

Elba Project

Dabne – Tecnologías de la información

Correction of the OCR

Software does not recognize the characters in an optimal form; therefore it is necessary to review texts and to correct or to verify that they are identical to the original one.

Elba Project

Dabne – Tecnologías de la información

The screenshot displays the ABBY FineReader 8.0 Professional Edition interface. The main window is titled "Untitled batch (Recovered) - ABBY FineReader 8.0 Professional Edition - [1 - Text]". The menu bar includes File, Edit, View, Batch, Image, Process, Tools, Window, and Help. The toolbar shows icons for Scan&Read, Scan, Read, Check Spelling, and Save. The interface is divided into three panes: Batch, Image, and Text. The Batch pane on the left shows a list of scanned pages, with page 1 highlighted. The Image pane in the center shows a document page with a green selection box around a text block. The Text pane on the right shows the extracted text. A "Check Spelling" dialog box is open, displaying the word "Ulla" as "Not in dictionary". The dialog box shows the word "Ulla" in a red box, and the text "Gómez Ulla atiende a me" is visible. The suggestions list includes "Aulla", "Bulla", "Ella", "Fulla", "Hulla", and "Illa". The dictionary language is set to "Spanish". The status bar at the bottom shows "Spanish (Traditional Sort)" and the system tray includes the Start button, taskbar, and system clock showing 12:42.

Batch

Image

Text

Check Spelling

tar Gómez Ulla atiende a me

Not in dictionary

Gómez Ulla atiende a menos del 50% de los pacientes civiles previstos

Suggestions:

Aulla	Ella	Hulla
Bulla	Fulla	Illa

Dictionary language: Spanish

Undo Options... Close

El hospital militar Gómez Ulla atiende

Start El GIMP Untitled... *Sin nom... *Sin nom... *Sin nom... *Sin nom... *Sin nom... *Sin nom...

INS Spanish (Traditional Sort) 12:42

Elba Project

Dabne – Tecnologías de la información

- **txt** is the extension of a flat text document or plain text, talks about to textual data in ASCII format . This format is most portable since it is compatible with practically any computer and any application.
- **rtf** is the acronym of Rich Text Format, language of description developed by Microsoft who uses marks to enrich the typesetter aspect of documents. Its use is a poor format standardized with diverse incompatibilities is not recommendable to keep documentation.

Elba Project

Dabne – Tecnologías de la información

html (Hypertext Markup Language) is a kind of text document that is used by browsers to visualize texts and images. HTML documents usually are called Web pages

pdf (Portable Document Format), devised by Adobe, is the public tool used in companies with world-wide standards for a safe and trustworthy distribution and interchange of electronic documents. In case of using ABBYY software (OCR application) is interesting the use of an option that is pdf with text under the image.

Elba Project

Dabne – Tecnologías de la información

In Elba project we will preferably use pdf and html formats, the reasons why we choose these formats is because:

pdf allows us to respect original presentation, typesetter, grammar writing rules that for many investigators is relevant information

html, though we are going to loose information from original texts edition, we will be able to make a wider difusion, to enable search engines to find them, to read them on our web browsers (by default nowadays every computer has a web browser).

Elba Project

Dabne – Tecnologías de la información

In the electronic edition we want to conserve the texts and to show them with all the typesetter wealth of the original ones, in some cases we must use other formats for their display , in general called of marks, labels or markers, with which we assigned some qualifying ones to describe the different peculiarities, typesetter or others, of texts