



# *cb2Bib User Manual*

*Bibliographic Reference Extracting and Managing Tool*

© 2004-2025 Pere Constans. Last updated on 2025-11-02.

# Contents

- Overview
  - **cb2Bib Description**
  - **Using cb2Bib**
    - **Procedure**
    - **Buttons Functionality**
    - **Additional Keyboard Functionality**
  - **Advanced Features**
  - **Requirements**
    - **Compilation**
    - **Deployment**
  - **Credits and License**
- **Installation**
  - **Installing on Unix systems (tarball)**
  - **Installing on Unix systems (RPM)**
  - **Installing on Debian systems (deb)**
  - **Installing on Windows systems**
  - **Installing on Mac OS X**
- **Configuration**
  - **Configuring Annote**
  - **Configuring BibTeX**
  - **Configuring Clipboard**
  - **Configuring Documents**
  - **Configuring Files**
  - **Configuring Fonts**
  - **Configuring Network**
  - **Configuring Shortcuts**
  - **Configuring Utilities**
- **Search BibTeX and PDF Document Files**
  - **Search Functionality**
  - **Contextual Search**
  - **Notes**
- **cb2Bib Command Line**
- **cb2Bib Annote**
  - **Annote Syntax**
- **cb2Bib Citer**
- **Examples**
  - **Predefined Formats**
    - **BibTeX**
    - **ISI Reference Format**
    - **RIS Reference Format**

- **Additional Features**
  - **Embedded File Editor**
    - **Bookmarks**
    - **Update Documents Metadata**
    - **Export Selected Entries**
    - **Right-Click Menu Functionality**
    - **Reference List Right-Click Menu Functionality**
    - **BibTeX Browser Right-Click Menu Functionality**
    - **Additional Functionality**
  - **Predefined Placeholders**
    - **Cite and Document ID Placeholders**
    - **Cite Command Placeholders**
  - **BiBTeX Entry Types Available as cb2Bib Fields**
    - **Main Fields**
    - **Other Fields**
  - **Reading and Writing Bibliographic Metadata**
    - **Reading Metadata**
    - **Writing Metadata**
  - **PDF Reference Import**
    - **Introduction**
    - **Automatic Extraction: Questions and Answers**
  - **Extracting Data from the Clipboard**
  - **Processing of Author Names**
  - **Processing of Journal Names**
  - **Field Recognition Rules**
  - **Regular Expression Editor**
- **Supplementary Notes**

# Overview

cb2Bib is a free, open source, and multiplatform application for rapidly extracting unformatted, or unstandardized bibliographic references from email alerts, journal Web pages, and PDF files.

cb2Bib facilitates the capture of single references from unformatted and non standard sources. Output references are written in BibTeX. Article files can be easily linked and renamed by dragging them onto the cb2Bib window. Additionally, it permits editing and browsing BibTeX files, citing references, searching references and the full contents of the referenced documents, inserting bibliographic metadata to documents, and writing short notes that interrelate several references.

**Current version: cb2Bib 2.0.2.90.** See **Change Log File** for a detailed list of changes and acknowledgments, and **Release Notes** for additional notes and information.

See **Release Note cb2Bib 2.0.1.**

## cb2Bib Description

cb2Bib reads the clipboard text contents and process it against a set of predefined patterns. If this automatic detection is successful, cb2Bib formats the clipboard data according to the structured BibTeX reference syntax.

Otherwise, if no predefined format pattern is found or if detection proves to be difficult, manual data extraction is greatly simplified by cb2Bib. In most cases, such manual data extraction will provide with a new, personalized pattern to be included within the predefined pattern set for future automatic extractions.

Once the bibliographic reference is correctly extracted, it is added to a specified BibTeX database file. Optionally, document files are renamed to a DocumentID filename and moved to a desired directory as a personal article library, and their metadata is updated with the bibliographic reference. See **Reading and Writing Bibliographic Metadata** section.

cb2Bib facilitates writing short notes related to bibliographic collections. Notes are written using a minimalist markup on a plain text editor, and can latter be converted to HTML. Related references and links become easily accessible on any browser or by the embedded cb2Bib viewer. See **Release Note cb2Bib 1.1.0.**

# Using cb2Bib

## Procedure

- **Select the reference to import from the email or web browser**

On Unix machines, cb2Bib automatically detects mouse selections and clipboard changes. On Windows machines, copy or Ctrl-C is necessary to activate cb2Bib automatic processing.

- **cb2Bib automatic processing**

Once text is selected cb2Bib initiates the automatic reference extraction. It uses the predefined patterns from file [regexp.txt](#) to attempt automatic extraction. See **Configuring Files** section for setting the user predefined pattern matching expression file. After a successful detection bibliographic fields appear on the cb2Bib item line edits. Manual editing is possible at this stage.

- **cb2Bib manual processing**

If no predefined format pattern is found or if detection proves to be difficult, a manual data extraction must be performed. Select, either using mouse or Shift+arrow keys, the reference fields from the cb2Bib clipboard area. A popup menu will appear after selection is made. Choose the corresponding bibliographic field. See **BiBTeX Entry Types Available as cb2Bib Fields**. If operating with the keyboard, first letter of the field is set as a menu shortcut. Then, typing 'A' sets selection to 'author', or '+A' to 'add authors'. Selection is postprocessed and added to the cb2Bib item line edit. cb2Bib field tags will show on the cb2Bib clipboard area. Once the manual processing is done, cb2Bib clipboard area will contain the matching pattern. The pattern can be further edited and stored to the [regexp.txt](#) file using **Insert Regular Expression**, Alt+I. See the **Extracting Data from the Clipboard** and **Regular Expression Editor** sections.

- **Download reference to cb2Bib**

cb2Bib has the built-in functionality to interact with publishers "Download reference to Citation Manager" service. Choose BibTeX format, or any other format that you can translate using **External Clipboard Preparsing Command**. See **Additional Keyboard Functionality**, **Alt C**. Click "Download" from your browser. When asked "Open with..." select cb2Bib. cb2Bib will be launched if no running instance is found. If already running, it will place the downloaded reference to the clipboard, and it will start processing. Make sure your running instance is aware of clipboard changes. See **Buttons Functionality**. For convenience, the shell script [c2bimport](#), and the desktop config file [c2bimport.desktop](#) are also provided.

- **Adding documents**

PDF and other documents can be added to the BibTeX reference by dragging the file icon and dropping it into the cb2Bib's panel. Optionally, document files, are renamed to a DocumentID filename and moved to a desired directory as a personal article library (See **Configuring Documents** section). Linked to a reference documents correspond to the BibTeX tag [file](#). Usual reference manager software will permit to retrieve and visualize these files. Download, copy and/or moving is scheduled and performed once the reference is accepted, e.g., once it is saved by pressing **Save Reference** button.

- **Multiple retrieving from PDF files**

Multiple PDF or convertible to text files can be sequentially processed by dragging a set of files into cb2Bib's PDFImport dialog. By starting the processing button, files are sequentially converted to text and send to cb2Bib clipboard panel for reference extraction. See **PDF Reference Import** for details.

- **Journal-Volume-Page Queries**

Takes input Journal, Volume, and first page from the corresponding edit lines and attempts to complete the reference. Additionally, queries consider [title](#), [DOI](#), and an [excerpt](#), which is a simplified clipboard panel contents. See **Configuring Network** section, the distribution file [netqinf.txt](#), and **Release Note cb2Bib 0.3.5** for customization and details.

- **BibTeX Editor**

cb2Bib includes a practical text editor suitable for corrections and additions. cb2Bib capabilities are readily available within the editor. E.g., the reference is first sent to cb2Bib by selecting it, and later retrieved from cb2Bib to the editor using 'right click' + 'Paste Current BibTeX'. Interconversions Unicode <-> LaTeX, long <-> abbreviated journal name, and adding/renaming PDF files are easily available. BibTeX Editor is also accessible through a shell command line. See **cb2Bib Command Line** and **Embedded File Editor**.

## **Buttons Functionality**

- **About**

About cb2Bib, bookmarks, and online help.

- **Configure**

Configure cb2Bib. See **Configuration** section.

- **Search references**

Opens the cb2Bib's search dialog. The search is performed either on the current BibTeX file, or for all BibTeX files on the current directory. Optionally, the search is extended to reference's files. Hits are displayed on an editor window. See **Search BibTeX and PDF Document Files**. See also **Configuring Utilities** section to configure the external to text converter.

- **PDFImport**

Launches cb2Bib's PDFImport window. Files dragged into PDFImport window are sequentially translated to text and sent to cb2Bib clipboard panel. cb2Bib automatic and manual capabilities are then easily available to extract and supervise reference extractions. See **PDF Reference Import**.

- **Exit**

Exits cb2Bib.

- **Dis/Connect Clipboard**

Toggles automatic cb2Bib and desktop clipboard connection. While the automatic cb2Bib-clipboard connection permits reducing keystrokes, the disconnected mode is needed in cases where multiple mouse selections or copies are required to complete a reference extraction. See also **Release Note cb2Bib 0.4.1** and **Release Note cb2Bib 0.2.1** if you experience problems with this feature.

- **Network Reference Query**

Starts Network Query. It usually takes input Journal, Volume, and first page from the corresponding edit lines and attempts to complete the reference. See **Configuring Network** network section to customize querying. See the distribution file [netqinf.txt](#) and also **Release Note cb2Bib 0.3.5** for the details.

- **View BibTeX Reference**

View current reference as will be output to the BibTeX file. Eventual manual changes should be done on the item line edit.

- **Save Reference**

Inserts the current bibliographic reference to the output BibTeX file. This action decides whether or not a reference is accepted. Scheduled actions such as PDF downloading, copying or renaming will be performed at this time.

- **Open BibTeX File**

Opens the current BibTeX output file. Right click within the BibTeX Editor window for its particular functionality. See also **Embedded File Editor**.

## **Additional Keyboard Functionality**

Most keyboard shortcuts are customizable. See **Configuring Shortcuts**. In the following, default shortcuts are used to describe functionality.

- **Alt A**

Starts **cb2Bib Annote**. Specify the note's filename in the dialog. A new note is created if the file name does not exist. The cb2Bib Annote is opened as a separate program. Exiting cb2Bib will not exit the note's viewer. On the viewer, pressing key E launches the default text editor. The viewer will track the editor, and will update the note's display each time the editor saves it. The viewer's functionality is disabled if cb2Bib was not compiled and linked against QtWebKit or QtWebEngine library. See **cb2Bib Command Line** to use Annote in command line mode.

- **Alt B**

Edits the Bookmarks and Network Query Info file [netqinf.txt](#).

- **Alt C**

Prepares cb2Bib's clipboard through a user specified external script or tool. Preparing is necessary to catch formatted references that can not be easily extracted using recognition patterns, or that are written in ambiguous formats. Many available scripts or specific user-written tools can be incorporated to cb2Bib through this external preparing capability. In addition, simple, one-line scripts can be used within PDFImport to provide, for instance, the journal name when missing from the PDF first page. The cb2Bib distribution contains the sample scripts [isi2bib](#) and [ris2bib](#) that convert ISI and RIS formatted strings to BibTeX. See **Configuring Clipboard** for details.

- **Alt D**

Deletes temporary BibTeX output file. This permits using cb2Bib output files as temporary media to transfer references to a preferred reference manager and preferred format. **Caution:** This feature is not intended for the users who actually store their references in one or several BibTeX files. Remember to import references prior to delete cb2Bib output file.

- **Alt E**  
Edits the regular expression file. It permits an easy access and modification of stored extraction patterns. New patterns are conveniently added to the regular expression file by using the **RegExp Editor** button functionality.
- **Alt F**  
Launches a file dialog for selecting the source file name for the BibTeX entry [file](#). Selected files are displayed either, as the actual source filename, or, as the target filename, depending on the file copy/rename/move settings. See **Configuring Documents**. Alternatively to **Alt F**, documents can be easily linked to a reference by dragging the document file and dropping it to the cb2Bib panel.
- **Alt I**  
Edits and optionally inserts the current regular expression pattern. See the **Extracting Data from the Clipboard** and **Regular Expression Editor** sections.
- **Alt J**  
Edits the Journal Abbreviations file.
- **Alt O**  
Opens the currently linked document for browsing. Documents can be easily linked to a reference by dragging the document file and dropping it to the cb2Bib panel, or with **Alt F**. Linked documents correspond to the BibTeX tag [file](#).
- **Alt P**  
Postprocess BibTeX output file. It launches a user specified script or program to postprocess the current BibTeX file. The cb2Bib distribution contains two sample scripts. One, [bib2pdf](#) is a shell script for running [latex](#) and [bibtex](#); this permits to check the BibTeX file for possible errors, and to easily produce a suitable output for printing. The other one, [bib2end.bat](#) is a batch script for running [bib2xml](#) and [xml2end](#), which converts references into Endnote format. See **Configuring BibTeX** for details.
- **Alt R**  
Restarts cb2Bib automatic engine. Takes input data not from the system clipboard but from the cb2Bib clipboard panel. This permits editing the input stream from poorly translated PDF captions, correcting for author superscripts, or helps in debugging regular expressions.
- **Alt W**  
Writes current reference to the source document file. This option is intended for writing and updating bibliographic metadata to document files without needing to use BibTeX files. Only local and writable files are considered.
- **Alt X**  
Check Repeated looks for existing references in the BibTeX directory similar to the current one. The search is done for exact cite ID, and for title and author field values, or, if empty, for booktitle and editor, using the **approximate string** search pattern. See also **Configuring BibTeX**.
- **F4**  
Toggles between Main and Other Fields reference edit tabs.
- **Esc**



Quits cb2Bib popup menu. The cb2Bib menu pops up each time a selection is made in the clipboard panel. This saves keystrokes in a normal bibliographic extraction. Press **Esc** or **Right Click** mouse button if you need to gain access to the editor cut/copy/paste functionality instead.

