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

Chapter 8 - 13

8. Experience of the use of the STANDAT system.

By now STANDAT has been in function for 6 years and it is possible to assess the results, the successes and the problems in the use of the concept. For this purpose we have had discussions with colleagues involved in the use of STANDAT, and we have used details from user surveys made previously.

The experience from the use of the STANDAT-concept can be divided into experience related to each of the four component elements: the file format, the code lists, the edp support programmes and the organisational set-up.

First it should be noted that an important general point is - the choice of solution disregarded - that though common standards may seem both appealing and necessary from a top-down point of view, this is not necessarily how it is regarded from a bottom-up point of view.

Many users tend to regard a common solution that is defined from above as an encumbrance. This is especially the case for organisations that have already developed their own solution to data transfer problems when the common solution is introduced (eg local solution based on common definitions of simple spreadsheets between two or more users). STANDAT was not embraced with unequivocal enthusiasm when in was launched - and there are still users that tolerate the fact that they have to use the common format, but who certainly do not like doing it.

This is one of the premises that should be taken into account when planning how to introduce such common systems.

Experience related to the use of the file format.

Although in theory the file format looks simple and straightforward, practice has demonstrated that it is not always easy to produce a correct STANDAT file. Apart from using the right reserved words, following the overall structural and syntactic requirements and positioning data on each line correctly, there have been and still are difficulties for the users in transforming a DEFINITION section into the corresponding DATA section. The problems are related to the embedding of subjects which may be rather complex, but also to the question of where you can omit subjects in the different levels of the hierarchical structure.

Therefore it is very important to communicate the syntactic and semantic rules of STANDAT to the users. And to present some explicit examples of correct files when describing the data required in STANDAT-format, as specified in chapter 7.

These problems related to the understanding of the file format also highlights the need for information and education, cf below. Some sort of formalised telephone 'hot-line' help facility for the users would also have been useful here.

Experience related to the use of the code lists.

When producing the code lists, it is necessary to find a balance between two opposing requirements: on the one hand, the code lists should be structured, they should reflect the state of the art of scientific knowledge, they should be comprehensive and without redundancy in the codified elements.

On the other hand, if the system is to be user-friendly and relatively easy to update, the code lists should also be set up in a way that is *pragmatic*. If one is *too* ambitious on the question of code lists, the process of development and updating will be very time consuming and there is a risk of the code lists becoming too complicated in structure and content for everyday practical use. As all codes are to represent the same phenomena continually, it is important not to put meaning into them, eg the codes should not reflect a hierarchic structure concerning the entity in question.

Practice has demonstrated that especially widely used value code lists such as the substance parameter list, the measuring unit list and the measuring method list has had a tendency to grow fast and not always in a non-redundant way. Accordingly an important experience is that the development of this kind of code lists should be watched closely and that existing international classification lists should be used as far as possible as the basis for codification.

Another consideration concerns the organisation of subjects in the subject code list. As described in chapter 3 all subjects in STANDAT are structured hierarchically. This implies that many-to-many relations can only be implemented by repeating the top-level subject the required number of times. As network structures are not uncommon in connection with environmental issues (eg monitoring networks) this restriction may in some cases cause inappropriate use of the format.

Experience related to the computer based support programmes.

The experience from the use and development of the SSP is first of all that this kind of support software is a necessity in a situation where the production of files is the responsibility of a very heterogenous group of people. It is necessary for the producers of STANDAT files to be able to get exact information on eg the precise code, type and format of an information type, based on the latest updated version of the set of code lists. Furthermore it is important to be able to get an overview of the STANDAT file produced. But most important is the possibility of making a test of the file before submitting it to the recipient of the data.

A crucial point in the design and development of the SSP is that the procedure for making syntax check must be as close as possible to perfect. The producer of a STANDAT-file must not risk getting an approvement that is not correct when submitting her / his file to a check via the SSP, as this leads to inconvenient use of resources when the recipient of the file returns it and the producer has to start all over. To avoid this situation a lot of time and effort has been put into making the best possible syntax check procedures of the SSP.

Nevertheless, one should be aware of the fact that identification of syntactic errors is only part of the problem. Experience has demonstrated that it is equally - if not more - important to ensure that the produced STANDAT file fulfils the requirements of the recipient "semantically" - eg that the DEFINITION section matches the description made by the recipient, that REF(erence) data corresponds with key data in the receiving database and that the value codes used are the right ones also taking the context into account[1]. At present the SSP does not cope with this aspect of testing STANDAT files, but the Danish EPA is planning to make a new version dealing with at least part of the semantic test task.

Other experience concerns the error and warning messages produced by the SSP. It is very important that they are understandable to all the users of the STANDAT format. In the existing version of the SSP the way of describing errors and the conditions causing them is rather technical and this has led to many misunderstandings. On the other hand it is - as generally recognized in connection with software development - not a simple task to produce relevant, precise and easily understandable computer-generated error messages.

Concerning the STANDAT Load System the fact has been recognised that the more general an edp-based solution the more complex the resulting code becomes and the more effort has to be put into making specifications for the specific load-procedures for eg an individual database. Nevertheless it has been worth the initial effort both because maintenance is limited to one and only one system, and because the addition of new test-and-load functions has proved to be fairly straightforward.

Experience related to the organisational set-up.

When launching a system as comprehensive as the STANDAT concept, the ideal solution is if possible to use existing organisational set-ups, in the way that the national data topic centre organisation was used in Denmark. In this way you can be sure of having a link to the most important users of the system, and you make sure that you have access to scientific expertise as well as knowledge about the administrative requirements.

*Information, education, work-shops and seminars*are all extremely important when introducing a concept like STANDAT to a large group of users. It is a question of both supplying information and training, *and* of ensuring consensus on the importance of using a common system and of the benefits of doing it.

When introducing STANDAT in Denmark resources for this task were not available. Ideally STANDAT would have been introduced at a large seminar for representatives for all the participants in the process of collecting data on the environment. This could have been complemented by educational work shops where the system could have been presented in details, and where test examples of STANDAT files could have been produced by the future users in a supportive learning-by-doing-environment.

[1] In any specific type of data transfer some combinations of substance parameter codes and measuring unit codes are valid (and reasonable) and some are not.

Equally important is of course written information, aimed at different user-groups. This includes short, general introductions in the form of booklets, a comprehensive introduction to all aspects of the format and technical guides for specialised user groups. Precise, written descriptions are also important in any specific data transfer - these should be produced by the recipient of the relevant data in a form that is comprehensible for all responsible for delivering data. A guidebook on how to describe data-files would be very relevant here.

Another important point in this context is that the difficulties in applying a standardised concept should not be underestimated - it *does* require resources, information and user support. Especially the very different background and circumstances of the users were a problem. There are great differences between the STANDAT users in technical basis, software applications, human and economic resources and education and experience. Especially the varied prioritising and resources in the field of environmental data in the different municipalities and counties posed a problem.

The conclusion to this problem is that the better you are able to supply the users with education, information, as well as resources, the easier you will get on your way.

In some cases in Denmark it has been attempted to supply the data producers with computer based programmes for registering data and producing appropriate STANDAT files (eg in the area of waste data). The national data focal point (in this case the Danish EPA) supplied a common registration system which was to be used at the waste treatment plants, either directly as a registration system or indirectly as a link between an existing system and the required implementation of STANDAT codes and file format in this specific area of use. This is a way of ensuring homogenous STANDAT input files, at least at a syntactic level. But by using such a programme as a link to existing registration systems, the difficulties with making the right translation of concepts and codes between the local system and STANDAT should not be underestimated. It is also a method that is resource consuming at the central level, and one that is not totally in accordance with the original concept of independence of software solutions.

An example: the Danish Aquatic Action Plan.

For most of the time that STANDAT has been used, annual user surveys have been conducted in connection with the Danish Aquatic Action Plan. The national data topic centres (that are the most important recipients of environmental data in this context) have been asked for an assessment of the data transfer in connection with the Danish Aquatic Action Plan for the previous year. Some of the conclusion from these surveys are:

- expect initial difficulties ! The results of the first year were problematic, but improvements were marked the year after when the necessary adjustments had been made from earlier mistakes.

