Software required for legacy taxonomic treatment digitisation: overview

ver. 1.2

Thomas Hamann

This project was subsidized in part by the EU project “pro-iBiosphere” (Grant agreement 312848).

Introduction

This document gives a short overview of all of the software required to digitize legacy taxonomic treatments. This list assumes a Windows environment. Sometimes alternative options are given.

Text preparation, processing, and XML proofreading

- Windows Explorer or equivalent file manager
- Microsoft Word 2003 or later
- Adobe InDesign CS2 or later if your source files are InDesign files instead of text files
- Adobe Acrobat X Pro or later if your source files are PDFs with a good OCR
- Abbyy FineReader Professional 12 or later if your source files are PDFs without a good OCR or simply without OCR
- Notepad ++(http://notepad-plus-plus.org/)[1], which is free
    - Compare Plug-in for Notepad++, to compare files
- Perl (http://www.perl.org) - Use the latest version of ActivePerl Community Edition (http://www.activestate.com/activeperl/downloads) that is suitable for your version of Windows (ActivePerl is also available for other operating systems)
- Altova XML Spy Professional 2010 or higher, or equivalent

[1] Also available as a Portable Apps version for installation on a USB stick or in a Windows Home Folder.

- Write access to your hard disk and the folder Perl is installed in
- Making back-ups of your scripts and using a version control system such as Git (http://git-scm.com/) is advisable. You can then use Github (https://github.com/) as an online repository for your files.

  Note: Git has a fairly steep learning curve, but works fine as long as nothing goes wrong.

  - Making back-ups on a USB stick is also advisable.
- The latest version of the FlorML XML schema, available at https://github.com/thoha/FlorML (most up to date) or https://github.com/ncbnaturalis/FlorML

Image processing

- Windows Explorer or equivalent file manager
- Microsoft Excel 2003 or later
- Adobe Photoshop CS2 or higher (not Photoshop Elements)
- Notepad++ (http://notepad-plus-plus.org/)
- EXIFTool (http://www.sno.phy.queensu.ca/~phil/exiftool/)
- Windows' batch file processing ability (this might require additional user rights)
- Write access to your local hard disk, to avoid problems with network connections while processing batch files and large quantities of images

Alternative programs

- Text editors:
  - Sublime (http://www.sublimetext.com/), freeware, also available for the Mac and Linux.
  - BBEdit (http://www.barebones.com/products/bbedit/), paid, Mac-only
- Version control system:
  - Mercurial (https://www.mercurial-scm.org/) with Hg-Git Mercurial plug-in (http://hg-git.github.io/)
- OCR programs:

- Tesseract (https://code.google.com/p/tesseract-ocr/), free but no graphical user interface.