A research data files & folders naming convention for the impatient

Abstract

We propose a naming convention for research data files and folders, intended as a practical tool to achieve reasonable quality without using too much time designing one's own. It is by no means mandatory and may be adapted freely if needed. Regardless of the present suggestion, it is advisable for all collaborators of a project or research group to follow a consistent practice.

This convention is intended for experimental, observational or computed data. We present how to build the full path (parent folders + file name) as a sequence of simple information elements, a few of which can be re-ordered if necessary for the project. We then demonstrate how to use the convention in practice though plausible fictitious examples from various research fields.

The convention is not directly applicable for source code and similar files, which usually follow their own rules and practices. However, it is possible to use it for parts of a dataset, see example 4 for a case that combines code and data files.

Recommended information elements and sequence

Elements used in the filenames or as folders if needed.

Elements normally present only within filenames

Project (or subproject) name or acronym.

Analysis/processing method (name or acronym) and/or type of data, source, or acquisition method.

Specific system, sample or object name, or acronym (if applicable).

Creator initals (esp. if multiple collaborators are involved in the project).

Date/time. Format: YYYYMMDD

(_hhmmss)

Numbering (using zeroes for padding if necessary : 01 not 1, more digits if the need is expected).
With possible prefix : none for iteration, V for version, ...

One essential external variable (temperature, pressure or similar) could be used here instead (with unit and padding zeros) if it is the single distinctive piece of information, but this might force to use a period for decimals => this would be better recorded in a CSV listing file.

File extension (natural for the file format)

Overview

HOW/WHAT (details)

WHO/WHEN

Depending on the project, WHERE can be used as part of **WHAT,** or treated as a subproject.

Use **HOW** and **WHAT** in the most appropriate order for your project, encoded in the filename or as parent folders.

WHO and WHEN might not be important in all projects but they can be difficult to add retrospectively => if unsure, include them right away.