## Advanced Features

Advanced features, and processing and extraction details are described in the following sections:

- **Automatic Extraction: Questions and Answers**
- **Extracting Data from the Clipboard**
- **Processing of Author Names**
- **Processing of Journal Names**
- **Field Recognition Rules**
- **Regular Expression Editor**

Configuration information is described in the following sections:

- **Configuration**
- **Predefined Placeholders**

Utilities and modules are described in the following sections:

- **Search BibTeX and PDF Document Files**
- **Embedded File Editor**
- **PDF Reference Import**
- **Reading and Writing Bibliographic Metadata**
- **cb2Bib Command Line**
- **cb2Bib Annote**
- **cb2Bib Citer**

## Requirements

### Compilation

To compile cb2Bib, the following libraries must be present and accessible:

- Qt 5.7.0 or later from **Qt Project**. On a Linux platform with Qt preinstalled, make sure that the [devel](#) packages and Qt tools are also present.
- QtWebKit or QtWebEngine library (optional) to compile cb2Bib Annote viewer. No special action or flag is needed during compilation.

- Compression libraries **LZ4** or **LZO** (optional). To chose a particular one, type `configure --enable-lz4` or `configure --enable-lzo`. On machines with SSE4 instruction set, the **LZSSE** compressor can be used in place of LZ4 and LZO, by typing `configure --enable-lzsse`. If none of the above compressors were appropriate on a particular platform, type `configure --enable-qt-zlib` before compiling.
- X11 header files if compiling on Unix platforms. Concretely, headers `X11/Xlib.h` and `X11/Xatom.h` are needed.
- The header files `fcntl.h` and `unistd.h` from `glibc-devel` package are also required. Otherwise compilation will fail with `'::close' undeclared`.

## **Deployment**

Although not needed for running cb2Bib, the following tools extend cb2Bib applicability:

- **MathJax**, available at <https://www.mathjax.org>, for displaying mathematical notation. Simply, download and unzip it in a desired directory. See **Configuring Annote**.
- **ExifTool**, version 7.31 or later, available at <https://exiftool.org>, for metadata insertion.
- **pdftotext**, found packaged as **xpdf**, and downloadable from <https://www.xpdfreader.com/download.html>.
- The **bib2xml** and **xml2end BibUtils**, for the postprocessing script `bib2end.bat` on Windows platforms.
- LaTeX packages, for checking BibTeX files correctness and for references printing through the shell script `bib2pdf`.

## **Credits and License**

The cb2Bib icons are taken from the *Oxygen*, *Crystal SVG*, and *Noia* icon sets, to be found at the **KDE Desktop Environment**. Several people has contributed with suggestions, bug reports or patches. For a detailed list of acknowledgments see the **Change Log File**.

The cb2Bib program is licensed under the terms of the **GNU General Public License** version 3.

***Last updated on 2025-11-02.***

*First released version 0.1.0 on 2004-06-29.*

© 2004-2025 Pere Constans

## **Installation**

## Installing on Unix systems (tarball)

The following is the general, platform independent install procedure.

- Unpack the distribution file:

```
tar -xzf cb2bib-2.0.2.90.tar.gz
```

- Move to cb2Bib directory:

```
cd cb2bib-2.0.2.90
```

- Type the following commands:

```
./configure --prefix /usr/local  
make  
make install
```

Installation is now complete.

**Note:** If the `./configure` step would fail while having the appropriate Qt libraries and utilities installed, try `qmake` instead of `./configure`, and configure manually the required file directories once cb2Bib first starts.

To uninstall type `make uninstall` from within the cb2Bib compilation directory.

## Installing on Unix systems (RPM)

To build an appropriate RPM for your platform, type, e. g.,

```
rpm -rebuild -target=i686 cb2bib-2.0.2.90-1.src.rpm
```

or a distro-dependend, equivalent command (perhaps `rpmbuild`). This will compile cb2Bib and build the required binary RPM (often placed at the `/usr/src/packages/RPMS/i686` directory). See also **Release Note cb2Bib 0.6.90** regarding `QTDIR` environment if having compilation problems. Once the binary RPM is build, installation is as follows.

To install your RPM binary, simply type

```
rpm -Uhv cb2bib-2.0.2.90-1.i686.rpm
```

To uninstall, type

```
rpm -e cb2bib-2.0.2.90
```

## Installing on Debian systems (deb)

To install cb2Bib, first make sure that you are actually using the packages for the proper Debian suite, as configured in the `/etc/apt/sources.list` file.

Next, issue the following commands as root,

```
apt-get update
```

```
apt-get install cb2bib
```

to resolve all required dependencies and install the program.

## Installing on Windows systems

On Windows platforms installation is simple. Just launch the Windows Installer

```
cb2bib-2.0.2.90-install.exe
```

and follow the installation wizard indications. To uninstall, click the 'Uninstall' icon.

## Installing on Mac OS X

To install cb2Bib from its sources, make sure you have the following build tools on your system:

- Qt toolkit version 5.7.0 or later (qt-mac-\*.dmg):

<https://www.qt.io/download-dev>

- XCode from Apple:

<https://developer.apple.com/xcode/>

- bin-utils via darwinports

- **Buiding with make/Makefile:**

Type on a shell window:

```
tar -xzf cb2bib-2.0.2.90.tar.gz
cd cb2bib-2.0.2.90
./configure --prefix /Applications/cb2Bib --qmakepath
```

```
/Developer/Tools/Qt/qmake  
make  
make install
```

- **Buiding with make/Makefile (no configure and no external compression):**

Type on a shell window if `configure` fails:

```
tar -xzf cb2bib-2.0.2.90.tar.gz  
cd cb2bib-2.0.2.90  
/Developer/Tools/Qt/qmake -config use_qt_zlib  
make
```

- **Buiding with XCode:**

Type on a shell window:

```
tar -xzf cb2bib-2.0.2.90.tar.gz  
cd cb2bib-2.0.2.90  
/Developer/Tools/Qt/qmake cb2bib.pro -spec macx-xcode
```

Open `cb2bib.xcodeproj` with XCode and build from there.

**Note:** It has been reported that qmake does not make usable XCode projects from subdirs. It is possible to produce one single `.pro` file for the whole project, by typing `qmake -project -r` to create a base `.pro` file. An example and detailed instructions can be found at `./qmake/cb2bib-osx.pro`.

See also **Configuration**.

## Configuration

### Configuring Annote

- **Annote Cascading Style Sheet (CSS)**  
This file contains the style sheet to personalize the appearance of the HTML notes generated by the cb2Bib. The cb2Bib distribution includes the `tex2html.css` file as a CSS template.
- **MathJax Header File**  
The mathematical notation in the text notes is displayed by **MathJax**, the successor of the **jsMath** Java Script library. Its location and configuration must be specified inside the HTML files in order to be known by the browser. Check and eventually edit the distribution file `tex2html_local_mathjax_header.html`. Should web script be

preferred set script source to

<https://cdn.mathjax.org/mathjax/latest/MathJax.js>.

- **Include CSS in HTML**

Styles for the notes will be included, if checked, into the HTML file. In this way, all the information, text and layout, is contained in one single file.

- **Use relative links**

If checked, linked local files will be set relative to the current HTML document.

- **Annote Viewer Fonts**

Selects default and monospaced fonts for the Annote viewer. Changes in the fonts might need restarting the viewer unless using some of the latest QtWebKit libraries. The viewer is disabled if cb2Bib was not compiled and linked against QtWebKit or QtWebEngine. Note also that fonts specified in the CSS prevail over this selection.

## Configuring BibTeX

- **Cite ID Pattern**

Specifies the pattern for formatting cite's ID. Predefined placeholders are available as a context menu, by right-clicking this edit line. Placeholders will be substituted by the actual reference field values. See **Cite and Document ID Placeholders** for descriptions.

- **Author and Editor Name Format**

Sets Authors and Editor names in abbreviated or full form, if the latter is available.

- **Journal Name Format**

Sets output journal names in abbreviated or full form. Note that this feature only works if journal names are found in the [Journal Abbreviation List file](#). See **Processing of Journal Names**.

- **Number Separator**

Sets number separator, e.g., ' - ' or ' – '. Spaces count. It applies to [pages](#), multiple [volume](#), [number](#), and [year](#) cases.

- **Cite Command Pattern**

Specifies the pattern for formatting cite command. Predefined command patterns for LaTeX and Markdown (see **Pandoc User's Guide**) are available in the line context menu. Other, customized command patterns are also available, see **Cite Command Placeholders** for descriptions.

- **Convert entry strings to LaTeX**

If checked, cb2Bib converts special characters to LaTeX commands. Most BibTeX import filters do not process LaTeX escaped characters. Therefore, keeping this box unchecked can be appropriate when using cb2Bib as a temporary media to transfer references to non BibTeX reference managers.

- **Set 'title' in double braces**

If checked, it writes extra braces in title. This will keep capitalization as is, when processed by BibTeX.

- **Postprocess 'month'**

If checked, cb2Bib elaborates the 'month' string on a BibTeX fashion. E.g., 'April 1' becomes "'1~" # apr'. No processing is done if the input string is not written in English.

- **Try Heuristic Guess if recognition fails**

If checked, when automatic recognition fails, cb2Bib tries to catch some of the fields of the reference through an heuristic set of rules. See **Field Recognition Rules**.

- **Check Repeated On Save**

If checked, cb2Bib looks for existing references in the BibTeX directory similar to the one being saved. The search is based on exact cite ID match, or on reference contents, by considering title and author field values, or, if empty, booktitle and editor, and using the **approximate string** search pattern. If similar references are found, the current reference is not saved, and the similar ones are displayed. Pressing the **save button one second time will proceed to actually saving the current reference**. Note that this feature is not applied in command line mode, when using `cb2bib -txt2bib` or `cb2bib -doc2bib`. See also **Additional Keyboard Functionality**.

- **External BibTeX Postprocessing**

Use this box to select a BibTeX postprocessing external tool. The name of the executable, the command arguments and the output file extension are required fields. Arguments, any number, are passed to the executable. For the sake of generality, it is required to specify the `%finput` and `%foutput` placeholders. The first one is later substituted by the current BibTeX filename. The second one is substituted by the current filename with the specified output extension. **Caution:** Be careful if using the same file extension for input and output, e.g., using `bib` because you want to use a beautifier or a sorting script. cb2Bib has no control on external disk modifications. Therefore, if the script failed, the input data would possibly be lost. See also **Additional Keyboard Functionality**.

## Configuring Clipboard

- **Replace/Remove from Input Stream**

If checked, input stream is preprocessed by performing a customizable set of string substitutions/removals. This option is mainly intended to remove image HTML `alt` tags. Although not visible, `alt` tags reach the clipboard when selecting and copying text. Author lists with email icons may contain `alt` strings that would interfere with the author names processing. In addition, this option is also appropriate to help translating special characters to Unicode and LaTeX. Use it carefully, as to avoid unwanted or unsuspected substitutions. See also **Extracting Data from the Clipboard**.

- **External Clipboard Preparsing Command**

Prepares input stream through an external, user-defined tool. Use the box bellow to specify its name and path. cb2Bib executes the command `tool_name tmp_inputfile tmp_outputfile`. You might consider a wrapper shell script to fulfill this particular syntax requirement. Two examples, `isi2bib` and `ris2bib` are provided. To test them, make sure the **BibUtils Package** is available on your machine. Otherwise, modify these scripts according to your needs. See also **Additional Keyboard Functionality**, **Extracting Data from the Clipboard**, and the examples **ISI Reference Format** and **RIS Reference Format**.

- **Perform always, as part of an automatic extraction**

Performs preparsing each time the recognition engine is invoked. **Caution:** cb2Bib, when not in disconnected mode, starts the recognition engine each time the clipboard

changes. Therefore, it might send arbitrary data to the external parsing tool. The tool might not be prepared to handle '**any data**' and might fall into a sort of '**infinite loop**'. cb2Bib kills the external tool after a reasonable waiting. But, if the tool is called through a wrapper script, killing the script will not end the tool itself. Therefore, **check this box only when needed**. If you write your own preparer, design it as to write no data to output file whenever it can not process an input properly. When the preparer produces no data, cb2Bib sends instead the input stream to the recognition engine. In this way, preparsing and normal cb2Bib functioning will work harmoniously.

- **Do not show log**

If unchecked, the external process messages, and the input and output streams are shown in a log window. Showing output logs is useful for debugging purposes.

- **Add document metadata to Input Stream**

When checked, if the document linked to a reference contains relevant metadata, then metadata will be added to the current clipboard contents. The metadata is included at the time of adding the document to the current reference, e. g., when dropping a file into the cb2Bib panel. If the document has BibTeX information, cb2Bib will automatically set the corresponding fields. If it has not, but relevant bibliographic information is found, this data is only added to the clipboard panel. To insert it in the edit lines, activate the Heuristic Guess (Alt+G). The option **Prepend** or **Append** to the clipboard contents is provided for conveniently writing regular expressions considering metadata contents. File documents are linked to the references by the BibTeX tag 'file'. See also **Reading and Writing Bibliographic Metadata**.

## Configuring Documents

- **Rename and Copy/Move document files to Directory**

If selected, each file 'drag and dropped' onto the cb2Bib main window is renamed to [DocumentID.pdf](#) (or DocumentID.ps, DocumentID.dvi, etc.) and moved to the storage directory. If unselected, the file URL is written to the [file](#) BibTeX keyword, without any renaming or moving of the file. The actual copy/move action is scheduled and performed once the reference is accepted, e.g., once it is saved.

- **Copy or Move document files**

Choose whether copy or move Network Files dropped onto the cb2Bib main window. See also **Use External Network Client**.

- **Set directory relative to the BibTeX File Directory**

If checked, the document file is copied/moved to the current BibTeX file directory. If the Documents Directory box contains a **relative directory** it will be added to the file name. For example, if it contains [articles](#), files will be copied to [/current\\_bibtex\\_path/articles/](#). An absolute path in the Documents Directory box will be ignored in this context. Note that the file dialog returns here relative file addresses. Consequently, only the necessary portion of the full name, instead of the fullpath filename, is written to the BibTeX entry. File retrieving from within the cb2Bib browser will be relative to the BibTeX file absolute location.

