Ceusters W, Bouquet L. Language Engineering and Information Mapping in Pharmaceutical Medicine. Journal of the Belgian Medical Informatics Association, 2000;7(1):26-34.

Language Engineering and Information Mapping in Pharmaceutical Medicine: Dealing Successfully with Information Overload

Dr. W. Ceusters [1], *L. Bouquet* [2]

1. *Director R&D, Language & Computing nv, http://www.landc.be*
2. *Director, ATEK bvba, http://www.atek.be*

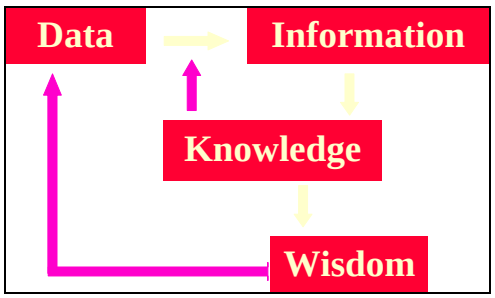
1. Introduction:

It has been claimed that the widespread use of computers in business, and the availability of digital networks, would rapidly lead to paperless offices. Surprisingly or not, the reality is quite the opposite. Document printing has become as easy as pushing a button. In the good old days, one tiny spelling error would force a secretary to retype a letter or memo completely. Now it just takes a few keystrokes to correct a document on screen, and to print out a new version of it… how easy to deal with language on a machine!

Another vision did not yet come true neither: the computer as a knowledgeable device that can take decisions on its own, or that can communicate with humans in a natural way. The bottleneck here is the representation of human knowledge in a format that can be processed by machines. Computers are very good at numbers and structured, coded data. Unfortunately, 99.99% of human knowledge is expressed in an unstructured, non-formal, ambiguous communication vehicle: natural language. And … how difficult to deal with language on a machine!

Are we getting confused here? Not really. Computers are extremely well suited to deal with the overt form of language as a stream of meaningless characters. But when it comes to understanding, much more steps are to be taken.

1.1 Living in the knowledge age



In business and manufacturing, three components have been considered extremely important: people, money and resources. Recently, a fourth component has been added: knowledge. In the consulting business, it has even become the most important component of all.

What is knowledge, and where does it come from? A traditional view is the knowledge production cycle. At the beginning, there are raw unprocessed data. Once collected and formatted, they can be processed in such a way that relationships become visible: data are turned into information that can be used to improve business or manufacturing processes. The more information that is disclosed, the more clever we become, until so much information (inside a domain) is disclosed, that a level of deep understanding is reached: the knowledge level. Having this knowledge, it becomes easier to derive more and better information from the data. We might even control the events that produce the data. Becoming knowledgeable is itself a matter of moving from one state into another (the “knowledge microprocess”): observation, understanding, prediction, application, justification. The ultimate goal (hopefully) is to reach the level of wisdom. This level can only be reached when sufficient knowledge from various domains has been acquired.

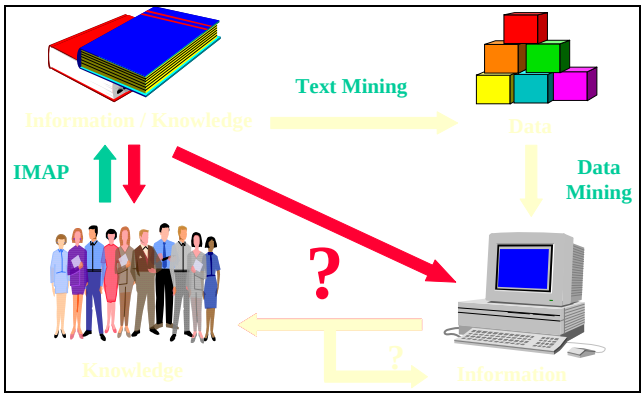
Information Technology can be defined as technical support to turn data into information. Knowledge engineering is a similar discipline at the level of information and knowledge. Language engineering can be seen as a special branch of knowledge engineering, dealing with knowledge in the form of language.

1.2 Knowledge management in pharmaceutical medicine

The pharmaceutical industry is well aware of the importance of effective knowledge and information management. Bringing a new drug to the market is a multi-stage process that typically takes between 7 and 15 years. Huge amounts of information have to be gathered, analysed and communicated. Between 100 and 1000 people intervene somewhere in the drug development life cycle: feasibility studies, planning, clinical trial monitoring, medical writing, regulatory affairs, post-marketing surveillance, pharmacovigilance, etc. Tens of thousands of documents are generated and have to be analysed.

