STANDAT - Experience from developing and implementing a standardised format for exchange of data

Chapter 1 - 7

European Environment Agency

INDEX:

1. Preface: objectives, structure and scope of report.

This report is part of a package of projects launched by the Danish Ministry of Environment and Energy for the support of the European Environment Agency (EEA). The scope of the report has been defined in a cooperation between the EEA and the Danish Environmental Protection Agency (Danish EPA). The project was initiated in August 1995 and completed in December 1995 / January 1996.

One of the core tasks of the EEA and the EIONET will be the establishment of a comprehensive, coherent and quality-ensured collection of environmental data. Data for this purpose must be collected from many different sources among the EU countries. For this reason it is important as soon as possible to establish the conditions for a non-problematic transfer of data between the many participants in the network of organizations related to the agency. Without this, the collection of data will be extremely time consuming and resource-requiring. It is also important to ensure the possibilities of combining data across subject-areas regardless of where and by whom data were collected.

The experience from the development and use of systems like the Danish STANDAT system is relevant in this connection. STANDAT is the Danish system for exchange of environmental information - a concept that includes a range of code lists, a standardised file format and some dedicated computer systems for the support of the STANDAT users as well as an organisational structure.

It is important to emphasize that the aim of the project is not an adoption of the STANDAT system by the EEA, but an attempt to utilise the experiences gained i Denmark from the use of such a standardised system.

Objectives.

The main objectives of this project are:

* To transfer knowledge and experience of the use of the Danish STANDAT system to the EEA

* In brief to examine a couple of other relevant formats for data transfer in operation, using a predefined set of parameters

* To contribute to the development of a data transfer system for the EEA that will ensure an uncomplicated exchange of environmental data in the EEA network.

Scope of report.

As described above the main point of the project is the utilisation of the experiences of the STANDAT format. Therefore, it has not been an aim of the project to go deep into other formats or concepts for data exchange. Two such other formats are discussed

briefly for reasons of comparison. Using these as references is especially important when discussing ideas for development of an exchange format like STANDAT, and as a background when presenting different scenarios and recommendations.

Data exchanged via written forms are not relevant in the context of this report.

Nor is it the intention of the report to go into any detail with different technical solutions based on the use of edp-based networks etc.

Structure of report.

First of all, it should be noted that the last chapter of the report (chapter 13) gives an executive summary, that provides a brief overview of the main points of the report.

The first chapters of the report concentrate on STANDAT itself. Chapter 2 is concerned with the background and history of STANDAT and answers such questions as: Why develop a standardized system for data transfer, how was STANDAT developed, what considerations were taken into account during the development process.

Chapter 3 describes the system of code lists - there are four different types of code lists. Chapter 4 presents the file format with the three sections: the HEADER section, the DEFINITION section and the DATA section.

Chapter 3 and 4 are rather technical in their content and should be skipped by readers not interested in these aspects of STANDAT.

Chapter 5 deals with edp support programmes for the STANDAT system. The STANDAT load programme for loading data into databases is presented together with the STANDAT support programme for the support of the users when producing and checking files.
Chapter 6 is about the organizational structure for administration, maintenance and development of the STANDAT system.

In chapter 7 the process of defining, creating and transferring a STANDAT file is described and the main principles are presented.

Chapter 8 analyzes the experience of the use of the STANDAT system.

Chapter 9 describes two other, similar interchange formats. The descriptions are mainly based on a predefined set of parameters, eg general concept, use of file format, use of code lists and organizational preconditions.

Chapter 10 introduces ideas for further development of an interchange format for environmental data like the STANDAT system based ia on some of the points in chapter 9.

Chapter 11 sets up different scenarios for data transfer and discusses in what situations each scenario is relevant. First a brief overview is presented of the differences between the EEA and the Danish environmental administration when it comes to organisational set up and needs for data transfer.

Chapter 12 presents conclusions and overall recommendations.

Acknowledgements and sources of information.

This project was carried out by a project group consisting of Kit Clausen and Annelise Ravn of the Danish EPA.

Apart from our own knowledge and experiences from the development and use of the STANDAT system, the inputs for the project has been extensive discussions with colleagues in the Danish EPA, experts from the EEA, Eurostat and SANDRE/France.

We have received invaluable help from colleagues working with or having worked with STANDAT, Sten Aabo Hansen, Niels Henrik Mortensen and Erling Lyager. From the EEA, especially Jef Maes and Sigfus Bjarnason have been involved in the project. Furthermore we have had discussions with Philippe Lebaube, Olli Janhunen, Chris Nelson and John Allen from Eurostat in Luxembourg, and Vincent Blanc (Office International de l'Eau) in France.

We would like to express our thanks to all the people we met in connection with the making of this report for the kind support and valuable information received.

2. Background and history.

During the 1980'ies environmental policy in Denmark gained momentum, and a need was recognized for information on which to build strategies and make priorities - and for information as a basis for assessing the effects of the actions taken.

In the same period the Danish Aquatic Action Plan was initiated. This plan included a monitoring programme that was the largest so far on a Danish scale. Large amounts of data were to be transferred from the Danish counties and municipalities to the then Ministry of the Environment. It was anticipated that this would require great quantities of manpower if nothing was done to facilitate the process of exchanging the necessary data.

For this reason it was decided to develop a Danish system for data transfer dedicated to environmental information. The name of the system was to be STANDAT, an acronym of *stan*dardized *da*ta *t*ransfer.

Before STANDAT, the then Danish Ministry of the Environment typically received environmental data either as spread-sheet files or as ordinary comma-separated files. This meant that agreements had to be made in each case for the structuring of data, use of codes, organisation of the file etc. Much time was spent on making agreements, converting files and checking them. In the new situation, this would mean chaos when the huge amounts of water related data were to be delivered to the ministry.

Potential strategies.

Before the decision was made to develop a standardised data transfer format, other strategies and concepts were taken into consideration.

One such strategy was to base the process of data transfer on standardised software, provided by the central ministry to all data suppliers throughout the country. This strategy guarantees that input files are homogeneous and that their structuring and content are in accordance with the requirements of the central database. But it also presupposes that the local collectors of data are able and willing to adopt the registration systems as they are designed and applied centrally. There is no room for individual needs and solutions or creativity at the local level and the strategy is not very flexible. Furthermore the need for resources at the central level would be very large.

Another way of exchanging data is to base the data-transfer on ordinary comma-separated files. This concept is on the one hand simple and easy to understand and it is furthermore supported as a standard output function in eg spread sheets. On the other hand it presupposes that the sender and the recipient in each new case of data-interchange make an agreement on the specific structuring and codification of the files to be transferred. The possibility of making ad hoc solutions instead of establishing a more common view of the world including a common set of code lists may be tempting, but poses new problems as the experience of the then ministry had proved - eg in the use of resources and in the lack of possibilities for making data work together across databases and subject areas.

Another consideration was the experience from the use of the EDIFACT standard that was by the end of the 1980'ies primarily oriented towards interchange of documents in relation to trade. A specialization of the standard to a form and a set of code lists more in compliance with the needs of environmental data transfer was not yet initiated. So the strategy of the Danish EPA was to develop an exchange format specifically oriented towards environmental issues but with the possibility for conversion to other more generalized formats such as EDIFACT kept in mind.

The development process.

In the development of STANDAT, it was necessary to balance several (sometimes opposing) requirements:

- the system was aimed at ensuring a non-problematic exchange of information
- the system was to be easy to understand and use
- the system had to secure an optimal use of resources
- the system had to secure the coherence between data from different environmental information systems and different subject areas where it was relevant
- the system had to secure unambiguity in the form and content of the data transferred
- the system was to ensure that exchange of environmental information could be independent of hardware and software solutions - that it would not be necessary for all users to utilise the same computer systems
- the system should be set up in a way that would support an easy, standardised loading of data into data bases, and make quality control easy
- the system had to be able to handle differences in the use of character sets / code pages etc.