Use this option if you plan to store in a same or a related directory the BibTeX and document files. This option is appropriate for storing bibliographic collections in



removal devices. Likewise, when cb2Bib is launched in USB mode, by means of the command line switch `-conf`, the alternate option is not available. See **Release Note cb2Bib 0.8.4** and **Export Selected Entries**.

- **Insert BibTeX metadata to document files**

If checked, cb2Bib will write bibliographic metadata to the linked document, once the current reference is accepted and saved. See also **Reading and Writing Bibliographic Metadata**.

- **Document ID Pattern**

Specifies the pattern for formatting the document's filenames. Predefined placeholders are available as a context menu, by right-clicking this edit line. Placeholders will be substituted by the actual reference field values. See **Cite and Document ID Placeholders** for descriptions.

- **ExifTool Metadata writer**

cb2Bib uses **ExifTool** for writing bibliographic metadata to the attached documents. Select here the ExifTool path name. On Windows, remember renaming `exiftool(-k).exe` to `exiftool.exe` for command line use. See also **Writing Metadata**.

## Configuring Files

- **Journal Abbreviation List File**

This file contains a list of journal names equivalences: a capital-letter acronym, standard abbreviated form, and full name of the journal. If an input journal name is recognized, cb2Bib will use the standard abbreviated form for the `journal` bibkey. If your usual journal were not within the distributed, default `abbreviations.txt`, you could edit this file, or point to a personalized abbreviation file. **Note:** Changes in the abbreviation file only take place after restarting cb2Bib. See **Processing of Journal Names**.

- **Regular Expression List File**

The cb2Bib distribution includes the file `regexps.txt` with a few set of rules for reference extraction. This includes most of the scientific literature. Extracting from email alerts or publisher abstract pages is a *volatile* task. Information does not follow a standardized structure. Extraction pattern may then change as often as the web design needs to. Besides, such extraction from the clipboard is system dependent, in a way that produces different formatting of the text copies on different systems. You can use your personalized `regexps.txt` file, for testing, debugging -regular expressions are reloaded each time the automatic recognition engine executes-, and fulfilling your particular extraction needs.

- **Bookmarks and Network Query Info File**

The cb2Bib distribution includes the file `netqinf.txt` that contains bookmarks data, and server related information for bibliographic querying. Note that cb2Bib treats bibliographic queries as generalized net bookmarks. This allows accessing almost any online bibliographic resource. Check this file for implementations details and customization.

- **Browser Cascading Style Sheet (CSS)**

This file contains the style sheet to configure the appearance of the bibliographic references when viewed in browser mode. The cb2Bib distribution includes the

[references.css](#) and [references-dark.css](#) file as a CSS examples.

- **Part Of Speech (POS) Lexicon**

This box must contain the address to the cb2Bib distribution file [lexicon.pos](#). This file contains a set of patterns and related POS information required for indexing documents, i. e., to extract keywords from documents for the c2bCiter module.

- **Search In Files Cache Directory**

Directory containing internal data for Search In Files functionality. If an existing directory is selected cb2Bib will write all internal data on it. If otherwise, cache data will be written on the same directory from where BibTeX are searched. It might be, therefore, convenient to group all this files in a separate directory that does not need to be backup, and that can easily be deleted whenever desired.

## Configuring Fonts

- **Font Selector**

Selects the main window and editor font family and size.

- **Context Colors**

Doubleclick on context color items to select syntax highlighter font color. Besides syntax highlighting, and to ease manual bibliographic extractions, cb2Bib has the following coloring convention. 'cb2Bib irrelevant text' colors non-word, non-digit, and cb2Bib's internal tags. 'cb2Bib relevant text' refers to the reference's year. 'cb2Bib highly relevant' attempts to guess text sectioning, highlighting 'abstract', 'introduction', and 'keywords'.

## Configuring Network

- **Use External Network Client**

cb2Bib manages local and network files in an equivalent manner. Network file retrieving, however, requires sometimes password and/or cookies administration. The KDE desktop incorporates [kfmclient](#) utility. A command [kfmclient \(copy|move|exec\) source \[destination\]](#) permits copying or moving files, with [kfmclient](#) taking care of advanced browsing preferences. By checking this box, cb2Bib will use the specified file manger client.

- **Use Proxy**

If checked, cb2Bib will access the network through a proxy host. Set the Host name, the Port, and the proxy Type. A login dialog will appear if the proxy requires authentication. Login data is not stored, it must be entered at each session.

- **Perform Network Queries after automatic reference extractions**

Network queries can be used to complete a partial reference extraction. For instance, provided a reference 'J. Name, 25, 103' and an appropriate pattern to extract it, cb2Bib will attempt to complete the reference automatically. No query is performed if automatic reference extraction was tagged as BibTeX.

- **Download document if available**

If checked, cb2Bib downloads document files to the directory specified in **Rename and Copy/Move document files to Directory**. See also the file [netqinf.txt](#) for details.

Download is scheduled and performed once the reference is accepted, e.g., once it is saved. Note that when document file is local, e.g., when PDFImport or switch `-doc2bib` is used, no document is downloaded.

- **Keep Query temporary files (Debug Only)**

cb2Bib may use up to three temporary files to perform a network query. If this box is checked, the temporary files are not deleted. This facilitates the testing and customization of the information file `netqinf.txt`.

## Configuring Shortcuts

- Customizes most key sequences for actions shortcuts. Concretely, cb2Bib specific actions are configurable, but not standard actions such as 'Open', 'Exit', 'Copy', or 'Paste', which are already predefined to the standard, specific key sequences for each platform. Shortcuts are customizable for the cb2Bib main panel, editor, and reference list actions. Single-key shortcuts, i.e., for manual reference extraction and shortcuts in c2bCiter, are non-configurable, since they they closely map non-translatable BibTeX keywords.

## Configuring Utilities

- **To plain text converter**

Selects the external `some_format_to_text` tool that cb2Bib uses to convert document files prior to reference extraction and searching. cb2Bib executes the line command `converter [options] inputfile tmp_output.txt`, where `[options]` are user defined arguments. As a default, cb2Bib launches `pdf2cb`, a modified PDF to text utility found in the XPDF package. Modifications are available at `xpdf/` directory in the cb2Bib sources. Default arguments are `-q -f 1 -l 1` to extract only the first, title page when used within PDFImport, and `-q`, to convert the complete document when used within Search in Files. Appropriate for PDFImport could also be a document metadata extractor. Often metadata contains structured information regarding document authors, title, and source. A simple shell script wrapper could be the following `any2text_search`:

```
#!/bin/csh
# Convert documents to text according to filename extension
# any2text_search input_fn.ext output_fn.txt
set ext = $1:e
if ( $ext == 'djvu' ) then
    /usr/bin/djvutxt "$1" "$2"
    if ($status) exit 1
else if ( $ext == 'chm' ) then
    (/usr/local/bin/archmage -c text "$1" "$2") >& /dev/null
    if ($status) exit 1
else
    # If using pdf2cb
    /path/to/pdf2cb -q "$1" "$2"
    # If using pdftotex
    # /usr/bin/pdftotext -enc UTF-8 "$1" "$2"
    if ($status) exit 1
```

```
endif
```

## Search BibTeX and PDF Document Files

### Search Functionality

- **Search pattern**

Patterns and composite patterns can be either **approximate strings**, strings, contexts, regular expressions, or wildcard filters. Patterns admit Unicode characters. The scope of each pattern can be the reference as a whole or be focused on a particular reference field. The fields [year](#), [file](#), and [journal](#) are treated specifically. The field [year](#) has the qualifiers [Exact](#), [Newer](#), and [Older](#). The field [file](#) can optionally refer to either the filename or the contents of such a file. Finally, for [journal](#), the input pattern is duplicated to the, if available, journal fullname, and they two are checked against the [journal](#) actual field contents and, if available, its expanded contents. For example, typing 'ijqc' retrieves all references with [journal](#) being 'Int. J. Quantum Chem.'. Or, typing 'chemistry' retrieves any of 'J. Math. Chem.', 'J. Phys. Chem.', etc. This expansion is not performed when the pattern scope is set to [all](#).

- **Search scope**

By default, searches are performed on the current BibTeX output file. If **Scan all BibTeX files** is checked the search will extend to all BibTeX files, extension .bib, present in the current directory. It might be therefore convenient to group all reference files in one common directory, or have them linked to that directory. When **Scan linked documents** is checked, and one or more pattern scope is [all](#) or [file](#), the contents of the file in [file](#) is converted to text and scanned for that given pattern. See **Configuring Utilities** section to configure the external to text converter.

- **Search modifier**

cb2Bib converts TeX encoded characters to Unicode when parsing the references. This permits, for instance, for the pattern 'Møller' to retrieve either 'Møller' or 'M{\o}ller', without regard to how the BibTeX reference is written. By checking **Simplify source**, the reference and the converted PDF files are simplified to plain ASCII. In this way, the pattern '\bMoller\b' will hit any of 'Møller', 'M{\o}ller', or 'Moller'. Additionally, all non-word characters are removed, preserving only the ASCII, word structure of the source. Note that source simplification is only performed for the patterns whose scope is [all](#) or [file](#) contents, and that and so far, cb2Bib has only a subset of such conversions. Implemented TeX to Unicode conversions can be easily checked by entering a reference. The Unicode to ASCII letter-only conversion, on the other hand, is the one that cb2Bib also uses to write the reference IDs and, hence, the renaming of dropped files. cb2Bib can understand minor sub and superscript formatting. For instance, the pattern 'H2O' will retrieve 'H<sub>2</sub>O' from a BibTeX string  $H_{2}O$ .

## Contextual Search

A convenient way to retrieve documents is by matching a set of keywords appearing in a close proximity context, while disregarding the order in which the words might had been written. cb2Bib considers two types of contextual searches. One flexibilizes phrase matching only at the level of the constituting words. It is accessed by selecting **Fixed string: Context** in the pattern type box. The other one, in addition, stems the supplied keywords. It is accessed by selecting **Context**. By way of stemming, the keyword *analyze*, for example, will also match *analyse*, and *aluminum* will match *aluminium* too.

The syntax for **Context** type patterns is summarized in the following table:

Operator	Example	Expansion
space	contextual search	contextual AND search
	contextual search matching	contextual AND (search match)
+	contextual search +matching	contextual AND (search \bmatching\b)
_	contextual_search	contextual.{0,25}search
-	non-parametric	non.{0,1}parametr
.	non.parametric	non.{0,1}parametr
Diacritics and Greek letters:		
	naïve search	(naïve naive) AND search
	kendall tau	kendall AND (tau τ)

In the above examples, operator space **AND** means match words in any order. Operator **\_** preserves word order, and operator **+** prevents stemming and forces exact word match. Operator **-** considers cases of words that might had been written either united, hyphenated, or space separated. Diacritics are expanded if the diacritic mark is specified. This is, *naïve* will not match *naive*. On the other hand, Greek letters are expanded only when typed by name.

## Notes

- cb2Bib uses an internal cache to speed up the search of linked files. By default data is stored as `current_file.bib.c2b`. It might be more convenient, however, to setup a temporary directory out of the user data backup directories. See **Search In Files Cache Directory** in **Configuring Files**. When a linked file is processed for the first time, cb2Bib does several string manipulations, such as removing end of line hyphenations. This process is time consuming for very large files.
- The **approximate string** search is described in reference <https://arxiv.org/abs/0705.0751>. It reduces the chance of missing a hit due to transcription and decoding errors in the document files. Approximate string is also a form of serendipitous information retrieval.

# cb2Bib Command Line

The complete listing of command line uses follows.

```
Usage: cb2bib
       cb2bib [action] [filename1 [filename2 ... ]] [--conf [filename.conf]]

Actions:

--configure [filename.conf]           Edit configuration

--bibedit [filename1.bib [filename2.bib ... ]] Edit/browse BibTeX files
--citer [filename1.bib [filename2.bib ... ]] Start cb2Bib citer
--import tmp_reference_filename       Import reference, usually from ad hoc websites

--doc2bib fn1.doc [fn2.doc ... ] reference.bib Extract reference from document file
--txt2bib fn1.txt [fn2.txt ... ] reference.bib Extract reference from text file

--index [bibdirname]                 Extract keywords from document files

--html-annotate filename.tex          Convert annotate file to HTML
--view-annotate filename.tex          Convert and visualize annotate file
--view-annotate filename.tex.html     Visualize annotate file

Switches:

--conf [filename.conf]               Use configuration file
--sloppy                             Accept guesses in automatic reference extraction

Examples:

cb2bib                               Start cb2Bib extraction panel
cb2bib --import tmp_reference_filename Import reference
cb2bib --bibedit filename.bib        Edit BibTeX filename.bib
cb2bib --conf                        Start cb2Bib in USB mode
cb2bib --doc2bib *.pdf references.bib Extract references from PDF title pages

Notes:

-Use switch --conf to particularize specific settings for specific actions.
-The file cb2bib.conf must be readable and writable. If it does not exist, cb2Bib will create one based on predefined defaults.
-If starting cb2Bib from a removable media, use the command 'cb2bib --conf' without configuration filename. Settings will be read from and written to /cb2bib/full/path/cb2bib.conf, being therefore independent of the mounting address that the host computer will provide.
-To import references from a browser select when asked c2bimport, which expands to 'cb2bib --import %f'. The browser will provide the temporary reference filename.
-A number of factors influence the reliability of automatic extractions. Consider writing customized regular expressions and network queries, and use metadata when available.

Important:

-The commands --doc2bib and --txt2bib do not append the references to the references.bib. They create a new file, or silently overwrite it if already exists.
```

- **Note:** On Windows use [c2bconsole](#) instead of [cb2bib](#). See [Release Note cb2Bib 1.3.0](#).
- **Note:** If using reference extraction command, see [Automatic Extraction: Questions and Answers](#).

# cb2Bib Annote

The cb2Bib Annote module is named after the BibTeX key [annote](#). Annote is not for a 'one reference annotation' though. Instead, Annote is for short notes that interrelate several references. Annote takes a plain text note, with minimal or no markup, inserts the bibliographic citations, and converts it to a HTML page with links to the referenced documents.

