original document. But that is not all. For after these parallel formulations, by the user and by the indexer, even if they use the same language, it cannot be taken for granted that the terms used by each person to define the same subject are both identical; our natural languages are too rich and too fuzzy! So it is necessary to resort to a new kind of translation (interlinguistic) in order to transfer the formulation of the subject from a plurivocal language to a biunivocal language. This is the raison d'être of information languages, the golden rule of which could be "One subject for an utterance, one utterance for a subject". So it can be stated that information languages have developed from the desire to reach the essential kind of information compatibility; in other words, conceptual compatibility (Schmitz-Esser, in Compatibility, 1996, p.11.21), which means the agreement of the requester and the indexer on the meaning of the words they use in common.

Unfortunately, the increase in the number of these languages has brought about a disorder that has been well known to humans since the Tower of Babel. For in the last few decades so many information languages were designed that their number exceeds that of natural languages used in information literature. And when a library merges external references into its catalog, it generally inherits indexing formulas that are incompatible with its own indexing formulas. From the standpoint of conceptual compatibility, this situation is worse than the initial one! This side effect of information languages is paradoxical, in that it goes against the following founding principles. As noted in (Lancaster, 1986, p.181) "Perhaps somewhat surprisingly, controlled vocabularies tend to promote internal consistency within information systems but reduce intersystem compatibility. Systems based on natural languages are inherently more compatible than those using controlled vocabularies." Paradoxical but logical. For any new language relies on specific features which distinguish it from other languages, and hamper intercommunication. A typical example is given by slang languages, which are designed by small social groups in order to have their own secret language.

As professionals of the information field, we have all seemingly underestimated the negative effects of this proliferation, welcoming every new information language as an overall enrichment. It would surely have been more judicious for specialists in every area to agree on specific terms in order to gradually lower the discrepancies. The reality has been somewhat different, in spite of the praiseworthy advice of standardization committees and of a few pioneers. According to one of them, I. Dahlberg, "In most cases, the proposed projects of harmonization were not granted the necessary financial help simply for reasons related to the huge amounts of money deemed necessary for their realization. Also, we may have lacked insight about the usefulness of the results to be achieved. And last but not least, as regards the systems involved, there may have been reservations about sharing with others that which one had stored ... sometimes even for reasons of competition." (Dahlberg, in Compatibility, 1996, p. 34-35). Today, worldwide information exchanges make us more sensitive to such drawbacks and many people understand that in this new environment, the harmonization of information languages has become the natural complement to their variety, in order to reach a universal conceptual compatibility. As we have seen above, this harmonization can be sought either between the languages themselves or between the indexing formulas.

2. A Radical Solution: Unifying All Information Languages

A radical solution for linguistic compatibility would be the general acceptance of a unique language, either from consensus or due to social pressure. As stated humorously by G. Riesthuis "There are two simple solutions. The first one is not to use a controlled information language at all, the other one is to get everyone to use the same information language" (Riesthuis in: Compatibility, 1996, p 24). For the moment, let us take these proposals in earnest.

Initially, the first one seems to be suicidal. However, it is not quite contemptible, at least as an auxiliary tool. It relies on choosing a common natural language (English, of course) as a universal indexing language and on making the best use of automatic indexing, which is regularly improving (homogenization of lexical forms, recognition of usual syntagms). As for the second one, it is a dream as beautiful and elusive as the search for a universal language. For information systems have become a gigantic network of knowledge shared by various (and often rival) organisms, each of which is keen on keeping its own idiosyncrasies, in particular its own information language. So the virtuous endeavors for such unification must carefully deal with existing conditions and are in practice reduced to only two types of projects:

- a universal solution, working for a progressive acceptance of a common, universally accepted indexing language as a secondary tool.
- a sector-based solution, which merges or links together several closely related information languages.

2.1. Acceptance of a Common Secondary Language

Is it unreasonable to imagine that information officers all over the world might agree on the same indexing language as a secondary indexing tool? Such a solution could optimize the internal habits of every information center as well as the possibilities for merging external databases. But what would be the ideal profile of that indexing language? Universal in its content, structurally simple, as remote as possible from any particular idiom and supported by some powerful and sound international institution. A century ago, these features corresponded precisely to the aims of the UDC. Unfortunately, historical developments and the attachment to specific characters triumphed over the spirit of unity. Nowadays, can the UDC still be a reliable candidate for such a role? Some assume that it is, because of its past, because of its broad geographical location and because of the support of the FID (Sosinska-Kalata in: Compatibility, 1996, p. 143-151). However, in my opinion the structural modifications with which it is currently confronted, as fruitful as they may be, do not favor the universal acceptance of the UDC.