At first it was decided to have a private consultant make suggestions for a standardized format. The result of this project was called STANDAT version 0, and it was specialised for water related data. This version had both global (system-defined) and local (user-defined) code lists. In the extreme case, these last code lists could be used by only two users - the sender and the recipient of a given file, and the code list could be transferred together with the file. The global code lists were very specific in STANDAT v.0 and the format was concentrated on parameter-data - each line of the STANDAT v.0-file had the format: parameter, measurement system, quantity..

The problem of this version 0 was that it was both too inflexible (in the file format) and not generalized enough (in file format and code lists). Furthermore the use of local code lists would have made it too chaotic when large amounts of data on many different issues were to be transferred between several senders and recipients. This version of STANDAT was for these reasons never put to use.

The further development process was carried out by staff-members of the Danish EPA, and the result was STANDAT version 1.1, issued in 1989. Apart from the considerations listed above, special care was put into two issues at this point of the development process: guaranteeing that the recipient would have the full key to interpreting the file *included in the file itself.* And making it as easy as possible to handle files in all conceivable computer-based ways. The final file format was based on the concepts of entities and relations from database theory.

The basis for STANDAT's code list system.

Kommunedata, the IT-centre and software house of the Danish municipalities and counties, had previously developed a set of code lists for their own environmental edp system. These code lists were therefore used by many municipalities and counties, and it was decided to use them as the core of the code list part of the new system. It was evident that the code lists needed to be developed and expanded, as STANDAT was to have a larger scope than the existing Kommunedata systems. This was to be taken care of via the organizational set-up for STANDAT (please refer to chapter 6).

Other considerations.

To be able to achieve the objectives given it was decided also to develop edp-based support programmes. They were not included when STANDAT was first issued, but they were developed in their first versions in the subsequent years. The STANDAT support programme is especially produced to meet the requirement for the system to be easy to use and to provide a basic test facility for user-generated files, whereas the STANDAT load programme supplies file-loading facilities together with a more complete test procedure.

In the next chapters the four component elements of the final version of the STANDAT-format are described: code lists, file format, organisational set-up and edp-based support applications.

Readers with no interest in the technical details of file format and code lists are advised to skip the next couple of chapters.

3. Code lists.

An important component of the STANDAT system is the code lists. In short the code lists define *what* you can transfer data on and the file format defines *how* to do it. It should be noted that in the descriptions and examples of the next chapters reserved words of the STANDAT vocabulary are printed in **bold-**faced types.

The set of code lists.

The use of codes is well known from many different fields. E.g. many countries have created a system of civil registration numbers which are assigned to you at birth and stay the same through your life. The civil registration number is typically used to identify persons in tax systems, in connection with social security etc. Another use of codes is known from the postal service where postal codes identify particular areas.

The primary aim of codifying systems is to ensure a unique identification of the specific objects in the systems. If you take the civil registration number this is used to identify individuals and to distinguish between people who may have the same name or address or whatever identification you normally use when referring to a specific person. In the same way a postal code enables you to discern between e.g. towns with identical names.

In short a common code list makes unambiguous reference possible with no further description of the object referred to and without further information than the code itself. These are exactly the objectives of using codes in STANDAT. And in this way the current set of STANDAT-codes defines the environmental issues it is possible to transfer data on in the system.

STANDAT is based on four different sorts of code lists viz the subject code list, the information type code list, the combination code list and a set of value code lists. The contents of each type of code list is explained below. Using the terminology of database theory the subjects define the entities of the data model, the information types are the attributes and the combination code list describes the connection between attributes and entities. Finally, the value code list defines the domains of specific attributes. The description of relations between the entities lies in the parent-ID part of the subject description.

The subject code list.

The subject code list defines on what subjects data can be exchanged and supplies the code for each subject in STANDAT. A subject is defined as a set of logically coherent pieces of information. E.g. the enterprise subject contains information on V.A.T identification number, address, phone number and the name of the enterprise's contact person, if any.

Every subject is part of a hierarchy, either as the top (or the root, depending on your point of view) of the tree structure or as a dependent node in the tree. In STANDAT the enterprise subject is the root of the entire hierarchy. This is not because enterprises are necessarily the basic element in environmental themes, but merely a heritage from

adopting the fundamental structure of the code lists of Kommunedata's MIS-system (an edp-based database system for the environmental administration of some of the Danish counties and municipalities).

The subject code list includes for each subject registered four pieces of information. Besides the subject code itself (an eight-figured unique number); a short textual description of the subject; a "lock" specification; and the code of the "parental" subject.

Subject ID:	Name:	Lock:	Parent ID:
0000 2300	Inspection of waste water discharge	F	0000 0000
0000 2310	Measurement of waste water	F	0000 2300
0000 2311	Result of measurement of waste water	F	0000 2310
0000 2312	Remarks on measurement of waste water	F	0000 2310
0000 2320	Samples of waste water	F	0000 2300
0000 2321	Analyses of waste water samples	F	0000 2320
0000 2322	Remarks on waste water samples	F	0000 2320

Table 3.1: Part of the subject code list.

The length of a subject description as a whole is 84 characters with the following division into fields:

Subject code: pos. 1 - 8,
Subject name: pos. 10 - 73,
Lock: pos. 75 - 75,
Parent id: pos. 77 - 84.

The lock field contains either an 'F' or an 'L' to indicate whether the subject is 'F' - free or 'L' locked for further development, e.g. a new association of an information type. This field was introduced to have the possibility of disabling a subject yet still obeying the fundamental rule of STANDAT of always keeping track of history.

The figure below illustrates how a small part of the STANDAT hierarchy of environmental subjects is set up. Each subject is identified by its code number with the enterprise subject starting at code no. 0000 0000. By the beginning of 1996 the total number of registered subjects was in the magnitude of 300.

The information type code list.

The information type code list defines what information can be exchanged on all the subjects - every piece of information which is part of and describes the contents of a subject is listed in the information type code list. Examples of types are spatial and temporal related information about address, UTM location, year, date, etc. And more specific information about e.g. analysis results described as substance identification, measuring method, unit, and the actual result of the analysis. The types are numbered in succession and identified by a unique eight-figured number. E.g. UTM x and UTM y values are registered in the information type codes 0000 0047 and 0000 0048. Below a short extract of the information type code list is presented.