From within cb2Bib, to write a note, type [Alt+A](#), enter a filename, either new or existing, and once in Annote, type [E](#) to start the editor. Each time you save the document the viewer will be updated. For help on Annote's syntax type [F1](#). If cb2Bib was compiled without Annote's Viewer, typing [Alt+A](#) will start the editor and HTML viewing will be committed to the default web browser.

From the command line, typing

```
cb2bib --html-annote annote.tex
```

will produce the HTML file [annote.tex.html](#).

See also **Configuring Annote** and **cb2Bib Command Line**.

## Annote Syntax

The resulting HTML file [annote.tex.html](#) can be seen at **cb2Bib Annote**.

```
% annote.tex

%\c2b_bibtex_directory{/home/constans/Documents/BibReferences}
%\c2b_makeindex

\newcommand{\RR}{\mathbb{R}}
\newcommand{\mnial}[3]{(#1 - #2)^#3}

\title{cb2Bib Annote}

\begin{abstract}
This documents describes cb2Bib Annote. It succinctly lists Annote's minimalists
syntax.
\end{abstract}

\section{cb2Bib Directives}

\subsection{Make Index}

\begin{verbatim}
%\c2b_makeindex
\end{verbatim}

\subsection{BibTeX Directory}

\begin{verbatim}
%\c2b_bibtex_directory{/home/constans/Documents/BibReferences}
\end{verbatim}

\section{Simple Markup}

\subsection{Uniform Resource Locator}
```

```

\begin{verbatim}
- URL: https://www.molspaces.com/cb2bib/doc/c2bannote/
- Named URL: https://www.molspaces.com/cb2bib/doc/c2bannote/[cb2Bib Annote]
- On a blank window: _https://www.molspaces.com/cb2bib/doc/c2bannote/[cb2Bib Annote]
\end{verbatim}

\subsubsection{Example}

- URL: https://www.molspaces.com/cb2bib/doc/c2bannote/

- Named URL: https://www.molspaces.com/cb2bib/doc/c2bannote/[cb2Bib Annote]

- On a blank window: _https://www.molspaces.com/cb2bib/doc/c2bannote/[cb2Bib Annote]

\subsection{Bibliographic Citations}
\begin{verbatim}
\cite {key}
\end{verbatim}

\subsubsection{Example}

Citing cb2Bib \cite{cb2bib_key}.

\section{LaTeX Markup}

\subsection{Document Sections}

\begin{verbatim}
\title{Title string}
\end{verbatim}

\begin{verbatim}
\section{Section string}
\end{verbatim}

\begin{verbatim}
\subsection{Section string}
\end{verbatim}

\begin{verbatim}
\subsubsection{Section string}
\end{verbatim}

\subsection{Document Environments}

\begin{verbatim}
% env = abstract, equation, itemize, and verbatim

\begin{env}
\end{env}
\end{verbatim}

- Note. equation rendering requires MathJax \cite{jsmath_key, mathjax_key}

\subsubsection{Examples}
\begin{verbatim}
\begin{itemize}
\item Description 1
\item Description 2
\end{itemize}
\end{verbatim}

\begin{itemize}
\item Description 1
\item Description 2
\end{itemize}

\begin{verbatim}
\begin{equation}
\int_D ({\nabla\!\cdot} F)dV=\int_{\partial D} F\cdot ndS
\end{equation}
\end{verbatim}

\begin{equation}
\int_D ({\nabla\!\cdot} F)dV=\int_{\partial D} F\cdot ndS
\end{equation}

```



```

\subsection{Mathematical Macros}

\begin{verbatim}
\newcommand{name}[number of arguments]{definition}
\end{verbatim}

\subsubsection{Example}

\begin{verbatim}
\newcommand{\RR}{\mathbb{R}}
\newcommand{\mnial}[3]{(#1 - #2)^#3}

... a subset of  $\mathbb{R}$  values ... the monomial is  $\mnial{a}{x}{2} > 0$ 
for  $x \neq 0$ , and  $\mnial{a}{x}{3} \mnial{c}{x}{3}$  for  $x < a$  \and  $x < c$  or
 $x > a$  \and  $x > c$  ...
\end{verbatim}

... a subset of  $\mathbb{R}$  values ... the monomial is  $\mnial{a}{x}{2} > 0$ 
for  $x \neq 0$ , and  $\mnial{a}{x}{3} \mnial{c}{x}{3}$  for  $x < a$  \and  $x < c$  or
 $x > a$  \and  $x > c$  ...

\section{MathJax Example}

\begin{verbatim}
% Example from https://www.mathjax.org/#demo

When  $a \neq 0$ , there are two solutions to  $(ax^2 + bx + c = 0)$  and they are
 $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ .
\end{verbatim}

When  $a \neq 0$ , there are two solutions to  $(ax^2 + bx + c = 0)$  and they are
 $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ .

```

## cb2Bib Citer

The cb2Bib Citer is a keyboard based module for inserting citation IDs into a working document. Conveniently, the command **cb2bciter**, or its expansion **cb2bib -citer**, can be assigned to a global, desktop wide shortcut key. This will provide an easy access to the citer from within any text editor. Pressing the shortcut turns on and off the citer panel. Once appropriate references are selected, pressing key C sets the citations either to the clipboard or to a LyX pipe, closes the citer panel, and returns keyboard focus to the editor.

By default, **cb2bciter** loads all references from the current directory, specified in the cb2Bib main panel. On the desktop tray, the cb2Bib icon indicates that the citer is running. Its context menu offers the possibility to load other files or directories, or to toggle full screen mode.

Search, filtering, navigation, and citation are keyword based. Pressing keys A, I, J, T, and Y sorts the references by author, included date, journal, title, and year, respectively. Key F initiates filtering, and Esc leaves filtering mode. References are selected when pressing enter. Key S toggles the current selection display, and Del clears the selection. The combination Shift + letter navigates through the rows starting by the letter.

Advanced filtering capabilities are available after indexing the documents. Document indexing, or term or key sentence extraction, is performed by clicking the tray icon menu

action Index Documents. Once indexing is done and after clicking Refresh, pressing K displays the document extracted keywords, and pressing G, the collection glossary of terms. On a keyword, pressing R display all documents indexed by the keyword. On a document, pressing R display related documents. Relatedness is set from keyword based similarity measures.

Key sequence Alt+C toggles clipboard connection on and off. When connection is on, the clipboard contents is set, each time it changes, as the filter string. This provides a fast way to retrieve a given reference while browsing elsewhere.

```
Usage:      cb2bib --citer [dirname1 [dirname2 ... ]]
           cb2bib --citer [filename1.bib [filename2.bib ... ]]
```

#### Display Keys

```
A      author - journal - year - title
I      included date - title
J      journal - year - author
T      title
Y      year - author - journal - title
```

#### Filter Keys

```
D      Delete last filter
F      Enter pattern filter mode
G      Toggle glossary of terms view
K      Toggle document keywords view
R      Display related documents
```

```
Left   Move to previous filtered view
Right  Move to next filtered view
```

#### Action Keys

```
C      Cite selected citations and close citer window
Del    Unselect all citations
E      Edit current citation's source
Enter  Select current citation
Esc    Exit filter mode or close citer window
O      Open current citation's file
S      Display the set of selected citations
V      Display document excerpts in keywords view
Shift+ Keyboard search navigation
U      Open current citation's URL
W      Write notes using Annote
```

#### Tray Icon Actions

```
F1      Citer help
Ctrl+O  Open BibTeX directory
Alt+O   Open BibTeX files
F5      Refresh
Ctrl+F  Search in files
Alt+L   Set Lyx pipe
F2      Toggle cb2Bib
Alt+C   Toggle clipboard
Alt+F   Toggle full screen
        Index documents
```

See also **Release Note cb2Bib 1.3.0**, **Release Note cb2Bib 1.4.0**, **Release Note cb2Bib 1.4.7**, **cb2Bib Command Line** and **cb2Bib Annote**.

# Examples

This set of examples considers only simple cases of predefined bibliographic formats, which at present are most frequent on the Web.

For complex cases that would require *ad hoc* regular expressions, see the cb2Bib's **Advanced Features**.

To test the examples, launch cb2Bib, and select the text within the boxes (type also Ctrl-C if running cb2Bib on a Windows machine). cb2Bib will extract the selected reference.

## Predefined Formats

- **BibTeX**
- **ISI Reference Format**
- **RIS Reference Format**

### BibTeX

```
@article{Efron,  
  title      = {{The Estimation of Prediction Error}},  
  author     = {Bradley Efron},  
  journal    = {Journal of the American Statistical Association},  
  pages      = {619 - 632},  
  volume     = {99},  
  number     = {467},  
  year       = {2004},  
}
```

Nowadays most authors and publishers websites export references to BibTeX format. This is the safest way to retrieve a reference with cb2Bib. Select from `@article` to the last `}`. cb2Bib imports the reference. Once imported, cb2Bib permits adding the abstract, keywords or renaming and relocating the document file.

### ISI Reference Format

```
PT Journal  
AU Kohn, W  
AU Becke, AD  
AU Parr, RG  
TI Density functional theory of electronic structure  
SO JOURNAL OF PHYSICAL CHEMISTRY  
BP 12974  
EP 12980
```

```
PG 7
JI J. Phys. Chem.
PY 1996
PD AUG 1
VL 100
IS 31
J9 J PHYS CHEM
ER
```

Example provided for testing [isi2bib](#) external preparsing script. See **Configuring Clipboard** for the setup details.

### **RIS Reference Format**

```
TY - JOUR
AU - A. J. Coleman
PY - 1963
TI - Structure of fermion density matrices
JO - Rev. Mod. Phys.
SP - 668
VL - 35
ER -
```

Example provided for testing [ris2bib](#) external preparsing script. See **Configuring Clipboard** for the setup details.

## **Additional Features**

### **Embedded File Editor**

cb2Bib contains a practical editor suitable to manipulate the files related to a cb2Bib session. Abbreviations, bookmarks, regular expressions, and BibTeX are readily available. On BibTeX mode, the editor contains a Reference List to ease file navigation, and to extend the editor functionality. The BibTeX editor can optionally be turned to a reference browser.

### **Bookmarks**

Bookmarks pointing to local or network files are available at the editor menu bar. They provide a fast access to external resources. Concretely, they are suitable for retrieving working documents, writing notes, or for accessing internet databases.

The editor bookmark functionality had been introduced in the cb2Bib version 0.9.3. Currently, bookmarks are set manually in the [netqinf.txt](#) file, see **Configuring Files**. Each bookmark is an entry line with the form

```
editorbookmark=Description|Target file name|Shortcut|Icon file name
```

having four fields, description, target file name, shortcut, and icon file name, separated with three (required) vertical bars |.

```
# Bookmark Examples:
# - A URL:
editorbookmark=URL Description|https://www.molspaces.com/cb2bib/doc/bibeditor||
# - A separator, which is a blank line:
editorbookmark=
# - A TeX document, which will be opened with its default application:
editorbookmark=Document Description|/home/doc/chapter1.tex||
```

## Update Documents Metadata

The Update Documents Metadata functionality is available at the Edit and context menus on the BibTeX editor mode. Documents referred in the BibTeX file tags are scanned for metadata. If the BibTeX reference does not coincide with the bibliographic metadata, the document metadata is updated. In this way, metadata is synchronized with the BibTeX file contents. A log window appears, and possible errors are reported. Reported errors are, non-existence of a document file, read-only files, mismatches between BibTeX references and the actual metadata (often due to HTML tags or other illegal characters in BibTeX), or that the writing to the document format is not implemented. Note that this process will update all documents referenced in the BibTeX file. **While this process is safe, it implies writing into the documents.** Therefore take the usual measures and backup your data. See also **Writing Metadata**.

## Export Selected Entries

Selected entries can be exported to a separate BibTeX document. Click on **File->Export Entries** menu option, and provide an export filename at the Export Dialog. Optionally, export will copy the document files attached to the citation. The copy of documents is similar to the cb2Bib 'rename/copy/move' general procedure. See **Configuring Documents**, on **Set directory relative to the BibTeX File Directory**, for copying options. Documents will not be overwritten: copying of existing documents is skipped. Possible issues are reported in the new document, as LaTeX comments.

## Right-Click Menu Functionality

Default Key	Action
Ctrl+F	Find in text
	Toggle word wrapping
	Selection to LaTeX

Default Key	Action
	Selection to Unicode
	Journals to full name
	Journals to alternate full name
	Journals to abbreviated name
	Journals to alternate abbreviated name
	Update documents metadata
Shift+Ins	Paste current BibTeX
Alt+P	Save and postprocess BibTeX file

### **Reference List Right-Click Menu Functionality**

Default Key	Action
Alt+C	Cite selected entries
	Open document file
	Browse by DOI
	Web search by Author
	Web search by Title
	Web search settings
	Clear entry selection
	Refresh list and browser

## **BibTeX Browser Right-Click Menu Functionality**

<b>Default Key</b>	<b>Action</b>
Alt+C	Cite selected entries
	Local search for selected text
	Web search for selected text
	Web search settings
	Clear entry selection
	Refresh list and browser

## **Additional Functionality**

Backspace	Deletes the character to the left of the cursor
Delete	Deletes the character to the right of the cursor
Ctrl+A	Selects all text
Ctrl+C	Copy the selected text to the clipboard
Ctrl+Insert	Copy the selected text to the clipboard
Ctrl+K	Deletes to the end of the line
Ctrl+V	Pastes the clipboard text into text edit
Shift+Insert	Pastes the clipboard text into text edit
Ctrl+X	Deletes the selected text and copies it to the clipboard
Shift+Delete	Deletes the selected text and copies it to the clipboard
Ctrl+Z	Undoes the last operation
Ctrl+Y	Redoes the last operation
LeftArrow	Moves the cursor one character to the left

Ctrl+LeftArrow	Moves the cursor one word to the left
RightArrow	Moves the cursor one character to the right
Ctrl+RightArrow	Moves the cursor one word to the right
UpArrow	Moves the cursor one line up
Ctrl+UpArrow	Moves the cursor one word up
DownArrow	Moves the cursor one line down
Ctrl+Down Arrow	Moves the cursor one word down
PageUp	Moves the cursor one page up
PageDown	Moves the cursor one page down
Home	Moves the cursor to the beginning of the line
Ctrl+Home	Moves the cursor to the beginning of the text
End	Moves the cursor to the end of the line
Ctrl+End	Moves the cursor to the end of the text
Alt+Wheel	Scrolls the page horizontally
Ctrl+Wheel	Zooms the text

## Predefined Placeholders

### Cite and Document ID Placeholders

- `<<author_all_abbreviated>>` Takes first three letters of the last word of all authors's last name in cite, and converts to lowercase.
- `<<author_all_initials>>` Takes capitalized initials of all authors in cite.
- `<<author_first>>` Takes first author last name.
- `<<author_first_lowercase>>` Takes first author last name in lowercase.
- `<<citeid>>` This placeholder is meant to be used **alone, and only for document IDs**. It takes the pattern defined for the cite ID. If the cite ID is modified manually, the document ID is synchronized automatically.
- `<<journal_initials>>` Takes capitalized initials of journal name.
- `<<pages_first>>` First page.