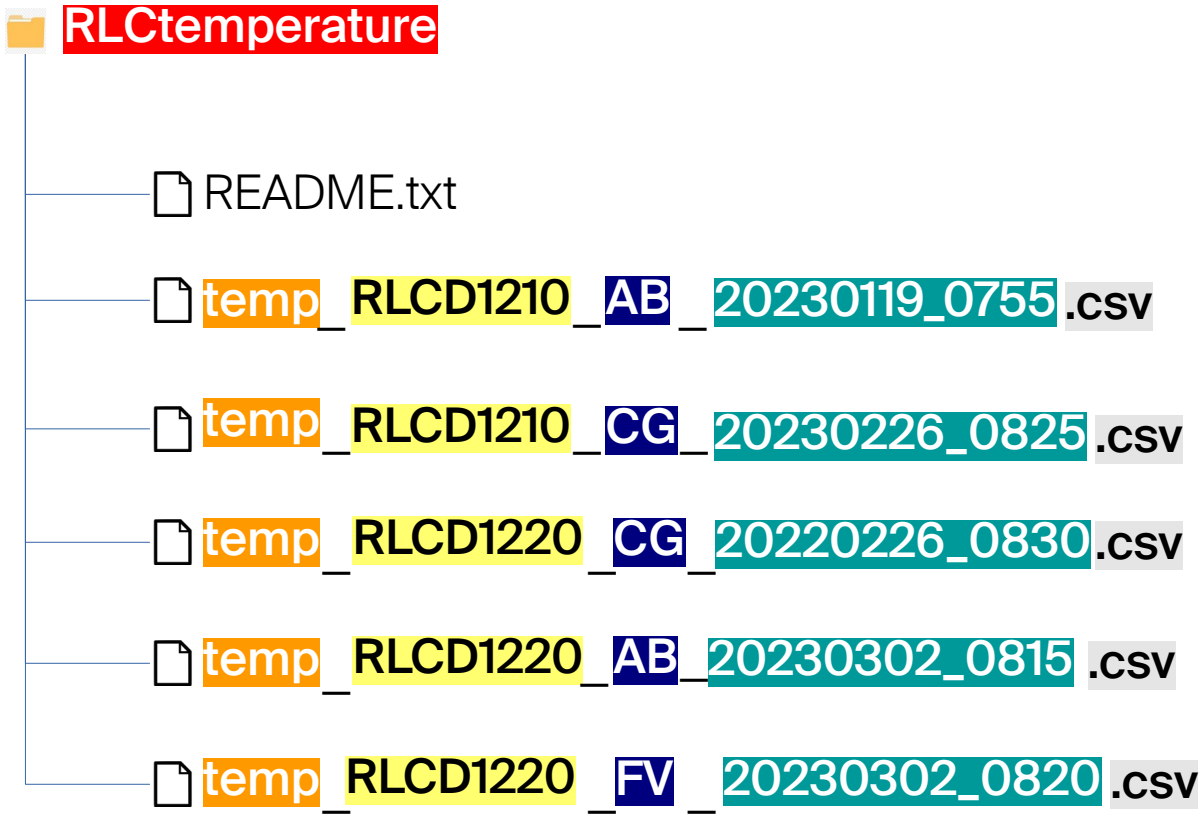
Basic rules

- Try to keep filenames reasonably short (32 characters or less) but understandable. The full path (folders and file name) should be less than 250 characters. Some information elements common to many files (for example a project name) can be conveniently used for folders instead of file names, which is helpful in this respect.
- We assume the user has full control over folder and file names, extensions excepted. If this is not the case, the user might be unable to include some elements - these should then be described in metadata or documentation files instead of filenames.
- All elements must be written using only letters and digits, plus one single period before the extension, plus underscores as separators between the elements. Dashes are acceptable as part of names of methods, systems, etc. We recommend to only use the ASCII character set (latin alphabet, no accents or other diacritics) unless there are strong reasons to do otherwise.
  To put it another way, only use the following:
  ABCDEFGHIJKLMOPQRSTUVWXYZabcdefghijklmnopqrtsuvwxyz0123456789_-
- Acronyms and collaborators initials should be defined explicitly in a README or other documentation files. Furthermore, you must include the names of all participating collaborators when preparing the metadata for a research dataset for archiving or dissemination. A contact e-mail or ORCID iDs is also strongly recommended.
- Include a README file in the root folder, and to any other folder where it can be useful.