- the mistakes could be related mainly to two factors: the problems of the receivers of data in describing the required data-file in a comprehensive and comprehensible way. And the resource problems of the senders of data when it came to understanding STANDAT, implementing changes to edp-systems, creating files etc.

- typical mistakes were:
  - reporting of non-existent value-codes
  - invalid combinations of codes
  - missing values for the identification of information (keys)
  - reporting of the same data twice or more
  - lack of consistency in the data transferred

- it is very important that great care is put into the quality control of data and data-file before sending it. This requires many resources, if not for the sender, then for the recipient.

- it is extremely important and of great help to the users if the agency responsible for the format supplies them with user support software.

- it is important to have staff with an expertise in the data-organisational aspects of computer science in the different parts of the reporting system

- when political decisions are made on *what* data are to be collected and *how* it is extremely important to use the know how of computer scientists to make sure that the decisions are implemented in a way that makes information computerisable.

Things not to do.

Another way of summing up the experience made in Denmark is to focus on things not to do:

- First of all one should not decide to give up on the task of ensuring coherence and comparability in the data collected. Even though it requires resources and even though there are always initial difficulties, it is worth while in the long run.

- One should not underestimate the resources needed for implementation

- One should not forget to supply the users with as many help facilities as ones resources allow

- One should not develop code lists, formats, etc. in ways that make new developments impossible/very difficult to implement

- One should not be over-ambitious in relation to code lists. There is a discrepancy between the ambitions of scientist and the requirements of monitoring with administrative-political aims. The discrepancy will typically be seen most clearly in relation to the time for development of new code lists, where the scientific ambition typically is to be exact and go into detail, whereas the administrative need is to have the code lists ready as soon as possible

- One should not underestimate the problems that arise when introducing new concepts in areas where solutions already exists.

9. Similar interchange formats - experience, advantages and drawbacks.

STANDAT is fairly unique in being a data transfer format that is dedicated to environmental information generally speaking, in being fairly simple and pragmatic in its concept, and in having been in use for several years. Other formats are as far as we have been able to ascertain either more general i.e. not oriented specifically towards environmental information, or more particularly developed to cope with a particular topic within the field of environmental data exchange.

This chapter is dedicated to a *brief* overview of a couple of these other data transfer concepts. A thorough study of the concepts is not a primary aim of this report, so to facilitate the process of comparison it has been done on the basis of a predefined set of parameters. The parameters and main points about the two concepts as compared to STANDAT are presented in table 1 in this chapter. The points in the text refer to this table.

The other concepts are the GESMES-concept of EUROSTAT, a development of the EDIFACT standard, dedicated to transfer of statistical data, even fairly complicated sets of data. And the SANDRE reference format, initiated and supported in France by the Ministry of the Environment, the six French water Agencies, the Fisheries Council of France, the French Institute of the Environment and the International Office for Water.

The 2 formats have been chosen because they are different in focus and therefore offers different kinds of inspiration for both the development of STANDAT and for the EEA considerations on data exchange.

Lastly this chapter will briefly present an example of an international set of code lists that are not attached to any specific file format - the code lists developed by NCC, Nordic Code Centre.

	STANDAT (STANdardized DATa exchange)	GESMES (GEneric Statistic MESsage)	SANDRE (Secrétariat Administratif National de Données Relatives à l'Eau)
RESPONSIBLE ORGANISATION	Danish EPA, Copenhagen	CEN/EBES/EEG 6 (Comité Européen de Normalisation / European Board for EDI standards /EBES Expert group 6)	The French Ministry of the Environment
RANGE/DEDICATION	Environmental data generally speaking including data on eg sources of pollution. Raw data and derived data.	Any kind of statistical information - typically multi-dimensional data sets and metadata such as footnotes, measurement units etc.	All data on water
GENERAL CONCEPT	File format based on the entity - relation model	Based on the EDIFACT standard	Format based on entity / relationship models completed by code lists and

	STANDAT (STANdardized DATa exchange)	GESMES (GEneric Statistic MESsage)	SANDRE (Secrétariat Administratif National de Données Relatives à l'Eau)
			data dictionaries
COMPONENT ELEMENTS	File format, code lists, edp-based support programmes, organisational set-up	Messages, segments and data elements.	Common data dictionary, national nomenclature, standards and exchange protocols.
FILE FORMAT	Header section, definition section and data section. Hierarchial structure, embedded subject groups. Simple ASCII file with line separated data.	A message consists of a contiguous sequence of segment type- identifiers, each followed by the required data elements. The hierarchical structure of data is reflected in the structure of a message.	Header section (sender, recipient etc) and data section. ASCII file with a relational structure. One object per line.
CODE LISTS	Common set of code lists on subjects, information types and value domains. Combination code list defines combinations between subjects and information types.	UN/ECE Edifact and EU/Eurostat for codes relating to structure definitions etc. Gesmes supports the identification and /or transmission of externally maintained code lists	Nationally valid codes on water related subjects eg water analysis parameters Also geographical reference system on hydrography etc.
ORGANIZATIONAL PRE-CONDITIONS AND SET-UP	Steering committee. Expert groups based on national data topic centres. Secretariat in Danish EPA, practical updating by Kommunedata. Subscription based.	GESMES is a specialisation of the general EDIFACT format and is based on the same organisational framework	Steering committee, follow-up committee, specialised working groups, correspondents and a permanent team at IOW. Free of charge.
MAINTENANCE OF COMMON CODE LISTS ETC.	Application from user, followed by expert and data manager assessment, biannual update and distribution of code lists (diskette)	UN/Edifact for structural lists and codes of relevant maintenance agencies for data dictionaries and domain specific code lists	Application from users, expert and data manager assessment, code lists updated at each application and accessible by a modem-linked server
IN USE SINCE	1989	1993/94	1994
CURRENT STATUS	Approximately 300 subjects, 1250 information types and about 170 value code lists. Approx 75 subscribers.	Will be implemented into statistical dataflows between Eurostat and European Economic Area member states concerning eg balance of payment data, short term indicators, national accounts etc. Also used by private companies	Approximately 250 objects, 1000 data-elements and 50 code lists. Approximately 170 users in France and abroad.

Table 9.1: Overview of the concept of the STANDAT-format, the GESMES EDIFACT message and the SANDRE reference format.

The GESMES EDIFACT protocol.

GESMES is a data exchange format conforming to the EDIFACT syntax. An EDIFACT interchange, a message, is composed of a sequence of segments. Each segment is identified by a unique 3 character code. Some segments are defined as part of the EDIFACT Syntax (described in the ISO standard 9735), while other segments called User Data Segments are defined in the UN Trade Data Interchange Directory (UNTDID). Segments may be grouped together to reflect the structure of the data set to be exchanged. The data model used is the entity-relationship model.

The smallest unit in an EDIFACT message is the data element. Each segment comprises one or more data elements which may be simple or composite. Eg the DAM (Date and Time) segment contains the following data elements:

Tag:	Name:	M/C:	Format:
C507	DATE/TIME/PERIOD	M	
2005	Date/time/period qualifier	M	an..3
2380	Date/time/period	C	an..35
2379	Date/time/period qualifier	C	an..3

To specify that the message date (qualifier code 137) is 24 December 1995 the actual segment would be:

```
DAM+137:951224:101'
```

the format code 101 meaning YYMMDD.

An example of a total GESMES file is included in annex IV.

At present many general and industry specific codes for EDIFACT messages are defined in the UN Data Element Directories and there is an organizational mechanism for identifying code responsible agencies that take care of the code maintenance tasks for specific areas. Regarding environmental matters the GESMES format at the time being includes no common, standardised segment types specifically oriented towards this subject area. Therefore the usage of the format for exchange of environmental data to a high degree depends on individual agreements between the involved partners on codification etc. This means that the data exchange partners must either agree in advance on the data dictionary of concepts (e.g. environmental concepts) and code lists, or send these definitions in the GESMES message itself.

Metaphorically speaking one could say that in EDIFACT the data elements are the words of a language, the segments are the sentences and a message equals a chapter in a book. The formal rules of the language, the syntax, is defined by the standard. But the semantic aspect of the language ie the generation of meaningful messages depends to a large extent on the elaboration of a detailed agreement between the sender and the recipient concerning structure, contents and codification of data.