- `<<ppages_first>>` First page, written as, e. g., 'p125'.
- `<<title>>` Title. To truncate titles exceeding a maximum length `l` use `<<title_l>>`, where `l` stands for an integer value.
- `<<title_underscored>>` Title with blanks set to underscores. To truncate title to `l` characters use `<<title_underscored_l>>`.
- `<<title_first_word>>` First word in title, in lowercase.
- `<<volume>>` Volume number.
- `<<year_abbreviated>>` Last two digits from year.
- `<<year_full>>` All digits from year.

**Note:** If `author` is empty, `editor` will be considered instead. On conference proceedings or monographs this situation is usual. Similarly, if `title` is empty, `booktitle` is considered.

**Note:** Only one placeholder of a given field, e. g. `<<author_first>>` or `<<author_all_initials>>`, should be used to compose the ID patterns. cb2Bib only performs one substitution per field placeholder.

**Note:** cb2Bib performs a series of string manipulations, such as stripping diacritics and ligatures, aimed to provide ID values suitable for BibTeX keys and platform independent filenames. Currently only ASCII characters are considered.

## Cite Command Placeholders

- `<<citeid>>` The `citeid` placeholder replicates the pattern for each citation in the selected citation list. For example, the pattern `\citenum{<<citeid>>}` expands to `\citenum{cid1} \citenum{cid2} ...`
- `<<prefix|citeids|separator>>` The `citeids` placeholder replaces the selected citation list by prepending `prefix` and appending `separator` within the pattern. For example, the markdown pattern `[<<@|citeids|;>>]` expands to `[@cid1; @cid2; ...]`, and the LaTeX pattern `\citeauthor{<<|citeids|,>>}` expands to `\citeauthor{cid1, cid2, ...}`.

**Note:** For additional information on cite commands see **LaTeX Bibliography Management** and **Pandoc User's Guide**.

## **BiBTeX Entry Types Available as cb2Bib Fields**

cb2Bib includes nearly all standard and extended BibTeX fields. The complete list is as follows. The field descriptions are taken from **The BibTeX Format** written by Dana Jacobsen.

### Main Fields

`abstract` An abstract of the work.

`author` The name(s) of the author(s), in the format described in the LaTeX book.

<b>file</b>	Usually, the PDF filename of the work.
<b>journal</b>	A journal name. Abbreviations are provided for many journals.
<b>keywords</b>	Key words used for searching or possibly for annotation.
<b>pages</b>	One or more page numbers or range of numbers, such as <b>42 - -111</b> or <b>7, 41, 73 - -97</b> or <b>43+</b> (the <code>`+'</code> in this last example indicates pages following that don't form a simple range). To make it easier to maintain Scribe-compatible databases, the standard styles convert a single dash (as in <b>7 - 33</b> ) to the double dash used in TeX to denote number ranges (as in <b>7 - -33</b> ).
<b>title</b>	The work's title, typed as explained in the LaTeX book.
<b>volume</b>	The volume of a journal or multi-volume book.
<b>number</b>	The number of a journal, magazine, technical report, or of a work in a series. An issue of a journal or magazine is usually identified by its volume and number; the organization that issues a technical report usually gives it a number; and sometimes books are given numbers in a named series.
<b>year</b>	The year of publication or, for an unpublished work, the year it was written. Generally it should consist of four numerals, such as <b>1984</b> , although the standard styles can handle any <b>year</b> whose last four non punctuation characters are numerals, such as <code>\hbox{(about 1984)}</code> .

### **Other Fields**

<b>address</b>	Usually the address of the <b>publisher</b> or other type of institution. For major publishing houses, van Leunen recommends omitting the information entirely. For small publishers, on the other hand, you can help the reader by giving the complete address.
<b>annotate</b>	An annotation. It is not used by the standard bibliography styles, but may be used by others that produce an annotated bibliography.
<b>booktitle</b>	Title of a book, part of which is being cited. See the LaTeX book for how to type titles. For book entries, use the <b>title</b> field instead.
<b>chapter</b>	A chapter (or section or whatever) number.
<b>doi</b>	The Digital Object Identifier is a unique string created to identify a piece of intellectual property in an online environment.

<code>edition</code>	The edition of a book---for example, ``Second". This should be an ordinal, and should have the first letter capitalized, as shown here; the standard styles convert to lower case when necessary.
<code>editor</code>	Name(s) of editor(s), typed as indicated in the LaTeX book. If there is also an <code>author</code> field, then the <code>editor</code> field gives the editor of the book or collection in which the reference appears.
<code>eprint</code>	Electronic document file.
<code>institution</code>	The sponsoring institution of a technical report.
<code>ISBN</code>	The International Standard Book Number.
<code>ISSN</code>	The International Standard Serial Number. Used to identify a journal.
<code>month</code>	The month in which the work was published or, for an unpublished work, in which it was written. You should use the standard three-letter abbreviation, as described in Appendix B.1.3 of the LaTeX book.
<code>note</code>	Any additional information that can help the reader. The first word should be capitalized.
<code>organization</code>	The organization that sponsors a conference or that publishes a <code>manual</code> .
<code>publisher</code>	The publisher's name.
<code>school</code>	The name of the school where a thesis was written.
<code>series</code>	The name of a series or set of books. When citing an entire book, the <code>title</code> field gives its title and an optional <code>series</code> field gives the name of a series or multi-volume set in which the book is published.
<code>URL</code>	The WWW Universal Resource Locator that points to the item being referenced. This often is used for technical reports to point to the ftp site where the postscript source of the report is located.

## Reading and Writing Bibliographic Metadata

### Reading Metadata

Metadata in scientific documents had been rarely appreciated and used for decades. For bibliographic metadata, no format specification had been widely accepted. cb2Bib adapted back in 2008 the PDF predefined metadata capabilities to set BibTeX bibliographic keys in

document files.

cb2Bib reads all XMP (a specific XML standard devised for metadata storage) packets found in the document. It then parses the XML strings looking for nodes and attributes with key names meaningful to bibliographic references. If a given bibliographic field is found in multiple packets, cb2Bib will take the last one, which most often, and according to the PDF specs, is the most updated one. The fields `file`, which would be the document itself, and `pages`, which is usually the actual number of pages, are skipped.

The metadata is then summarized in cb2Bib clipboard panel as, for instance

```
[Bibliographic Metadata
<title>arXiv:0705.0751v1 [cs.IR] 5 May 2007</title>
/Bibliographic Metadata]
```

This data, whenever the user considers it to be correct, can be easily imported by the build-in 'Heuristic Guess' capability. On the other hand, if keys are found with the prefix `bibtex`, cb2Bib will assume the document does contain bibliographic metadata, and it will only consider the keys having this prefix. Assuming therefore that metadata is bibliographic, cb2Bib will automatically import the reference. This way, if using PDFImport, BibTeX-aware documents will be processed as successfully recognized, without requiring any user supplied regular expression.

See also **Release Note cb2Bib 1.0.0**, **Configuring Clipboard**, and **PDF Reference Import**.

## **Writing Metadata**

Once an extracted reference is saved and there is a document attached to it, cb2Bib will optionally insert the bibliographic metadata into the document itself. cb2Bib writes an XMP packet as, for instance

```
<bibtex:author>P. Constans</bibtex:author>
<bibtex:journal>arXiv 0705.0751</bibtex:journal>
<bibtex:title>Approximate textual retrieval</bibtex:title>
<bibtex:type>article</bibtex:type>
<bibtex:year>2007</bibtex:year>
```

The BibTeX fields `file` and `id` are skip from writing. The former for the reason mentioned above, and the latter because it is easily generated by specialized BibTeX software according to each user preferences. LaTeX escaped characters for non ASCII letters are converted to UTF-8, as XMP already specifies this codec.

The actual writing of the packet into the document is performed by ExifTool, an excellent Perl program written by Phil Harvey. See <https://exiftool.org>. ExifTool supports several document formats for writing. The most relevant here are Postscript and PDF. For PDF

documents, metadata is written as an incremental update of the document. This exactly preserves the binary structure of the document, and changes can be easily reversed or modified if so desired. Whenever ExifTool is unable to insert metadata, e.g., because the document format is not supported or it has structural errors, cb2Bib will issue an information message, and the document will remain untouched.

See also **Configuring Documents** and **Update Documents Metadata**.

## PDF Reference Import

### Introduction

Articles in PDF or other formats that can be converted to plain text can be processed and indexed by cb2Bib. Files can be selected using the Select Files button, or dragging them from the desktop or the file manager to the PDFImport dialog panel. Files are converted to plain text by using any external translation tool or script. This tool, and optionally its parameters, are set in the cb2Bib configure dialog. See the **Configuring Utilities** section for details.

Once the file is converted, the text, and optionally, the prepared metadata, is sent to cb2Bib for reference recognition. This is the usual, two step process. First, text is optionally preprocessed, using a simple set of rules and/or any external script or tool. See **Configuring Clipboard**. Second, text is processed for reference extraction. cb2Bib so far uses two methods. One considers the text as a full pattern, which is checked against the user's set of regular expressions. The better designed are these rules, the best and most reliable will be the extraction. The second method, used when no regular expression matches the text, considers instead a set of predefined subpatterns. See **Field Recognition Rules**.

At this point users can interact and supervise their references, right before saving them. Allowing user intervention is and has been a design goal in cb2Bib. Therefore, at this point, cb2Bib helps users to check their references. Poorly translated characters, accented letters, 'forgotten' words, or some minor formatting in the titles might be worth considering. See **Glyph & Cog's Text Extraction** for a description on the intricacies of PDF to text conversions. In addition, if too few fields were extracted, one might perform a network query. Say, only the DOI was catch, then there are chances that such a query will fill the remaining fields.

The references are saved from the cb2Bib main panel. Once Save is pressed, and depending on the configuration, see **Configuring Documents**, the document file will be either renamed, copied, moved or simply linked onto the [file](#) field of the reference. If **Insert BibTeX metadata to document files** is checked, the current reference will also be inserted into the document itself.

When several files are going to be indexed, the sequence can be as follows:

- **Process next after saving**

Once files are load and Process is pressed, the PDFImport dialog can be minimized (but not closed) for convenience. All required operations to completely fill the desired fields

(e.g. dynamic bookmarks, open DOI, etc, which might be required if the data in document is not complete) are at this point accessible from the main panel. The link in the [file](#) field **will be permanent**, without regard to which operations (e.g. clipboard copying) are needed, until the reference is saved. The source file can be open at any time by right clicking the [file](#) line edit. Once the reference is saved, the next file will be automatically processed. To skip a given document file from saving its reference, press the Process button.

- **Unsupervised processing**

In this operation mode, all files will be sequentially processed, following the chosen steps and rules. **If the processes is successful**, the reference is automatically saved, and the next file is processed. **If it is not**, the file is skipped and no reference is saved. While processing, the clipboard is disabled for safety. Once finished, this box is unchecked, to avoid a possible accidental saving of a void reference. Network queries that require intervention, i.e., whose result is launching a given page, are skipped. The processes follows until all files are processed. However, it will stop to avoid a file being overwritten, as a result of a repeated key. In this case, it will resume after manual renaming and saving. See also **cb2Bib Command Line**, commands [-txt2bib](#) and [-doc2bib](#).

## **Automatic Extraction: Questions and Answers**

- **When does cb2Bib do automatic extractions?**

cb2Bib is conceived as a lightweight tool to extract references and manage bibliographies in a simple, fast, and accurate way. Accuracy is better achieved in semi-automatic extractions. Such extractions are handy, and allow user intervention and verification. However, in cases where one has accumulated a large number of unindexed documents, automatic processing can be convenient. cb2Bib does automatic extraction when, in PDFImport mode, 'Unsupervised processing' is checked, or, in command line mode, when typing [cb2bib -doc2bib \\*.pdf tmp\\_references.bib](#), or, on Windows, [c2bconsole.exe](#) instead of [cb2bib](#).

- **Are PDFImport and command line modes equivalent?**

Yes. There are, however, two minor differences. First, PDFImport adds each reference to the current BibTeX file, as this behavior is the normal one in cb2Bib. On the other hand, command line mode will, instead, overwrite [tmp\\_references.bib](#) if it exists, as this is the expected behavior for almost all command line tools. Second, as for now, command line mode does not follow the configuration option 'Check Repeated On Save'.

- **How do I do automatic extraction?**

To test and learn about automatic extractions, the cb2Bib distribution includes a set of four PDF files that mimic a paper title page. For these files, distribution also includes a regular expression, in file [regexps.txt](#), capable of extracting the reference fields, provided the [pdftotex](#) flags are set to their default values. Processing these files, should, therefore, be automatic, and four messages stating [Processed as 'PDF Import Example'](#) should be seen in the logs. Note that extractions are configurable. A

reading of **Configuration** will provide additional, useful information.

- **Why some entries are not saved and files not renamed?**

Once you move from the fabricated examples to real cases, you will realize that some of the files, while being processed, are not renamed and their corresponding BibTeX data is not written. For each document file, cb2Bib converts its first page to text, and from this text it attempts to extract the bibliographic reference. By design, when extraction fails, cb2Bib does nothing: no file is moved, no BibTeX is written. This way, you know that the remaining files in the origin directory need special, manual attention.