2.2. Unifying Similar Languages

2.2.1. Difficulties in Unifying Languages

Unifying two languages is a difficult process. The languages must first of all have semantic and structural similarities. On the other hand, unifying can take place at several levels, the ultimate one being fusion or complete identity.

As regards natural languages, the proximity of a pair depends on vocabulary, phonetics, type of writing, syntactical structures and genealogical relatedness. The harmonization of some features is often based on practical needs, either technical or economical. For instance, the widespread use of English vocabulary throughout the world in air navigation or the use of Roman characters in writing Chinese and Japanese. As for unifying information languages, the desired similarity is chiefly lexical and structural. Paradoxically, the discrepancies are in fact still stronger in information languages than in natural languages. Roughly speaking, one can tell that in spite of huge structural and semantic discrepancies, all human languages are combinations of elementary concepts which can be denoted by the casual term "words". On the contrary, information languages divide fields of knowledge according to two incompatible categories, whose basic units have different templates : classifications, in which the basic unit is the subject, and post-coordinated languages, in which the units are concept-terms (or, roughly speaking, words). The former are 'enumerative', or in other words mere lists of ready-made subjects, coded with class numbers and treelike (the subjects are superimposed from the most general to the most specific). On the other hand, in post-coordinated languages, as in natural languages, the subjects are made up of a combination of combined concept-terms. Here the syntax, as crude as it may be, is compulsory, while the hierarchical relationships are optional.

We may illustrate the onus of these discrepancies with a fictive but plausible example. Let us assume that a language designer intends to reconcile five information languages belonging to the same field, namely 'Information Science':

Two classification schedules: that of our periodical 'K.O.' (approximately 120 subject headings) and that of the French bibliographical Bulletin of the INIST-The Institute of Scientific Information – (approximately 45 subject headings).

The alphabetical list of authorized keywords used for the index of the French periodical 'Documentaliste' (approximately 250 terms).

Two thesauri dealing with the field of documentation: the first published by the INIST, made up of 1000 terms, and an experimental faceted thesaurus made up of 1500 terms.

It is immediately obvious that we cannot unify a set of languages that are so different and that we must divide it into two subsets. A first criterion could be the number of semantic units, according to which we would begin by selecting the first three languages, and then grouping together the two thesauri. However, the structural incompatibility described previously prevents us from automatically associating the two classification schedules with the list. Therefore, the only reliable division would be to separate the two classification tables on one hand and on the other hand to isolate the remaining three post-coordinate languages. Now if we look more closely at the two classification schedules, we can see that each of them follows the logic of the content of the periodical it is designed to organize and that each content is different from the other; more general in the INIST bulletin, and focused on the organization of concepts in K.O.

INIST Bulletin	Knowledge Organization
1 General studies, library science	0 Form divisions
2 Documentation institutions	2 Theoretical foundations
3 Documentation sources Document processing	3 Classification systems and thesauri: Structure and construction
Information sources	4 On Universal Classification systems and thesauri
4 Information analysis	

It is clear that the main classes of the two schedules are far from being similar, and that an attempt to unify them would be doomed to failure. Let us now return to the three post-coordinate languages. In this case it is possible to unify the terms expressing the same concepts in each language, all the more so as some of them are already identical (thesaurus, classification, library,...). But these languages remain very divergent in their semantic structure. The index of 'Documentaliste' is but a list of controlled words, whereas the terms of the thesauri are meant to establish canonical semantic relationships. Moreover, the macrostructures of the two thesauri are different. The first one divides the whole field of documentation into microfields, while the second structures through the use of facets.

In short, the harmonization of information languages is subjected to radical conditions of similarity and one can easily understand why such attempts are so infrequent.

2.2.2. Methods of Unification

The most encouraging results have been reached in harmonizing the post-coordinate languages of narrow fields, of which the famous UMLS (Unified Medical Language System) by the National Library of Medicine offers a particularly impressive example (Hoppe in: Compatibility, 1996, p 97-100). This system integrates and unifies more than 200,000 concept-terms from 30 different sources! But there are several steps in the process of unification, and the integration of languages often remains at the level of lexical compatibility, which is attained if every concept common to several languages is expressed by one and the same term. The inventory of similarities and discrepancies between different languages can be made far more easily with a computer if one disposes of electronic lists of the terms (Soergel, in Compatibility, 1996, p. 47-75).

