Chapter 4.Informational Technologies

Course:

Information Technology in Research

Mat. Gabriela MAILAT
Prof.Ph.D.Eng Elena HELEREA

1. Basic concepts as regards informational technology
    1.1. Information
2. Informational technology
    2.1. Internet
    2.2. OCR (Optical Character Recognition)

Information has been identified as one of the basic requirements of human existence.

It holds three principal characteristics:

- ✓ information are the most important economic resource,
- ✓ the information consumption is intense,
- ✓ the development of the global informational infrastructure is primordial.

The common informational area consists in several indivisible levels, which are:

1. Information as such, in electronic format (databases, document bases, image bases etc.);
2. Hardware and software components available for the users to process these information;
3. Physical infrastructure (wires, radio and satellite communication networks);
4. Basic services of telecommunications; especially e-mail, data electronic transfer, interactive access to databases and interactive transmission of the digital image;

5. Applications that offer the users the specific services they need in order to make use of the previously mentioned levels.
6. The users who have not been trained to the purpose of using the applications, but who are aware of the potential of using the information and communication technology, and likewise of the conditions they must fulfill in order to benefit from its advantages.

Information

1. *information as product* (category of „information intermediaries" that deal with arranging and distributing the information through books, magazines, radio and television)

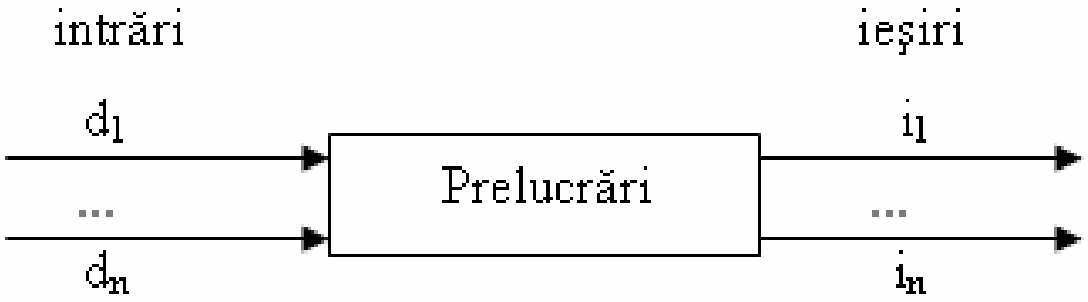
2. *Information as production mean* (the number of those dealing with information as production mean has greatly risen during the last years, they participating „in making up the internal collections of information, necessary for the continuous and efficient activity of each and every institution.

Datum/a = concrete form of expression of the information**.**

Datum/a = a number, a value, a relation that contributes to solving an issue or that is obtained following a research, which is to undergo some processing.

Datum/a may be considered the raw matter for information

Schematically, the relation between data and information may be represented this way:



unde $d_1 \div d_n$ date

$i_1 \div i_n$ informaţii

Relation data information

Assessment of information

For the qualitative assessment of information, there may be considered **three** aspects:

1. temporal dimension,
2. content and
3. form of information.

Assessment of information

1. ***Temporal dimension***

Information - *opportune*

certain actual frequency.

The period whereto information relates constitutes a temporal attribute of high importance.

Assessment of information

2. ***Content*** – the most important dimension of information.

Information: *exact*
pertinent
complete
exhaustive
concise.

Assessment of information

3. Form

It defines the modality of presenting the information, being the one that makes information attracting, easy to use and understand.

The information must be clear, precise, orderly presented in an adequate manner (text, graphs, schemes etc.) and on an accessible support (paper, see-through, the monitor of a computer etc.).

Informational Technology

In order to become useful in any field of activity, the information must be collected, stored, processed and transmitted to those in need.

The use of the informational technology improves the opportunities in enhancing information.

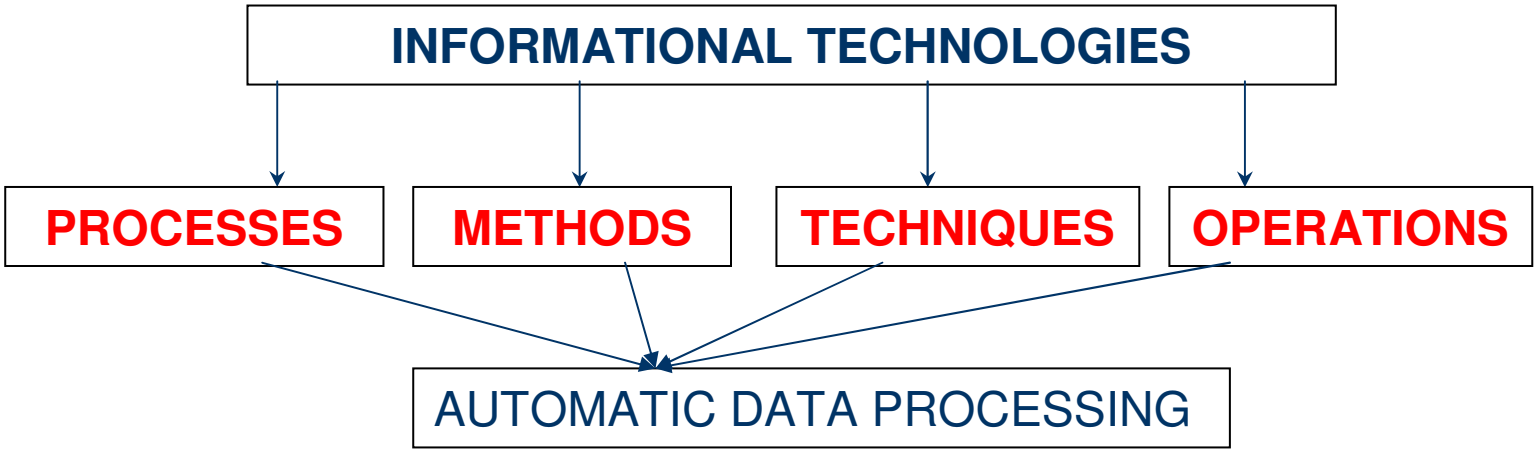
Informational Technology

Informational technologies enable collecting, processing, storing and transmitting the information as voice, image, text and in numerical form, based on microelectronics, through combining informatics and telecommunications.