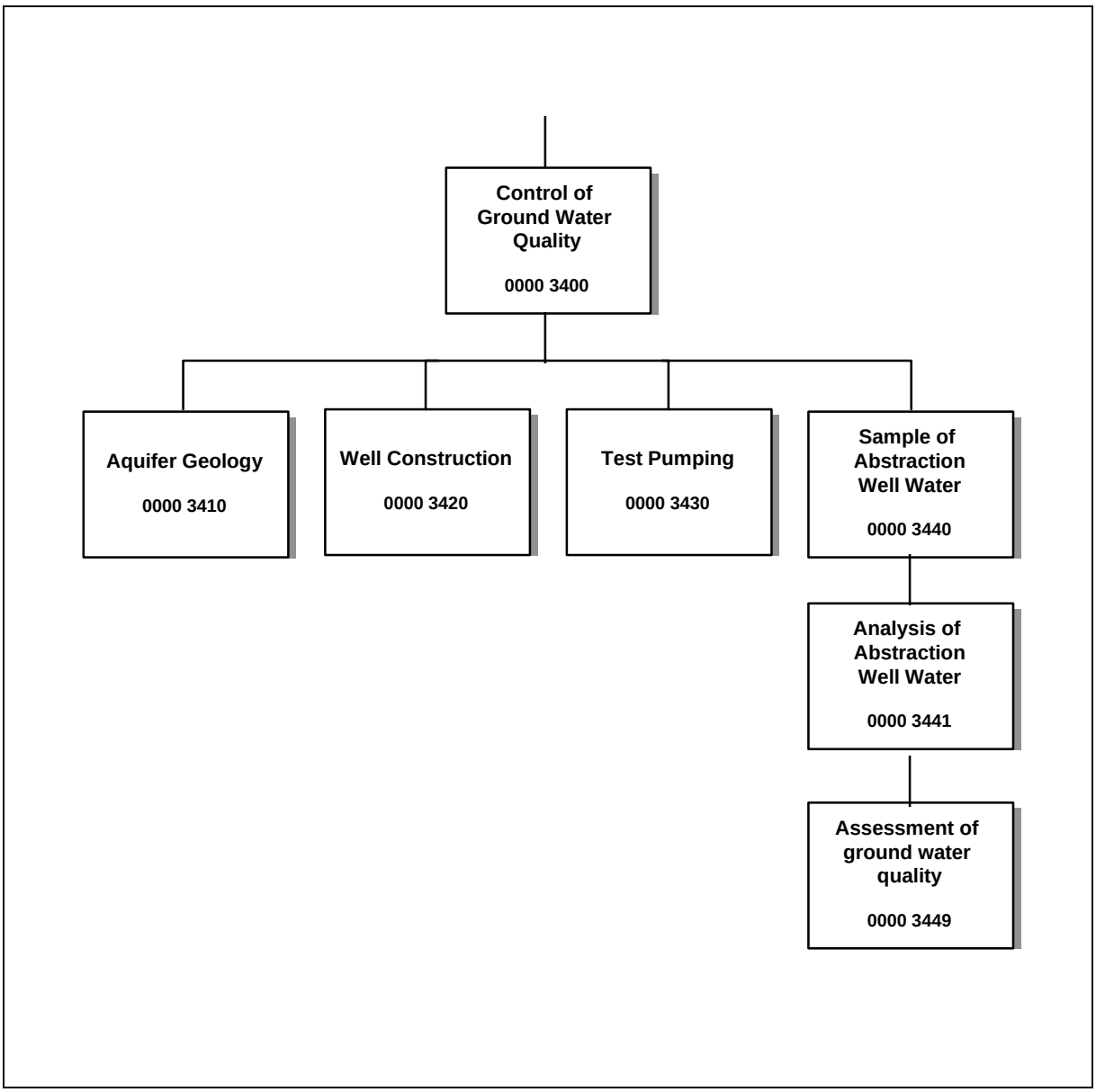

Figure 3.2: Part of the STANDAT subject hierarchy.

Data type ID	Format	Value code list	Description
0000 1792	N 2.0	STD00153	Method applied for treatment of soil
0000 1793	N 7.0		Amount of soil for treatment
0000 1794	N 2.0	STD00154	Method for immobilisation

Figure 3.3: Extract of the information type code list.

Beside the data type ID the information type code list contains the following information: the format; a reference to the attached value code list, if any; and finally a short textual description of the information type.

The description of an information type in the STANDAT system has a total length of 247 characters composed as follows:

Information type code:	pos. 1 - 8	
Data type:	pos. 10 - 10	'D' for Date, 'N' for Number or 'S' for String
Data format:	pos. 12 - 21	For data type 'N' the format 'x.y' where x is the maximum number of digits before the decimal point and y is the maximum number after the decimal point. In the case of integers y has the value zero. As for data type 'S' the format indicates the maximum number of characters allowed. The format of the data type 'D' is defined in the first part (the so called 'Header') of the STANDAT file.
Ref. to value code list:	pos. 23 - 30	If a value code list is referred to this code list enumerates the allowed values of the type in question.
Description:	pos. 32 - 247	A textual description of the information type.

The combination code list.

The connections between the subjects and the types of STANDAT are defined in the combination code list - for every subject in the subject code list the associated information types are listed. There are two fields in this code list, namely subject codes and information type codes:

Subject code:	pos. 1 - 8
Inf. type code:	pos. 10 - 17.

Below a short extract of the actual combination code list is shown. The subject with code number 80000006 concerns general information on a bathing water control station.

80000006	00000103	Number of samples per year
80000006	00000621	Year of report
80000006	00001568	Remarks
80000006	00001600	Year of abolition of the station
80000006	00001680	Year of establishment of the station

Table 3.4: Extract of the combination code list.

The value code lists.

The last element to be described in the system of code lists connected with the STANDAT concept is the value code lists.

A value code list enumerates the allowed values of a specific information type. E.g. one value code list describes the set of substances which it is possible to transfer measuring results on. Another one lists the codes for valid fish species in STANDAT transfers.

An example of part of a value code list is presented below. Every value code list is uniquely identified by an 8-character id composed of the characters 'STD' followed by a 5-figure number.

00	Not reported	
01	Recycling/sorting	
02	Incineration	
03	Land filling	
04	Special treatment	
05	Transported from plant	
06	Exportation	

Table 3.5: Part of the value code list STD00087: Methods of waste management

Please note the column to the right. This column (or field) is common for all value code lists in STANDAT. It is called the "out-of-date" mark and for some value codes it indicates that it is recommended not to use this specific value in the code list any more. This field was introduced after some years of use of STANDAT because of an increasing need to be able to signal that specific values have been deleted or replaced. The need arises if eg a measuring method is to be substituted by a new and better one.

On the other hand it is - as mentioned before - a basic principle of STANDAT not to delete any code. It must always be possible to transfer data referring to outdated codes. So instead of deleting value codes it has been decided to solve the problem in this way.

The format of the various value code lists differs depending on the needs for code length and description fields. E.g. the substance code list has a code length of 4 digits and a single field of textual description with a maximum of 20 characters. Whereas the code list concerning species, which is based on Nordic Code Centre´s RUBIN-system (cf chapter 10) has a code length of 7 characters and no less than 14 description fields, including i.a. the latin names of the species. The description of the specific formats of the actual set of value code lists is distributed together with the semiannual update package which is sent to the subscribers.

The description file

The description file identifies for each code list in STANDAT (including the various value code lists) the format of the actual files. It is used in connection with a.o. the user support programme SSP (which is described in more detail in chapter 5) to generate a

database which mirrors the structure and contents of the STANDAT code list system. An example of a format specification is depicted below.

FILE std00002
RELATION std00002
DESCRIPTION Postal code list
FIELD code **INTEGER** 1 4
FIELD postal region **STRING** 6 25
FIELD out-of-date mark **DATE** 27 36

Table 3.6: An example of a description file.

In short this part of the description file communicates that the value code list STD00002 concerns postal codes and is composed of three fields with a total length of 36 positions, namely a 4-numbered integer code (pos. 1 - 4), a 20-character description of the postal regions (pos. 6 - 25) and finally a 10-character date field (pos 27 - 36) identifying the "out-of-date" mark, if any.

The code list system - a world view.

Altogether the system of code lists describes a specific "world view" concerning the structuring, contents and connections between pieces of information on environmental subjects. It must be emphasized though, that the resulting "data model" is not based on a top-down analysis but is the result of an on-going "bottom up" based addition of new elements. The world view is static in the sense that no code once established is ever deleted[1]. On the other hand the system is dynamic because new subjects, information types, connections and value codes/value code lists are continuously being added.

[1] Of course erroneous codes or descriptions resulting from errors or misunderstandings in the semiannual update process are excepted.

4. The file format.

Just as the system of code lists describes the spectrum of environmental information dealt with, the file format describes the structural frame for the actual data transfers.

A STANDAT file is an ASCII-file composed of three parts: a HEADER, a DEFINITION section and a DATA section. This chapter is a short description of the syntax and contents of the three elements.