In order to unify several thesauri you generally choose one of them as a pilot language, in principle the one which offers the richest and most balanced semantic grid, with a reasonable rate of pre-coordinate terms. Progressive harmonization begins with the comparison of synonymous and quasi-synonymous terms between languages, and in each case on the choice of one preferred term. As soon as a decision has been taken, the common term is poured into the new cumulative language, and for each old one an equivalence is chosen between the original and the new term, as a provision for concordance tables. This painstaking work of lexical harmonization is in itself time consuming and expensive. And should you also choose to reach the level of relational compatibility, the harmonization of semantic relationships can raise inextricable problems. So it seems reasonable to keep the original structure of every thesaurus and to select the most appropriate semantic structure for the new one.

Another tool for harmonization is the Macrothesaurus, which was devised in the 70s by Ms Wolff-Terroine and her team (Institut, 1979). It is based on the principle that all sciences use a pool of common concepts around which the specific terms of every discipline are articulated. The aim is to go from the conceptual community to the lexical community. So sector-based thesauri which adopt the elements of the macrothesaurus can thus communicate on the upper levels thanks to that superstructure of general terms.

In conclusion to this chapter it is obvious that the general attempts at documentary reconciliation are paved with as many pitfalls as good intentions, and that such projects are more manageable if their goals are moderate. Unifying different information languages is time consuming and expensive. Moreover, while this solution may provide coherent indexing formulas for the future, it does not dispense one from creating concordance tables for these formulas. Which leads us to the second solution.

3. An Easier Solution: Converting the Indexing Formulas

Theoretically it is always possible to translate an indexing formula from a language A into a language B. Therefore, if we assume that the original database is indexed with A and the imported records are indexed with B, it would suffice to reindex them with B in order to unify the database, without unifying the languages. Quite a simple solution, but impossible for obvious financial reasons if done manually. Fortunately, it is feasible to computerize the greater part of this process at a reasonable cost. And if this works for two languages, why would it be impossible to imagine a universal converting system which would convert any information language into any other information language via a conversion center ?

Before exploring the possibilities and the theoretical limits of such ambitions, let us briefly mention some lessons drawn from the secular experience of human translation and the more recent but nonetheless impressive results of computer-aided translation.

3.1. Some Lessons from Human Translation

3.1.1. The distance between two languages varies greatly according to the specificity of every pair. But even in the best of cases, there is never a perfect concordance between their ways of representing the real world (an axiom known as the Sapir-Whorf hypothesis).

3.2.2. Consequently, translation is as impossible as necessary, or in other words a perfectly faithful translation is in principle out of reach, and every translated text implies some amount of unfaithfulness. As quoted by a famous Italian proverb 'A translator is a traitor'.

3.2.3. The fact that there is no perfect equivalence between two languages implies that the translation process is not reversible (we all know that from bilingual dictionaries).

3.2.4. Human translating is not a mechanical conversion of words but an interpretation of the source text followed with its formulation in the target language.

3.2. Some Lessons from Automatic Translation

The old dream of a machine that could compete with a real translator was the inspiration for a number of ambitious projects in the early years of computers. Nowadays, after decades of extensive research, successful attempts and disillusions, we can roughly speaking distinguish three levels in the applications of CAT (computer aided translation) :

3.2.1. Translators of Set Phrases.

These first-level linguistic tools are nothing more than improved bilingual dictionaries, which add a lot of common phrases to lexical items about weather, jobs, etc. Therefore their efficiency is very limited.

3.2.2. Automatic Translation Programs.

These second-level packages have much more in common with human translators as they are designed in such a way that they can theoretically process any text from a source language S to convert it into a text in a target language T. (In fact, they simulate the situation of a translator who translates a text in which he understands only the grammatical structure). Each of these systems necessarily involves the following linguistic tools and programs:

Linguistic tools :

* a bilingual, well-documented dictionary from S to T.
* a generative grammar system, matching the structures of S with the structures of T.

Programs :

* a morpho-syntactical analyzer of language S which converts a sentence into abstract grammatical structures.
* a program of conversion of the S structures into the T structures (via the rules of the grammar system).
* a program of conversion of words S into words T (via the dictionary).
* a program of morpho-syntactical adjustment of the T sentence (via the grammar system).