**Extractions are seen as failed, unless reliable data is found in the text.**

- **What is *reliable data*?**

Note that computer processing of natural texts, as extracting the bibliographic data from a title page, is nowadays an approximated procedure. cb2Bib tries several strategies: **1)** allow for including user regular expressions very specific to the extraction at hand, **2)** use metadata if available, **3)** guess what is reasonable, and, based on this, make customized queries. Then, cb2Bib considers extracted **data is reliable if i)** data comes from a match to an user supplied regular expression **ii)** document contains BibTeX metadata, or **iii)** a guess is transformed through a query to formatted bibliographic data. As formatted bibliographic data, cb2Bib understands BibTeX, PubMed XML, arXiv XML, and CR JSON data. In addition, it allows external processing if needed. Other data, metadata, guesses, and guesses on query results are considered unreliable data.

- **Is metadata reliable data?**

No. Only author, title, and keywords in standard PDF metadata can be mapped to their corresponding bibliographic fields. Furthermore, publishers most often misuse these three keys, placing, for instance, DOI in title, or setting author to, perhaps, the document typesetter. Only BibTeX XMP metadata is considered reliable. If you consider that a set of PDF files does contain reliable data, you may force to accept it using the command line switch `-sloppy` together with `-doc2bib`.

- **How successful is automatic extraction?**

As it follows from the given definition of reliable data, running automatic extractions without adhoc `regexprs.txt` and `netqinf.txt` files will certainly give a zero success ratio. In practice, scenario 3) often applies: cb2Bib guesses several fields, and, based on the out-of-the-box `netqinf.txt` file, it obtains from the web either BibTeX, PubMed XML, arXiv XML, or CR JSON data.

- **What can I do to increase success ratio?**

First, set your favorite journals in file `abbreviations.txt`. Besides increasing the chances of journal name recognition, it will provide consistency across your BibTeX database. In general, do not write regular expressions to extract directly from the PDF text. Conversion is often poor. Special characters often break lines, thus breaking your regular expressions too. Write customized queries instead. For instance, if your PDFs have DOI in title page, set the simple query

```
journal=The Journal of Everything|
query=https://dx.doi.org/<<doi>>
capture_from_query=
```



```
referenceurl_prefix=  
referenceurl_suffix=  
pdfurl_prefix=  
pdfurl_suffix=  
action=htm2txt_query
```

then, if it is feasible to extract the reference from the document's web page using a regular expression, include it in file `regexps.txt`. Note that querying in cb2Bib had been designed having in mind minority fields of research, for which, established databases might not be available. If cb2Bib failed to make reasonable guesses, then, you might consider writing very simple regular expressions to extract directly from the PDF text. For instance, obtain title only. Then, the posterior query step can provide the remaining information. Note also, especially for old documents, journal name is often missing from the paper title page. If in need of processing a series of those papers, consider using a simple script, that, in the cb2Bib preprocessing step, adds this missing information.

- **Does successful extraction mean accurate extraction?**

No. An extraction is successful if reliable data, as defined above, is found in the text, in the metadata, or in the text returned by a query. Reference accuracy relies on whether or not user regular expressions are robust, BibTeX metadata is correct, a guess is appropriate, a set of queries can correct a partially incorrect guess, and the text returned by a query is accurate. In general, well designed sets of regular expressions are accurate. Publisher's abstract pages and PubMed are accurate. But, some publishers are still using images for non-ASCII characters, and PubMed algorithms may drop author middle names if a given author has 'too many names'. Expect convenience over accuracy on other sources.

- **Can I use cb2Bib to extract comma separated value CSV references?**

Yes. To automatically import multiple CSV references you will need one regular expression. If you can control CSV export, choose `|` as separator, since comma might be used, for instance, in titles. The regular expression for

```
AuthName1, AuthName2 | Title | 2010
```

will simply be

```
author title year  
^([^\|]*)\|([^\|]*)\|([^\|]*)$
```

The reference file `references.csv` can then be split to single-line files typing

```
split -l 1 references.csv slineref
```

and the command



```
cb2bib --txt2bib slineref* references.bib
rm -f slineref*
```

will convert `references.csv` to BibTeX file `references.bib`

## Extracting Data from the Clipboard

Clipboard contents is processed according to the following rules:

- Perform external, user-defined preparsing on input stream. See **Configuring Clipboard**.
- Perform user-defined substitutions on input stream. See **Configuring Clipboard**.
- Check if input stream is already a BibTeX entry. If so, process entry.
- Check if input stream is, in this order of preference, a PubMed XML, arXiv XML, CR JSON, or Medline entry. If so, process entry.
- Preprocess author names: PI JOAN III -> Pi III, J. (care of name prefixes, suffixes, and removal of ambiguities).

If otherwise,

- Extract DOI  
(DOI, URL and FILE/PDF are preprocessed, performed before the automatic recognition takes place.)
- Extract URL
- Remove leading and trailing white spaces, TABs and CRs.
- `"\r\n"`, `"\n"` and/or `"\r"` replaced by the line indicator tag `<NewLine>`.
- Replace `"\t"` and ten or more consecutive `"\s"` by the tabular tag `<TabN>`.
- Simplify white spaces
- Start the automatic recognition engine.

If the automatic recognition engine fails, optionally, a heuristic guessing will be performed.

See also **Field Recognition Rules** and **Reading and Writing Bibliographic Metadata**.

## Processing of Author Names

cb2Bib automatically processes the author names string. It uses a set of heuristic rules. First, the authors separator is identified. And second, it is decided whether or not author names are in natural or reverse order, or in the 'Abcd, E., F. Ghij, ...' mixed order.

Cleanup author string:

- Escape BibTeX to Unicode
- Remove digits from authors string
- Remove any character except `- ' , ; & \ . \s \w`
- Simplify white spaces

- Consider composing prefixes (da|de|dal|del|der|di|do|du|dos|el|la|le|lo|van|vande|von|zur)
- Consider composing suffixes (II|III|IV|Jr)
- Some publishers use superscripts to refer to multiple author affiliations. Text clipboard copying loses superscript formatting. Author strings are clean from 'orphan' lowercase, single letters in a preprocessing step. Everything following the pattern **[a-z]** is removed. Fortunately, abbreviated initials are most normally input as uppercase letters, thus permitting a correct superscript clean up.  
*Caution:* Lowcase, single, a to z letters are removed from author's string.  
*Caution:* Superscripts **will be added to author Last Name** if no separation is provided. Users should care about it and correct these cases.

Rules to identify separators:

- Contains comma and semicolon -> ';'
  - Contains pattern '^Abcd, E.-F., ' -> ','
  - Contains pattern '^Abcd, ' -> 'and'
- Contains comma -> ','
- Contains semicolon -> ';'
  - Any other -> 'and'

Rules to identify ordering:

- Contains comma and semicolon -> Reverse
- Pattern '^Abcd, ' -> Reverse
- Pattern '^Abcd EF Ghi' -> Natural
- Pattern '^Abcd EF' -> Reverse
- Pattern '^Abcd E.F.' -> Reverse
- Any other pattern -> Natural

## Processing of Journal Names

cb2Bib processes journal names according to its editable database, stored at [abbreviations.txt](#). This file contains a list of journal names equivalences: a capital-letter acronym, the abbreviated form, and the title of the journal, all three on one single line.

The [abbreviations.txt](#) file has the following structure:

```
JA|J. Abbrev.|Journal of Abbreviations
AN|Am. Nat.=Amer. Naturalist|American Naturalist=The American Naturalist
```

The first field, the capital-letter acronym, is a user-defined shorthand to access a journal title by typing it at the extraction panel.

The second field is the abbreviated form of the journal. To adapt to multiple abbreviations

in use, cb2Bib allows one alternate version of the abbreviation, indicated with an equal sign =. In the above example, the ISO 4 abbreviation 'Am. Nat.' is the primary one and 'Amer. Naturalist' is the alternate one.

Finally, the third field is the full title of the journal. As for the abbreviations, the full title also admits one alternate form.

Abbreviated and full title alternates serve two purposes: journal recognition and citation styling. The former is performed internally by cb2Bib as part of a bibliographic reference extraction, and the latter is accomplished in the embedded BibTeX editor by replacing back and forth abbreviated-full forms, in order to set journals in accordance to the guidelines of a particular publication.

Journal names processing is performed whenever a string is recognized as 'journal', and, additionally, when pressing [Intro Key](#) at the journal edit line.

- Retrieves Journal name in **abbreviated form** if found.
  - If Journal name is not found in the database, returns input Journal name.
  - Search is case insensitive.
  - **Warning:** Journal codes can be duplicated. If duplicated, returns input Journal name.
- 
- Retrieves Journal name in **full form** if found.
  - If Journal name is not found in the database, returns input Journal name.
  - Search is case insensitive.
  - **Warning:** Journal codes can be duplicated. If duplicated, returns input Journal name.

See [Configuring Files](#), [Configuring BibTeX](#), and [Right-Click Menu Functionality](#).

## Field Recognition Rules

- **Abstract**
  - If [Abstract](#) is found.
  - If [Summary](#) is found.
- **Author**
  - Check capitalization patterns. See [A Simple Extraction Procedure for Bibliographical Author Field](#).
- **Keywords**
  - If [Key\s{0,1}words](#) is found.

- **Volume**

- If `Volume:{0,1}` is found.
- If `Vol.{0,1}` is found.
- If `\b(\d+)[, :]\s*\d+W\d+` is found.
- If `\b(\d+)\s*(\d+)\b` is found.
- If `\b(\d+)[, :]\s*\d+\b` is found.

- **Number**

- If `Numbers{0,1}:{0,1}\s*(\d-)+` is found.
- If `No.{0,1}\s*(\d+)` is found.
- If `Issue\:{0,1}\s*(\d+)` is found.
- If `\d\s*(\d+)\b` is found.

- **Pages**

- If `\bPages{0,1}[:\.\.]{0,1}(\d\s-)+` is found.
- If `\bp{1,2}\.{0,1}\s+(\d+)` is found.
- If `\b(\d+)\s*-{1,2}\s*(\d+pp)\b` is found.
- If `\b(\d+)\s*-{1,2}\s*(\d+)\b` is found.

- **Year**

- If `\b(19|20)(\d\d)\b` is found.

- **Title**

- If `\bTitle:{0,1}` is found.

- **ISBN**

- If `\bISBN\b(?:-\d+){0,1}:{0,1}(?:-\d+){0,1}\s*(\d+[-\d-]+\d+)` is found.
- If `\bISBN\b(?:-\d+){0,1}:{0,1}(?:-\d+){0,1}\s*(\d+)` is found.

- **Journal**

- Check cb2Bib internal database.

## Regular Expression Editor

Once a manual processing is done, cb2Bib clipboard area contains the extraction tags, plus, possibly, some other cb2Bib tags introduced during the preprocessing (see **Extracting Data from the Clipboard**). The **RegExp Editor** will generate a guess regular expression or matching pattern usable for automated extractions.

The cb2Bib matching patterns consist of four lines: a brief description, the reference type, an ordered list of captured fields, and the regular expression itself.

```
# cb2Bib 2.0.2.90 Pattern:
American Chemical Society Publications
article
journal volume pages year title author abstract
^(.+), (\d+) \(.+\), ([\d\-\s]+), (\d\d\d\d)\.+.+<NewLine3>(.)<NewLine4>
(.+)<NewLine5>.+Abstract:<NewLine\d+>(.)$
```

The Regular Expression Editor provides the basic skeleton and a set of predefined suggestions. The regular expressions follow a Perl-like syntax. There are, however, some slight

differences and minor limitations. Information about the basics on the editing and working with Regular Expressions as used by cb2Bib can be found at the Qt document file **Qt Documentation's QRegExp Class**.

### Remember when creating and editing regular expressions:

- Switch the clipboard mode to 'Tagged Clipboard Data', using the clipboard panel context menu.
- Extract the bibliographic reference manually. On the clipboard panel will appear some cb2Bib tags that indicate which fields are being extracted. Once done, type Alt+I to enter to the regular expression editor. In the editor, there are the four line edits that define a cb2Bib pattern, one copy of the clipboard panel, and an information panel. The information panel displays possible issues, and, once everything is correct, the actual extracted fields. The clipboard panel highlights the captures for the current regular expression and current input text.
- Patterns can be modified at any time by typing Alt+E to edit the regular expression file. Patterns are reloaded each time the automatic pattern recognition is started. This permits editing and testing.
- cb2Bib processes sequentially the list of regular expressions as found in the regular expression file. It stops and picks the first match for the current input. **Therefore, the order of the regular expressions is important.** Consequently, to avoid possible clashing among similar patterns, consider sorting them from the most restrictive pattern to the less one. As a rule of thumb, the more captions it has the most restrictive a pattern is.
- **The cb2Bib proposed patterns are general, and not necessarily the most appropriate for a particular capture.** E.g. tag `pages` becomes `([\d|\-|\s]+)`, which considers digits, hyphens, and spaces. It must be modified accordingly for reference sources with, e.g., `pages` written as Roman ordinals.
- **Avoid whenever possible general patterns (.+).** There is a risk that such a caption could include text intended for a posterior caption. This is why, sometimes, the cb2Bib proposed pattern is not hit by the input stream that originated it. **Use, whenever possible, cb2Bib anchors like <NewLine1> instead of <NewLine\d+>. They prevent (.+) captions to overextend.**
- To debug a large regular expression it might be useful to break it to the first capturing parenthesis. For instance, the above pattern will be

```
# cb2Bib 2.0.2.90 Pattern:
American Chemical Society Publications
article
journal
^(.+),
```

- Then, check if anything is captured and if this corresponds to `journal`.
- Add on successive steps your set of captions and BibTeX fields.