The HEADER section.

The HEADER contains administrative and technical information on the sender and receiver of data, the ASCII code set and actual STANDAT version used, etc. As a whole the HEADER is structured as follows with every text line starting in first position[2]:

Specification:		An example:
HEADER		**HEADER**
STANDAT Version number	V1.1	
Code set		DS/ISO 646
Date format		YYYYMMDD
Sender Institution		Roskilde County
Sender Municipality No.		025
Sender Name		Lise Hansen
Recipient Institution		Danish EPA
Recipient Municipality No.		101
Recipient Name		Dept. of Chemistry
Date of extract		19951201
Hour of extract		09
Minute of extract		30
Coordinate System		UTM
Geographical Zone		32
Remarks		Data on Bathing Water Quality, 1995.
END HEADER		**END HEADER**

Table 4.1: The HEADER section of a STANDAT file.

The reason that the HEADER includes information on date format and geographical reference system is that this gives both the sender and the receiver of data freedom to choose the most convenient representation for their use.

It should be noticed that every significant line of information is obligatory. I.e. no line of information, except for the remarks part, is allowed to be omitted in the HEADER of a STANDAT file. This is both because the information in the lines are important, and

[2] The rule of positioning textual data in the beginning of the line is general throughout the STANDAT file.

because each piece of information is connected with a specific position (a line number) and not identified and delimited by e.g. a reserved word.

The DEFINITION section.

The DEFINITION section of a STANDAT file defines the structure and contents of the data to be transferred in the terms of the STANDAT code list system described in chapter 3.

Any definition section should reflect the hierarchical structure of the subject code list. Subjects are embedded according to the "tree"-structure defined by the child/parent-ordering of the subject code list (cf chapter 3, figure 3.2).

There are three elements of description in the DEFINITION section. The first element concerns the identification and mutual ordering of the subjects to be transferred; the second one specifies the selection of information types; and the third element defines whether the data transferred are referential or substantial (scope of data).

Let us take a look at an example regarding the ordering of subjects:

```
DEFINITION
 GROUP <Subject Code 1> <Scope>
...
END GROUP
GROUP <Subject Code 2> <Scope>
...
END GROUP
END DEFINITION
```

Table 4.2: A DEFINITION section for non-embedded subjects.

In this example subject code 1 and 2 have no relationship:

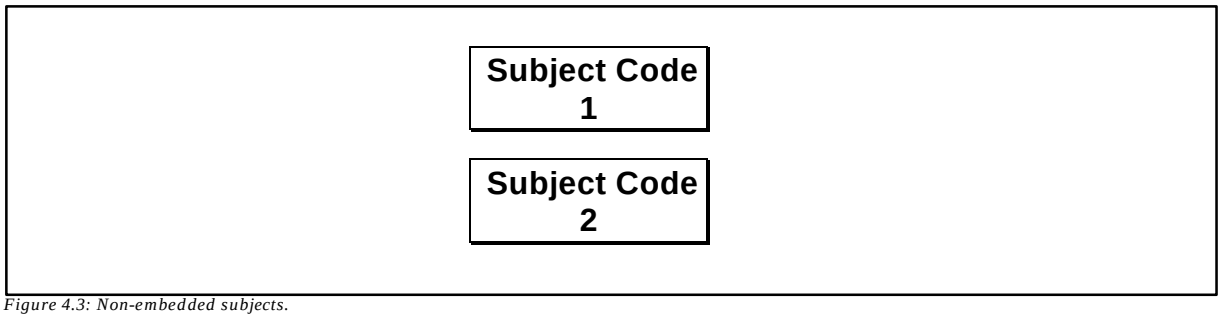

Figure 4.3: Non-embedded subjects.

If on the other hand subject code 2 is subordinated to subject code 1 then the DEFINITION section would have this structure:

DEFINITION
GROUP <Subject Code 1> <Scope>
...
GROUP <Subject Code 2> <Scope>
...
END GROUP
END GROUP
END DEFINITION

Table 4.4: A DEFINITION section for embedded subjects.

In this case the corresponding Entity-Relation-diagram would be:

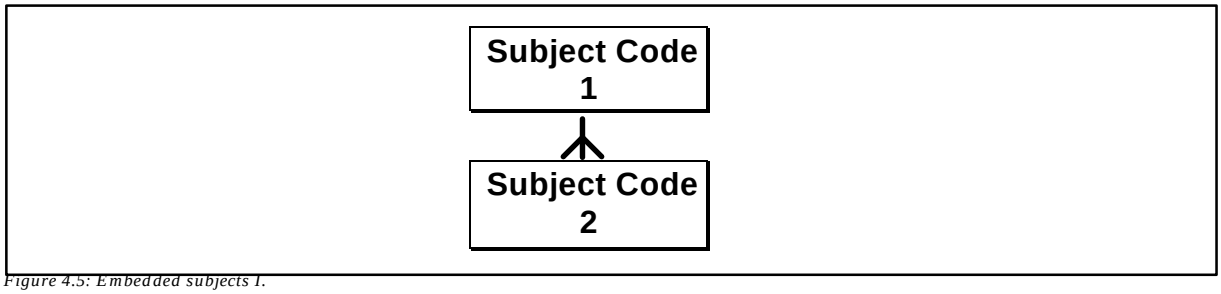

Figure 4.5: Embedded subjects I.

It should be noted that speaking in terms of relational databases STANDAT is only capable of defining and transferring entities (subjects) with a one to one or one to many relationship. Furthermore it is required always to refer to every in-between-subject in an embedded structure except in the case where it is only the innermost subject in the embedment that you want to transfer data on. An example: The subjects with code numbers 0000 0000, 0000 0200 and 0000 0201 have the hierarchical ordering:

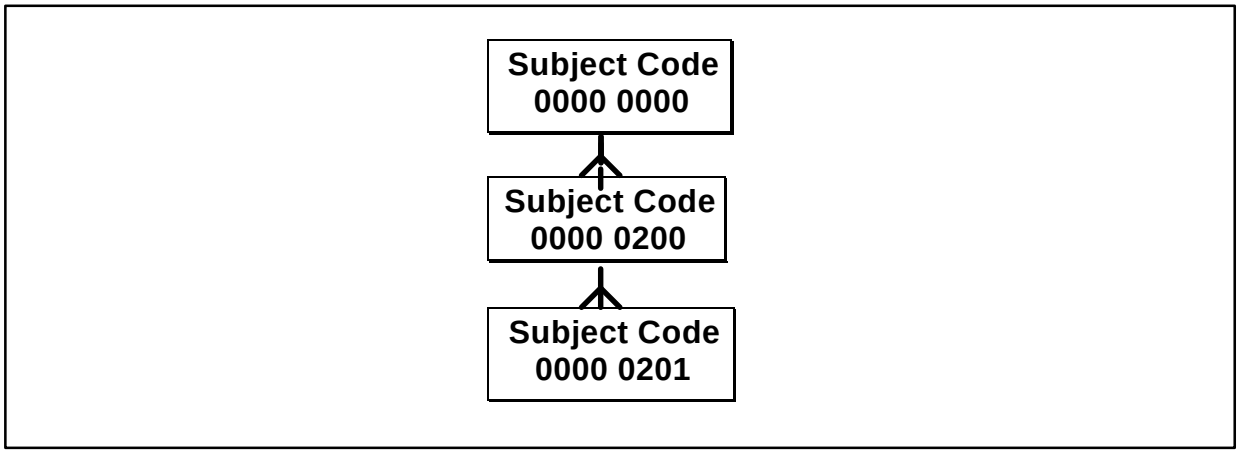

Figure 4.6: Embedded subjects II.