3.2.3. Automatic Translators Using a Pivot Language

Every package of the previous category requires huge investments. Assuming that an international organization wants every text produced in one of the six official languages to be translated into the other ones, it would be necessary to design 30 packages (5 x 6), since the translating process is unidirectional. Therefore, it is tempting to imagine an alternative method using a pivot language as an intermediary between any pair of languages. Such a language should be an abstract set of rules and concepts overlapping all of the possibilities of real languages but independent from any given language. With this solution the number of packages fitting the previous example would decrease from 30 to 12 (two for every language).

Up until now these translators have remained experimental, as the construction of an artificial and universal pivot language creates many problems. But the model is of interest to us, as we will be confronted with a similar problem with information languages.

3.3. Problems of Conversion in Information Languages

There are many common features between natural and information languages in the process of translation. However, the idiosyncrasies of information languages are so numerous that we cannot directly transfer the methods of analysis of human and automatic translation into that of information language conversion. We may state that these idiosyncrasies are the following :

3.3.1. The utterances produced with information languages are not statements but mere subjects; namely, topics likely to stimulate the interest of an author or a searcher. Therefore, the linguistic objects to be translated are cruder than in natural languages.

3.3.2. Unlike computer-aided translation in natural languages, automating the conversion of indexing formulas is not only a desirable aim but an absolute economical necessity, since the cost of manual processing would be too great. So, the relative simplicity of information languages, when considered along with the financial limits of most structures, argue in favor of first-level algorithms, rather than the sophisticated processes used for automatic translation.

3.3.3. Last but not least, we have previously mentioned the discrepancy between two types of information languages : *concept based information languages (e.g. thesauri) and subject based information languages (enumerative classifications).* As a result of this discrepancy the possibility of a bi-directional automated conversion between , for example, the DDC and a thesaurus, is highly unlikely because one cannot foresee the countless combinations of descriptors that are feasible with a thesaurus. The sole bridge that can be built between the two is a one-way bridge from the DDC to the thesaurus. This structural incompatibility must imperatively be taken into consideration when developing a general converting system.

3.4. Achieving Indexing Compatibility via Concordance Tables.

Except in cases of structural incompatibility, it is possible to achieve the harmonization of indexing formulas using a simple model which can be reduced to two elements : a set of one-way concordance tables and a compatibility software for the imported records.

3.4.1. The Concordance Tables

In their most basic form they are two-column tables that establish the closest possible corresponding index terms between A and B. Examples:

THESAURUS A	THESAURUS B
Agricultural training	Agricultural education
Employment	Job
DDC	UDC
621.39 (computers)	681.3

The construction of a concordance table between languages of the same kind can to some extent be computer-aided via a program that detects identical forms in the indexes of classifications and identical or equivalent terms in thesauri. If incompatibility comes from a difference in natural language, the multilingual thesauri of the field or even one good bilingual electronic dictionary can be helpful. The main difficulties in establishing concordances in information languages are well-known: discrepancies in the extent of the fields of knowledge, in the level of specificity, in the level of pre-coordination. The most frequent one, variation in the level of specificity, is common to both information languages and natural languages and can be demonstrated through the following example : many languages, such as Greek, use three different terms to express the new relationships of brotherhood following a marriage (the reader can easily find them out by him/herself) while English has only one term (brother-in-law). In this case, translating any term from Greek into English presents no problem, but the opposite situation leads the translator to a dead end or to a random solution if he/she does not know the context.

Such cases of semantic discrepancy are commonplace in information languages. Thus, some designers of concordance tables take particular care in refining concordance relations. For example, Zimmermann in (Compatibility, 1996, p. 138-142) splits them between lexical equivalence, synonymy and partial synonymy, so that the user can choose a kind of relationship before asking a question.

3.4.2. The Program of Harmonization of Index Terms.

Let us assume that a database indexed with language X is regularly used to integrate records indexed with language Y. The compatibility of indexing formulas can be pursued in different ways according to the choice of the preferential language and the place of the substitution process.

3.4.2.1. Choice of the Preferential Indexing Language.

In principle, the host database keeps its own information language, unless that of the imported records is far better. So it will suffice to establish a concordance table from Y to X along with a program that translates the imported indexing formulas from Y to X. This implies that either both languages have got the same structure or that Y is a classification and X a post-coordinate language. At the query phase the users will go on using X, but if they prefer to use Y their queries can be converted into X.

3.4.2.2. The Choice of the Input versus Output Stage.

In the previous example harmonization is achieved at the input stage when records are imported. On the other hand, it is possible to skip this stage and to harmonize the system at the output stage, thanks to a query using the OR operator.