The pharmaceutical industry is also a multi-national business. This means not only that multiple languages have to be addressed, but also that effective communication channels have to be set up across the language borders

1.3 Towards a paradigm shift in knowledge exploitation



Traditionally, information and knowledge is stored on paper in documents and text books, and expressed by means of natural language. Paper is still the most widely used medium to instruct or educate people. The Information Mapping® method (short IMAP®) has been developed to write documents that are easier to understand by the intended audience. Its aim is to disclose information to people.

However, a paradigm shift is noticeable today. To cope with information overload, knowledge should also be exploitable by machines. This is not the same as educating people by using the computer as a new medium that replaces paper and books. It means representing and storing knowledge and information in the machine so that the machine itself can manipulate it, at least for tasks for which it is better suited than humans. As such, modern text mining algorithms decompose text in meaningful chunks that can then be used for true data mining purposes.

The net result is that the computer can assist people more effectively in discovering new knowledge, while at a later stage, it is perhaps possible to have this also done by a machine.

2. Claims of this paper

In this paper, we offer a simplified view on the main principles that are to be used by systems aiming to make computers understand natural language. We will show that many of these principles are shared by the Information Mapping method, especially within pharmaceutical medicine.

As a consequence, we will try to convince you of the following:

- effective use of knowledge requires information to be processable by machines, and not just by humans;
- the knowledge production and exploitation cycles should be merged;
- Information Mapping does for humans what Language Engineering does for machines;
- Information Mapping and Language Engineering share some major principles which may lead to very effective technical implementations;
- the knowledge-intensive nature of pharmaceutical medicine makes it an ideal object for exploitation.

We will (hopefully) achieve this by first explaining what is language engineering and how it relates to terminology and formal ontologies. We will look at the main principles and techniques that are used, and we will also show some typical problems that can be solved.

In a second stage, we will briefly explain the main principles of the Information Mapping method.

Next, we will have a close look at the commonalties amongst language engineering and information mapping, “divide and conquer” being the most important one.

Finally, we will demonstrate that a language engineering oriented technical implementation of the Information Mapping method will eventually pay off in knowledge intensive environments.

3. Identifying some problems

3.1 Inaccuracy in information retrieval

Numerous examples exist in which even extremely simple applications for medical natural language understanding can improve the behaviour of existing software applications. Here is an example: every year, the Pharmaceutical Industry Association of Belgium issues a CD-ROM containing the patient information sheets of the drugs of its members available on the market. The information on CD-ROM has some obvious advantages over the same information in a book, especially when it comes to search specific bits of information. Ironically, the improved searchability discloses also its major weakness: the system is completely unaware of simple medical terminology, leading to an underperformance that is not easily accepted by the users. When the CD-ROM is searched by a Dutch speaking physician for drugs that may be used to treat diabetes, it should not make any difference whether he refers to diabetes by the strings “diabetes”, “suikerziekte” or “diabetes mellitus”. The answers he retrieves are however different in each case! In technical terms: rather than getting all and only true hits, the responses contain false positives as well, whereas some drugs that should be retrieved, are not (false negatives). As a consequence, recall and precision (two measures that are used to describe the performance of information retrieval packages) are lower than one would expect.

The major reason for the underperformance of the system, is the exclusive use of “(sub)string search” instead of “conceptual search”. Some basic linguistic knowledge, e.g. that “suikerziekte” is a synonym of “diabetes mellitus”, is not available. Also purely pragmatic knowledge is not exploited by the system. Historically, the word “diabetes” is used for any disease in which a chemical compound leaves the body through the kidney, associated with increased loss of water. ”Diabetes mellitus” is an example in which glucose (sugar) is lost in large quantities, while “diabetes insipidus” or “fosfaatdiabetes” (Dutch) refer to other problems. Diabetes mellitus, however, is so common that physicians just use “diabetes” as short for that particular disease only. If they would mean another disease, they would say so explicitly! The search system on the CD-ROM is not aware of this. As a consequence, when “diabetes mellitus” is used as a query, the system will not find those entries where “diabetes” is used as short for “diabetes mellitus” (false negatives), while it will retrieve entries related with “diabetes insipidus” or “fosfaatdiabetes”, when queried for “diabetes”.