The SANDRE reference format.

The SANDRE format is dedicated to making all water data in France compatible, homogenous and comparable. This does not mean that ideas, concept and experience may not be utilised in other environmental subject fields, but at present the format is oriented towards data on surface water (quality and flow), drinking water, sewage, ground water and marine water.

A key objective of the SANDRE concept has been to create a common data dictionary covering the environmental issues mentioned above. The entries of the data dictionary have been created in a cooperation between the users of the format and specialised working groups with both data managers and experts in the relevant field.

Much effort has been put into defining in detail the supplementary pieces of information needed for a precise specification of each subject dealt with in the data dictionary. Eg one of the so called 'trames' describing data on the results of measuring water quality is abbreviated 'OPP'. It comprises a (unique) identification of the measuring station, but also information on the exact date and time of the start and the end of the actual measurement. In this way the SANDRE format suggests a set of information types necessary for an exhaustive description of the subject matters in question. It is not part of the SANDRE concept to require the use of the total set of information types - you are free to make a selection adequate for the actual transfer of data.

The exchange format is composed of 2 sections. The first section contains administrative information on sender and recipient. The second section contains the data to be transferred in terms of the nomenclature. If you do not want to transfer data on a specific piece of information you just omit it.

In addition to using the subjects and value code lists of the common data dictionary it is also allowed to define and use local codifications.

An example of a complete SANDRE file is given in annex III.

Other sets of code lists.

The Nordic Council of Ministers in 1985 set up a network organisation with the task of developing code lists on a scientific foundation. The network had nodes in Denmark, Sweden, Norway and Finland, and was called Nordic Code Centre (NCC).

NCC has produced code lists on organisms and chemical/physical parameters for the use in research, environmental administration and the like. Examples of code lists are:

- phytoplankton
- vascular plants
- mollusca
- Baltic invertebrates
- pisces (fish)
- mammalia
- threatened species
- analytical determinants
- water research

- vegetation and terrain types

Each code list has a list identification consisting of 2 characters. Each specific code list has a version number and signature because the code lists are updated by insertions into the system.

The biological code lists typically has 3 components:

NAME	MNEMONIC RUBIN CODE	NUMBER
Felis Clausensis	FELICLAU+D1 (mammals)	198748937847

Table 9.2: The component elements of a NCC code lists - example not authentic.

The names in the biological code lists are the ordinary latin names, a genus name and a species epithet. The chemical / physical parameters are most often in English. This provides a possibility for using this part of the code lists as a tool for controlling names in databases.

RUBIN is an acronym for *Rou*tine for *B*iological *In*formation. The RUBIN codes were made to meet the need for short names for use in forms and for storing and searching in computers, where the long latin names are problematic. The RUBIN codes are mnemonic on the basis of the latin names so that they are recognizable to experts and scientists. The codes consist of 8 characters and a list identification, eg D1, the mammal code list.

The number code allows for hierarchial sorting as they are not alphabetic like the name and the mnemonic RUBIN codes. Therefore the NCC code lists also have a number part that is ordered hierarchical in a sequential series of numbers. The numbers supply the rank and place in the hierarchical structure according to the biological classification in classes etc.

The number codes have 12 digits allowing for changes in the systematics. The first digit differentiates between biological and other parameters, and the last digit supplies the version of the number.

This reflects the fact that the number codes can be changed according to changes in classifications etc. This is both the strength and the problem of the RUBIN codes. It is a strength because it allows for flexibility in a scientific area where classifications *do* change. But it is a fact that is difficult for computer-based systems to handle.

There is not a file format attached to the NCC code lists.

At present the development of the NCC code list system is no longer subsidised by the Nordic Council, and this poses a threat to the continuation of the NCC work.

Summary

The three tools described in this chapter are quite different in their aims and way of handling the task of exchanging environmental data.

The GESMES standard is very much oriented towards defining a general frame for transferring statistical data as such. Ie the predefined elements of the format concern aspects such as message administration, identification of sender, reporting period etc. and specification of the dimensions and data in the array to be transferred. The agreed set of code lists comprises no dedicated environmental codifications. This very crucial part of any exchange of environmental information is as mentioned before left to the participants in the data transfer process.

SANDRE on the other hand is specifically oriented towards environmental data or more precisely: data on water. For each relevant type of data in this area of expertise much effort has been put into defining as precisely as possible the necessary supplementary information types and describing the "life cycle" of the data. Furthermore very specific code lists on a.o. water analysis parameters, aquatic organisms and methods of analysis have been elaborated. Although a well defined file format is also part of the SANDRE concept, focus has primarily been put on structuring and codifying relevant pieces of environmental information.

Finally the NCC system is exclusively oriented towards codification. There is no file format connected with the list, and the aim of producing the code lists has been to produce an exhaustive set of unambiguous "domain descriptions" reflecting the "state of the art" in the various fields of scientific expertise.

Altogether the three concepts have chosen to focus on different and within their respective spheres very important aspects of the process of exchanging (environmental) data:

- generality and flexibility in the descriptive, data structuring part of the exchange format
- identification and exhaustive description of relevant types of information
- scientifically correct and unambiguous codification of the allowed values regarding a specific set of information types.

Seen from the point of view of the EEA all these aspects should be taken into consideration when designing the actual model and guidelines for dataflow and sharing of data in the EIONET. And of course existing code lists etc. should be used whenever feasible for the relevant purpose.

10. Ideas for further development of an interchange format for environmental data like the STANDAT system.

When a concept has been tried through some years of use, you get an idea of its strong points and shortcomings, and you get an idea of what features should be changed. This is of course also the case with STANDAT. In this chapter some of the ideas for further development and new designs are sketched.

It should be emphasised, that the development prospects presented in this chapter are only *ideas*. Their possible implementation will be a question of resources and a question for discussion in eg the STANDAT steering committee.

The relation to international standards.

STANDAT has so far been used for national purposes only. The Danish Ministry of Environment and Energy typically receives its data in STANDAT-format from the different national data-sources. The Ministry is then responsible for delivering the relevant subsets of data (typically highly aggregated) to international bodies and organisations, eg the EU, OECD, PARCOM. These organisations have their different formats for delivery of data - in surprisingly many cases the format is still a predefined paper-form.

There is no doubt that the demand for data to be exchanged internationally on edp-based formats will increase rapidly in the future. There are fundamentally two different ways of handling this when you have a fairly well-functioning national format: one way is some degree of adaption of the national format, that makes it possible to convert files from the national format into one or more international formats. The other way is a total adoption of the relevant international format at the national level.

The problem of the first solution is that you have to employ at least two different formats at one and the same time, and that you have to develop the conversion software. On the other hand, it is not likely that *one* global format for the whole environmental area will be decided on for several years, so it may be argued that by having *one* well functioning national format, and *one* national organisation (in eg Denmark the Ministry for Environment and Energy) responsible for converting national data into any relevant international format, you are quite well-endowed.

The problem of the second solution is as suggested above that a global international format has yet not been decided on. Furthermore, for any country that has a fairly robust and well-established exchange-format, any new concept has to be very convincing and easy to apply to offer an alternative at the national level.

The answer to this question should for any country be based on an individual assessment related to a set of parameters:

- does the country have its own solution
- how well functioning is this solution
- the relative user-friendliness of the common solution

- the applicability of the common solution in the country in question (dependent on ia organisational set-ups and traditions for hardware and software use).

A development-strategy for STANDAT on this point has not yet been decided on, but there is an awareness in Denmark of the urgency of this question.

Ideas related to the code list system.

One obvious need of the present set of code lists is a thorough assessment, revision and updating of their contents. They are the part of STANDAT that least effort has been put into, the starting point was not flawless and the development process has at times been somewhat erratic.

A hazard in the way the code list system is handled in STANDAT is that the number of codes may become so large that it becomes difficult to maintain and use the code list system. This is both because a code once established[2] is never disposed of, and because the code list system does not have the possibility for distinguishing between general and specialized codes.