Informational Technology

Informational technology = a paradigm of the technical-economic development, which includes communications, photonics, manufacturing systems, networks, software, information storage equipment and memories.

Informational Technology



Informational Technology

Several aspects and activities from within informational technologies may be grouped in the categories presented in the scheme below:

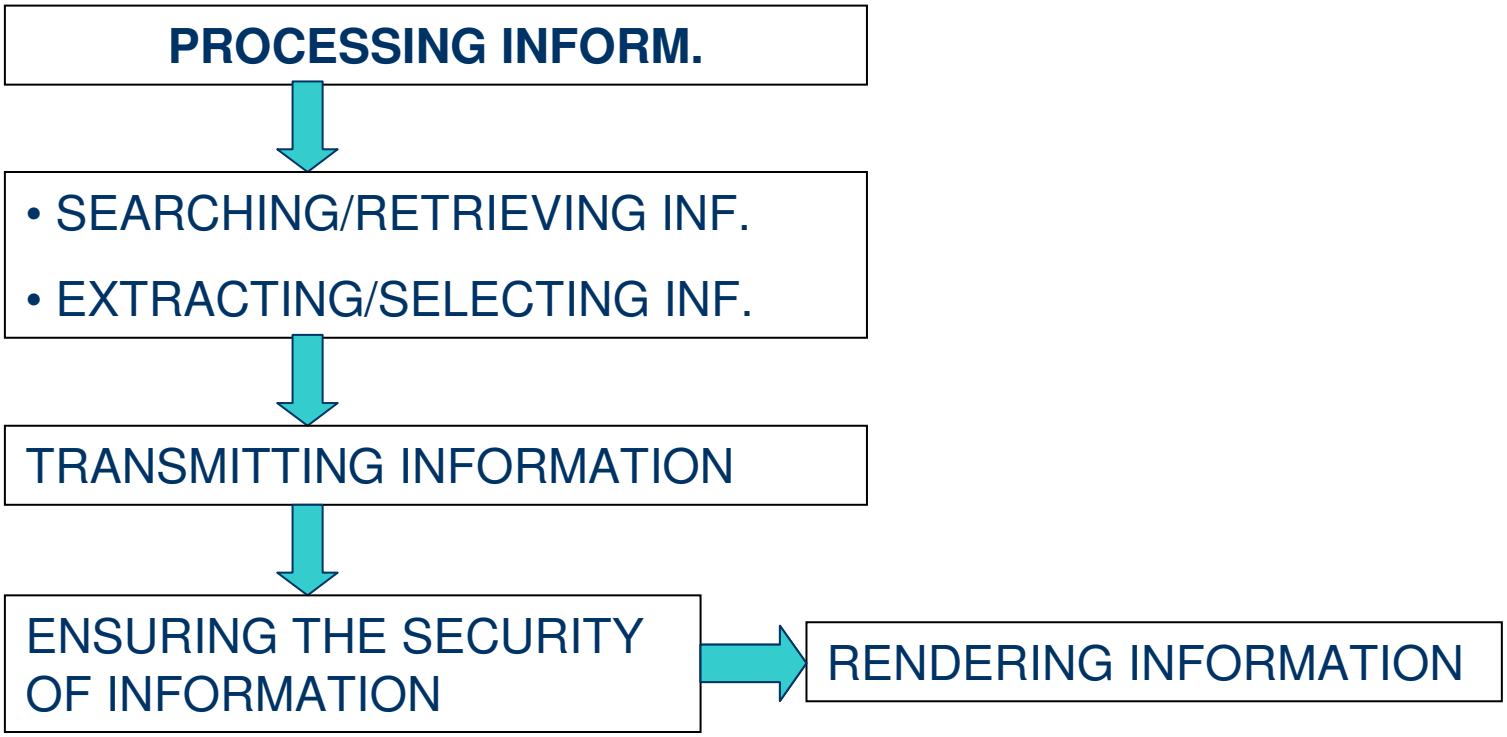
Informational Technology

- COLLECTING INFORMATION
- REPRESENTING INFORMATION
- REGISTERING INF. (WRITING)
- IDENTIFYING INF. (READING)

- ORGANIZING IN MEMORY AND
- KEEPING INFORMATION

PROCESSING INFORM.

Informational Technology



Informational Technology

Global characteristics:

- Ensuring the accessibility of whatsoever information under verbal, symbolical form, through the intermediary of the computer;
- Making available high memorizing capacities for the information processing systems ;
- The possibility of using human language so as to perform the querying of the information processing systems;

Informational Technology

- Any information that has proved its usefulness **may be transmitted in another** point within the same system, at low costs;
- The information processing systems are increasingly capable of informing, of supporting the decisional process and even more, of learning.

Informational Technology

Two essential changes :

- Descentralization of the informatic power and of the data storage, and shifting towards the user;
- Utilization to a large extent of the telecommunications electronically inter-connecting the different components of the informational systems.