Finally, there is also a lack of contextual knowledge. This problem, however, is far more difficult to resolve than the two previous ones.

3.2 Patient selection for clinical trials

Another example where language engineering can pay off quickly, is the selection of patients to enter clinical trials. Patients must meet specific eligibility criteria. Selecting patients is a time consuming activity (from the clinical investigator’s point of view) while missing patients that could have entered a study, will increase study duration, which causes a delay in marketing and selling.

An attractive solution is to integrate clinical study protocols in electronic healthcare record systems. As such, routinely entered data during a patient’s visit (not specifically for study eligibility but for prevention, treatment or follow-up of other medical conditions), could trigger the attention of the physician on possible eligibility. As this system may not interfere (or at least as little as possible) with the regular use that is made of i, the language used by the physician must be “translated” into the language used in the study protocol. This requires a deep understanding of medical terminology, not by the user of the system, but by the machine itself. Language engineering makes this possible!

4. Natural language processing

4.1 To be precise

Natural language processing applications come in many flavours. At the heart of the technology, there is a specific discipline of science called *computational linguistics*, aiming to develop computational models of language that explain how language works in human beings, and how this insight can be used to allow computers to work with language. If the focus is more on the development of practical applications rather than on theoretical studies, the term *language engineering* is preferred.

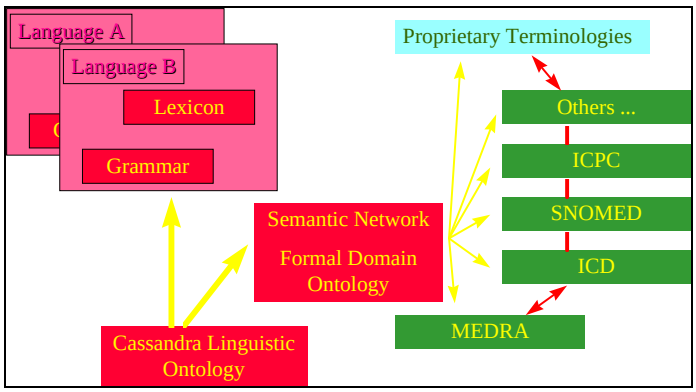
A major distinction is to be made between *language processing* and *speech processing*. The basic aim of *speech processing* is to turn the sound wave generated by a human speaker into a digitally represented text without understanding the meaning of what was said. *Language processing* on the other hand sets out with the verbal representation of —say— an ASCII text, and uses this format to do some further useful processing. Another distinction has to do with the direction of processing. Whereas natural language understanding is generally called *natural language analysis* (going from a text to its meaning), the opposite (going from a meaning representation to a text) is called *natural language generation*. *Machine translation* can be seen as a combination of both.

Natural Language Understanding is considered one of the most complex problems in artificial intelligence. However, under some specific circumstances it is possible to make a computer understand natural language: when used in a closed domain, and for a well-defined task.

4.2 Conceptual and linguistic ontologies

NLP applications must have sufficient knowledge about the domain of discourse (e.g. medicine) and the (grammatical) rules related to specific languages.

Dictionaries play an important role here. Dictionaries are usually large books meant to be used by humans to look up the meaning of unknown words. Most electronic dictionaries currently available differ only from paper dictionaries in that they are published on a digital medium. For medical natural language understanding purposes, dictionaries have to be fundamentally different in nature: they are primarily meant to be used by machines!



The important thing in these dictionaries, is that entries are not related to each other directly, but through a language independent formal medical concept system. An important characteristic of this system is the clear separation of different kinds of relationships that hold between the concepts. “Language independence” does not mean ignoring language, a mistake quite often committed by people working in that field. If the concept system is solely intended to be used as a knowledge base for internal processing, without any communication being needed in natural language, then there are some arguments for such an approach. If not, it will definitely lead to unsatisfactory behaviour. Our approach is to keep the concept system separate from any linguistic knowledge. But in addition, for each specific language (English, Dutch, French…) a linguistic ontology is maintained, capturing the relationships between the grammars of these languages, and the language independent concept system.