Take the case of an information type concerning "address". In a specific case of use ie referring to a specific subject in STANDAT it may be necessary to add a specification of the *sort* of address in question. Is it eg the address of a waste water treatment plant itself or is it the address of the contact person at the plant. In the first case the STANDAT secretariat will receive an application for an information type with the description "Address of waste water treatment plant" and in the second case "Address of contact person of the plant". There are many such examples of needs for a specific definition of an address in connection with subjects in STANDAT. All these information types could probably without any problem be defined the same way: as a string with the length of e.g. 80 characters.

A way of solving this problem would be to introduce a third field in the combination code list. This field should contain the "specialized" part of the information type description. E.g. referring to a general address information type and for one combination supplying the specification "location of plant" and for another combination supplying the specification "contact person".

A solution of this type could reduce the number of necessary information type definitions significantly, but of course it would also introduce other demands on e.g. the support programmes connected with STANDAT.

Another aspect concerning the code list system is the possibility of having a set of 'free-for-use' subjects, information types and value code lists. At present this facility does not exist in a formalised way in STANDAT. In practice it *is* possible to define your own local codes by using the SSP user support programme. But to transfer these local codes you have to make a specific agreement with the recipient of the data - otherwise her / his test programme is going to reject the contents of the file.

[2] The "out-of-date" marking of value codes was partly introduced to cope with this problem.

A possible general solution could be defining a specific range of code values to be free-for-use. The users of STANDAT have often put forward a wish for a facility of this type, also taking into account the procedure for getting new codes acknowledged in STANDAT. A problem would of course be the risk of an uncontrolled development of a sub-set of STANDAT codes that does not have official approvement of its form and contents.

Another possibility for improving STANDAT code lists concerns the structuring of subjects. At present the only type of possible ordering is hierarchical. I.e. subjects can only reflect a one-to-one or one-to-many relation between entities. In most cases this is sufficient, but of course network structures are also a reality in the wide span of environmental data. An example is a monitoring network composed of a set of monitoring stations each reporting data concerning different environmental issues.

This aspect of the structure of the STANDAT code list system has not yet been fully addressed, but a possible solution might include the introduction of key fields. An example is key data concerning identification of bore-holes and the related samples. A unique and unambiguous identification of these entities presupposes key data concerning geographical location, date, depth etc. If the transferred STANDAT file does not contain this information it will be impossible to use the data in eg a national database. At present identification and use of key data is a matter of agreement between sender and recipient of data. But it might as well be part of the code list system to identify the necessary set of key information types and make their use compulsory.

One feature that STANDAT does not take care of in a systematic way is the question of other geographical references than those related to geographical points. Examples are references to demarcated areas such as catchment areas and string areas such as rivers. This could be taken care of by having a new information-type with a new format, having not one number, but several, connected numbers.

Ideas related to the file format.

The possibility mentioned earlier in this chapter of transferring local or temporary value code lists would require a facility to define the relevant code list and to list the allowed value codes and their definition. A proper place to put this kind of specification would probably be in the DEFINITION part of the file format, also to ensure that the specified value domain could be recognized by a load system before testing the data part of STANDAT files.

Ideas related to edp support programmes.

One idea related to the support programmes would be a streamlining of STANDAT to Internet-use. This would encompass development of a specialized mail-server for the STANDAT-users, that could initiate a load-program, taking care of an automatised loading of data into the relevant database at the recipient end including semantic and syntactic control. The system should forward a return-message to the sender, notifying her if the transfer has been accepted or not, and if not, what the problem is.

Another technical development that would be relevant is a compressing-feature. Within complicated subject-areas, STANDAT-files can become very large, and thereby use much storage capacity - as well as taking longer time for network based transfers. A compressing-feature in the SSP programme would solve this problem.

When the DATA section of a STANDAT file expands beyond a certain size it becomes difficult to get an overview of the actual contents of the data to be transferred. To help the user the SSP programme should include the possibility for producing a view of data with translations of codes in a design close to that of a spreadsheet. The mere display of data in rows and columns would enhance the possibility for identifying diverging values. At present the report part of the SSP is very simple and it does not include this kind of "viewer-function".

The SSP and the STANDAT Load System have been developed separately. They have different hardware / software platforms and different functionality and interfaces. In the future it would be relevant to merge the two software packages into one, both to ensure total consistency in the testing procedure (taking into account both the syntactic and the semantical aspects), but also to minimize the effort needed for re-programming when introducing new facilities in the STANDAT format.

This strategy implies that the process of making an agreement on transferring a specific type of data via STANDAT includes elaboration of an exact description of the data to be transferred, including a specification of syntactic and semantic requirements. This description is to be edp-based and follow the set-up rules of such descriptions to be used as input to the common test module of a merged support-and-load program. In this way the producers of STANDAT files will immediately be able to carry out a test exactly matching the test procedure of the recipient and thus ensuring a significantly more error-proof transfer of data.

Ideas related to the organisational structure.

The one most important issue at the organisational level is promoting STANDAT and supplying information and education on its use. The SSP programme is an important element in this connection, but there still is a need for a better user guide, and for offers for the users for seminars and courses. A hot-line function and an offer for in-situ education by a STANDAT expert would be very relevant, but this would probably be too resource consuming to be feasible.

11. Scenarios for data transfer.

According to the Master Plan for EEA and EIONET[3] the EEA has been established with the aim of "provid(ing) objective, reliable and comparable information for those concerned with framing, implementing and further developing the European environmental policy and to ensure, that the public is properly informed about the state of the environment" (p. 1).

According to this definition the potential users of information from the EEA include the European Commission Directorates, the Council of Ministers, the European Parliament, other union bodies, national environmental authorities, international organisations, non-governmental organisations, representatives from sectors (such as industry, commerce and agriculture), the media and the general public (ibid, attachment 1, p. 3).

On the other hand the data that are going to provide the basis for this task are to be collected from a wide-spread network, comprising an increasing number of different nations with very heterogenous organisational set-ups in the area of environmental management.

The task of establishing an efficient structure for data collection and data flow in the EIONET is a matter of great importance - but also a task of great complexity and a task where the needs are not yet totally defined.

It is not the aim of this report to discuss the data needs of the various levels in the EIONET. But it should nevertheless be emphasized that an overall discussion of and decision making on this subject is of crucial importance if the EIONET is not to be dominated by ad hoc solutions producing inconsistent, redundant and useless information. The solution must of course be developed continuously to correspond to upcoming, new demands.

At present many aspects of the data flows in the EIONET are still uncertain. European Topic Centres on several areas still need to be set up, and it is not finally decided how many data the Agency itself is going to have in-house, and at what level of aggregation.

In making recommendations for the data transfer it will therefore be useful to operate with some different scenarios for the way of exchanging data.

Finally it should again be stressed that the recommendations are based on the experience of the STANDAT system, not on a generalized discussion of different ways of transferring edp-based information.

Differences between the Danish system for collecting environ mental information and the EIONET set-up.

To use the experience from the STANDAT system in a constructive way, it is necessary to clarify the differences between the Danish system for data transfer, and the Agency EIONET set-up.

[3] EIONET - Master plan for EEA and EIONET, april 1995.

The differences between the two set-ups can be narrowed down to three points: size/magnitude, complexity and mandate regarding requisition of data.

Some of the points will be made clear by a comparison of a model of the EIONET system and the Danish system (for the latter please refer to chapter 6, figure 6.1).

The EIONET organisation is complicated because it is not only defined/divided by subjects (water, air pollution etc, taken care of by the European Topic Centres) but also by nations (and their National Reference Centres, National Focal Points and national networks).

Figure 11.1 demonstrates the complexity of the EIONET system. In Denmark the principal component elements of the system are the Ministry, the National Topic Centres and the counties and municipalities. In the EIONET system there is the Agency itself, the European Topic Centres, the National Focal Points as well as the National Reference Centres and *their* individual national networks. It should be noted, that the national networks can be set up in different ways with different levels of centralisation. These national differences add to the complexity of the system.

The differences in size are obvious: not only is the system more complicated and has more layers, it also comprises not only one country with fourteen counties and about 270 municipalities, it comprises all the EU countries with their individual regions and local levels as well as several other European countries. Even more so, the members/users can be expected to grow in number as the EU accepts new members and as more non-EU countries apply for member-ship of the Agency network, most notably the Eastern European countries.