Query = formula X **OR** formula Y

This kind of solution has been tested by the Battelle laboratories of Columbus under the name Vocabulary Switching System or VSS (Lancaster, 1989, p. 195). Here, concordance tables are replaced by a dictionary of synonyms which are collected from different databases by a largely automated program and then controlled by specialists. The user can then use these equivalents to extend her/his requests to several bases.

4. Searching for the Panacea: Intermediate Lexicon or Reference Language ?

4.1. The Concept of Switching Language.

In the previous section we assumed that the harmonization of indexing formulas or of search formulas had to be reached through the simplest and least expensive processes. This is why we did not even evoke the problems of harmonizing the possible syntactic tools of information languages (facets, relational indexing, auxiliary tables of classifications,...) and restricted the solutions to the semantic equivalence of index terms. It is understandable that this limitation, as well as the drawbacks related to every kind of translation, increase the amount of noise and silence when compared to the original indexing formula. But taking syntactic structures into account in indexing formulas would require sophisticated and expensive programs similar to automatic translation software, for a worthless result.

Even though the method of establishing concordance tables may be quite elementary, their creation remains nonetheless painstaking and costly, which explains why so few tables are commercially available (Scott, 1993). Now if a database imports data from numerous external databases, harmonizing the indexing formulas requires a different table for each external information language. Therefore, several decades ago, the designers of information systems conceived a universal tool capable of transforming any indexing formula into the equivalent formula of any other language, in a far less expensive way than the one-way method. This concept, very close to that of a pivot language, was called *intermediate language* or *switching language*. It consists in a particular information language capable of representing any concept or conceptual combination so that it can express any indexing formula from an information language and conversely translate it into any other language. In this way, we switch from a bilateral and uni-directional process to a trilateral and bi-directional process, as shown below:

Conventional process
Thesaurus A --> Thesaurus B
Thesaurus B --> Thesaurus A
Process using a switching language
thesaurus A <<---> switching language <<---> thesaurus B
Example tumor <<----> 17904 <<----> neoplasm

The economy comes from the fact that it suffices to create two concordance tables for every information language. For example, if we wish to establish a general compatibility between six different languages using the conventional method, we need 30 concordance tables (6x5), whereas we only need 12 tables (2x6) with a switching language. This model is so fascinating that a project was launched in the 70s by UNISIST (UNISIST, 1971) and some prestigious members of the Information Science field were associated with it (J. C. Gardin, E. J. Coates, I. Dahlberg). But we must admit that the results of the study were never encouraging.

4.2. Theoretical Difficulties Connected with the Concept of a Switching Language.

In our opinion, in order to achieve the objectives for which it is built, a switching language (SL) should fulfill three main conditions:

a - to allow for bi-directional convertibility with all information languages. : Lx > SL and SL > Lx

b - to reach the same level of specificity as the most specific information language, so that the loss of information should be minimized in the course of the double translating process.

c - as a standard, to react neutrally to natural languages, and consequently, to be artificially coded.

These conditions are not of equal importance. The first condition is fundamental, because this model is essentially based on the concept of intermediacy. The second condition itself is mandatory as we saw in 3.4.1. The last condition is desirable but not essential. Now if our previous view of two irreconcilable structures within the information languages has some reliability, no information language can provide bi-directional convertibility with all other information languages, which drastically condemns the project for a universal switching language between information languages.

In our opinion, it is because this condition was underestimated that the UNISIST project finally met with indifference. In a recent theoretical and historical article (Dahlberg, 1996, p. 155.163), I. Dahlberg narrates the episodes of this project which eventually ended up in the Coates's BSO - Broad System of Ordering - (Coates et al., 1978). It is a coded classification of about 4 000 subjects which «was used in some applications, mostly for demonstration purposes», but could not be used as an intermediate language, in spite of the recognized competence of the authors. While I. Dahlberg attributes this failure to the poor structure of the BSO classification, we think that the reason is rather to be found in the inadequacy of the theoretical approach.

4.3. The Solution of a Reference Language.

As a general switching language is not viable between information languages, an alternative solution can be imagined with what we shall refer to as a *reference language;* a language which has to fulfill all the requirements of the switching language except the first condition of bi-directional convertibility. For as we have seen in 3.3.3., it is possible to convert all information languages, either enumerative or post-coordinate, into a post-coordinate language, but the opposite cannot be achieved with simple automatic devices. Using a reference language, the process of conversion is bilateral and unidirectional, as with the concordance tables, but the number of tables and conversion programs amounts only to that of the different languages (12 for 12 languages, instead of 12x11= 131).