4.3 Understanding clinical definitions

For humans, it is sufficient to define *Zenker’s diverticulum* as a *diverticulum of the esophagus caused by intraluminal pressure*, to make the term meaningful. An electronic dictionary intended to help human

readers, may be implemented as a 2-column table, the first containing the terms, the second containing the definitions.

For a machine, this format is totally unacceptable. Definitions need to be dissected completely, while each building block must have a meaning on its own. Various representations are possible. We developed one that is close to the grammar rules used in medical sublanguage, and that is easily translatable into language-independent concept representations.

Meaning is added to dictionary entries through explicit context definition. A number of knowledge building blocks have been defined, of which “concepts” and “links” are the most important ones. Concepts refer to things that may be instantiated in the real world, while links relate concepts amongst each other.

A clear distinction is maintained between IS-links (subsumption relation) and other links. This guarantees that automatic classification of newly defined concepts can be achieved.

Once the necessary components of a language engineering system are in place, it is possible to develop various applications.Simple ones are keyword or word-spotting based and can be used for automated encoding, an activity that is extremely important in Europe where in many countries, the revenues of hospitals depend on coding medical diagnoses and procedures.Other applications allow medical staff to highlight sentences or paragraphs in patient documents such as discharge letters to access bibliographic services, even in other languages.

Translating portions of electronic patient records is also feasible today, as well as analysing the contents of surgical procedure reports.

4.4 Clean separation of knowledge

A key factor for success is the clean separation of various kinds of knowledge.

Conceptual knowledge is the knowledge of sensible medical concepts and how they are related to each other. It is this kind of knowledge that tells us that hypertension is the measurement of a blood pressure that is too high for a given patient

This knowledge by itself does not allow a physician to state whether or not a given patient suffers from hypertension. To do so, additional knowledge related to clinical definitions and criteria is required. Criteria might be that at least 2 measurements have been taken, and that “too high” means 14/9 for a patient younger than 30 years, and 17/11 for somebody older than 70.

Both types of knowledge are sufficient for a computer to make it understand what goes on in the world. For communication, especially in natural language, more is needed. Surface linguistic knowledge is required to tell the computer how concepts are expressed in language, while pragmatic knowledge can help the machine make more adequate use of it.

Because many different conceptualisations of (parts of) medicine have been developed over the years, knowledge related to these particular coding and classification systems has become mandatory.

5. Information Mapping

Having explained the main underlying principles of language engineering, it is now time to look more closely at the Information Mapping method.

The Information Mapping method is a proprietary, research-based methodology for the analysis, organisation and presentation of information. The Information Mapping method was designed in the late sixties by Robert Horn, a psychologist specialised in memorisation and learning. The method is the result of research at Harvard and Columbia Universities, but also from practical business applications.

The initial aim of the method was to make paper based information better understandable for the human reader. However, recent practice has revealed that the method has quite some interesting "side effects":

- it enables collaborative writing, i.e. several writers contributing to the same document (e.g. an intranet)

- it produces highly structured documents that are perfectly fit to be integrated within document management systems
- it can serve as a kind of gateway between human knowledge and formal domain knowledge. This is what this paper intends to demonstrate, at least for the medical domain.

5.1 The four pillars of the Information Mapping method

The information Mapping method rests on four pillars:

- the units of information
- user- and task orientation
- the seven information types.
- the research-based principles.

5.2 Units of Information

The Information Mapping method sets out with defining the units of Information. The basic units of information are the information block and the information map. The information block is the smallest unit of information. Information blocks are placeholders for information and are limited in size. A block can contain text, a list, a table, a graphic, an audio sequence, a video sequence etc. The content of one block will never exceed the limits of human short-term, working memory. For example, a bulleted list within a block will never contain more than 7±2 bulleted items.

Information maps can contain up to 7±2 information blocks.

5.3 User- and task orientation

IMAP-writers always make an analysis of the target population in terms of action and knowledge. First, they identify the actions they expect from the reader, based on the document. Then, for each action, they investigate whether additional knowledge should be provided, so that the reader can perform the expected action.

