Information Management Resource Kit

Module on Management of Electronic Documents

UNIT 5. DATABASE MANAGEMENT SYSTEMS

LESSON 6. TEXTUAL DATABASES AND CDS/ISIS BASICS

NOTE

Please note that this PDF version does not have the interactive features offered through the IMARK courseware such as exercises with feedback, pop-ups, animations etc.

We recommend that you take the lesson using the interactive courseware environment, and use the PDF version for printing the lesson and to use as a reference after you have completed the course.

Objectives

At the end of this lesson, you will:

- understand the functionalities offered by CDS/ISIS, a textual DBMS;
- understand the technical work needed by developers to implement these functionalities;
- understand when you should use CDS/ISIS.

Introduction

Imagine you need a system to store, retrieve and disseminate data describing textual resources such as books, projects, papers, etc.
In this case, textual databases containing bibliographies, webliographies, project descriptions, etc., can match your needs.

CDS/ISIS (Computerised Documentation Systems/Integrated Set of Information Systems), is a **textual database management system** designed to build and manage textual databases.

What does CDS/ISIS offer?

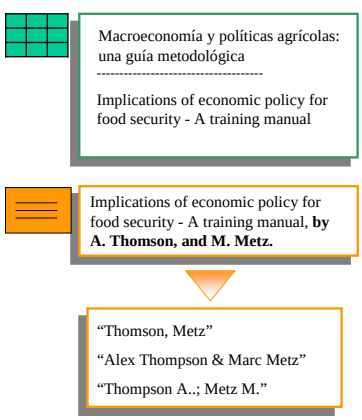
CDS/ISIS was designed in order to provide some important functionalities for document management.

There are many different versions of CDS/ISIS, which share following common features:

- Handling the structure of textual databases
- Text-oriented formatting
- Fast and powerful retrieval
- Handling different languages and scripts

Let's review together the importance that these functionalities have for the user…

What does CDS/ISIS offer?



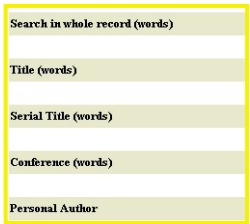
Textual databases come with…

1) Elements with a highly variable length (like titles or abstracts).
2) Elements that come with an unknown number of occurrences (like authors).
3) Groups of elements that should be processed as a group (like author's initials and author's surnames);
for example, you may want to render author's names in different ways, like "*Renard, Guyon"*, or *"Claude Renard & Jean Guyon*", or "*Renard C.*; *Guyon J."*, etc.

CDS/ISIS satisfies these needs, as…

1) It does not reserve a fixed length for fields or records, although there is a maximum.
2) It allows a field to be defined as repeatable.
3) If the names have been stored appropriately, it can render them in different ways.

What does CDS/ISIS offer?



How can users search data with CDS/ISIS?

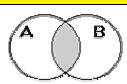
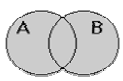
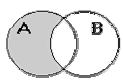
Normally you search all data that has been indexed, but with CDS/ISIS searches can be restricted to certain fields: for example, users can search the titles, the author's names, the keywords, etc.

Users also can truncate to search for words with a stem.

This technique allows a search on **leading sequences of characters**. CDS/ISIS will automatically include all search terms having the specified root. Right-truncation is indicated by placing a **dollar sign ($)** immediately after the last root character.

What does CDS/ISIS offer?

Also, users can combine terms using ISIS "logical" or " boolean" operators.
The three most important ones are:

Operator	Action	ISIS Syntax	Example
AND	 Intersection	*	a query *goats* * *sheep* retrieves records where both *goats* and *sheep* occurs
OR	 Addition	+	a query *goats* + *sheep* retrieves records where either *goats* or *sheep* occur, or both
NOT	 Exclusion	^	a query *goats* ^ *sheep* retrieves all records where *goats* occurs, unless *sheep* occurs in the same record

Note: Be careful with the "NOT" operator. You would exclude works that are both on goats and sheep, thus miss useful information on goats.