# Supplementary Notes

## Release Notes

### Release Note cb2Bib 2.0.1

To optimize search on PDF's contents, cb2Bib keeps a cache with the extracted text streams, that are compressed to reduce disk space and reading overhead. Nowadays, compressors with extremely high decompression speed are available. Two of them are LZSSE, for SSE4 capable architectures, and LZ4, for a broader range of CPUs. These two compressors can now be used by cb2Bib, with the latter set as the default compression library in cb2Bib builds. When upgrading to version 2.0.1, the **first search** on the document collection will recreate the cache, and this step **will be noticeably slow**.

Additionally, cb2Bib 2.0.1 includes original, optimized text matching code for AVX2 capable architectures that is used for search matching and BibTeX parsing. This code is **not set** in default builds and needs to be explicitly enabled at compilation time.

Finally, it is important mentioning the inclusion in version 2.0.1 of stemmed context search, see **Contextual Search** for details, and contributed feedback in handling citations and extending cite commands to markdown syntax, see **Predefined Placeholders**.

### Release Note cb2Bib 2.0.0

Throughout the 1.9.x series, the cb2Bib sources were updated to the improved string processing capabilities of Qt5 and PCRE libraries. This update has brought a remarkable speedup for in-document searches and full search indexing.

Alternate normalization of journal titles and abbreviations, upgrading jsMath to MathJax, extending network queries syntax, and a PDF user manual are the additional enhancements in cb2Bib 2.0.0.

Back in version 0.3.3, cb2Bib introduced network queries to obtain the data for a citation. While convenient, queries to publishers' websites were difficult to setup and fragile. Nowadays, fortunately, arXiv, PubMed and Crossref offer structured APIs. These interfaces provide to the end user an easy setup for completing bibliographic citations.

### Release Note cb2Bib 1.9.0

The cb2Bib sources have been ported to Qt5. To highlight this major update in library requirements the version number is set to 1.9.0. Later, once stabilized and new functionality related to Qt5 enhancements are applied, version number will be set to 2.

At this point cb2Bib has exactly the same functionality as its preceding version 1.5.0. To build the program, however, only qmake and its related config procedure are available. The cmake scripts have not yet been ported.

Qt5 brings important enhancements related to regular expressions and string processing. Some careful updates to the cb2Bib sources are needed to fully benefit from them. They will be implemented through the 1.9.x series. We expect by then a performance boost on full text, regular expression based searches.

### **Release Note cb2Bib 1.5.0**

Included in version 1.5.0 sources there is a patch for XPDF 3.0.4, the default tool to convert PDF documents to plain text. The modified code separates superscripts to avoid words being joined to reference numbers and author names joined to affiliations' glyphs. Interested users will need to download the package, apply the patch, and compile it.

Additionally, this version improves converted text postprocessing. This step normalizes character codes, reverts ligatures, restores when possible orphan diacritics and broken words, and undoes text hyphenation.

Conversion to text and postprocessing is important for reference extraction, and document indexing and searching. It is therefore recommended to delete cached document-to-text data to benefit from the present improvements. cb2Bib stores cached texts in \*.c2b files in an user specified directory. After that, by performing a search or initiating indexing an updated cache will be created.

### **Release Note cb2Bib 1.4.7**

Approximate and context searches effectively locate our references of interest. As collections grow in size, and low performance devices, netbooks and tablets, start being used, complete document searches become demanding. Besides, it is often not clear what to query for, and then a glossary of terms provides guidance. Often too, interest lies on subsetting documents by being similar to a given one.

Version 1.4.7 adds a pragmatic term or keyword extraction from the document contents. Accepted keywords are set as the substrings appearing at least twice in one document, appearing at least in three documents, and conforming to predefined part-of-speech (POS) sequences. Keyword extraction is performed by either clicking on [Index Documents](#) at the [c2bciter](#) desktop tray menu, or, by typing `cb2bib -index [bibdirname]` on a shell. During extraction, the [Part Of Speech \(POS\) Lexicon](#) distribution file must be available and readable. On termination, indexing files are saved on the [Search In Files Cache Directory](#). Simply copying this directory will synchronize keyword indexing to a second

computer.

After refreshing `c2bciter` module, pressing key G displays the glossary of terms. On a reference, pressing K displays its list of keywords. Pressing R on a keyword lists the references related to that keyword. Pressing R on a reference lists similarly related references. Similarity is assessed based on keyword occurrences. Left and Right keys provide previous and next navigation. Pressing V on either a reference keyword, or a keyword reference, visualizes the keyword excerpts from the reference's document. To close excerpt dialog press Esc or Left keys.

See also **cb2Bib Citer**, **Configuring Files**, and **cb2Bib Command Line**.

### **Release Note cb2Bib 1.4.0**

The `c2bciter` module was introduced in version 1.3.0. Its name, as it was described, states its purpose of being "aimed to ease inserting citation IDs into documents". In fact, it does have such functionality. And, it has also another, equally important one: it provides a very fast way to retrieve a given work from our personal collections.

Retrieving is accomplished through pre-sorted views of the references and filtering. Both, views and filtering, scale on the (tens of) thousands references. Usually, we recall a work from its publication year, a few words from its title, or (some of the letters of) one of its authors names. Often, what we remember is when a reference was included into our collection. Therefore, having such a chronological view was desirable.

The implementation of this sorted-by-inclusion-date view was not done during the 1.3.x series, but postponed to version 1.4.0; somehow, to indicate that some sort of 'proprietary' BibTeX tag might be required to specify inclusion timestamps. I have been reluctant through the cb2Bib's life span to introduce 'cb2Bib-only' tags in the BibTeX outputs. I believe that there is little gain, and it costs, possibly, breaking interoperability.

In the end, the choice was to not write any 'timestamp' tag in references. Instead, `c2bciter` checks for the last modified date of the linked documents to build an approximated chronological view. The advantage is that all, not just 'version 1.4.0 or later', references are sorted. Furthermore, if a reference is later corrected, and the document metadata is updated too, the modification date is reflected in the view. The obvious inconvenience is that no such sorting can be done for references without an attached document.

See also **cb2Bib Citer**.



## **Release Note cb2Bib 1.3.0**

When version 0.2.7 came up, it was mentioned in **Release Note cb2Bib 0.2.7** that cb2Bib 'doesn't have the means to automatically discern an author name from a department or street name'. I forgot mentioning, that I did not expect cb2Bib would have had such a feature. Since the last **Release Note cb2Bib 1.1.0**, the cb2Bib internals had changed significantly. Some changes, such heuristic recognition for interlaced authors and affiliations, get easily noticed. Other changes, however, do not, and need additional explanation.

From version 1.2.3, the switches `-txt2bib` and `-doc2bib` set cb2Bib to work on console mode. The non-exact nature of the involved extractions makes logging necessary. On Windows, graphic or console modes must be decided not at run time, but when the application is built. So far, logging and globing were missing. This release adds the convenience wrapper `c2bconsole`. Typing `c2bconsole -txt2bib i*.txt out.bib`, for instance, will work as it does in the other platforms.

Lists of references are now sorted case and diacritic insensitive. For some languages such a choice is not the expected one, and some operating systems offer local-aware collation. Due to usual inconsistencies and inaccuracies in references, this decision was taken to group together 'Density Matrix' with 'Density-matrix', and Møller with Moller, which, in a personal collection, most probably, refer to the same concept and to the same person. Additionally, document to text converted strings are now clean from extraneous, non-textual symbols. Therefore, recreating cache files is recommended.

Finally, this release introduces a new module, named `c2bciter`, and aimed to ease inserting citation IDs into documents. The module should ideally stay idle at the system tray, and be recalled as needed by pressing a global, desktop shortcut. This functionality, while desirable, and usual in dictionaries, is platform and desktop dependent. On KDE there are currently known issues when switching among virtual desktops.

See also **cb2Bib Citer**, and **cb2Bib Command Line**.

## **Release Note cb2Bib 1.1.0**

A frequent request from cb2Bib users has been to expand the command line functionality. So far few progress has been seen in this regard. First, the addition of in-document searches and reading/inserting metadata were priorities. Second, cb2Bib is not the tool to interconvert among bibliographic formats. And third, cb2Bib is designed to involve the user in the search process, in the archiving and validation of the discovered works and references.

For the latter reason, and for not knowing a priori how would such a tool be designed, the cb2Bib internals had been interlaced to its graphical interface. At the time of version 0.7.0, when the graphical libraries changed, and a major refactoring was required, the code started moving toward a better modularization and structure. The current release pushes code

organization further. As a result, it adds two new command line switches: `-html-annotate` and `-view-annotate`.

The new cb2Bib module is named after the BibTeX key 'annotate'. Annote is not for a 'one reference annotation' though. Instead, Annote is for short notes that interrelate several references. Annote takes a plain text note, with minimal or no markup, inserts the bibliographic citations, and converts it to a HTML page with links to the referenced documents.

From within cb2Bib, to write your notes, type Alt+A, enter a filename, either new or existing, and once in Annote, type E to launch your default text editor. For help, type F1. Each time you save the document the viewer will be updated. To display mathematical notations, install **jsMath** locally. And, remember, code refactoring introduces bugs.

See also **cb2Bib Annote** and **cb2Bib Command Line**.

### **Release Note cb2Bib 1.0.0**

Approximately four years ago the first cb2Bib was released. It included the possibility of easily linking a document to its bibliographic reference, in a handy way, by dragging the file to the main (at that time, single) panel. Now, in version 1.0.0, when a file is dropped, cb2Bib scans the document for metadata packets, and checks, in a rather experimental way, whether or not they contain relevant bibliographic information.

Publishers metadata might or might not be accurate. Some, for instance, assign the DOI to the key Title. cb2Bib extracts possibly relevant key-value pairs and adds them to clipboard panel. Whenever key-value pairs are found accurate, just pressing Alt+G imports them to the line edits. If keys with the prefix `bibtex` are found, their values are automatically imported.

The preparsed metadata that is added to the clipboard panel begins with `[Bibliographic Metadata` and ends with `/Bibliographic Metadata]`. Therefore, if you are using PDFImport together with a set of regular expressions, such that they contain the begin (^) or end (\$) anchors, you can safely replace them by the above tags. In this manner, existing regular expressions remain useful with this minor change. And, with the advantage that, if recognition fails for a given document, metadata might give the hardest fields to extract from a PDF article, which are author and title.

See also **Reading and Writing Bibliographic Metadata**.

### **Release Note cb2Bib 0.8.4**

The previous cb2Bib release added the command line option `-conf`

`[full_path]cb2bib.conf` to specify the settings location. This feature was intended, mainly, as a clean way to run the program on a host computer from a removable drive. The work done focused on arranging the command line and settings related code. It was left for a later release to solve some requirements regarding the managing of file pathnames and temporary files.

This release addresses these two points. Now, when cb2Bib is launched as `cb2bib -conf` – without a configuration filename– it treats filenames as being relative to the cb2Bib actual location. Temporary files, if needed, will be placed at this location as well. Therefore, no data is being written on the host, and cb2Bib works independently of the actual address that the host assigns to the removable drive.

The Windows' un/installer cleans/sets configuration data on the registry. Being aware of this particular, it might be better not to install the program directly to the USB drive. Just copy the cb2Bib base directory from a home/own computer to the removable drive, and then run it on the host computer as `cb2bib -conf`.

### **Release Note cb2Bib 0.8.3**

cb2Bib accepts several arguments on its command line to access specific functionality. So far, the command `cb2bib tmp_ref` permits importing references from the browser, whenever a *download to reference manager* choice is available. In addition, the command `cb2bib -bibedit ref.bib` directly launches the BibTeX editor for file browsing and editing.

This release adds the command line option `-conf [full_path]cb2bib.conf` to specifically set a file where all internal settings are being retrieved and stored. This has two interesting applications. On one hand, it easily permits switching from several sets of extraction rules, since the files `abbreviations.txt`, `regexps.txt`, and `netqinf.txt` are all stored in the cb2Bib's settings. And, on the other hand, it allows installing the program on a USB flash drive, and cleanly running it on any (e. g., library) computer. Settings can be stored and kept on the external device, and therefore, no data will be written on the registry or settings directory of the host computer.

So far, however, this feature should be regarded as experimental. The Qt library to which cb2Bib is linked does read/write access to system settings in a few places (concretely, in file and color dialogs). On Unix and Mac OS systems this access can be modified by setting the environment variable `DAG_CONFIG_HOME`. No such workaround is presently available in Windows.

See **cb2Bib Command Line** for a detailed syntax description.

## **Release Note cb2Bib 0.8.1**

Several changes in this release affect installation and deployment. First, the cb2Bib internals for settings management has been reorganized. Version 0.8.1 will not read previous settings, as user colors, file locations, etc. On Unix, settings are stored at [~/.config/MOLspaces/cb2Bib.conf](#). This file can be removed, or renamed. On Windows, it is recommended to uninstall previous versions before upgrading.

Second, cb2Bib tags are not shown by default. Instead, it is shown plain, raw clipboard data, as it is easier to identify with the original source. To write a regular expression, right click, on the menu, check 'View Tagged Clipboard Data', and perform the extraction from this view.

And finally, cb2Bib adds the tag <<excerpt>> for network queries. It takes a simplified version of the clipboard contents and sends it to, e.g. Google Scholar. From there, one can easily import BibTeX references related to that contents. Therefore one should unchecked in most cases the 'Perform Network Queries after automatic reference extractions' box.

## **Release Note cb2Bib 0.7.2**

cb2Bib reads the clipboard contents, processes it, and places it to the main cb2Bib's panel. If clipboard contents can be recognized as a reference, it writes the corresponding BibTeX entry. If not, the user can interact from the cb2Bib panel and complete or correct the reference. Additionally, this process permits to write down a regular expression matching the reference's pattern.

To ease pattern writing, cb2Bib preprocesses the raw input data. This can consider format conversion by external tools and general substitutions, in addition to including some special tags. The resulting preprocessed data is usually less readable. A particularly illustrating case is when input data comes from a PDF article.

cb2Bib now optionally presents input data, as raw, unprocessed data. This preserves the block text format of the source, and thus identifying the relevant bibliographic fields by visual inspection is more straightforward. In this raw mode view panel, interaction works in a similar manner. Except that, no conversions or substitutions are seen there, and that no regular expression tags are written.