Informational Technology

Informational technology comprises, beside the elements ensuring the collection, processing, storage and transmission of information (as voice, image, text and likewise numerically, based on microelectronics, through combining informatics and telecommunications); likewise the theoretical and methodological elements upon the development of informatics systems.

Informational Technology

The main informational technologies:

- Internet
- OCR (Optical Character Recognition)
- ATM (Automat Teller Machine)
- Online technologies for achieving gatherings (conferencing)

Informational Technology

- Multimedia technology
- Technology of electronic document management
- Other informational technologies

INTERNET

Internet = web of all networks

Computer networks – **Advantages:**

a) Resource sharing – the equipment and especially the data are available for anyone within and throughout the network, whatsoever the user's physical localization

INTERNET

b) The access to information from a distance – may take different forms, from the access to programs/various types of software to the access to databases from a distance (booking, banking operations etc.)

c) Interpersonal communications – they may take the form of e-mail messages, virtual gatherings, instruction/training at a distance

d) Shared virtual reality.

INTERNET

Internet is based on computer networks dispersed worldwide, which communicate among themselves through an Internet **Protocol** – Protocol (IP).

Every host on the Internet has an IP address encoding its network address and host address, the combination being unique.

INTERNET – Electronic mail

The e-mails consist in the message in itself and an envelope.

To: xxx@yahoo.com

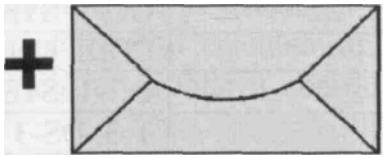
From: zzz@gmail.com

Date: sat 17th nov

Subject: Conferinta

Avshvchvjhdcvjhdvjh

Attachments:

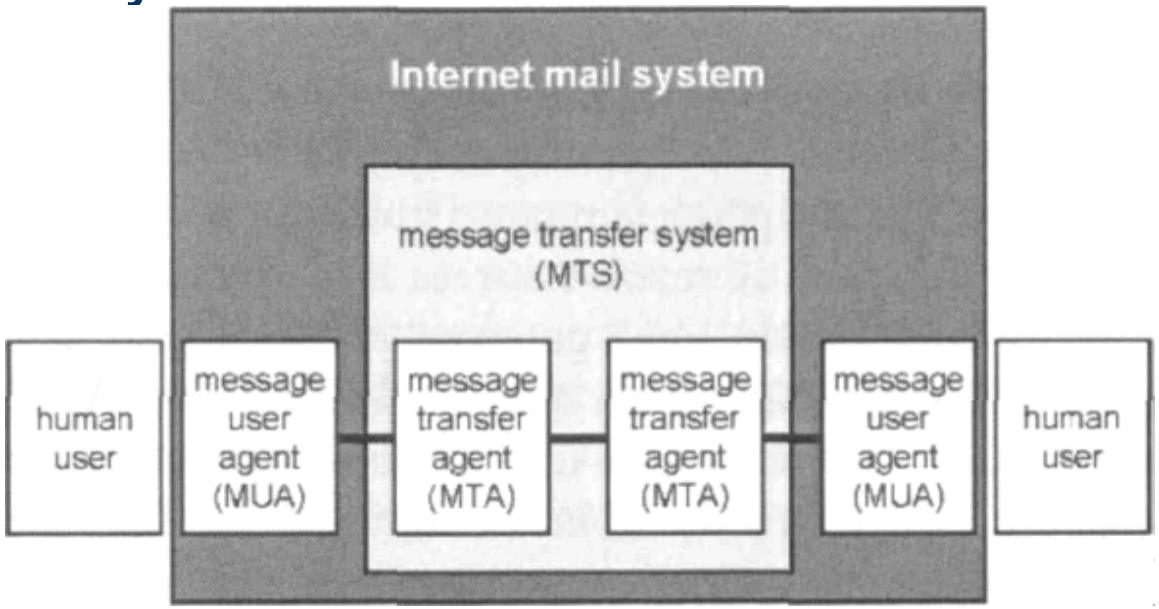


INTERNET – Electronic mail

- MTS (Message transfer system) from the Internet allows the transportation of the messages through the Internet network based on “store and move forward” or “store and download”.
- The information may be transmitted at any time