The table below shows an example of a simple task analysis, done in order to write an instruction leaflet for an electronic blood pressure meter. The left 'Action'-column lists the tasks that the reader should be able to perform with the leaflet. For each task, the right 'Knowledge'-column lists the knowledge that should be provided to the reader, so that he/she can perform the task properly:

Action	**Knowledge**
(Re)placing the batteries	• What kind of batteries are allowed? • Where is the battery compartment?
Calibrating the electronic blood pressure meter	• Why is calibrating important? • When do you have to calibrate? • What do you need to calibrate?
Using the electronic blood pressure meter	• How does it work? • Under which circumstances is it allowed to use it? • What are the differences compared to the traditional sphygmomanometer?

5.4 The seven information types

The actions in the example shown belong to the 'Procedure' information type. The 'knowledge', however can be of various information types.

With the information types, we are touching the very basics of the Information Mapping method. It maps information onto knowledge categories. The table below lists the seven information types, along with a brief explanation and some examples.

Information Type	**Description**	**Examples**
Procedure	A set of steps that the reader has to perform to obtain a specified outcome	• How to measure a patient's blood pressure • How to fill in an on-line medical record
Process	A series of events or phases that take place over time with an identifiable purpose or result. A process can have several actors who are not necessarily human actors	• The development of a new drug • Data flow within a Hospital Information System • The Krebs cycle
Structure	A physical object or something that can be divided into parts and has boundaries	• The anatomy of the human brain • The chemical composition of aspirin • The composition of a surgical team • The toolkit needed to service an electronic blood pressure meter
Concept	A group of items that share a unique combination of attributes, not shared by other groups, and that can be referred to by the same generic name or symbol	• Definition of 'Internal Medicine' • What is an 'Insulin Shock'?
Principle	A statement that expresses what should or should not be done or what seems true in light of evidence	• The Hippocratic oath • Conditions to obtain reimbursement of medical cost
Fact	Result of an observation	• The HDL-cholesterol level of a given patient • The number of beds in a specific hospital
Classification:	Sorting of items into categories	• Classification of blood groups • Overview of technologies within medical imaging

As an example, we show again the table sub 5.3, this time with an additional column, listing the information type of each knowledge topic in the right column:

Action	**Knowledge**	**Information Type**
Procedure: (re)placing the batteries	What kind of batteries are allowed?	Principle
	Where is the battery compartment?	Structure
Procedure: calibrating the electronic blood pressure meter	Why is calibrating important?	Principle
	When do you have to calibrate?	Principle
	What you need to calibrate?	Structure
Procedure: using the electronic blood pressure meter	How does it work?	Process
	Under which circumstances is it allowed to use it?	Principle
	What are the differences compared to the traditional sphygmomanometer?	Classification

Determining the information types is only an intermediate step. For each information type, the method gives one or more presentation modes that are best fit to present the information contained in a given information type. These presentation modes are called 'key blocks'.

The table below lists the seven information types, along with some key blocks that can be used to present that type of information:

Information Type	**Key Block(s)**
Procedure	• Procedure table (step-action table) • Flowchart
Process	• Process table (stage-description table) • Cycle chart
Structure	• Illustration • Element-function table
Concept	• Definition • Analogy
Principle	• Policy • Guideline

Information Type	**Key Block(s)**
Fact	• Statistics • Facts
Classification	• Classification table • Classification tree

5.5 The Research-based principles

The method is complemented by seven research-based principles. These are principles that are normally adhered to by every technical writer or medical writer. The table below lists the seven principles, along with a brief explanation:

Principle	**Explanation**
Chunking	Group information into 'digestible' chunks
Relevance	• Place 'like' things together • Exclude unrelated items from each chunk
Labelling	Provide the reader with a label for each chunk of information
Consistency	Use consistent terminology, organisation and formats
Integrated graphics	Use tables, illustrations, and diagrams as an integral part of the writing
Accessible detail	Write at the level of detail that will make the document useable for all readers
Hierarchy of chunking and labelling	Group small chunks around a single relevant topic and provide the group with a label

6. Shared features in Language Engineering and Information Mapping

Not surprisingly, both language engineering and information mapping have a strong cognitive basis, as they have to do with perception. Information Mapping has been developed by psychologists, and language engineering is often studied in centres for cognitive psychology or philosophy.

Both are strongly task-oriented. For IMAP, the task is fixed: producing more readable documents. In LE, task-orientation is a basic requirement to make a system work. The “machine polyglot” or “general language solver” is not expected within the next 20 years.