## **Release Note cb2Bib 0.7.0**

This release moves forward cb2Bib base requirement to Qt 4.2.0. Compilation errors related to rehighlight() library calls, kindly reported by Bongard, Seemann, and Luisser, should not appear anymore. File/URL opening is carried now by this library, in a desktop integrated manner. Additionally, Gnome users will enjoy better integration, as Cleanlooks widget style is available.

All known regressions in 0.6.9x series have been fixed. Also, a few minor improvements have been included. In particular, file selection dialogs display navigation history, and BibTeX output file can be conveniently selected from the list of '\*.bib' files at the current directory. Such a feature will be specially useful to users that sort references in thematic files located at a given directory.

### **Release Note cb2Bib 0.6.91**

This release fixes a regression in the cb2Bib network capabilities. Network, and hence querying was erratic, both for the internal HTTP routines and for external clients. In addition to this fix, the `netqinf.txt` has been updated. PubMed is working again. Queries are also extended to include DOI's. A possible applicability will be for indexing a set of PDF articles with PDFImport. If the article contains its DOI number, and 'Perform Network Queries after automatic reference extractions' is checked, chances are that automatic extractions will work smooth.

### **Release Note cb2Bib 0.6.90**

cb2Bib has been ported from Qt3 to Qt4, a migration in its underlying system library. Qt experienced many changes and improvements in this major release upgrade. Relevant to cb2Bib, these changes will provide a better file management, word completion, faster searches, and better desktop integration.

Upgrading to Qt4 it is not a "plug and recompile" game. Thorough refactoring and rewriting was required. The resulting cb2Bib code is cleaner and more suitable to further development. As one might expect, major upgrades introduce new bugs that must be fixed. The cb2Bib 0.6.90 is actually a preview version. It has approximately the same functionality than its predecessor. So, no additions were considered at this point. Its use, bug reporting, and feedback are encouraged. This will help to get sooner a stable cb2Bib 0.7.

To compile it, type `./configure` as usual. The `configure` script calls the `qmake` tool to generate an appropriate `Makefile`. To make sure the right, Qt4 `qmake` is invocated, you can setup `QTDIR` environment variable prior to `./configure`. The `configure`'s call statement will then be `'$QTDIR/bin/qmake'`. E. g., type `'setenv QTDIR /usr'` if `qmake` happens to be at the directory `/usr/bin`.

### **Release Note cb2Bib 0.6.0**

cb2Bib uses the internal tags `<<NewLine_n>>` and `<<Tab_n>>` to ease the creation of regular expressions for reference extraction. New line and tabular codes from the input

stream are substituted by these numbered tags. Numbering new lines and tabulars gives an extra safety when writing down a regular expression. E. g., suppose field title is 'anything' between '<<NewLine1>>' and '<<NewLine2>>'. We can then easily write 'anything' as '.+' without the risk of overextending the caption to several '\n' codes. On the other hand, one still can use '<<NewLine\d>>' if not interested in a specific numbering. All these internal tags are later removed, once cb2Bib postprocesses the entry fields.

The cb2Bib identified so far new lines by checking for '\n' codes. I was unaware that this was a platform dependent, as well as a not completely accurate way of detecting new lines. McKay Euan reported that '<<NewLine\_n>>' tags were not appearing as expected in the MacOSX version. I later learn that MacOSX uses '\r' codes, and that Windows uses '\r\n', instead of '\n' for new line encoding.

This release addresses this issue. It is supposed now that the cb2Bib regular expressions will be more transferable among the different platforms. Extraction from plain text sources is expected to be completely platform independent. Extraction from web pages will still remain browser dependent. In fact, each browser adds its peculiar interpretation of a given HTML source. For example, in Wiley webpages we see the sectioning header 'Abstract' in its source and in several browsers, but we see, and get, 'ABSTRACT' if using Konqueror.

What we pay for this more uniform approach is, however, a **break in compatibility** with previous versions of cb2Bib. Unix/Linux users should not expect many differences, though. Only one from the nine regular expressions in the examples needed to be modified, and the two contributed regular expressions work perfectly without any change. Windows users will not see a duplication of '<<NewLine\_n>>' tags. To update previous expressions it should be enough just shifting the '<<NewLine\_n>>' numbering. And, of course, any working regular expression that does not uses '<<NewLine\_n>>' tags will still be working in this new version.

Finally, just to mention that I do not have a MacOSX to test any of the cb2Bib releases in this particular platform. I am therefore assuming that these changes will fix the problem at hand. If otherwise, please, let me know. Also, let me know if release 0.6.0 'break' your own expressions. I consider this release a sort of experimental or beta version, and the previous version 0.5.3, will still be available during this testing period.

### **Release Note cb2Bib 0.5.0**

Two issues had appeared regarding cb2Bib installation and deployment on MacOSX platforms.

First, if you encounter a 'nothing to install'-error during installation on MacOSX 10.4.x using the cb2Bib binary installer available at [naranja.umh.es/~atg/](http://naranja.umh.es/~atg/), please delete the cb2bib-receipts from [/Library/Receipts](#) and then rerun the installer. See also M. Bongard's clarifying note 'MACOSX 10.4.X "NOTHING TO INSTALL"-ERROR' for details.



Second, and also extensible to other cb2Bib platform versions, if PDFImport issues the error message 'Failed to call *some\_format\_to\_text*' tool, make sure such a tool is installed and available. Go to Configure->PDFImport, click at the 'Select External Convert Tool' button, and navigate to set its full path. Since version 0.5.0 the default full path for the MacOSX is already set, and pointing to `/usr/local/bin/pdftotext`.

### **Release Note cb2Bib 0.4.1**

Qt/KDE applications emit notifications whenever they change the clipboard contents. cb2Bib uses these notifications to automatically start its 'clipboard to BibTeX' processing. Other applications, however, does not notify about them. Since version 0.2.1, see **Release Note cb2Bib 0.2.1**, cb2Bib started checking the clipboard periodically. This checking was later disabled as a default, needing a few lines of code to be uncommented to activate it. Without such a checking, cb2Bib appears unresponsive when selecting/copying from e.g., acroread or Mozilla. This release includes the class `clipboardpoll` written by L. Lunak for the KDE's Klipper. Checking is performed in a very optimized way. This checking is enabled by default. If you experience problems with this feature, or if the required X11 headers aren't available, consider disabling it by typing `./configure --disable-cbpoll` prior to compilation. This will disable checking completely. If the naive, old checking is preferred, uncomment the four usual lines, `./configure --disable-cbpoll`, and compile.

### **Release Note cb2Bib 0.3.5**

Releases 0.3.3 and 0.3.4 brought querying functionality to cb2Bib. In essence, cb2Bib was rearranged to accommodate copying and opening of network files. Queries were then implemented as user customizable HTML posts to journal databases. In addition, these arrangements permitted defining convenience, dynamic bookmarks that were placed at the cb2Bib's 'About' panel.

cb2Bib contains three viewing panels: 'About', 'Clipboard' and 'View BibTeX', being the 'Clipboard' panel the main working area. To keep cb2Bib simple, only two buttons, 'About' and 'View BibTeX', are set to navigate through the panels. The 'About' and 'View BibTeX' buttons are toggle buttons for momentarily displaying their corresponding panels. Guidance was so far provided by enabling/disabling the buttons.

After the bookmark introduction, the 'About' panel has greatly increased its usefulness. Button functionality has been slightly redesigned now to avoid as many keystrokes and mouse clicks as possible. The buttons remain switchable, but they no longer disable the other buttons. User is guided by icon changes instead. Hopefully these changes will not be confusing or counterintuitive.

Bookmarks and querying functionality are customizable through the `netqinf.txt` file,

which is editable by pressing the **Alt+B** keys. Supported queries are of the form 'Journal-Volume-First Page'. cb2Bib parses [netqinf.txt](#) each time a query is performed. It looks for [journal=Full\\_Name|\[code\]](#) to obtain the required information for a specific journal. Empty, '[journal=](#)' entries have a meaning of 'any journal'. New in this release, cb2Bib will test all possible queries for a given journal instead of giving up at the first *No article found* message. The query process stops at the first successful hit or, otherwise, once [netqinf.txt](#) is parsed completely (in an equivalent way as the automatic pattern recognition works). This permits querying multiple -and incomplete- journal databases.

Users should order the [netqinf.txt](#) file in a way it is more convenient. E.g., put PubMed in front of JACS if desired an automatic extraction. Or JACS in front of PubMed and extract from the journal web page, if author accented characters are wanted.

So far, this querying functionality is still tagged as *experimental*. Either the querying itself or its syntax seem quite successful. However, downloading of PDF files, on windows OS + T1 network, **was found to freeze** once progress reaches the 30-50%. Any feedback on this issue will be greatly appreciated. Also, information on [kfmclient](#) equivalent tools for non KDE desktops would be worth to be included in the cb2Bib documentation.

### **Release Note cb2Bib 0.3.0**

cb2Bib considers the whole set of authors as an author-string pattern. This string is later postprocessed, without requirements on the actual number of authors it may contain, or on how the names are written. Once considered author-string patterns, the extraction of bibliographic references by means of regular expressions becomes relatively simple.

There are situations, however, where several author-strings are required. The following box shows one of these cases. Authors are grouped according to their affiliations. Selecting from 'F. N. First' to 'F. N. Fifth' would include 'First Affiliation' within the author string. Cleaning up whatever wording 'First Affiliation' may contain is a rather ill-posed problem. Instead, cb2Bib includes an [Add Authors](#) option. The way of operation is then to select 'F. N. First, F. N. Second, F. N. Third' and chose [Authors](#) and right after, select 'F. N. Fourth and F. N. Fifth' and chose [Add Authors](#).

Journal Name, 10, 1100-1105, 2004

AN EXAMPLE WITH MULTIPLE AUTHOR SETS

F. N. First, F. N. Second, F. N. Third  
First Affiliation

F. N. Fourth and F. N. Fifth  
Second Affiliation

Abstract: Select from "Journal Name ..." to "... second author set.". The 'F. N. First, F. N. Second, F. N. Third' author string is automatically processed as one author set, while 'F. N. Fourth and F. N. Fifth' is processed as



another, second author set.

At this point in the manual extraction, the user was faced with a red `<<moreauthors>>` tag in the cb2Bib clipboard panel. The `<<moreauthors>>` tag was intended to warn the user about the fact that cb2Bib would not be able to consider the resulting extraction pattern as a valid, general regular expression. Usual regular expressions are built up from an a priori known level of nesting. In these cases, however, the level of nesting is variable. It depends on the number of different affiliations occurring in a particular reference.

So far the `<<moreauthors>>` tag has become a true FAQ about cb2Bib and a source of many confusions. There is no real need, however, for such an user warning. The `<<moreauthors>>` has therefore been removed and cb2Bib has taken an step further, to its 0.3.0 version.

The cb2Bib 0.3.0 manual extraction works as usual. By clicking **Authors** the Authors edit line is reseted and selection contents moved there. Alternatively, if **Add Authors** is clicked, selection contents is added to the author field. On this version, however, both operations are tagged as `<<author>>` (singular form, as it is the BibTeX keyword for Authors). The generated extraction pattern can now contain any number of `<<author>>` fields.

In automatic mode, cb2Bib now adds all **author** captions to Authors. In this way, cb2Bib can treat interlaced author-affiliation cases. Obviously, users needing such extractions will have to write particular regular expressions for cases with one set of authors, for two sets, and so on. Eventhough it is not rare a work having a hundred of authors, it would be quite improbable that they were working on so many different institutions. Therefore, few regular expressions should actually be required in practice. Although not elegant, this breaks what was a cb2Bib limitation and broadens its use when extracting from PDF sources. Remember here to sort these regular expressions in decreasing order, since at present, cb2Bib stops at the first hit. Also, consider **Any Pattern** to get ride of the actual affiliation contents, as you might not want to extract authors addresses.

### **Release Note cb2Bib 0.2.7**

The cb2Bib 0.2.7 release introduces multiple retrieving from PDF files. PDF documents are becoming more and more widely used, not only to transfer and printing articles, but also are substituting the personal paper files and classifiers for the electronic equivalents.

cb2Bib is intended to help updating personal databases of papers. It is a tool focused on what is left behind in database retrieving. Cases such as email alerts, or inter colleague references and PDF sharing are example situations. Though in an electronic format, sources are not standardized or not globally used as to permit using habitual import filters in reference managers. cb2Bib is designed to consider a direct user intervention, either by creating its own useful filters or by a simple copy-paste assistance when handtyping.

Hopefully someday cb2Bib will be able to take that old directory, with perhaps a few hundreds of papers, to automatically index the references and rename the files by author, in a consistent manner. The required mechanism is already there, in this version. But I guess that this new feature will manifest some present limitations in cb2Bib. For instance, most printed and PDF papers interlace author names and affiliations. cb2Bib doesn't have the means to automatically discern an author name from a department or street name. So far one needs to manually use the 'Add to Authors' feature to deal with these situations. Also, the managing of regular expressions needs developing, specially thinking in the spread variety of design patterns in publications.

In summary, this current version is already useful in classifying and extracting the reference of that couple of papers that someone send right before submitting a work. A complete unsupervised extraction is still far away, however.

### **Release Note cb2Bib 0.2.1**

The cb2Bib mechanism 'select-and-catch' failed in some cases. Acrobat and Mozilla selections were not always notified to cb2Bib. Indeed, this 'window manager - application' connection seems to be broken on a KDE 3.3.0 Qt 3.3.3 system.

The cb2Bib 0.2.1 continues to listen to system clipboard change notifications, whenever they are received and whenever cb2Bib is on connected mode. Additionally, the cb2Bib 0.2.1 periodically checks for changes in the system clipboard. Checks are performed every second, approximately. This permits cb2Bib to work as usual, although one could experience 1-2 seconds delays in systems where the automatic notification is broken.

If the 'select-and-catch' functionality appears 'sticky', possibly happening while using non KDE applications from where text is selected, check the source file [c2bcclipboard.cpp](#), look for '[Setting timer](#)', and set variable [interval](#) to 1000. This is the interval of time in ms that cb2Bib will use to check for clipboard changes.