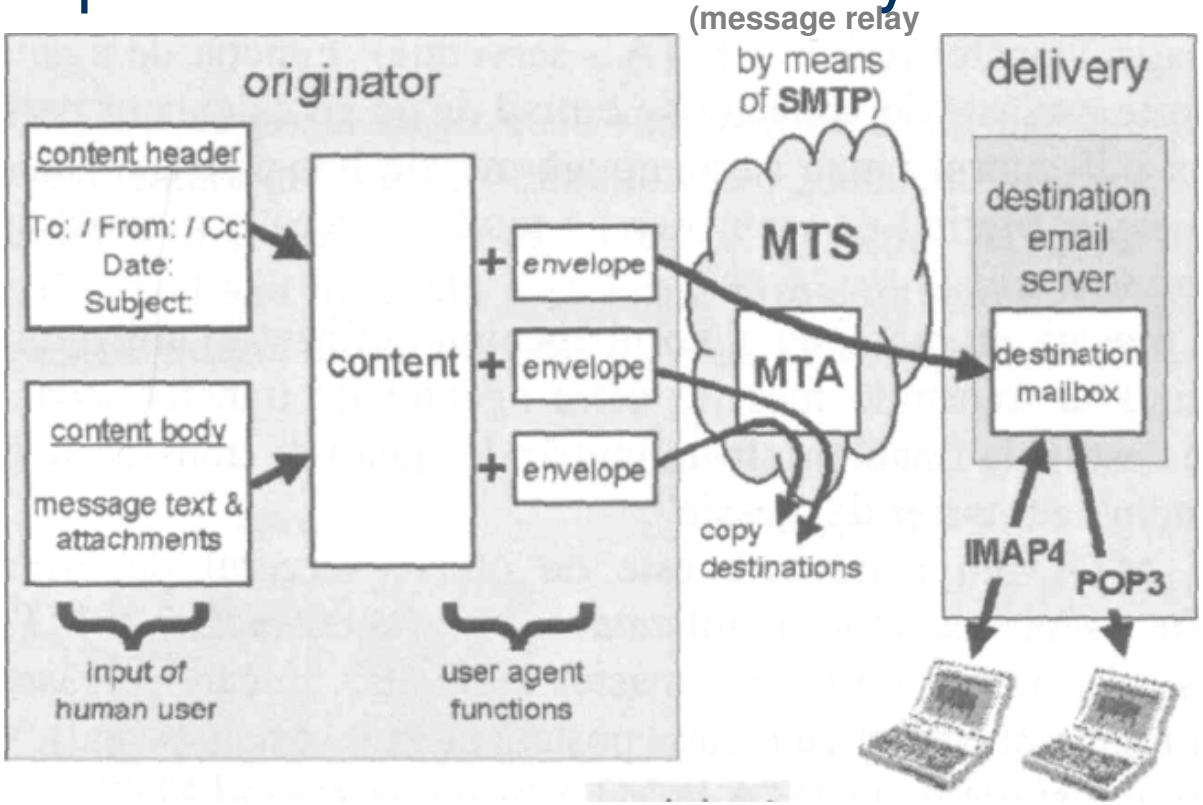
INTERNET – Electronic mail

- The elements of an e-mail system and of the transfer system



INTERNET – Electronic mail

- The operations of the e-mail system



INTERNET – PROCESSING A SEARCH /QUERYING FROM THE CLIENT

The steps taken by the web server in order to process a search received from the web client:

1. The web browser or another web client sends forth towards the web server a querying, asking for certain resources. The search is transmitted under HTTP form, while the address of the required resource is specified under Uniform Resource Locator (URL) format. The querying is usually made resorting to the HTTP Get order.

INTERNET – PROCESSING A SEARCH /QUERYING FROM THE CLIENT

2. After having received the querying from the client, the web server settles the existence of the resource within the resources controlled by the respective server;
3. In case the resource is available, the web server determines the access rights, and if these rights have not been infringed, it sends back to the client the content of the desired resource;
4. In case the access rights have been infringed, the web server rejects the querying, sending back to the client the well-deserved warning/notification;

INTERNET – PROCESSING A SEARCH /QUERYING FROM THE CLIENT

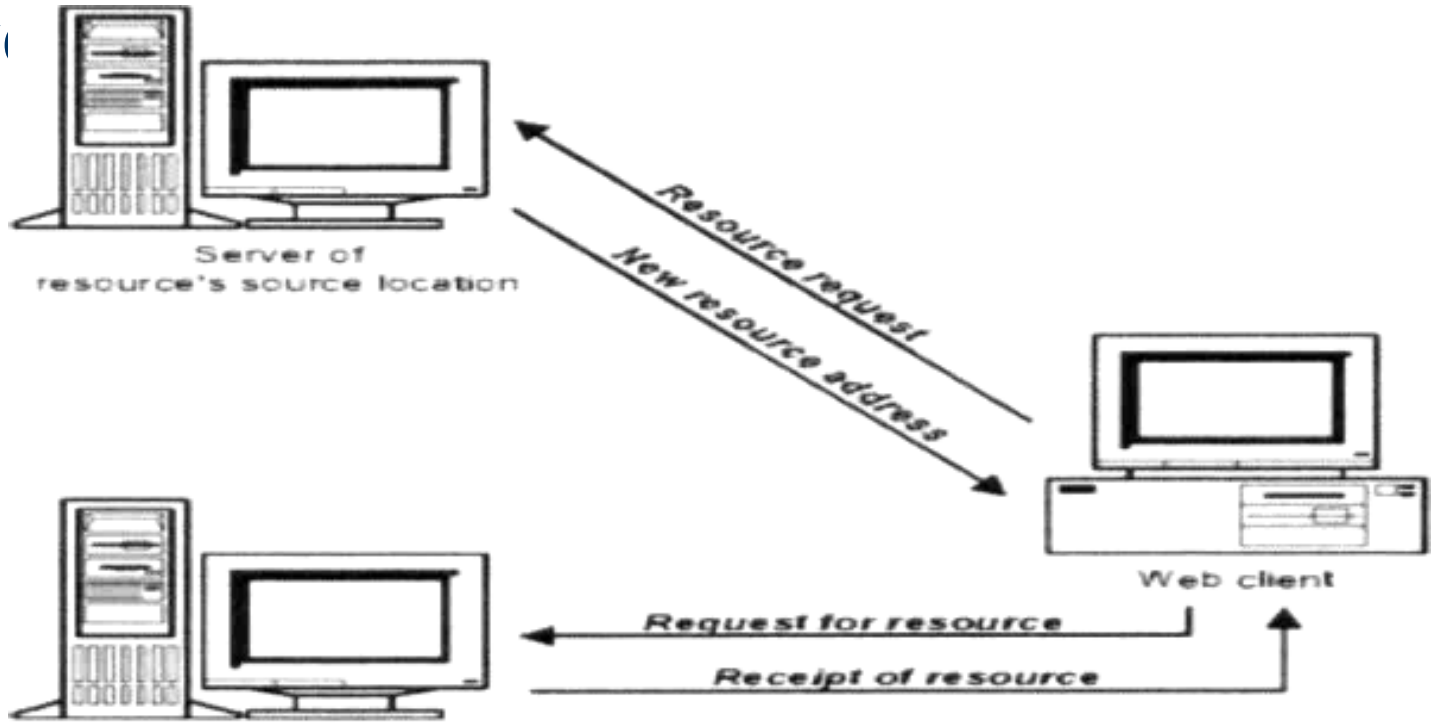
5. In case the resource is not to be found on the web server, the server establishes the information upon the resource from the configuration files, these ones including even a possible relocation within the web. If the resource has been allotted to the server, however temporarily redirected towards another location, the server informs the client upon this fact;
6. If the web server supports a virtual tree made up of other web servers, the search will be redirected towards the necessary resources;