Boolean operators

"I'm looking for documents on fish diseases".

What is the best expression for this search?

- ○ fish * diseases
- ○ fish + diseases
- ○ fish ^ diseases

Click on your answer

What does CDS/ISIS offer?

The ways of searching also depend on the database design, so these are defined by the database developers.

For example, for some fields the developer may have decided that each word is a separate entry. To search for adjacent words like compound keywords user can then use adjacency operators.

→ **PLANT . BREEDING** searches for the two words next to each other

PLANT .. BREEDING there may be one word in between

However, the database designer may have chosen that only those phrases in a certain field will be indexed that are between slashes or between < > (square brackets).

→ If such a field contains **<Plant breeding>** the record can be found by searching **PLANT BREEDING**.

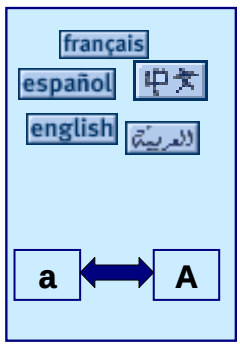
More sophisticated things are possible.

→ The database can be designed in such a way that searches can be restricted to certain fields by using prefixes, like AU=PLATO or TI=Dialogues

What does CDS/ISIS offer?

¡pªG¡³¿ ®Ñ©

Finally, another important functionality is the ability to handle **different languages and scripts**. In fact, you need to be aware about character encoding, especially with non-Latin scripts.



A database management system should not just display the characters correctly, but also be aware of the sequence of these characters in a script, especially when it sorts data and builds indexes.

It should also understand which upper case character corresponds with which lowercase character.

ISIS has solved this by using two tables:
- ISISUC.TAB, that defines the correspondence of upper case and lower case, and
- ISISAC.TAB, that defines the alphabetic characters and their sequence.

Even advanced developers of ISIS applications will seldom use these features, but it is useful to know that CDS/ISIS can be adapted.

What does CDS/ISIS offer?

We think CDS/ISIS is a good solution for us, but some features should be adapted to better match our needs… It is possible to do this?

Developers can adapt CDS/ISIS depending on the required features.

They can personalize the system in order to match your organization's needs.

But not all the adaptations can be made, and not all involve the same amount of work.

To better understand these capabilities and the work required, let's design a CDS/ISIS database.

Designing a CDS/ISIS database

Developers have to create a series of files in order to design and build a **CDS/ISIS Database**.

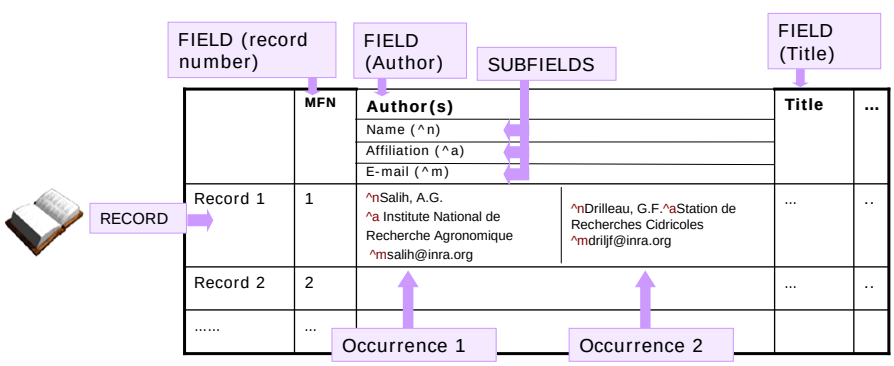
Developers must define:	To do so they create following files:	With following extension:
Which kind of fields there are.	Field definition table	.fdt
How to display the data.	Display formats	.pft (written in the formatting language)
How to search the data.	Field select table	.fst (also used to print sorted output)
How to input data.	Worksheets or web forms	.fmt (needed in a stand-alone Application, not in a web environment)

Let's have a look at them…

Defining fields

CDS/ISIS databases are organised collections of records each of them describing a resource (book, paper, project etc…).

