



Deliverable Report

Client : European Commission

Project : FRESH

Project N°: FP6-51605

Project Number: FP6-516059
Document number: DR_FRESH_WP2_2.6.3
Document Title: Semantic knowledge database populated with relevant information
Document status: Confidential

Abstract

This report contains the description of the semantic knowledge database populated with relevant information.

Keyword List WP2, symbol, recognition, integration, specifications, documentation, semantic knowledge database.



Deliverable Report

Client : European Commission

Project : FRESH

Project N°: FP6-51605

INTRODUCTION AND ANALYSIS OF NEEDS

The aim of this report is to enumerate all information which is present in wiring diagrams and which is of use for building up a semantic knowledge base for recognition purposes. The data to taken into consideration is either graphical or textual, and deal with representation purposes. A complete semantics knowledge base should have to capture all physical, electrical and pictural knowledge, it would be a huge undertaking and we restrain ourselves in this work to the most useful information for recognition purposes. We will also concentrate first and foremost on the most frequent elements.

ANALYSIS OF NEEDS

To answer the needs of the symbol recognition processes (D.2.4.2), the following information must be available in the knowledge base, for retrieval and editing:

- geometric composition,
- relative positioning of a symbol part with respect to another one,
- list of points where equipment is connected through wires,
- symbol drafting constraints (i.e. orientation, variable size, symbol which can be split into several discontinued parts...)

Text recognition, on the other hand, has to use information about the syntax used in annotations of diagrams and more specifically of the symbols.

The association rules between graphical parts and textual parts also have to be represented, for helping the segmentation and recognition processes. This is all the more important because text recognition can improve and guide symbol recognition, and vice versa.

The various processes described above can all include a *training phase* where the recognition parameters are learned through processing of ground-truthed data.

CHAPTER 2. AVAILABLE INFORMATION

At the final outcome of the project, all the information will be represented in various formats within the subprocesses led by each partner.

It is important to note that each component can be associated with

- one or several graphical representations,
- physical properties, which also have to be represented graphically, usually by some kind of annotation,
- syntactical rules for combining the component with the other components of a diagram.

2.1 INFORMATION AVAILABLE IN THE SEGMENTATION SUBPROCESSES

The software is an interactive tool by which an operator can quickly identify and segment a diagram into wire lines, symbol bounding boxes and text bounding boxes (see deliverable D.2.2.1). The system automatically computes the connections of the symbols and records them in a CSV file. It is also able to produce subimages of the isolated symbols and characters.

2.2 INFORMATION AVAILABLE IN THE SYMBOL RECOGNITION SUBPROCESSES

In developments prior to this project, reference databases of symbols have been designed, although not for aircraft wiring diagrams. Although only about 50 symbols are common to this base and to aircraft wiring diagrams, it can be useful to build upon existing know-how. In the existing reference bases, the following information is associated with each symbol:

- a vectorial representation of the geometric composition of the symbol and of the relative position of its textual parameters,
- the name of the symbol and the default syntax of the way the physical parameters of the component are represented,
- geometric constraints, expressed as a multiple choice list for size and angle.

While analyzing ATA Specification No. 100 (see deliverable D.2.6.1), LORIA has developed a method for describing visual information (deliverable D.2.4.2) and built up a table, also stored in CSV format, describing the graphical features of all symbols to be found in aircraft wiring diagrams. This table provides a basic representation of the shapes present in symbols and their composition rules.

NUMBER OF RECTANGLES —

To detect a rectangle, the system looks for 4 corners aligned as they should in a rectangle, through template matching, and then checks if the rectangle is closed or open by tracking the lines joining these corners.

NUMBER OF OCCLUSIONS —

Occlusion detection relies on a compacity test for small white connected components.

NUMBER AND TYPE OF CONNECTIONS WITH OCCLUSIONS —

Once a small occlusion has been detected, the recognition system checks for the number of connections on the contour of this occlusion

NUMBER OF CONNECTED COMPONENTS —

The system counts the number of the following types of connected components:

thick: this is done through a mathematical morphology operator

black without included white components: this is done using the inclusion tree of connected components built by the process

white without included black components: done in the same way...

NUMBER OF END POINTS —

this is the number of points “going out” from the symbol to connect to the wires.

2.3 OTHER POSSIBLE SOURCES OF INFORMATION

A source of information which we have not looked into yet is all that can be retrieved from the title block of the diagrams and the notes made by draftsmen. These pieces of information contain handwritten notes which are probably too hard to process automatically. But if necessary, an operator could probably extract useful information.

CHAPTER 3 CONTENT OF THE KNOWLEDGE BASE

We will now propose a set of characteristics to be stored in the knowledge base.

3.1 CHOICE OF MODEL DESCRIPTION

The geometrical description of the representation of each possible wiring diagram component has to be as precise and complete as possible, to enable non ambiguous recognition of graphical symbols in the documents to be analyzed while leaving the possibility of adding new symbols when they appear.

We must take into account the noise which is present when dealing with wiring diagrams. In our case, we will only have few learning samples for most symbols, given the huge amount of work that would be necessary to segment hundreds of diagrams. Therefore, we choose this solution: separately describing the ideal symbol on one hand, and modeling the noise as well as we can on the other hand. This allows us to modify independently assumptions about deformations or adjust the matching tolerance, in a batch of documents to be processed, without reconsidering the geometrical symbol models themselves.

Our current proposal is therefore to represent each model symbol in vectorial format, which has the following advantages:

- a distortion-free pixel-based representation can be derived from the vectorial model at any requested resolution;
- this representation is easily compatible with the expression of specific deformations or of the relative positioning of elements;
- complex symbols can be composed by geometrical assembly of simpler symbols, using their vectorial description;
- new symbols can be easily added to the database without having to recompute all the base with its signatures.