INTERNET – PROCESSING A SEARCH /QUERYING FROM THE CLIENT

7. If the web server is also used as proxy server, it acts on one hand as web server for the client having transmitted the querying and on the other hand as web client in order to query another web server. This a simple retransmission agent which retrieves and stores in cache web pages for the persons from within, however not allowing the visitors' access to internal resources. So as to use a proxy server, each and every browser must be configured in order to:
   a. send forth all GET searches/querying towards the proxy server and not towards the host specified in the URL;
   b. include the entire URL, inclusively the name of the server and the port within the GET querying.
8. After providing the information to the client, the server closes the connection with this one.

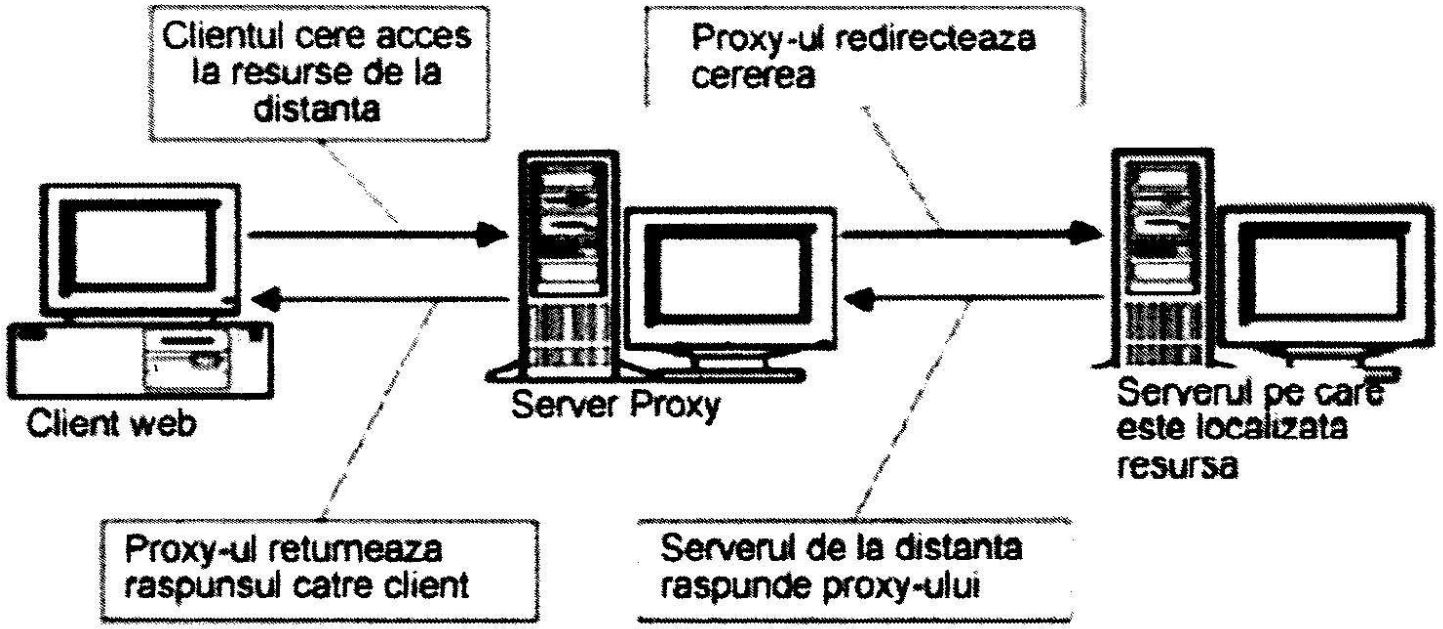
INTERNET – PROCESSING A SEARCH /QUERYING FROM THE CLIENT

- Interactions between the client and the serv



INTERNET – PROCESSING A SEARCH /QUERYING FROM THE CLIENT

- **Web server as retransmission agent (proxy)**



INTERNET – Search Engines

- The web engines on the Internet are specialized web sites, created in order to help people find information stored in other sites.
- They execute three basic tasks:
  - They search the Internet or „select" parts of the Internet, based on the important words;
  - They keep an index of the words they retrieve and of their place;
  - They allow the users to search for words or phrases (combinations of words) found within this index.