Records contain different data elements: fields and subfields, which represent attributes of the described resource, such as title, author, abstract etc…



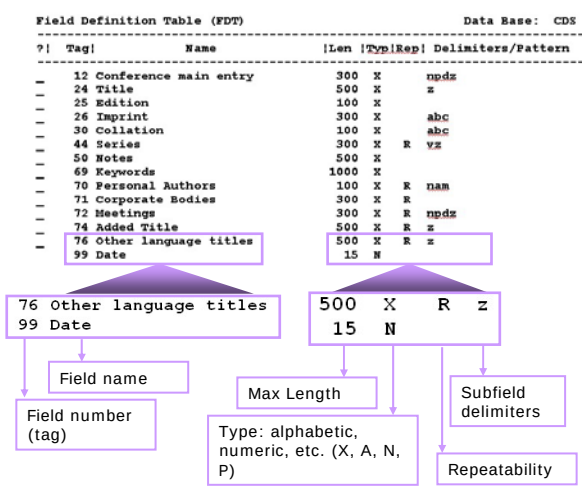
Defining fields

The fields and subfields may have variable length, and each of them may have any number of occurrences.

In this example, you have a **repeatable field** (Author) with **subfields** (name, affiliation, e-mail) **for each occurrence**. Subfields are delimited with subfield delimiter (^).

CDS/ISIS can have a maximum of two levels of data hierarchy (father-child) within a record (fields and subfields).

Defining fields



Fields can be defined in different ways depending on the kind of resources and on how you want to use the database.

Developers create the **Field Definition Table** which describes:

- the **record structure** (e.g. Title, Date, Authors, etc.), and
- the **characteristics** (maximum length, subfields, etc.) of fields and subfields.

Defining fields

MFN: 2 ← Record number
44: Methodology of plant eco-physiology
50: Incl. bibl.
69: Paper on: plant evapotranspiration
26: ^c1965
70: ^nBosian, G. ^mBosian@yahoo.com
70: ^nSmith, J.

For example, this bibliographic record follows a specific predefined structure.

Can you classify the following elements?

	Field number	Subfield delimiter	Data (occurrence 1)	Data (occurrence 2)
^n	○	○	○	○
70	○	○	○	○
Bosian	○	○	○	○
Smith	○	○	○	○

Click on your answers

Displaying data

Developers can define **how the data will be displayed** by writing some lines in the ISIS formatting language.

For example, let's look at some ways the following data can be displayed:

10: Of war and peace
20: ^aTolstoy^bLeo

The format:	Will result in:	Because:
v10	Of war and peace	**v**10 displays the field 10
v10.4	Of w	**.** Precedes the number of characters (in this case, it displays the first 4 characters)
v10*8.3	and	* precedes the offset (in this case, it displays 3 characters starting from the eighth character)
UC(v10)	OF WAR AND PEACE	**UC** = Upper Case (converts all letters to upper case)
v20	^aTolstoy^bLeo	**v**20 displays the field 20
v20^a	Tolstoy	**^**a displays only the subfield "a"
mhl(v20)	Tolstoy, Leo	**mhl** = Mode Heading Lowercase (separates subfileds with a comma; it leaves case untouched)

Also, fixed texts ("literals") can be inserted: *"Title: "v10* will result in *Title: Of war and peace*.