3.2 SPECIFICATION OF THE DESCRIPTION

In the following, we propose a framework for the description of all graphical elements' 2D geometry and semantic properties. The reference base allows us to represent a large variability of symbol representations, and also to include related semantical aspects. However, it does not include representational means for more complex entities such as panels or the data flow between equipment components.

It must be noted that the proposed data structure does not enforce any format (PIVOT, CSV...) but only gives a global view of the data to be handled and of the constraints attached to them.

3.2.1 SIMPLE SYMBOLS

The following symbols are considered as being simple:

- the symbols which cannot be composed by assembling other symbols already referenced in the base;
- connections which are not included in the description of a symbol;
- references of wire lines;

A symbol has a graphical representation and may have connections. All symbols are not components, but we can choose to use some symbols which are just subparts of bigger symbols, such as ~ ↗ × √. They can also be associated with text.

The geometric description of each symbol is based upon the following graphical elements: segment, circular arc, elliptical arc, polyline, text. The coordinate framework for each symbol has its origin at the upper left corner of the symbol's bounding box.

3.2.1.1 Segments

A segment is identified by its coordinates and by the symbol to which it belongs. For a given symbol, there can only be one segment having the same coordinates. Segments are the only basic features which can be subject to the following distortions: important size change independently of that of other features, superimposition with wire lines and partial representation.

3.2.1.2 Circular and elliptical arcs

An arc is described by the following features:

- name of symbol to which the arc belongs;
- coordinates of center. In addition, we have for a circular arc
 - radius

and for an elliptical arc

- values of maximum and minimum radius;
- angle of maximum radius.
- the starting and ending angle of the arc;
- can be continuous (by default) or dashed.

3.2.1.3 Polylines

A polyline is made of at least 2 segments. It is identified by the following features:

- name of symbol to which the polyline belongs;
- number of segments of the polyline;
- coordinates of each point of the polyline;
- can be continuous or dashed.

3.2.2 TEXT

The anchor point for a text zone is at the top left corner of its bounding box. At a given location of a symbol, there can only be one text block around the anchor point, but it can have various formats.

The text often complies with a specific format, which can be defined by a dictionary, a regular expression or a grammar. In the present proposal, we assume that the specific format of a text zone can be specified through a regular expression.

3.2.3 COMPOSED SYMBOLS

Some symbols are said to be composed because they are built by the assembly of symbols already present in the base.

The position of their component parts is not always specified in advance, it can vary depending on the diagram they appear in.

A choice which may not always be simple is that of deciding to which level if detail symbols have to be decomposed, especially for families having a common graphical base and differing only by a few graphical features (see Fig. 3.1.)

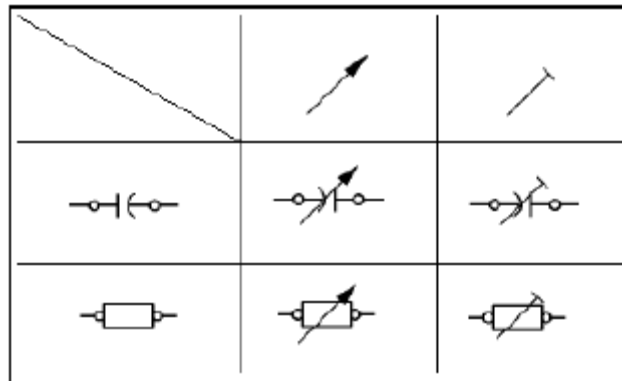


Figure 3.1: An advanced decomposition.

The advantages of such a decomposition are:

- It is possible to identify basic graphical features which can be the vocabulary for a higher-level description language, to express compositions in a more powerful way than just by assembling segments and arcs;
- it is possible to identify symbols which are graphically similar, which can be useful in the choice of the best way to differentiate them;
- the reference base can be described in a homogeneous way—a recurring graphical feature keeps the same representation, independently of the symbol in which it appears;
- the decomposition can be directly used in a structural recognition method.

However, there are some drawbacks to such a choice:

- Such a decomposition would not be directly usable by the current method as there is no composition operator which applies to the descriptors used.
- When creating a new symbol, there is a risk that the user may unknowingly recreate graphical features which already exist in the reference base. To avoid that, it is however possible to use methods to assist the user in retrieving symbols already known in the database.

3.2.4 CONNECTIONS AND WIRES

A *connection* has no representation per se, it is only a label to position the connection point with respect to the graphical representation of equipment. It must be noted that this positioning is very important to discriminate within some families of symbols.

Wires are free-length polylines, whose segments' orientation is constrained by the angles $\{0^\circ, 90^\circ, 45^\circ, -45^\circ\}$. The only other information to be present in the reference database is the connection points and/or extremities they link.



Deliverable Report

Client : European Commission

Project : FRESH

Project N°: FP6-51605

CHAPTER 4 CONCLUSION AND PERSPECTIVES

This report gives a view centered on WP2, i.e. symbol and text recognition. It may have to be complemented on knowledge around aircraft electricity, if the associated semantical knowledge is deemed to be necessary for recognition tasks.

As described in D.2.4.4, the reference base contains around 450 labeled symbols, extracted from Airbus manuals, as small images associated with a text field containing the name of the symbol. If we restrict ourselves to wiring diagrams, about 350 symbols are available.

The physical properties must be added by the various partners, so that each symbol can be represented by:

- a vectorial representation with all the attributes presented in this deliverable,
- an image of the symbol at a chosen resolution,
- the name of the symbol,