If you want to transfer information referring only to subject code 0000 0200 (and this makes sense without having data in the "parent" subject 0000 0000) then the DEFINITION section should consist only of a reference to subject code 0000 0200. But if the data in question is referring both to subject 0000 0201 and subject 0000 0000 it is required also to refer to the in-between subject 0000 0200 in the DEFINITION section.

Regarding the selection of information types this is simply done by enumerating the relevant information type codes for each subject in the DEFINITION. Of course it is

only allowed to pick out information types which are explicitly related to the subject in question in the combination code list. An example:

```
DEFINITION
GROUP <Subject Code 1> <Scope>
FIELD <Information Type Code 1>
FIELD <Information Type Code 2>
.
.
.
END GROUP
END DEFINITION
```

Table 4.7: Selection of information types.

The final element in the DEFINITION section concerns the scope of data. This qualifier is used to indicate whether the data are referential (Scope = **REF**) or substantial (Scope = **DAT**). I.e. whether the subject and the data are only used to identify a parent subject and a set of relevant key data. Or whether the subject contains the data which are carrying the essential in the transfer.

There are no predefined keys in STANDAT. This implies that it is of great importance that the sender and the recipient agrees specifically on the relevant set of key information types before a transfer. Otherwise it may become impossible for the receiving part to make a correct load of the data. As for the actual load of data it is the responsibility of the recipient to ensure that the loading programme only makes an update of the **DAT**-marked subjects of the STANDAT-file. If not there is a risk of overwriting relevant data in the **REF**-parts of the recipient database.

The DATA section.

The DEFINITION section of a STANDAT file specifies in detail how the actual data to be transferred are structured and interrelated. The DATA section contains the actual, relevant information and is delimited by the reserved words **DATA** and **END DATA**.

There are a few rules concerning the interpretation of the DEFINITION section. Firstly the number and sequence of information types enumerated in the DEFINITION section must be exactly mirrored in the DATA section.

I.e. with a DEFINITION section like this

```
DEFINITION
GROUP 00000000 DAT
FIELD 00000001
FIELD 00000002
FIELD 00000003
END GROUP
END DEFINITION
```

Table 4.8: Order of succession of information types.

the information types 00000001, 00000002 and 00000003 are to be repeated in exactly this sequence for each occurrence in the DATA section of the subject 00000000.

Another rule is that it is allowed to omit subordinate subjects carrying no data. But it is not allowed to omit parent subjects whether or not they are carrying data. An example:

```
DEFINITION
GROUP 00000000 DAT
FIELD 00000001
FIELD 00000002
GROUP 00000200 DAT
FIELD 00000043
FIELD 00000045
END GROUP
END GROUP
END DEFINITION
```

Table 4.9: Example of a DEFINITION section I.

defines a frame of which the following is a correct implementation:

```
DATA
GROUP 00000000
257
3
END GROUP
GROUP 00000000
257
3
GROUP 00000200
04222323
Hugo Rasmussen
END GROUP
END GROUP
GROUP 00000000
257
3
GROUP 00000200
04222323
Hugo Rasmussen1
END GROUP
GROUP 00000200
04222324
Hugo Rasmussen2
END GROUP
END GROUP
END DATA
```

Table 4.10: A DATA section corresponding to the DEFINITION section in table 4.9.

In this example the subordinate subject 00000200 is omitted once and afterwards repeated first one time and secondly twice embedded in the subject 00000000.

A third rule is that the enumeration in the DEFINITION section of subjects at the same level in the hierarchy is not determining for the sequence of these subjects in the DATA section.

I.e. if the subjects 00000200 and 00000300 are both at the same level of subordination to e.g. the subject 00000000 then this DEFINITION section

```
DEFINITION
GROUP 00000000 DAT
FIELD 00000001
GROUP 00000200 DAT
FIELD 00000043
END GROUP
GROUP 00000300 DAT
FIELD 00000093
END GROUP
END GROUP
END DEFINITION
```

Table 4.11: Example of a DEFINITION section II.

provides the possibility for repeating the subjects 00000200 and 00000300 inter-changeably in the corresponding DATA section as many times as necessary.

Table 4.12 presents an example of a complete STANDAT file with data on an analysis from a water supply plant. The left column is the STANDAT file itself; the right column is an explanation of each line of the file. This would not be part of an ordinary STANDAT file.

```
HEADER                              ; Start HEADER
V1.1                                ; Version number
DS/ISO 646                          ; Code set
YYYYMMDD                            ; Date format
Gundløse County                     ; Sender institution
899                                 ; Sender municipality number
Annelise Ravn                       ; Sender name
Danish EPA                          ; Recipient institution
101                                 ; Recipient municipality number
Kit Clausen                         ; Recipient name
19950315                            ; Date of extract
09                                  ; Hour of extract
30                                  ; Minute of extract
UTM                                 ; System of coordinates
32                                  ; Zone
Extract of data on analysis         ; Remark line
END HEADER                          ; End of HEADER
DEFINITION                          ; Start definition
GROUP 00000000 REF                  ; Institution
FIELD 00000033                      ; Municipality number
FIELD 00000039                      ; Name of institution
GROUP 00003200 DAT                  ; Water supply plant
FIELD 00001158                      ; Serial number
FIELD 00001236                      ; Name of water supply plant
FIELD 00001238                      ; Address
GROUP 00003210 DAT                  ; Circumstances of analysis
FIELD 00000143                      ; Date of analysis
FIELD 00001239                      ; Type of analysis
FIELD 00000601                      ; Laboratory
GROUP 00003211 DAT                  ; Analysis
FIELD 00000101                      ; Method
FIELD 00000095                      ; Parameter
FIELD 00000622                      ; Amount
END GROUP                           ; End of analysis
END GROUP                           ; End of circumstances of analysis
END GROUP                           ; End of water supply plant
END GROUP                           ; End of institution
END DEFINITION                      ; End of definition
DATA                                ; Start data
GROUP 00000000                      ; Start institution data
899                                 ; Municipality number
GUNDLØSE WATER SUPPLY PLANT         ; Name of institution
GROUP 00003200                      ; Start water supply plant data
1058                                ; Serial number
GUNDLØSE WATER SUPPLY PLANT         ; Name of water supply plant
BYVEJ 5 9999 GUNDLØSE               ; Address
GROUP 00003210                      ; Start data on circumstances of..
19950315                            ; Date of analysis
AN                                  ; Analysis type code
0112                                ; Lab. code
GROUP 00003211                      ; Start analysis data
0999                                ; Method of analysis code
0377                                ; Parameter code
0                                   ; Measured quantity
END GROUP                           ; End of analysis data
END GROUP                           ; End of data on circumstances..
END GROUP                           ; End of water supply plant data
END GROUP                           ; End of institution data
END DATA                            ; End of data
```

Table 4.12: Example of a complete STANDAT file with data on water analysis.

5. Computer based support programmes.

Edp support programmes for STANDAT comprises both software intended for the producers and for the recipients of STANDAT files. The two software programmes were developed by the Danish EPA to facilitate the implementation and use of STANDAT both at the EPA itself and for the users outside the Ministry of Environment and Energy.

SSP - The STANDAT Service Programme.

The SSP has been designed and developed with the producers of STANDAT files in mind. This is a very varied group when it comes to experience with the use of edp, when it comes to hardware and software platforms and knowledge of the STANDAT format as such. Therefore the primary aim of the development process has been to produce a PC programme with the following features:

- a user-friendly interface
- no special hardware and software requirements
- facilities for loading the STANDAT code lists and new versions of them
- user-friendly search-and-find facilities for identifying subjects, connected information types and value codes
- a complete syntactic test of the relevant STANDAT files
- easily understandable error and warning messages
- functions for converting a STANDAT file from one code-page to another
- generation of simple tabular reports on STANDAT files.