Defining searches

Another important thing to decide is…

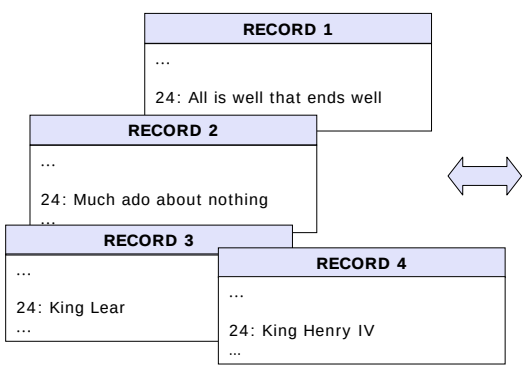
In order to provide fast retrieval in a library it is necessary to catalogue documents in the most appropriate way.
Therefore, librarians need to reflect on what type of catalogues they want to create.
Then developers will design and build a permanent index, called an "inverted file". To do this, they need to reflect, like librarians, on which data need to be indexed.

Let's look at an example of an inverted file…

Defining searches

Imagine we have a database with records containing title fields (n.24). We can invert these data by creating an index.

The inverted file contains extracted search terms, together with links to the records from which they were extracted.



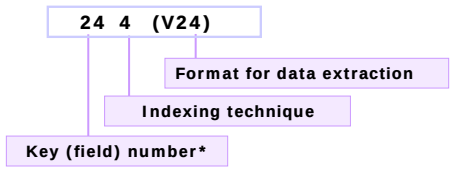
INDEX (INVERTED FILE)

This word:	is in record:
about	2
ado	2
all	1
ends	1
henry	4
is	1
IV	4
much	2
king	3,4
lear	4
nothing	2
that	1
well	1

Defining searches

Developers control what goes into the inverted file by defining a **Field Select Table**.

Field Select Table



In this example, the Field Select Table contains a line saying:

- which key number assign to the extracted term (24);
- which indexing technique must be used (4); and
- the formatting language used to extract a string from a field (V24 extract content of the field 24).

*It is good practice to let key 24 correspond to field 24.

By choosing the Indexing technique developers can decide to extract the whole field, each occurrence of a field, everything between text markers like/ / or <>, each word in a field.

By using the formatting language, they can format terms in the inverted file.

Defining searches

For example:

In a database there are records from Senegal and Burkina Faso. Their record id's are:

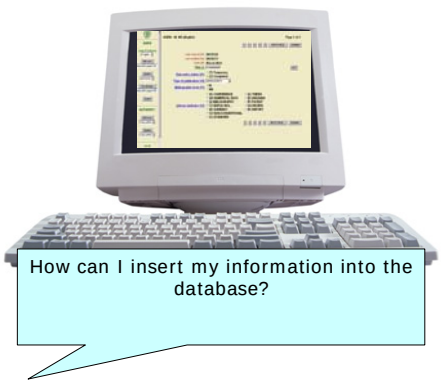
SE20030201004
BF20030605002
SE20030731005

If ISIS indexes the whole field, the index would be:
BF20030605002
SE20030201004
SE20030731005

But by using the formatting language to format only the first two characters, the index would just be:
BF
SE

Now an index on the code for country of origin has been created.

Defining data input



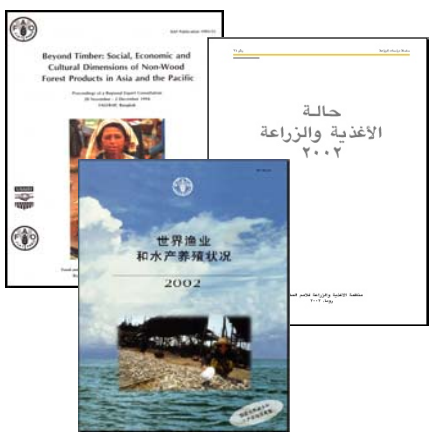
For web versions web pages are used to input or modify data, for other versions Worksheets have to be defined for that purpose.

They can be defined in such a way that they help to ensure data consistency.

Fields can have a default value, be defined as alphabetic or numeric, or the data must be according to a certain pattern.

Worksheets cannot enforce that the user picks values from a predefined list, or fills in certain mandatory fields.