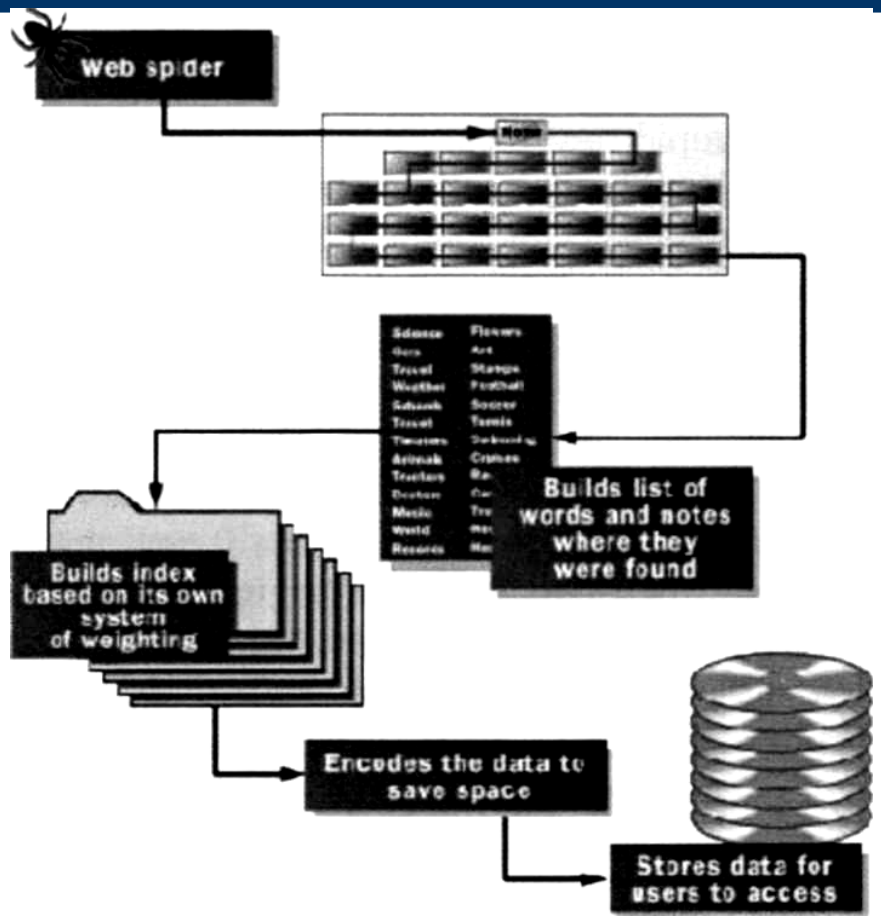
INTERNET - Search Engines

- The search engines on the Internet put together separate data so that the user might find what he/she needs.
- In order to find information from the billions of web pages, a search engine resorts to a special application, called „search robot" or „spider", in order to build a list of words found in the web pages.

INTERNET - Search Engines

- The process thereby a spider carries through its list is called „web crawling", and so that a search engine/spider might build an efficient list of words, it has to carry through the search through a multitude of pages.

A "Spider" obtains the content of a web page and creates a list of key words enabling the users to find the information they want



INTERNET - Search Engines

- A spider begins its search through the web usually starting from a list of servers intensely used and from very popular web pages. The Spider begins with a popular site, indexing the words from the pages and following all links found in the respective site, being this way to go through and to index the most widely used part of the web.

INTERNET - Search Engines

- Google.com started as an academic search engine.
- The system has been constructed so as to use several spiders, usually three, each spider being able to keep open 300 links/connections towards web pages at a given moment.
- At the highest performance, using four spiders, the system could perform the search over more than 100 pages on a second, generating 600 kilobytes of data in every second.

INTERNET - Search Engines

- There was necessary to build a system that should feed the spiders with information.
- Google.com initially had a server dedicated for offering URL-s to the spiders.
- Google also had its own DNS server, the translation of the names into addresses being significantly faster, diminishing at the same time the delays caused by the networks.

INTERNET - Search Engines

- At the moment a Google spider visited a HTML page, this one considered two things:
  - the words found in the page;
  - the position of these words in the page.
- The words found in the title, sub-title, metatags and other positions of relative importance were marked with special signification during the searches initiated by the users.
- The spider was built so as to index all significant words, leaving aside the connecting words.

INTERNET - Search Engines

- Other spiders use other procedures for indexing, allowing, for instance, the spiders to operate faster or allowing the users to search more efficiently; or both. For instance, some spiders maintain a list of words from the title, sub-title and links, together with the most widely used 100 words on each page and every word from the first 20 lines of text. Lycos uses this modality of indexing the content of the web pages.

INTERNET - Search Engines

- Other systems, such as AltaVista.com, go into another direction, indexing all words on a page, inclusively the connecting or „insignificant“ words. This pushing towards completion also has other functioning modalities, especially through the use of the meta-tags.
- Meta-tags allow the owner of a page to specify the key words and the concepts under whom the respective page will be indexed.

INTERNET - Search Engines

- After the spiders have finished the task of retrieving the information in the web pages, the search engine must store the collected information in a usable modality.
- There are two components which render the collected data accessible for the users:
  - Information stored with the data;
  - Method for indexing the information

INTERNET- Search Engines

- A search engine will only store the words and the URL where they were found.
  - Search engine with limited usages
- The search engines store much more data than the word and the URL wherein it was found.
- An engine could store the number of apparitions of the word in the page, being likewise able to assign a “weight" to each and every entry, with higher values attached to the words which appear towards the beginning of the document, in the sub-titles, connections/links, meta-tags or the title of the page.

INTERNET - Search Engines

- Every search engine has different formulae or modalities of assigning the weight for the words in the index. This is one of the reasons whereof a search according to the same word in different search engines will bring forth lists of different results, with the pages presented in different orders, even if there are indexed the same pages.