The SSP programme was developed in CLARION, and first issued in 1992. It is delivered free of charge to the subscribers of STANDAT. Figure 5.1 provides an overview of the facilities in the SSP.

The STANDAT LOAD System.

The test and load of STANDAT files into databases can be handled in two ways: either you develop a specific check and load procedure for each type of transfer / each database.

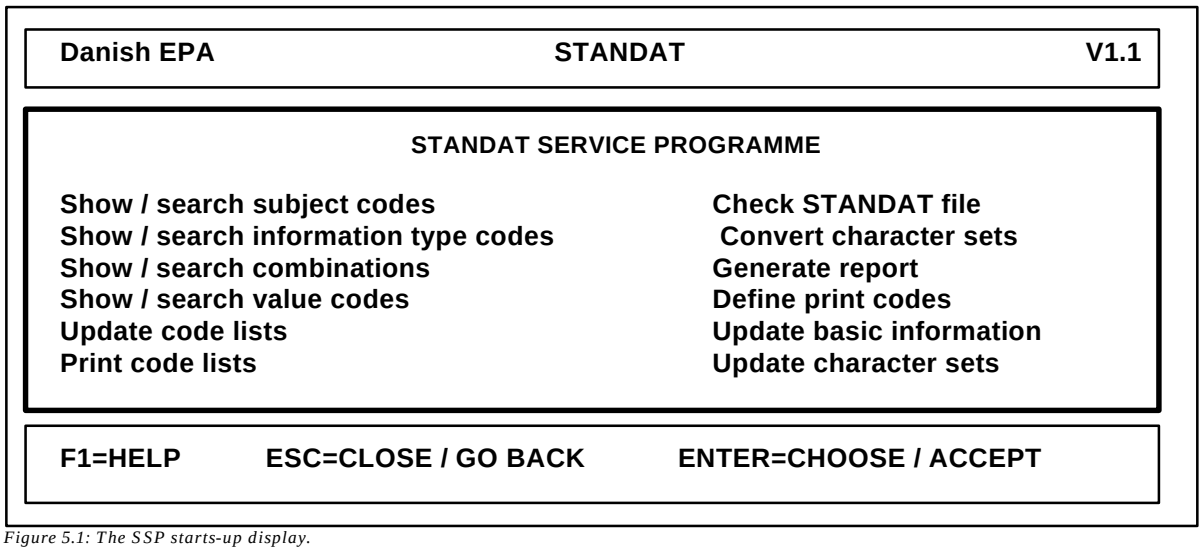

Figure 5.1: The SSP starts-up display.

Or a general solution for all types of transfers is applied. The Danish EPA has chosen the latter solution primarily because the agency receives an extensive and continuously expanding set of expert data transferred via STANDAT. At present the transfers are typically annual or bi-annual and concern data on ia bathing water quality, solid waste, waste water treatment, fish farming and contaminated sites.

The demands for the development of this general load system were ao aspects that it should be able to:

- go through a complete syntactic test of any kind of STANDAT file
- use a generalized specification of "semantic" requirements that could with a few specifications be used for any file
- perform a complete "semantic" check of any set of STANDAT files on the basis of the specification mentioned above
- produce the relevant error and warning messages
- have a general frame for describing the "object database" ie the database into which the relevant data are to be loaded
- perform the actual load of the data from a STANDAT file into the relevant (parts of a) database.

The STANDAT Load System of the Danish EPA has been developed to fulfil these requirements and the first version was implemented in 1993. It is primarily programmed in Pascal and SQL and it has been adjusted and taken into use at GEUS - The Geological Survey of Denmark and Greenland.

Figure 5.2 provides an overview of the elements of the STANDAT load programme.

The SSP and the STANDAT load system is further discussed in chapter 8 (experience) and chapter 10 (ideas for further development).

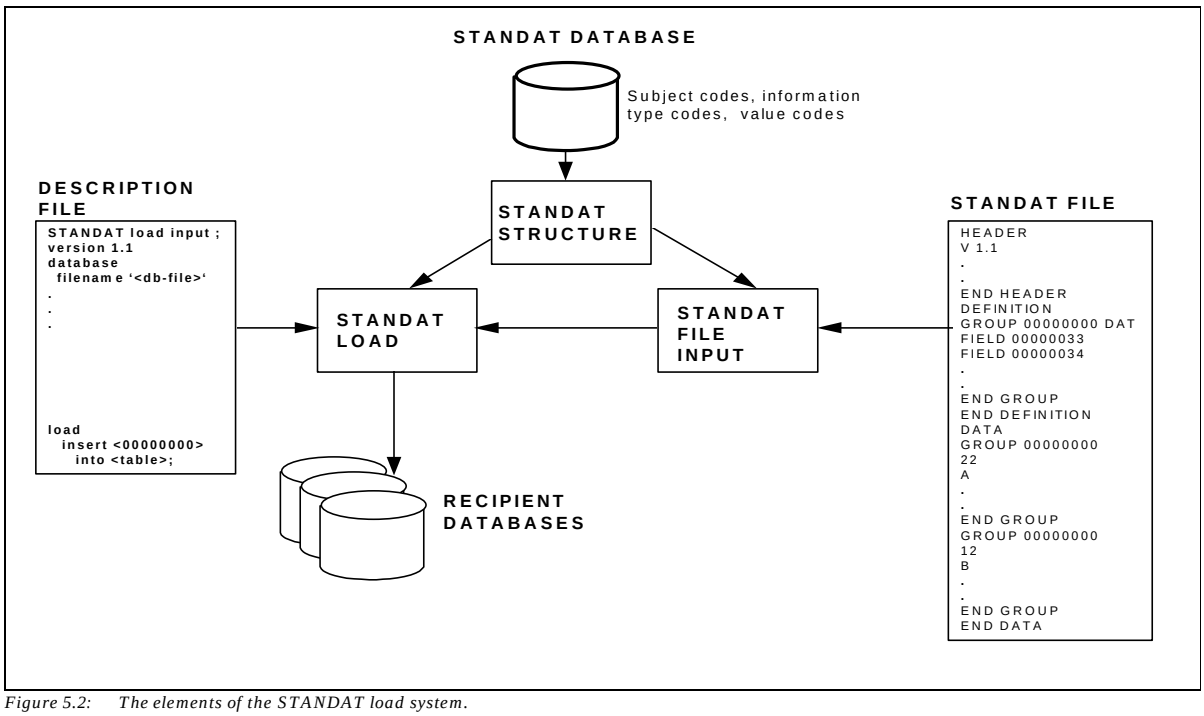

Figure 5.2: The elements of the STANDAT load system.

6. Organisation.

To maintain and develop a system like STANDAT, the technical components are not enough. It is important also to have an organisational set-up, that ensures a smooth cooperation between all the users of the system, and that makes sure that all those concerned are aware of the distribution of competence when using the file format and the code lists.

The organisational set-up for collecting environmental data in Denmark .

The Danish organisational concept for collecting data on the environment is decentralized and makes a point of giving the responsibility for any issue to the unit that is closest to the real problems and most knowledgeable about it.

The responsibility for collecting data on any given subject is assigned to specialized environmental data topic centres by the Ministry of Environment and Energy. These topic centres either get their data from counties and municipalities, or they conduct the collection of samples, surveys etc themselves. The topic centres are ia responsible for

- defining the data that are needed for the ministry to perform its tasks of planning, prioritising and assessing effects of measures taken
- assessing the quality of the data collected
- setting standards for the reporting of data from other parts of the organisational structure
- defining the guidelines for processing and using data, eg in models
- being up to the state of the art concerning methods of measuring and analysing data.

Topic centres are mainly placed in the different units of the Ministry of Environment and Energy.