But the choice of the type of information language for this role of reference tool will differ depending on wither it takes place at the starting point or at the final point of the conversion, as shown below.

Case 4.3.1.	*Case 4.3.2.*
(Classification scheme)	*(Thesaurus)*
From: RL	From IL1 IL2 IL3 IL4 IL5
To: IL1 IL2 IL3 IL4 IL5	To: RL

4.3.1. The concordance tables are oriented from the reference language to the various information languages, in which case the RL can be nothing other than a *classification scheme*. And it should be used at the searching stage only, as a universal search language. Thanks to the concordance tables RL----> IL, a query can be automatically converted into the appropriate formulas. We assume that the ICC (information coding classification) proposed by I. Dahlberg is supposed to play this role. This solution could be well-adapted to Internet databases, which require simple and quick search processes.

4.3.2. The concordance tables are oriented from the information languages to the RL, in which case the RL must be a *universal thesaurus*. Then all the indexing formulas are automatically converted into the general language, thanks to the tables IL---> RL. And the queries should themselves be expressed in the RL.

We are thus close to the solution of a second common indexing language, suggested in 2.1., except that the indexing process is automated. This solution would offer all the advantages of coordinate languages: flexibility, combinability and a high level of specificity.

5. A Widely Open Field of Research

5.1. Searchingfor a Universal Integrating Tool

Exploring the problems of linguistic compatibility in information systems leads us to a review of all the dimensions of their actors and facilities and to focus on the basic notion of conceptual compatibility between the actors, which may be overshadowed by the complexity of the mental tools. On the practical level, the comparative study of information languages and natural languages, as well as that of the different categories of information languages, allows us to develop viable solutions and to exclude illusive projects.

Nowadays, local solutions are already in use, as shown in (Compatibility, 1996), but it is time to switch to more ambitious projects and to search once more for a universal integrating linguistic tool, as in the 70s. Such a plan cannot be developed outside of international institutions, and the choice of the Reference Language will inevitably raise the perpetual questions of power, linguistic leadership and acceptance of a general rule, including technical problems!

5.2. Evaluating the Different Solutions

In the meantime, local solutions are still necessary. Unfortunately, it is difficult to compare the cost effectiveness of the various solutions because feasibility studies are still very limited (Roulin, in Compatibility, 1996, p. 123-135). Due to the lack of numeric data, we can only try to roughly appraise the plausible effectiveness of three solutions applied to databases indexed with several heterogeneous information languages, taking the 100 % rate of conceptual compatibility as a target. (We know that this is the inaccessible, ideal rate of all information languages, that can be expressed with the maximum rate of recall and precision.)

5.2.1 Homogenization of languages.

Several examples of this method are shown in Compatibility, 1996. It amounts to creating one common language for the whole database. Then it is likely that the rate of effectiveness will be close to 80%, as with most information languages. But the cost of the operation is impressive, if not totally discouraging.

5.2.2 Homogenization of Indexing Formulas

This method offers less guarantees than the method described above, since it relies on the use of concordance tables, which are bound to be lacking in some precision. Moreover, the conversion programs are necessarily poor, and cannot take into account any syntactic tool. Therefore, it is likely that the rate of precision and recall should decrease in a significant way. With regard to the cost, the main expense comes from the creation of the concordance tables, which are worthwhile only if the target language is frequently used.

5.2.3 Homogenization through Automatic Indexing of Titles.

This method was slightly evoked in section 2. Its advantages are well-known: low cost and testing through extensive use. It may be a stopgap method, provided that the database has a strong linguistic homogeneity. It is well known that the efficiency of this method is poor. But it would be interesting to test this efficiency compared with that of the second method.

In short, the future of theoretical and practical studies of the problems of compatibility is still rich in promises.

Appendix A: A Review of the Research Seminar Proceedings of the TIP/ISKO Meeting Warsaw, 13-15 September, 1995

Compatiblity and Integration of Order Systems - Research Seminar Proceedings of the TIP/ISKO meeting , Warsaw, 13-15 September, 1995. Warsaw: SBP, 1996, 242 p.

The Warsaw seminar on the compatibility of information languages, a joint initiative by Ingetraut Dahlberg and the Polish Society for Professional Information, gathered more than sixty experts from 12 countries. The proceedings reflect the wealth and diversity of the debates about this challenging theme. Although we are unable to review the 25 papers presented there, this review will attempt to present a representative overview of the proceedings.