INTERNET - Search Engines

- The data will be stored in an encoded manner, in order to economize the storing space.
- After the information has been compacted/encoded, it is ready to be indexed.
- ***Purpose***: is allows the very rapid retrieval of the information. There are several modalities to build an index, however one of the most efficient modalities is the use of a hash table. Through **hashing**, there will be applied a mathematical formula so as to attach a numerical value to every word, the formula being constructed so as to equally distribute the entries along a pre-determined number of divisions.

INTERNET - Search Engines

- The combination of efficient indexing and storage renders possible to obtain quickly the results, even if the user has been creating a complex querying.
- The querying/search through an index supposes having constructed a querying by a user and transmitting it towards the search engine. The querying may be simple, consisting in minimum a word, or may be complex, requiring Boolean operators, which allow the refining and the extension of the search.

INTERNET - Search Engines

The most frequently used Boolean operators are the following:

- AND – all terms separated through „AND" must appear on the page or in the document. Some search engines may use „+" instead of „AND";
- OR – at least one of the terms separated through „OR" must appear on the page or in the document;
- NOT – the term or the terms appearing after „NOT" must not appear in the document. Some search engines may use „-" instead of the word „NOT";
- FOLLOWED BY – one of the terms must necessarily be directly followed by the other term;
- NEAR – one of the terms must be at a distance specified in words from the other term;
- Inverted commas – the words between the inverted commas are dealt with as a phrase, and this phrase must be found within the document or the page.

INTERNET

Another application of the Internet is represented by the INTRANET and the EXTRANET.

- An **Intranet** is a network within an organization that connects multiple users through the intermediary of the Internet technologies.
- Intranet-s limit the unlimited territory of the Internet, establishing sectors with controlled access wherein the users may freely communicate and interact. The networks are based on the World Wide Web, allowing the users to communicate among different platforms in real time.

- Intranet functions on the basis of the Internet technologies, however within an organization.
- It allows several persons to interact in real time, to store and to search for document archives, to collaborate so as to create documents, to modify graphs, images, audio and video documents and not in the least to converse in real time through the intermediary of the chat. Moreover, depending on the construction manner of the intranet, the users may navigate on the Internet, without differentiating between the access from the intranet towards the Internet.

- Intranet-s offer a wide range of benefits which fit within two great categories: efficiency and efficacy.
  - **efficiency** stands for the improvement of the mechanisms of information exchange, removing the logistic obstacles so as to collect and/or distribute the necessary information in due (adequate) time.
  - **efficacy** supposes the organizational aspect upon the improved collaboration and upon the decision-making process.

- EXTRANET = INTRANET for the others in their turn
- Extranet = a Web site with controlled access, wherein part of the visitors come from outside the organization.
- Extranet-s are used for:
  - Several types of business applications
  - B2B or of electronic trade
  - Project management or
  - Collaborative extranet-s allow the exchange of documents, planning, associated to a certain project or partner.

Extranet allows :

- Sharing the updated documents, files or images with partners or clients in separate locations;
- Working in collaboration through rendering available to the purposes of editing, reviewing, updating and storing, the documents and the digital assets;
- Managing the projects in a centralized workspace;
- Offering current versions of the frequently updated documents.

Technology of Optical Character Recognition OCR

- It is used in the operations digitizing the text type data.
- To commercial purpose, there is made use of within shops, so as to read the barcodes on different products.

Technology of Optical Character Recognition OCR

Principles of the OCR Technology

- The systems of optical character recognition (OCR) only recognize the machine printed materials.
- Resorting to the ***pattern-matching technology,*** **OCR translates the specific shapes and type/font** of the printed characters into the corresponding ***computer codes***. Although the most advanced systems are capable to recognize multiple fonts, they may only process standard fonts, such as *Times Roman and Arial.*
- Once *all characters within a word have been recognized*, the ***word*** is ***compared*** against a ***vocabulary*** of ***potential answers*** for the final result.

Technology of Optical Character Recognition OCR

- Character recognition ***segments*** afterwards ***lines of text*** or words into ***separate characters***, which are recognized by the ***makeup*** of their component shapes. The *characters* of the machine-printed letters are evenly spaced across, and up and down, a given page, allowing the OCR **to read the text one character at a time**.
- **Segmentation into single characters** represents a critical recognition failure point for form processing organizations, because OCR recognition technology requires *high quality images*, with excellent *contrast*, characters and clarity.

Technology of Optical Character Recognition OCR

- Any text that is less than perfect will cause even the most sophisticated OCR systems to return significant reductions in accuracy, when processing degraded images.
- For instance, when characters break apart due to *poor image quality;* or if *multiple characters merge* due to *blurred or dark backgrounds* between them, ***recognition accuracy*** may be reduced by as much as **20** percent.

Technology of Optical Character Recognition OCR

- **The assessment** of the accuracy – applied to OCR – is represented by the ***percentage of characters correctly read*** on a ***given page of text***, and the systems vary widely, achieving 95 to 99 percent accuracy.
- But accuracy rates at anything below 100 percent can translate into huge productivity losses.
- An entire application or verification process could be compromised if even 5 percent of the data are either entered incorrectly or misread.

Technology of Optical Character Recognition OCR

- ***OCR* systems** must have ***the ability*** to „***proofread***” results:
  - *marking characters* that *the system does not recognize*
  - *sending rejected text to human operators*, for *manual processing*.

Technology of Optical Character Recognition OCR

Image analysis uses ***a set of pattern recognition techniques***, to enable ***computer systems*** *to recognize* and *to interpret* the images.

Image Analysis consists of ***two stages***:

- Image Processing and Analysis and
- Pattern Recognition

1. ***Image processing and analysis*** includes steps, such as:
   - ***Image repair*** to improve *quality and usability issues*
   - ***Feature extraction*** for the subsequent pattern recognition stage.