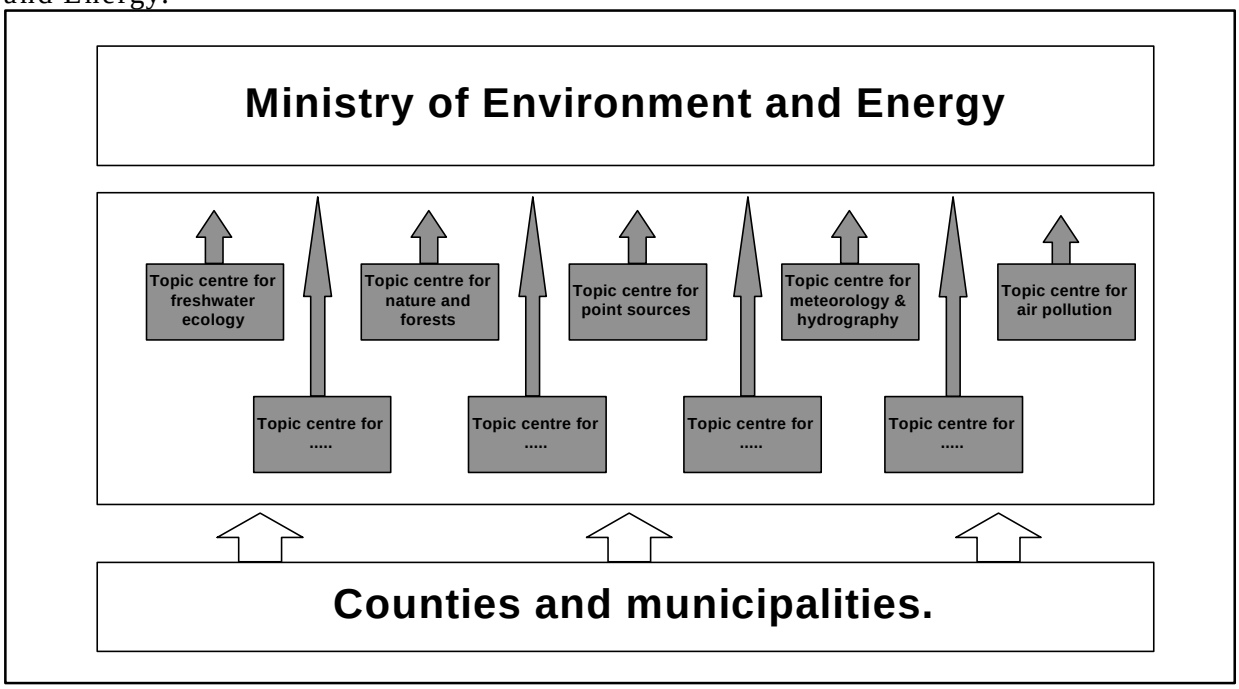

Figure 6.1: The organisational structure for collecting data on the environment in Denmark - the national data focal points.

As can be seen in figure 6.1, the data flow typically goes from the counties and the municipalities to the national data topic centres, that are responsible for aggregating this kind of data to supply a national / nationwide overview. In other cases, data are collected by the data topic centres themselves, in some cases via their own networks of measuring stations or via other kinds of measurements or surveys.

Before the stage of publication of national data, the final recipient of data is most often the Ministry of Environment and Energy. The ministry is on the national level responsible for all reporting of nationwide environmental information to the public, to the EU, to other international fora etc. The ministry is also responsible for putting together the information across counties and municipalities so that the data can be used for prioritizing and comparison. National databases on a large range of subjects are therefore placed in the ministry and / or its national data topic centres.

The organisational structure of the STANDAT system.

This decentralised structure is part of the organisation for the administration and development of the STANDAT system. In this way

- the interests and wishes of the users are taken into account
- questions and difficulties are solved by the relevant experts and on the relevant level of the organisational set-up
- there is a correspondence between the coordination system on the substantial side on the one hand and on the data technical side on the other hand.

The component elements in the organisational structure associated with STANDAT are presented in figure 6.2.

The steering committee.

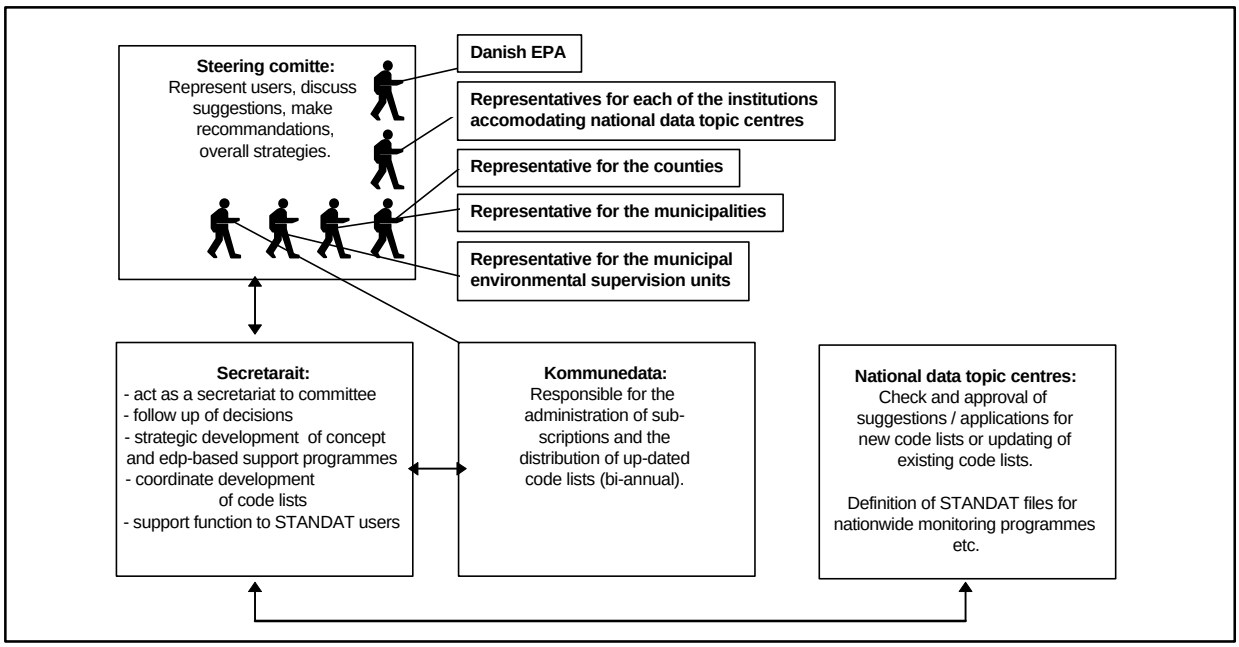

Figure 6.2: The organisational set-up for the administration and development of STANDAT.

In the steering committee all the principal participants and users are represented: the topic centres, the municipalities and counties, the Association of Environment and Food Control Units, the Ministry of Environment and Energy and Kommunedata.

The steering committee convenes twice a year. It makes recommendations, discusses strategic questions and acts on the questions and proposals put forward by other users through eg the secretariat.

The secretariat.

The secretariat is placed in the Department of Development and Environmental Information of the Danish EPA. One staff of the department is assigned to this task, but questions of a more strategic character are discussed in the entire data unit of the department. This unit consists of 5 persons, data experts and social and natural science experts. When it comes to technical questions related to eg development of edp-based tools etc, the IT department of the Danish EPA is consulted.

The tasks of the secretariat are:

- to act as a secretariat to the steering committee
- to take care of the follow-up on decisions made at the meetings of the committee
- to coordinate the handling of applications for new code lists or additions to existing code lists.
- to act as a support to the STANDAT-users if problems or questions arise.
- to take care of the strategic development of STANDAT itself and of the edp-based tools related to STANDAT.

Kommunedata.