Theoretical Background

Winfried Schmitz-Esser (Germany), *Language of general communication and concept compatibility*, places his study on the most general level, that of concept compatibility. How can we create a system in which the same word-concept is interpreted in the same way by the author, the reader and/or the searcher? Relying on examples from the vocabulary of politics, he analyses the paradox of controlled languages' which, in order to establish precise communication inside an information system, are condemned to reducing the signification of every word and thus hamper the communication between different systems. «In these broad and open domains, the organization of knowledge is possible, but only at the price of non-compatibility». The author appears skeptical regarding the prospect of general compatibility, but thinks that a general agreement can be reached in small professional circles. He also puts some hope in a more precise formulation of semantic relationships and in advances in linguistic engineering.

Gerhard Riesthuis (Netherlands), *Theory of compatibility of information languages,* starts from the same paradox and stresses the fuzziness which often pervades the notion of compatibility. He states that two information languages are fully compatible if an utterance denoting a subject in the language A can be translated directly (without reindexing and without a loss of information) into an utterance of the language B. Within this framework, he distinguishes three levels : term compatibility (ex Peking and Beijing), sentence compatibility (for example, the exact equivalence of two class numbers in two classifications) and subject compatibility (the convertibility of any subject between two languages). Then he shows that the third type, which is the most important, is often elusive in enumerative classifications but that faceted systems are the most efficient in that respect. He also states that it could be possible to automatically convert the subject headings of an enumerative classification into a set of descriptors, using an algorithm of syntactic analysis.

Ingetraut Dahlberg (Germany), *The compatibility guidelines - A reevaluation,* draws a historical picture of works related to compatibility over the last forty years, works in which the author herself was strongly involved. More specifically, she mentions her *Guidelines for the Establishment of Compatibility between Information Languages in the Social Sciences* that she prepared for UNESCO (1980) and the compatibility matrices, a semantic table she finalized for the comparison of concepts between several languages of the same field. She stresses that interesting results in harmonization have been reached in the fields of energy and medicine, but at the price of huge financial efforts. Lastly, she points out the efficiency of facets as a tool of semantic stucturation of a field, and proposes the development of a universal, faceted classification, which she created according to the theory of integrated levels : the ICC (information coding classification). She shows that it could be used as a switching language between many different languages, a less expensive solution than a series of bilateral concordance tables.

Thesauri Integrity - Structures and Software

Dagobert Soergel (USA) - *Data models for an integrated thesaurus database* - proposes some software models which can integrate several thesauri into the same database. Instead of accumulating all the relationships which link a term to other related terms in every item, his models have only three basic fields: Term A/Type of relation/Term B. A fourth field identifies the thesaurus to which the terms belong. This scheme (which in our opinion can be easily reproduced) allows one, through a simple sorting on one term, to produce an inventory of all the thesauri using that term and thus to facilitate the choices for the integrated thesaurus.

Ewa Chmielewska-Gorkzyka (Poland) - *Compatibility of indexing tools in a multidatabase environment* - had to reconcile ten thesauri of the same department, each of which was adapted to a special database. Instead of merging them all into an integrated language, she instead created a source-thesaurus leading the user to the terms adapted to the chosen database.

Eugeniusz Scibor (Poland) - *Establishment of concordances between a universal classification system and an interdisciplinary thesaurus* - shows the limitations of concordance tables in automatically reindexing an imported database. In this case, the translation of the subject-headings into a set of descriptors from the thesaurus often generates poor equivalents.

Discipline-Oriented Thesauri

Stephan Hoppe (Germany) - *The UMLS: a model for knowledge integration in a subject field* - describes the famous *Unified Medical Language System* of the Library of Medicine as the most prestigious achievement of information language harmonization. This task nonetheless required ten years and a considerable budget. It consists of four tools called *Knowledge sources.* The *metathesaurus* (200 000 terms!) is organized according to concepts and includes more than thirty medical vocabularies. Every entry term leads to a preferential descriptor linked to a univocal concept. The *semantic network* is made up of more than one hundred semantic fields, and each term of the metathesaurus is connected to one or more semantic fields. The *information sources map* is an interface between the user and the available information. Lastly, the *Specialist lexicon* gives the definition of 65000 medical terms.

Stella Dextre Clarke (UK) - *Integrating thesauri in the agricultural sciences* - reports on another ambitious project : merging two big agricultural thesauri (CAB and AGROVOC) into a single thesaurus allowing for the built-in management of three databases. The approach is original in that the harmonization was first applied to the hierarchical structure of the thesaurus.