Important aspects to take into account regarding this very heterogenous network are also matters of confidentiality and ownership of data, matters that are likely to be handled in different ways by the different member-states.

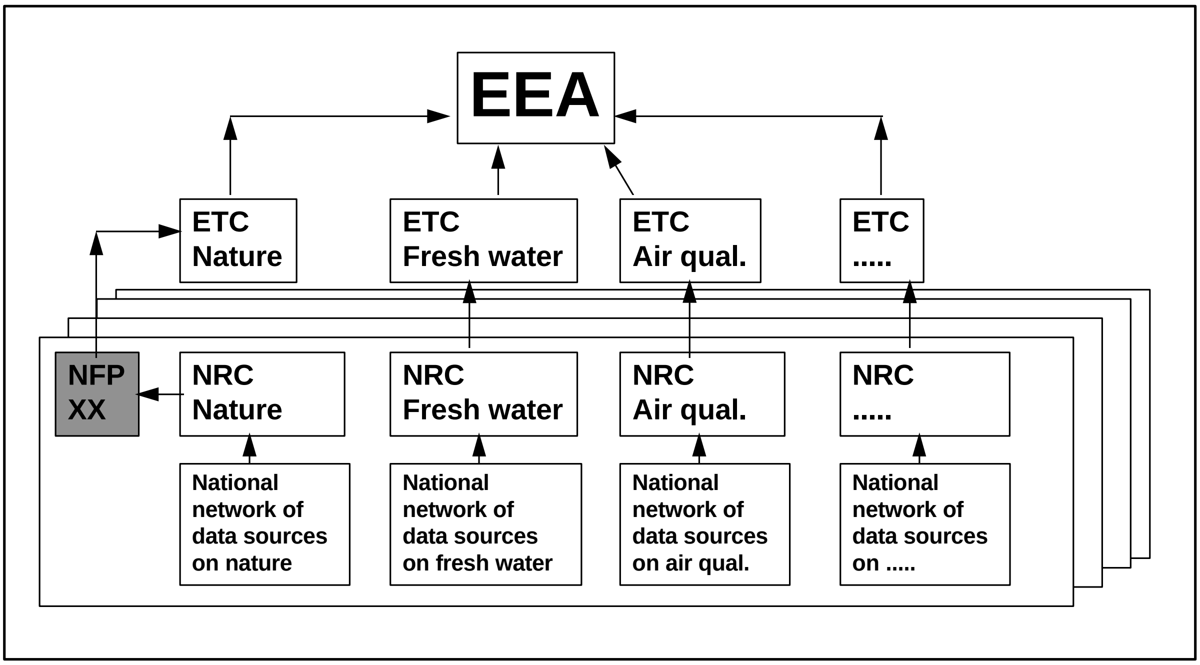
EEA
ETC Nature
ETC Fresh water
ETC Air qual.
ETC .....
NFP XX
NRC Nature
NRC Fresh water
NRC Air qual.
NRC .....
National network of data sources on nature
National network of data sources on fresh water
National network of data sources on air qual.
National network of data sources on .....

Figure 11.1: The main components of the EIONET.

As for the mandate concerning data collection, the Danish Ministry of Environment and Energy has two mechanisms at its disposal: the Ministry can specify requirements for the form and content of a specific data set in the legislation relating to the subject area in question. It should be noted that in Denmark such legislation is typically carried through after hearing of the parties concerned. Furthermore the Ministry has in special cases the option of negotiating compensation for the most important data collectors.

On the EU level directives are the only mechanism similar to the legislative tool of the Danish Ministry. This is not a tool directly available to the EEA itself, as directives are the responsibility of the EU administration in Brussels. There is a memorandum of understanding between the Agency and the National Focal Point of each member state on procedures for information flow. Here it is stated that 'Member Country X will actively participate in the realisation of the EEA Workprogramme, specifically to meet the information requirements emerging from the EEA Workprogramme' (article 3). Questions of comparability and joint information strategy is also mentioned, but not in very definite terms, and therefore not in a form that is particularly operational when obtaining data.

It is only to a limited degree possible for the EEA to subsidize data collection and transfer. The transfer of data to the European Topic Centres and the EEA may in this way to some extent be a question of goodwill seen from the angle of the potential data suppliers. And it must be foreseen that the different countries may have differing views regarding this matter.

The three points described above have to be taken into account when envisaging scenarios for the data transfer processes in the EIONET based on the STANDAT experience. The differences in magnitude, complexity and mandate make it more difficult for the EEA and the European Topic Centres to define, require and collect data in a standardized way when compared to the Danish Ministry of Environment and Energy.

Nevertheless, at the conceptual level the crucial questions are alike in the two network systems, and the experience gained from the smaller system will therefore still be useful also at a larger scale.

The rôle of the Agency and the European Topic Centres.

At present it is not decided how or at what level of aggregation the Agency itself will have environmental data. As far as the Agency is going to have data from the other levels of the network, it could be in the form of copies of data bases (or parts of data bases) from the European Topic centres. The need for standardisation of the practical data flow at this level would in this case mainly be related to the choice of data base tools and set up and organisation of databases.

This kind of solution does on the other hand not exclude or reduce the need for a common data model and codification across the various subject areas of the European Topic Centres. That is, if the Agency is interested in having the possibility for combining data, e.g. to calculate the total environmental pressure subdivided on different substances, across a division in societal sectors and environmental recipients.

Furthermore the fact that the EEA could get its data as copies of databases would not solve the problem of data-transfer for those responsible for establishing the necessary databases at the level of the European Topic Centres. Large amounts of data would still have to be transferred at the level below the Agency itself.

It should be noted that different subject areas may have different scenarios, so that the level of data transfer is different in eg the areas of data on air and coastal waters. One important factor deciding the most adequate solution for any area is the amount and expected frequency of data to be transferred and the needs for quick modifications in the scope and contents of data.

The scenarios.

The set-up of the scenarios has been based on the fact that the way of doing things can be built on different kinds of common solutions. These common solutions can be either related to software (and hardware) or they can be related to conceptual frameworks, data models and codes and to different ways of utilising network- and data-share-technology.

All scenarios have their strengths and weaknesses, and some of them are certainly more adequate than others seen from a top-down point of view. Based on the experience of the STANDAT concept, at least one of the scenarios is hardly recommendable, as shall be discussed (cf scenario 5). The idea is to present some different models from a range of possible solutions. The end-solution may very well be a combination of different scenarios.

Scenario I: The centralised model / standardised hardware and software.

In the pure form of the centralised model for data transfer, all standardisation is related to choice of hardware and software. The central recipient of data provides all other participants in the network with the software necessary for storing and retrieving the relevant data - and if necessary with the required hardware. The relevant software is a registration system with a predefined database, including an output facility that produces exactly the required data-file with the relevant format and codifications of data.

As far as the authors of this report are aware, this scenario has been partially used in the Finnish set-up for the collection of environmental information.

This model would be most relevant in cases where no great flexibility is needed - where the data collected and exchanged are the same over a longer period. And in cases where it is feasible and possible to require that all participants use the same (hardware and) software and where there are resources at the central level to provide the necessary (hardware and) software.

An estimation of the resources needed compared to the present Danish situation is that it would require more resources at the central level for software development - up to 3 or 4 times more resources. At the other levels less resources would be required, most work would have to be put into implementing the software solution in the local computer systems.

Scenario II: The decentralised model / standardised format (and code lists).

The use of the STANDAT-concept in Denmark is one example, and the SANDRE-format mentioned in chapter 9 another example of this set up. The model is based on the assumption that the partners in data-exchange choose their own hardware and software solutions and base the exchange of information on a common data model and file format, and most often on common code lists.

As outlined in this report (cf chapter 9) there are different ways of implementing such a model, and they all have their different advantages.

This model is more flexible than the first one and it is therefore one that would be adequate in most cases where large amounts of data have to be exchanged, where flexibility is needed, and where importance is attached to the possibility of combining and sharing data in all possible ways across subject matters and areas of competence (between European Topic Centres).