Kommunedata is responsible for the technical part of the updating of the code lists. This includes the insertion of the approved new codes into the code list system (see below); modification of the description file; and distribution of the updated codes etc. to the subscribers of STANDAT. Kommunedata is also responsible for registration and administration in relation to the subscription part of the STANDAT concept. A STANDAT-subscription costs 2000 dkk (1995-prices).

The national data topic centres.

The data topic centres are among the most important users of STANDAT. Furthermore, they have the expert knowledge about the subjects for data collection and they handle much of the environmental data at the national scale in Denmark.

They are therefore responsible for passing or rejecting suggestions for new code lists or additions to existing code lists within their expert areas. The secretariat coordinates this activity: when the secretariat receives requests for updating of the code list system, it forwards the request to the relevant topic centre(s) for assessment and approval.

Any user of the STANDAT system can make requests for new codes and new value code lists, but the topic centres and the secretariat are responsible for guaranteeing that the additions are logical, coherent with the rest of the system and in accordance with scientific / professional practice.

7. Defining, creating and transferring a STANDAT file - the main principles.

The understanding of the actual process of defining, creating and transferring a STANDAT file is closely linked to the understanding of the file format (chapter 4) the code lists (chapter 3) and the organisational set-up (chapter 6). Also the use of the STANDAT support programme SSP and the STANDAT load programme (chapter 5) is important in this context.

First of all it is not necessary to have your data stored in a specific database system or to use special hardware to be able to use STANDAT as a data exchange format. It is of course easier to produce a STANDAT file if your data are organized in a regular database system. This gives you the possibility for applying a proper retrieval-routine - a possibility not supplied with data stored in a set of spreadsheets.

The first thing to do before a data exchange is to make an explicit agreement on which data to transfer and how the exact structure and contents of the data-file is going to be. Preferably this agreement should be in writing and contain at least the following elements of specification:

- a general description of the data to be transferred
- an exact description of the STANDAT file to be produced including the contents of HEADER and DEFINITION sections and line-by-line examples of DATA blocks
- if key data (REF subjects) are to be transferred a detailed specification of the structuring and allowed contents of these subjects and the connected information types
- for any value code list in use an exact description (eg by stating the precise / relevant code numbers) of the allowed value codes. If it is relevant to restrict combinations of values from different value code lists the allowed combinations should be enumerated
- the time and if necessary specific media for delivery.

This can be done on an ad-hoc basis, but in Denmark it is typically done via the national data focal point organisation, defining the data sets for a large range of users and for several consequent deliveries of data at one and the same time.
There may be a need for parameters or the like in the data file that does not exist in the STANDAT code lists. A request will in that case be made to the secretariat for an extension of the code lists. Or perhaps a whole new code list needs to be established. If the request is urgent, an interim code or code list will be made. If not, the new codes / code lists will be included in the next biannual updating of the code list system, that is supplied to all subscribers by Kommunedata. The extensions will first of all be accepted or rejected by the relevant topic centre on the basis of their expert assessment of the request.

Typically, up till this stage it is the future *recipient* of data who is the most active part: defining the data-content, setting up the structure of the file and making requests for new codes. But in Denmark it is most often done in some kind of cooperation with the future supplier of data (please refer to figure 7.1).

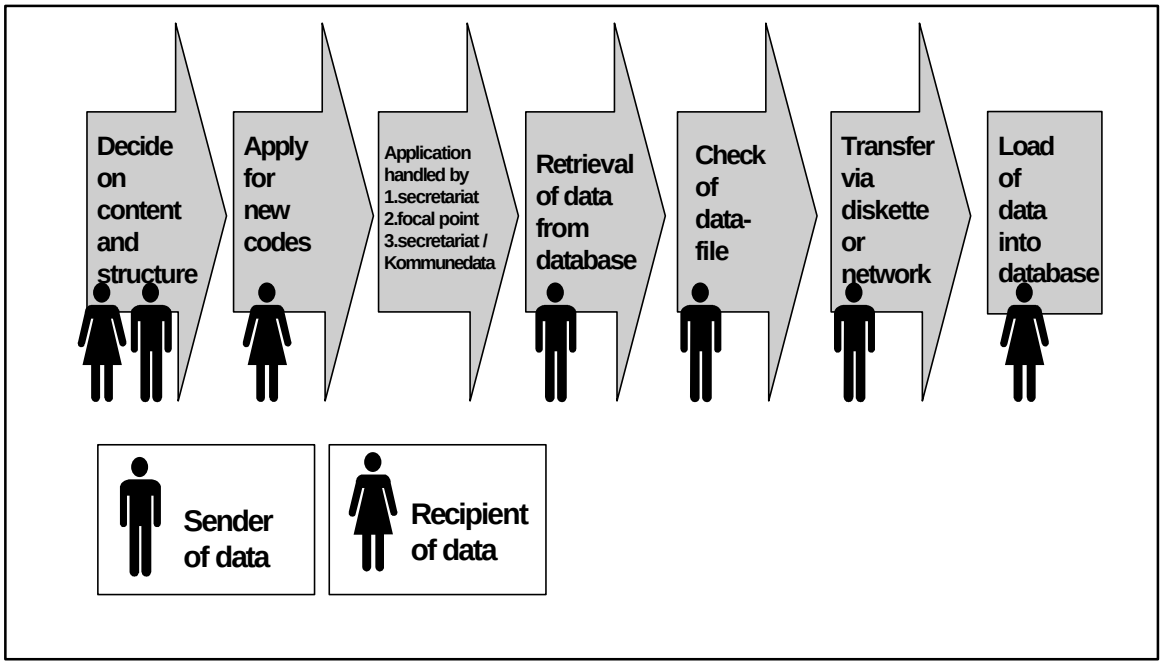

Figure 7.1: The main phases of creating and transferring a STANDAT-file. Steps 2 and 3 are only necessary of the relevant codes do not exist already.

The next thing to do as a sender of STANDAT data is to make the appropriate retrieval from your collection of data according to the specifications mentioned above. If your database has its own local codification it is crucial to make a correct translation of these codes into STANDAT value codes. If there is any doubt concerning a translation it is recommendable to contact the recipient and make a specific agreement about the interpretation[3].

When the STANDAT file has been produced it is recommended to test the file by using the STANDAT Support Programme. However, at the time being this software only conducts a formal, primarily syntactic test. The test procedure concerns erroneous usage of reserved words, incorrect structuring of the different parts of the STANDAT file, non-existing subject or information type codes, illegal combinations of subjects and information types, references to value codes not registered in the actual set of code lists etc.

It does not test whether the data transferred correspond in structure and specific contents to the data actually required by the recipient. E.g. whether the key information matches the existing key data in the object (recipient) database, or whether only the allowed subset of value codes and combinations of these are used.

The recipient of the STANDAT file has the task of making the final check before loading the data into her / his local database. In this process there are many possible degrees of universality in the check-and-load procedure. One can choose to develop a piece of software dedicated to testing and loading a specific STANDAT file. Or in the other extreme to make both the check and the load function totally general and describe the

3 In the years of usage we have noticed a tendency to use (parts of) the STANDAT value code lists in local systems. This of course makes the conversion process easier, but on the other hand it may cause applications for registering codes in the central value code list system that are mostly relevant at the local level.

specific set up and prerequisites by supplying the software with a specific set of parameters.

In the Danish EPA the latter type of solution has been chosen (please refer to chapter 5).

It is our experience that the more effort you put in making a precise specification of the data to be transferred beforehand the less time is wasted in sending erroneous STANDAT-files back and forth between the sender and the recipient. An important aim of the future development of the support software connected with STANDAT is to further formalise and integrate this specification so that the sender and the recipient of a particular STANDAT file go through exactly the same testing procedure. This would be a time-saving feature in the process of exchanging environmental data via STANDAT.