Corentin Roulin (Belgium) - *Bringing multilingual thesauri together: a feasibility study* - explains the method and the results of a feasibility study on the integration of two multilingual thesauri: *Eudised* and the *Vocational training thesaurus*. After a overall comparative survey he enumerates the four different needs that should be fulfilled by this project: one for the users and three for the system managers. He then defines five methods of 'rapprochement' between the thesauri: terminological harmonization, establishing 'cross-links', weak compatibility (concordance tables), strong compatibility (unification of concept-terms) and 'association' (harmonization of the structures and the relationships). In conclusion, a table with double entries enables one to test six hypotheses at each of these levels. each corresponding to a specific needs level. The conclusions of the feasibility study are prudent : the author recommends that the system be experimented at the most ambitious level in selected domains and that the second level be reached in the short term.

Library Classifications: Compatibility Issues

Harald Zimmermann (Germany) - *Conception and application possibilities of classification concordances in the OPAC environment* - states that nowadays libraries using an OPAC (Online Public Access Catalog) often practice shared cataloging, but that the classification systems are not unified. The author proposes the use of a particular thesaurus called a 'classification thesaurus' as a concordance tool. It sets equivalence relationships between index numbers of the different classifications which express similar subject headings.

Barbara Sosinska-Kalata (Poland) - *The UDC as an international standard for knowledge organization in bibliographic databases and library catalogues* - defends the return to a renewed UDC as the best international standard, the problem being to wait until the new version is completed.

Jadwiga Wozbiak (Poland) - *Kaba subject authority file* - reports an original unification process used for the creation of a new authority file in the library. of the Warsaw University. The list was complemented with English and French counterparts, so that the English counterpart is derived from the LCSH and the French counterpart from RAMEAU (the authorized list of the French National Library).

Annotated Bibliography

The last part of the book is not the least interesting. It consists in a copious bibliography on compatibility problems, annotated by Ingetraut Dahlberg. 450 references covering the last forty years are listed in alphabetical order by author, but each reference is also indexed with a classification of 13 class numbers which develops the class number 28 of Knowledge organization (Compatibility and Concordance between Information Languages).

Note

This article is the English version of a paper published in *Documentaliste*, 34(4/5), July / October, 1997, p. 212-226.

References

Coates, E. J. (1970, June). Switching languages for indexing. In: *Journal of Documentation, 26(2)*, 102-110.

Coates, E. J., Lloyd, G., & Simandl, D. (comp.) (1978). *BSO. Broad System of Ordering* . Schedule and index, 3rd rev.. The Hague-Paris: FID and UNESCO, XIV, 183p.

Compatibility and Integration of Order Systems (1996). Research seminar proceedings of the TIP/ISKO meeting. Warsaw, 13-15 September, 1995. Warsawa: Wydaw, 242 p, Annotated bibliography

Dahlberg, I. (1980). *Guidelines for the Establishment of Compatibility between Information Languages in the Social Sciences.* Prepared for the UNESCO Div. for the Int. Development of Social Sciences. Paris, 25p. Unpublished Document.

Dahlberg, I. (1996). Library Catalogs in the Internet: Switching for future Subject Access. In R. Green (Ed.). *Knowledge Organization and Change: Proceedings of the 4th International ISKO Conference.* Frankfurt: Indeks Verlag. 155-164.

Institut Gustave Roussy. (1979). *Le Macrothésaurus des Sciences et des Techniques. [The Macrothesaurus of Sciences and Technologies].* Paris: CILF.

Lancaster, F. W. (1989). *Vocabulary Control for Information Retrieval.* (2nd ed.). Arlington, Virginia: Information Resources Press, 270 p.

Saussure, F. de. (1949) *Cours de Linguistique Générale (4th ed.). [Course of General Linguistics].* Paris: Payot.

https://doi.org/10.5771/0943-7444-1997-4-213 https://www.inlibra.com/de/agb - Open Access -

Scott, M. L. (1993). *Conversion Tables: LC-Dewey, Dewey-LC*. Englewood: Libraries Unlimited, 365p.

UNISIST (1971). Study Report on the Feasibility of a World Science Information System. Paris: UNESCO, 161p.

Acknowledgment

I wish to thank Mrs. Martha Melter for helping me to revise the English text.

Jacques Maniez 61 rue de la Corvée F-21000 DIJON
Phone (033)03 80 41 35 40
E-mail: jmaniez@compuserve.com