Technology of Optical Character Recognition OCR

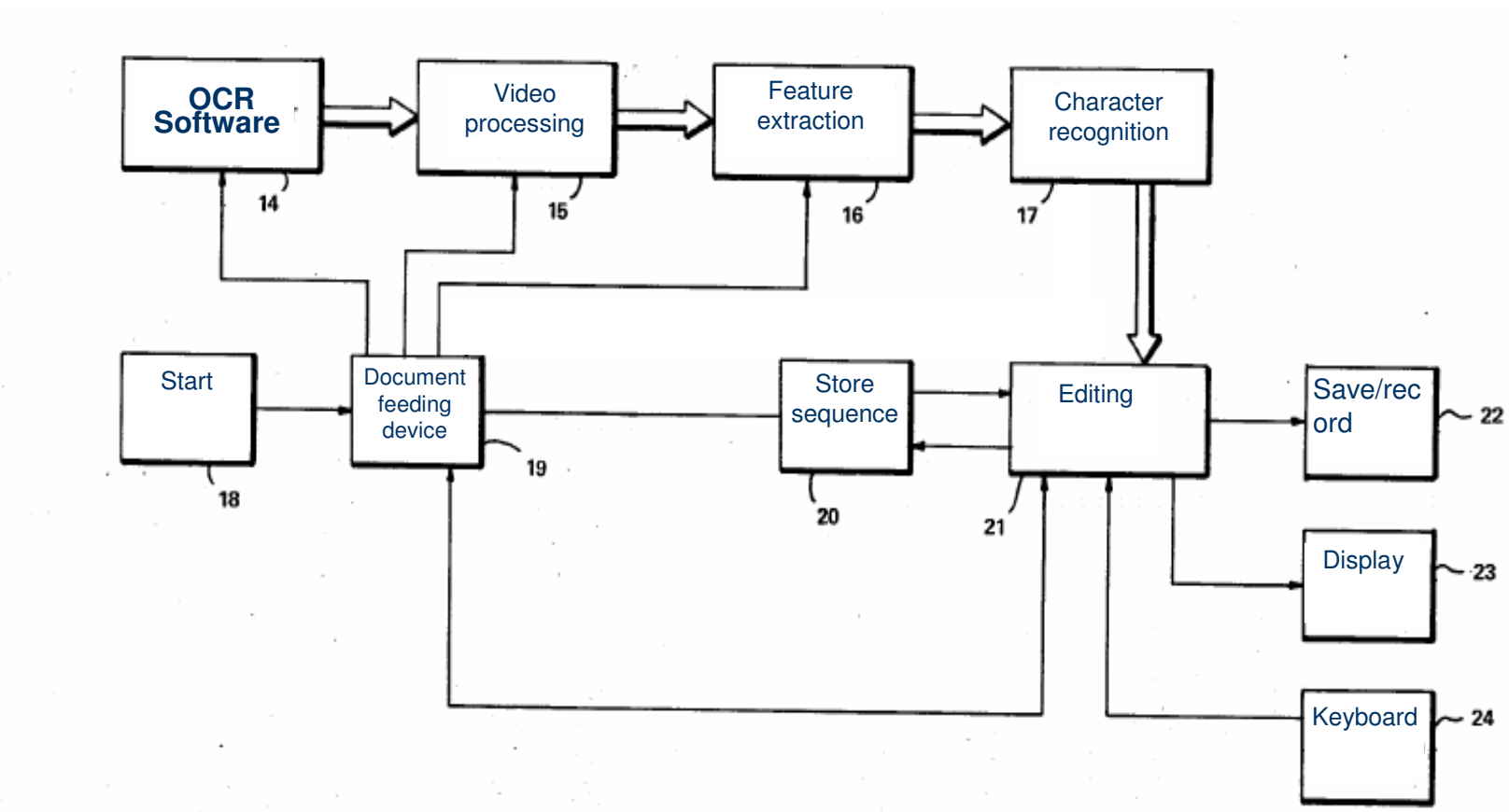
2. Pattern recognition includes in its turn:

- ***Area of interest location*** within the original image
- ***Data extraction***
- ***Validation of an existing representation or context***

Examples of image analysis:

- Region of interest location on letters, flats and parcels
- Automatic indicia location and detection on envelope images
- Check stock verification and signature authentication

Technology of Optical Character Recognition OCR



Technology of Optical Character Recognition OCR

General information

In basic terms, ***OCR software*** *examines* ***a scanned bitmap image*** and translates the text within it in a ***file*** *that can be* ***edited***.

The first OCR systems translated the text into a single font and size only.

Today's advanced programs attempt at duplicating not only the ***fonts***, but also ***complex*** layout ***features***, such as:

- ***columns, tables, headers and footers***
- even ***graphics***

Technology of Optical Character Recognition OCR

- ***OCR Software*** reads the text ***one character at a time***
- There are ***many different types*** of ***pattern recognition schemes***, and each OCR software
  - uses a ***different set of models*** and
  - implements them in different ways.

Technology of Optical Character Recognition OCR

Components of the OCR software

OCR systems frequently used include ***three components***:

- ***An image scanner***
- ***OCR Software and Hardware***
- ***An output interface***

Technology of Optical Character Recognition OCR

The process involves ***three operations***:

- ***Image analysis*** (Extracting individual character images from document)
- ***Image recognition*** (Recognizing these images based on shape)
- ***Contextual image processing*** în funcție de context
  - either to correct misclassifications made by the recognition algorithm
  - or to limit recognition choices