Divide et impera was the preferred strategy of the Roman Emperors. In IMAP, this is achieved by identifying different information types while in LE, we dissect knowledge into a limited number of basic building blocks.

Both rely heavily on principles and standards while, finally, many studies have shown immediate or short-term pay-off.

We will discuss some of these issues in more detail.

6.1 Cognitive principles in IMAP and LE

It is generally known that memorising facts is not what human brains are good at. Memory requires strong impressions and language alone is too weak an information carrier to be successful for that purpose. Only those who mastered language completely have succeeded in making people remembering their words. Other information carriers such as pictures and sounds, have proven to be more effective in learning and memorising. A combined use of these carriers has more effect than the expected sum of the individual results. Isn't this what IMAP exploits?

As explained earlier, language engineering is often used to develop models that explain how language works in humans. On the other hand, understanding how humans manipulate language and, more important, what representations are derived from that in our conscious minds, helps to develop better language applications.

6.2 Task orientation

IMAP-writers identify the actions that readers need to perform after reading a document. In a second stage, they identify the knowledge required to perform the actions. Identified actions belong to the procedure type of information. The knowledge required to perform the action can belong to one of the 6 remaining information types.

In language engineering, task orientation is also mandatory. It is not yet possible to analyse texts and represent the meaning of it without having any clue on what must be looked for. Please understand this correctly: it does not matter what the goal is exactly. Many tasks can be achieved, but not (yet) at the same time, and through the same system design.

In language engineering, source information (derived from the texts) and task information must also share some common representational features in order to make an LE system behave adequately.

6.3 Divide et impera

IMAP identifies different information types, while also language engineering takes into account various types of knowledge and information.

It is not difficult to see that information types from the IMAP method have equivalent components in the language engineering machinery. A specific example is the classification information type. A classification is one of the 7 information types identified in Information Mapping. In medical language engineering, it has turned out to be the most important one.

Formal classifications (in the form of graphs operated upon by formal classification rules) are used to represent the relationships that hold between concepts. Each concept has a fixed position inside that classification and inherits characteristics from its parents. Maintaining such classifications is a labour intensive, but mandatory endeavour.

6.4 Need for standards

Both Information Mapping and Language Engineering follow a principle-driven approach, processing of information along similar tracks.

While Information Mapping is a standard on its own, language engineering has to take several standards into account. It is strange to see that this is a relatively new insight. The most plausible reason is that language engineering applications are rather young. Reusability is an issue to which academic centres have paid little attention in the past. As applications are brought to the market, standards become mandatory.

7. Towards a combined approach for language engineering and Information Mapping

Having identified the common features in Information Mapping and Language Engineering, it is now time to investigate how the combination of these two disciplines might lead to powerful systems.

Information Mapping has a sound basis for a technical implementation. However, a key principle should be that the original goal, producing readable documentation, is not directly and exclusively looked for. A true pay-off will come when the methodology is integrated within existing applications such as document management systems and workflow systems.

But that on itself is not sufficient because current versions of these systems do not exploit knowledge to its full extent. Additional components must be developed such as language-, terminology- and knowledge servers.

The various kinds of information that are identified in Information Mapping, can easily be mapped (or interdigitated) with the basic knowledge building blocks exploited in language engineering systems, specifically in the medical domain.

Representing knowledge in such a way, may lead to various kinds of applications of which producing readable documents according to the IMAP principles is only one.

Not only may IMAP benefit from language engineering. Also the other way around, benefits are to be expected. Storing textual information in precisely identified blocks, accompanied by a meaningful label, may be used as a means to lead language engineering modules more efficiently to a solution. It helps language understanding systems a lot if they know beforehand that a given sentence or paragraph expresses a definition or a procedure!

In Pharmaceutical Medicine, one can think of a system that supports the complete life cycle of a drug development project. Conceptual knowledge about medicine and pharmacology, as well as pragmatic knowledge related to the procedures that are to be followed in such a project, can be stored in formal domain knowledge servers. Instantiations (phenomena observed in the real world, patient data…) can be stored in an IMAP object repository.

Language servers can be used for various tasks (controlled language checking, automatic translation, classification, information retrieval, mapping to terminology standards, content verification) while many activities to be done can be guided by means of a workflow system. The various documents that are to be generated in the lifetime of the project, can again by composed for human readers using templates of the IMAP layout.