On the other hand it requires that a central institution has both the ability and the agreement from the other network partners to decide on the data model, code lists etc. to be used. And it requires that there are resources at a central level to maintain these elements.

The resources needed for one country are in the same magnitude as the resources used in Denmark for the development and implementation of STANDAT. It is not easy to estimate the resources needed for an EIONET solution based on this scenario. Although there are more participants in the exchange of data, the subject areas are not all that different from the ones in the STANDAT code lists today, and the file format is the same whether it has a hundred or a thousand users. At a rough estimate the resources needed for development of the concept would be three times the ones used in Denmark (because more work would have to be put into the development of code lists), whereas the resources needed for implementation would be larger because of the larger number of users.

Scenario III: The open model / flat files / flat files and common code lists.

In the open model there is no common file-format, but data are exchanged in the form of simple, ordinary files. The exact structure and content of the file has to be agreed upon from case to case by the partners in the data-exchange.

A version of this model has code-lists that are common for at least the most important parameters etc. In this way some possibility is open for putting together part of the data collected on the different subject areas.

This model has obvious weaknesses in its lack of universality. Much effort has to be put into making specific agreements between the sender and the recipient in each case of data transfer. The model is most relevant in cases where few data are to be exchanged and where it would therefore be overkill to define common file formats etc. On the other hand at least it provides the possibility for codifying similar data elements in a uniform way.

The resources needed in the initial stages are far less with this model. The problems and the needs for resources arises at a later stage, when data are to be combined and compared and no common formats (and codes) exist.

Scenario IV: The all-data-are-shared-data model / network based model.

This model is based on the use of network technology and data-share tools and to work properly it should include elements from scenario II or the whole scenario.

With present-day network technology it is possible to store data in a part of eg the Internet with public access. It is also possible (though not without complications) to give different access-rights to different users, so that the relevant persons get the rights to up-date the central database and retrieve information from it, while other users have read-only access.

The advantage of this solution is that you can give public access to data at the same time as making it possible for all partners in a data-collection network to store and retrieve data at one central data-base. A solution of this kind requires an agreement on organizational set up to ensure that the state of the database regarding updating is unambiguous. Another central point is that for this scenario to work it is still necessary to have a common codification and a common format for the data files to be down-loaded into the database.

Therefore the use of network-technology is relevant in cases where there is a wish for common and equal access to data, and where you want to open for public access to data in the most direct way.

At a rough estimate, the resources needed are in the same magnitude as for scenario 2.

Scenario V: The ad-hoc-model.

The last model - and together with scenario III the least feasible at least from a top-down point of view - is the ad hoc model, where there are no standards and no common concepts whatsoever. This model can be seen as an extreme form of scenario III where the participants apply any kind of solution that they like, often the solution that is easiest to implement in the short run and the solution they happen to have used or met before.

The obvious problem is the lack of coherence in codification and data formats that makes it difficult to put data together and get an integrated assessment of the state of the environment and of the pressures on it.

The resource estimate would be the same as for scenario 3.

In the opinion of the authors of this report this model is not recommendable in any cases, cf below. For obvious reasons it is not a model that can be theorised about, but it is a way of doing things that will easily develop by itself in cases where no attempts are made at some kind of central management at an early stage, or where such attempts fail.

Lastly it should be made clear that both from a practical and from a data scientific view it is of the outmost importance, that data work and exchange of environmental data of common interest and importance in the Agency network should be based on a common data model, and on a common set of codes for all generally used types of information.

Without these prerequisites it will in the end prove extremely difficult to establish any kind of connection between data across the different subject areas and the different European Focal Points.

Starting points.

The precise extent and level of coordination is of course a question which has to be decided on by the Agency and the participants of the EIONET. Theoretically speaking a choice of starting point concerning exchange of environmental data has to be located in a continuum with at least three dimensions, viz data models, code lists and exchange format:

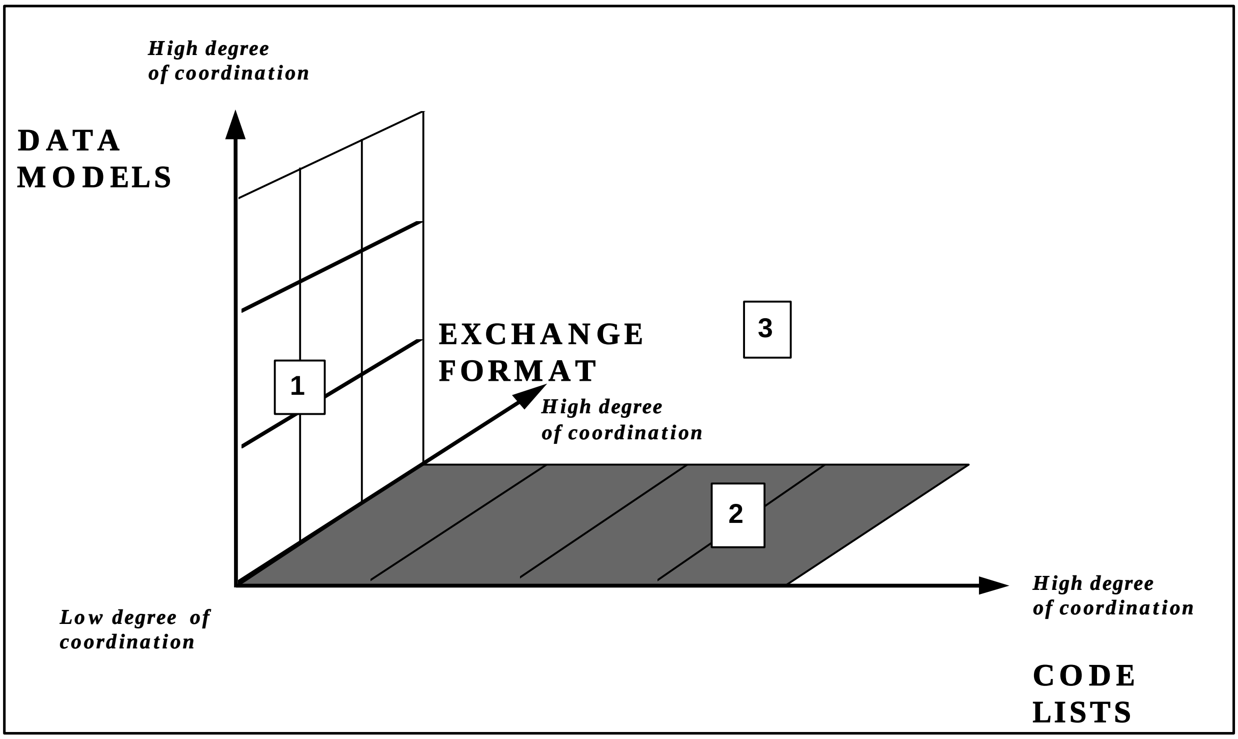
High degree of coordination
DATA MODELS
EXCHANGE FORMAT
3
1
High degree of coordination
2
High degree of coordination
Low degree of coordination
CODE LISTS

Figure 11.2: The three dimensions for choosing data transfer solution.

In figure 11.1 three imaginary solutions are placed in the continuum. Solution 1 has no coordination on code lists, but some degree of coordination on data models and exchange format. Solution 2 has a relatively high degree of coordination on exchange format and code lists but none on data models, and solution 3 has all three dimensions.

A fourth parameter concerns the software and hardware used in connection with registration etc of the data in question (cf scenario 1 above).

As mentioned before the starting point in this conceptual continuum may differ between the various topic centres / environmental themes, but it is nevertheless important to agree on a common denominator, at least for data models and code lists to ensure a minimum of coherence in core data.

When developing and implementing a common solution for data transfer, it can be done gradually or in one move. Either way, the basic steps of the process are the following. It should be noted that this process of analysis and decisions is not a simple step-by-step procedure, but a series of questions to be answered / a series of actions to be taken which are part of an iterative process.