When to use CDS/ISIS



Before ending, let's focus on the strong and the weaker points of CDS/ISIS.
This could be useful in deciding if this system matches your needs.

The following are the main **strong points** of CDS/ISIS:

• fast retrieval in data with large pieces of **unstructured texts**; and

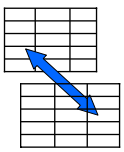
• managing of textual data in **non-Latin scripts** or languages with specific uses of accented characters.

When to use CDS/ISIS

On the other hand, **weaker points** of CDS/ISIS are:

• reformatting of **numerical data**: e.g., there are limitations if you want to convert integers into real numbers or floating-point numbers.

• managing **data that is being changed all the time**: if a record is deleted or modified, special reorganization procedures must be carried out to remove old data.



• data input from **standardized lists**: such links between tables are not a standard feature, so if you have the same name stored in different records, and you want to change it, you have to do it in each individual record.

However, the program offers some facilities for standardization, like the ability to define default values in a worksheet.
Special applications and plug-ins have been developed to enable, for example, data input from a thesaurus.

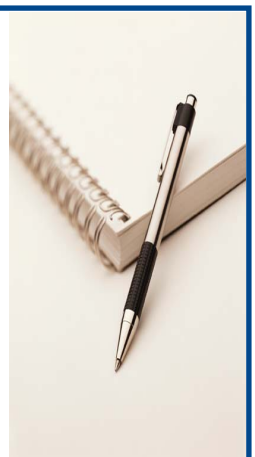
Summary

• CDS/ISIS as a **textual DBMS** is used for developing and managing free-structured textual databases and can be tailored for different applications.

• The system manages:
- the structure of textual databases,
- text-oriented formatting,
- fast and powerful retrieval, and
- the usage of different languages and scripts.

• Through specific files, developers can define:
- the structure of fields,
- how to display the data,
- how to search the data, and
- how to input data in the database.

• CDS/ISIS is particularly effective for retrieval in data with big pieces of **unstructured texts**, and for textual data in **non-Latin scripts** (or languages with specific usage of accented characters).



Exercises

The following five exercises will allow you to test your understanding of the concepts covered in this lesson.

Good luck!

Exercise 1

What is CDS/ISIS?

- ○ A set of tools for relational database management
- ○ A textual database
- ○ A set of tools for textual database management

Click on your answer

Exercise 2

What is the function of the Field Definition Table?

- ○ It is a list of the different elements that can be distinguished in a piece of information, and their properties.
- ○ It contains extracted search terms together with links to the records from which they were extracted.
- ○ It selects data from fields or subfields and formats the information for display.

Click on your answer

Exercise 3

Let's consider this fragment of a Field Definition Table.

Can you identify the following elements?

	Repeatability	Field name	Field number	Subfield delimiters
Imprint	○	○	○	○
30	○	○	○	○
R	○	○	○	○
Series	○	○	○	○
abc	○	○	○	○

Click on your answers

Exercise 4

What are the features of...

1	Field Select Table	a	contains extracted search terms together with links to the records which they were extracted from.
	Inverted File		defines rules for extracting key terms from a record and storing them in the index.

Click on your answer

Exercise 5

In which of the following situations could CDS/ISIS be the appropriate choice?

- ○ to store, retrieve and disseminate administrative data that change on a regular basis.
- ○ to store, retrieve and disseminate books and articles in different languages.

Click on your answer

If you want to know more...

CDS/ISIS originates from Unesco

Their ISIS site (http://www.unesco.org/isis) provides information about their work on ISIS including links to websites from the user community.

Bireme is an important developer of versions of ISIS.
Their product catalogue gives access to information on these products (under tools). See:
http://productos.bvsalud.org/html/en/home.html

Some of the products on this CD-ROM have been produced by the Institute for Computer and Information Engineering (ICIE), Warsaw, Poland.
On http://www.icie.com.pl/ you can learn more about their products and development work (see "products").