1. Analyze output requirements	Identify 'the questions to be answered' - what data are needed to produce the relevant set of reports, maps etc.
2. Decide on level of ambition	On the basis of an estimation of the available data sources, resources for data tasks etc define an appropriate level of standardisation
3. Decide on organisation for the task at hand	It is important to have an organisational set-up with a clear division of labour, with one organisation that has the overall responsibility, and with the necessary resources
4. Define a common data model	According to the 'view of the world' of the data collector(s) develop an appropriate data model describing the component elements and their coherence
5. Decide on common code list and develop them	Identify a common set of code lists taking into account the current and future needs for comparison and combination of data
6. Decide on exchange format	Taking into account the requirements listed in chapter 12.
7. Decide on computer based support programmes and design of these	Taking into account the requirements listed in chapter 12.
8. Implementation	Implementation of code lists in existing systems Implementation of in-put and out-put facilities for data format in existing systems Distribution of edp-based support programmes

Seminars, guidelines, hot-line facility

An ambitious solution would include all these steps, while a less ambitious solution would not involve steps 6 and / or 7, and parts of step 8. A stepwise solution would start with the development and gradual implementation of steps 1 to 5 (and the relevant parts of step 8), introducing step 6 and 7 at a later stage. A one-step-solution would entail a need to start work on steps 1 - 7 simultaneously, and putting extra care into step 8.

12. Conclusions and overall recommendations.

This chapter presents the overall conclusions and recommendations of this report, based on the experiences of developing and implementing a common, standardised format for exchange of environmental information in Denmark. The chapter should be seen in close connection with the previous chapter (11), where different scenarios for data-transfer in the EIONET were discussed and different solutions for data transfer in the different scenarios were presented.

The main point was that solutions can be based on different kinds and degrees of coordination (either related to software and hardware or related to formats and codes) and different ways of utilising network and data-share technology. And that more than one solution can be used in the different parts of the EEA network. But that a central premise for an optimal use of the data compiled is that the exchange of data of common interest in the Agency network should be based on a common data model and a common set of codes.

1. Common, global solutions are preferable.

By having a common, global solution for the whole EIONET and for all subject areas, it is easier to ensure connections between data across the different subject areas and topic centres, and you have a more adequate use of resources as only one solution has to be implemented, and information, education and edp-based support programmes are the same for the whole system.

No matter what is concluded on this question it is of the utmost importance to have a common definition of basic data, how they are connected and how they are to be codified. Furthermore, it is an urgent task to decide on these matters as the different European topic centres are pushing on with their work and may thereby come up with different, local solutions both on ways of exchanging data and on definitions and codifications of basic data.

2. Elements / experience from existing environmental data exchange concepts should be utilised in the Agency's development of a common solution.

This report is an attempt to help this process of extracting experience. From the other formats / code lists described in chapter 9 it can eg be deduced that these elements are important:

- generality and flexibility in the descriptive, data- structuring part of the exchange format
- identification and exhaustive description of relevant types of information
- scientifically correct and unambiguous codification of the allowed values regarding a specific set of information types.

3. When developing a global format for exchange of environmental information for use in the EIONET, some important requirements are:

- the format should be simple, easy to understand and use
- the system should secure an optimal use of resources
- the system should secure unambiguity in the form and content of the data transferred.
- the system should ensure that exchange of environmental information can be independent of hardware and software solutions
- 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

4. *Solutions should use - or at least be based on - suitable, existing code lists.*

In many subject areas international nomenclatures or code lists exist already. Such code lists - possibly with some adaption - should be used as much as possible.

The main problem in this context is to find the right code lists / nomenclatures. In some areas de facto standards have been set already, but in (many) other cases this is not so. In these cases it is important to find the balance between scientific requirements and user friendliness and to have a fairly pragmatic approach. Code lists should be exhaustive and correct, but the way of looking at environmental problems changes just as scientific nomenclatures can change. Therefore code lists should also be set up in a way that does not make it difficult to make additions, and does not require difficult changes in computer programs etc. when additions are made.

An important point in this context is that codes should not carry information in themselves (eg. a description of a hierarchical ordering) or subsidiarily they should do it in a way that is not a hindrance for further development.

The organisational set-up is important here, as it should be perfectly clear to all partners how to apply for new codes, and perfectly clear who has the responsibility for approving new codes or code lists and who has the responsibility for their implementation / distribution.

Examples of existing code lists that may be relevant are the NCC code lists, and possibly a subset of the value code lists connected with SANDRE and STANDAT. At present the NCC code lists are in danger of lapsing because the Nordic Council is no longer subsidizing the maintenance and further development of this code list system. Continuing this work may be a way for the EIONET to ensure an important contribution to a pool of common code lists.

In general it would be a relevant task for a working group on data exchange, CDS and codification to look into the existing code lists in detail and make recommendations for use of and / or changes in such code lists. The thesaurus part of the CDS may as a starting point describe environmental subjects at a high level of abstraction. In the course of time a more detailed level in the form of code lists concerning specific subject areas (micro thesauri) might prove to be both necessary and appropriate. Considerations concerning this matter is an urgent task cf recommendation 11.

5. *When developing / deciding on a set of common code lists, some important requirements are:*

- the set of code lists should be the same for all areas / all European Topic Centres. Only in this way is it possible to make sure that data can be used across subject areas

- the code lists should be up to the best scientific standards while yet ...

- ... being pragmatic in their set-up

- the codes should not in themselves carry signification as this makes the code lists less flexible and more difficult to maintain

- the code lists should be easy to develop and an organisational structure should be set up to make sure that the development is based on user requirements.

6. *The development of user friendly high-quality edp-based support programmes is a necessity when introducing a transfer format.*

No matter how simple and user-friendly, edp-based formats and large collections of code lists are not easy to handle for any user or any organisation. Therefore user support software is extremely important. The software should (possibly in different software packages for different user groups):

- ensure overview of code list system and search facilities
- have a user-friendly interface
- entail no special hardware and software requirements
- have facilities for loading the code lists and new versions of them
- offer a complete syntactic test of the relevant files
- supply easily understandable error and warning messages
- supply functions for converting a file from one code-page to another
- have facilities for generation of simple tabular reports on files.

For a load programme also:

- be able to perform a complete "semantic" check of any set of files on the basis of a formal specification
- 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 file into the relevant (parts of a) database.

7. *Information, education, seminars, guide books and hot line services are extremely important when introducing a concept for data transfer.*

Furthermore, the resources needed for these tasks should not be underestimated.

8. *The EEA has in its EIONET a suitable organisational set-up for the development and implementation of a common exchange format for the whole network.*

EIONET binds together all the potential users of a standardised solution for the exchange of Agency-related data. But it is important to hold on to the point, that *one* and only one organisation in the network should be made responsible for the development and implementation of the relevant solution.

The set-up of the EIONET paves the way for a decentralised organisational set-up. While still having one organisation that has the overall responsibility for development and implementation, also supplying a network for all the relevant users and thereby giving them the opportunity to influence the development of the format.

9. *The questions of need for resources is important to take into account.*

By having a common exchange format there is no doubt that resources can be used in a more economic way in the long run and it is ensured that there is possibilities for putting together data in the system across subject areas etc. But it should be kept in mind that this kind of solution still leads to requirements for resources in other functions - in developing and implementing the format, in administering and developing code lists, in supplying information and education and in coordinating the effort of all the users.

The need for resources is far largest in the initial stages.

10. *Pilot projects should be applied to test recommendations and possible solutions.*

When the outlines for a solution has been decided, it is important to test it and to make the necessary adjustments based on the experience of the test. If not, you risk having a solution that works in theory, but is difficult to handle in practice for the users.

11. *It is important to set up a working groupto make recommendations on the exact solution to apply.*

This working group should include representatives for *all* member-countries and it should work in close cooperation with the relevant persons in the Agency itself.

Some of the important tasks are:

- ensure consistence between global CDS and code lists etc
- develop format for data exchange
- use of / choice of code lists.

Again, this work is urgent, as the results are needed if local solutions are not to be applied for the different topic areas.

12. The global format should respect the individual national solutions that exist already.

As well as utilising the experience from the different national formats, a common, global EEA-format should respect the continuing existence of such national solutions. The member-countries can adopt one of two solutions:

- they can either gradually change their national, 'internal' format into the global one, or

- they can commit themselves / the organisations responsible for delivering data to the Agency network to supply the relevant translation facilities necessary for delivering data according to EEA / EIONET-requirements.

13. Executive summary.

This report is part of a package of projects financed by the Danish Government for the support of the European Environment Agency. The aim of the project is to utilise the experiences from the use of the Danish STANDAT system for exchange of environmental edp-based data.

Main principles of STANDAT.

STANDAT is a standardised data exchange format, developed in Denmark in the late 1980'ies to facilitate the exchange of large amounts of environmental information. The STANDAT concept has four main component elements: the code list system, the file format, the organisational set-up for the administration and development of STANDAT, and the edp-based support programmes.

STANDAT is a dynamic system in being under constant development as to the contents of the code lists. This development is user-driven via the organizational set-up.

STANDAT has a relatively simple and pragmatic set-up and is relatively easy to understand and use.

STANDAT ensures unambiguity in form and content of the data transferred, and ensures independence of hardware and software solutions between the different users.

Code lists.

There are four different kinds of code lists that together form the code list system.

The subject code list defines the subjects on which data can be transferred and supplies a code for each subject. The subject code list is hierarchical (for an example please refer to figure 3.2). This fact is related to the way that the file format is structured (see below).

The information code list defines what information can be exchanged on each subject and supplies the relevant codes (for an example please refer to figure 3.3).

The combination code list defines the possible connections between the subjects and the information types. This makes it possible to have a relatively small set of information types, as an information type (eg measurement method) can be associated to more than one subject.

Finally the value code lists supplies the predefined values for some of the information types (other information types are numbers, text strings or date-information).

Together the code lists define a 'view of the world' with regards to structure, content and connections between the different pieces of information on the environment.

File format.

There are three component elements of the file format:

- the HEADER section that contains administrative information on sender and recipient etc.

- the DEFINITION section that defines what data are to be transferred and how they are to be structured. This section is the key to interpreting the DATA section.

- the DATA section supplies the relevant information as specified in the DEFINITION section. The different subjects can be embedded in one another so that you can refer to the same parent-information for several subsets of data.

Edp-based support programmes.

STANDAT has two kinds of related edp-based support programmes:

The STANDAT Service Programme (SSP) that was developed for the support of the producers of STANDAT files. This programme provides an overview of the code list system and has facilities for loading new code list versions as well as search facilities. Even more important, it offers complete syntactic test features for STANDAT files with warning and error messages and it can generate simple tabular reports on STANDAT files.

The STANDAT load system is used in parts of the Danish Ministry of Environment and Energy and it uses a generalized specification of semantic requirements that can be used with very few specifications for any relevant file transfer. Files are controlled before they are loaded into the relevant databases.

Transferring information via STANDAT.

When you want to have data delivered in the STANDAT format you provide the suppliers of data with a general description of the data to be transferred, an exact description of the STANDAT file with examples, exact description of KEY data, description of value codes to be used and specification of the time for delivery.

If needed, new codes and code lists can be established via the STANDAT secretariat. New codes and code lists have to be assessed by the national Danish data topic centres.

When the supplier of data has retrieved the relevant data from her / his database it should be tested via the SSP, STANDAT Support Programme. Data are then transferred to the recipient on diskette or via network.

The recipient should make a final check of the file before down-loading it into his / her database. Here the STANDAT load programme is used for data delivered to the Danish EPA.

Organisational set-up.

The organisational set-up uses the organisation for collecting data on the environment in Denmark. This comprises a set of national data topic centres, that are some of the most important users of the STANDAT format. In the administration of STANDAT the topic centres are responsible for assessing requests for new codes and code lists in STANDAT.

The whole administration is coordinated by the secretariat placed in the Danish EPA. There is a steering committee with representatives for all the main user groups, eg counties, municipalities, Kommunedata and the topic centres. Kommunedata is responsible for the technical part of the updating of the code lists.

Scenarios for data transfer.

The conclusions of this report has ao been based on a set of scenarios for the process of data transfer within the EEA network. It is quite feasible that more than one solution will be necessary, as different solutions may be necessary for the different areas of work of the EEA. The scenarios envisaged in this report are:

Scenario I: *The centralised model / standardised hardware and software.*

Scenario II: *The decentralised model / standardised format (and code lists).*

Scenario III: *The open model / flat files / flat files and common code lists.*

Scenario IV: *The all-data-are-shared-data model / network based model.*

Scenario V: *The ad-hoc-model.*

Conclusions and recommendations.

Based on the experience of developing and using the STANDAT system and based on other points from this report some of the conclusions and recommendations are:

* *Common, global solutions are preferable.*

* *Elements / experience from existing environmental data exchange concepts should be utilised in the Agency's development of a common solution.*

* *A set of requirements for the development of an EIONET exchange format.*

* *Solutions should use - or at least be based on - suitable, existing code lists.*

* *A set of requirements for developing / deciding on a set of common code lists*

* *The development of user friendly high-quality edp-based support programmes is a necessity when introducing a transfer format.*

* *Information, education, seminars, guide books and hot line services are important when introducing a concept for data transfer.*

* *The EEA has in its EIONET a suitable organisational set-up for the development and implementation of a common exchange format for the whole network.*

* *The questions of need for resources is important to take into account.*

* *Pilot projects to test recommendations and possible solutions are important.*

* *It is important to set up a working group to make recommendations on the exact solution to apply.*

* *The global format should respect the individual national solutions that exist already.*

ANNEX I: Acronyms etc.

ASCII	American Standard Code for Information Interchange
CDS	Catalogue of data sources
CEN	Comité Européen de Normalisation
Danish EPA	Danish Environmental Protection Agency
EDIFACT	United Nations Electronic Data Interchange Administration for Commerce and Trade
EEA	European Environment Agency
EIONET	The network connected to the EEA for the collection of environmental data
EPA	See: Danish EPA
ETC	European Topic Centre
EU	European Union
EUROSTAT	The Statistical Office of the European Union
GESMES	Generic Statistical Message, the Eurostat format for exchange of statistical information (cf. chapter 9)
GEUS	Geological Survey of Denmark and Greenland
ID	Identification, ID-number is identification number
IOW	Information Office for Water (in France)
ISO	International Organisation for Standardisation
Kommunedata	The IT-centre and software house of the Danish municipalities and counties
NCC	Nordic Code Centre
NFP	National Focal Point
NRC	National Reference Centre
OECD	Organisation for Economic Cooperation and Development
PARCOM	Paris Commission (prevention of marine pollution from land-based sources)
Rubin	Routine for Biological Information
SANDRE	Secretariat d'Adminsitration National des Donées Relatives à l'Eau - SANDRE is the acronym for the French data echange format for water related information (cf chapter 9)
SQL	Structured Query Language, an Edp tool
SSP	STANDAT Service Programme (cf chapter 5)
STANDAT	Format for STANdardised DATa exchange
UNTDID	United Nations Trade Data Interchange Dictionary
UTM	Universal Trans Mercator, map projection.

ANNEX II: References and litterature.

Format d'échange SANDRE des données - example d'utilisation. SANDRE, Limoges 1995.

GESMES 93 Guidance to Users. Eurostat, Luxembourg 1993.

GESMES - the International Standard for the Exchange of Array Data. Eurostat, Luxembourg 1995.

GESMES/ECOSER User Guide. Eurostat, Luxembourg 1995.

MD6 Annual Report. Eurostat, Luxembourg 1994.

NCC Coding System. The Nordic Code Centre, Copenhagen 1990, Ulla Pinborg and Thorbjørn Paule.

SANDRE. National Secretariat on Water Related Data, International Office for Water, Limoges, (no date).

SANDRE - The Reference Format of Data about Water. National Secretariat of Water Related Data, Limoges (no data).

STANDAT v.1.1. Danish EPA 1994, Sten Åbo et al.

Important sources for information on other formats were talks with

GESMES: Philippe Lebaube, Olli Janhunen, Chris Nelson and John Allen, Eurostat in Luxembourg November 15 1995.

SANDRE: Vincent Blanc, Office International de l'Eau in Paris December 5 1995.

ANNEX III: Example of a SANDRE file.

ANNEX IV: Example of a GESMES file.