Fictitious example 0

Temperature measurements in two EPFL rooms

Basic example : dataset with just a few files and a README

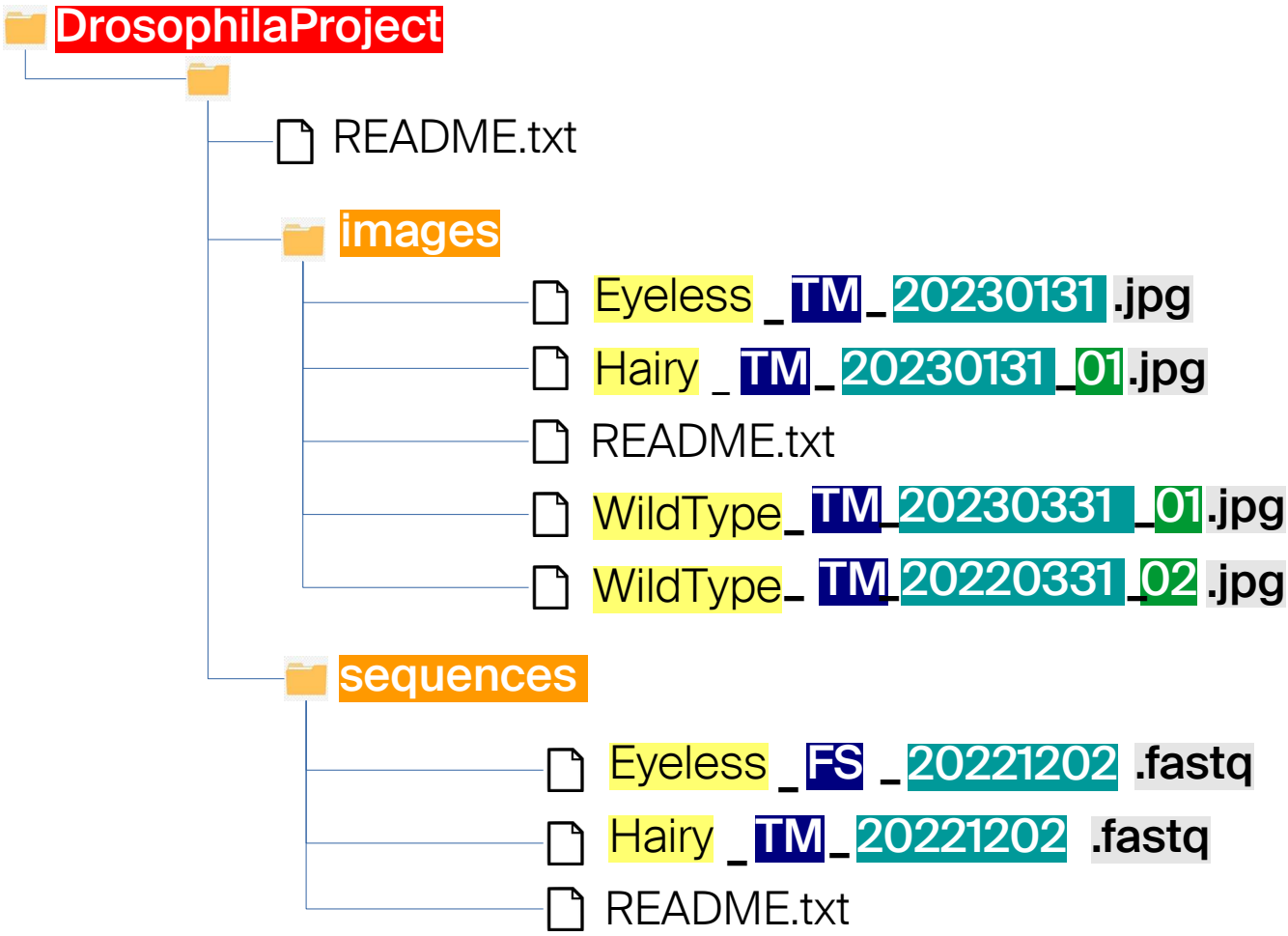


Fictitious example 1

Genotypes and phenotypes of Drosophila fruit fly mutants

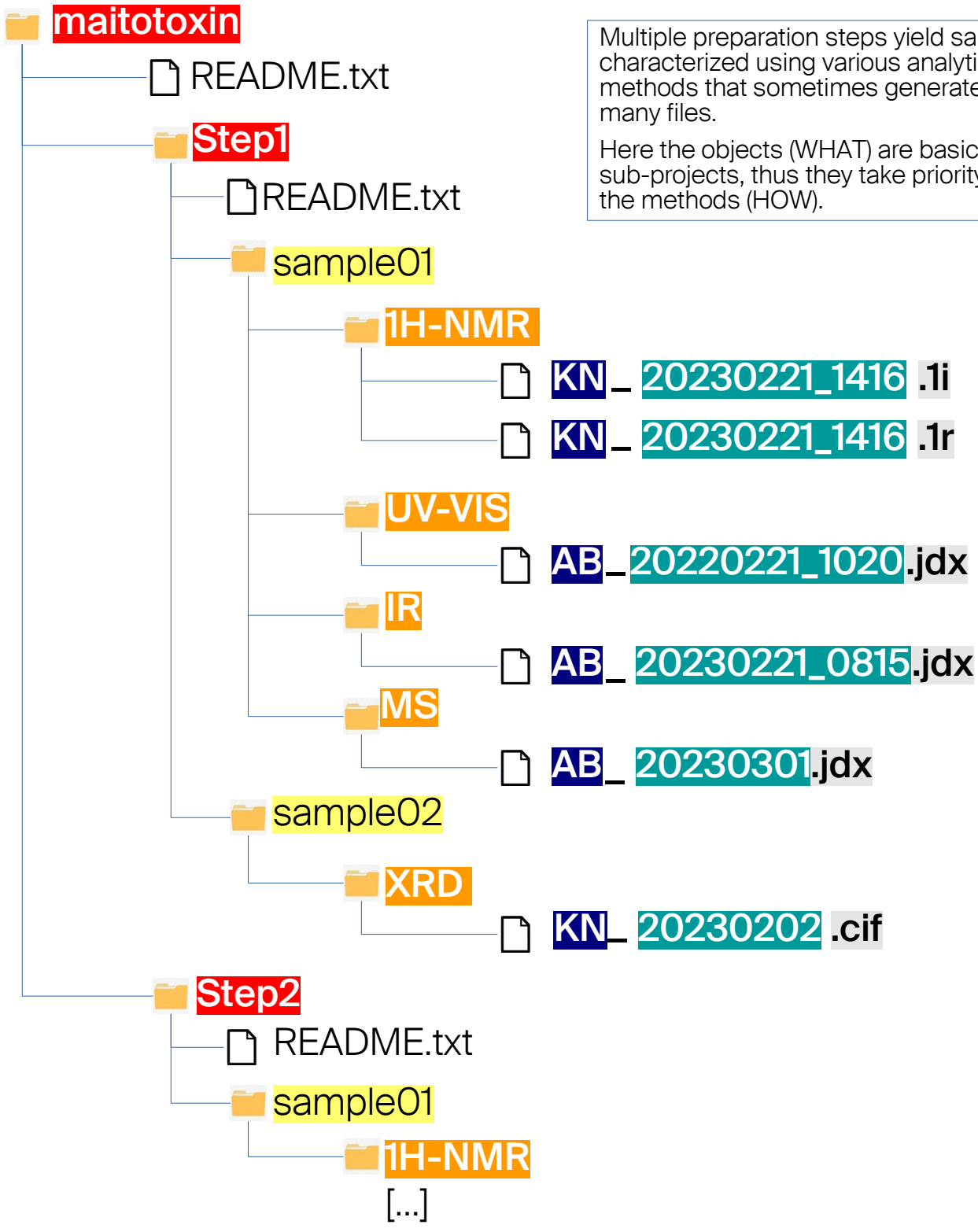
2 data types (pictures and genome sequences), organized using one folder for each.

Multiple images of flies with some given phenotype/mutations have been recorded.



Fictitious example 2

Characterization data for a multi-step chemical synthesis



Multiple preparation steps yield samples, characterized using various analytical methods that sometimes generate many files.

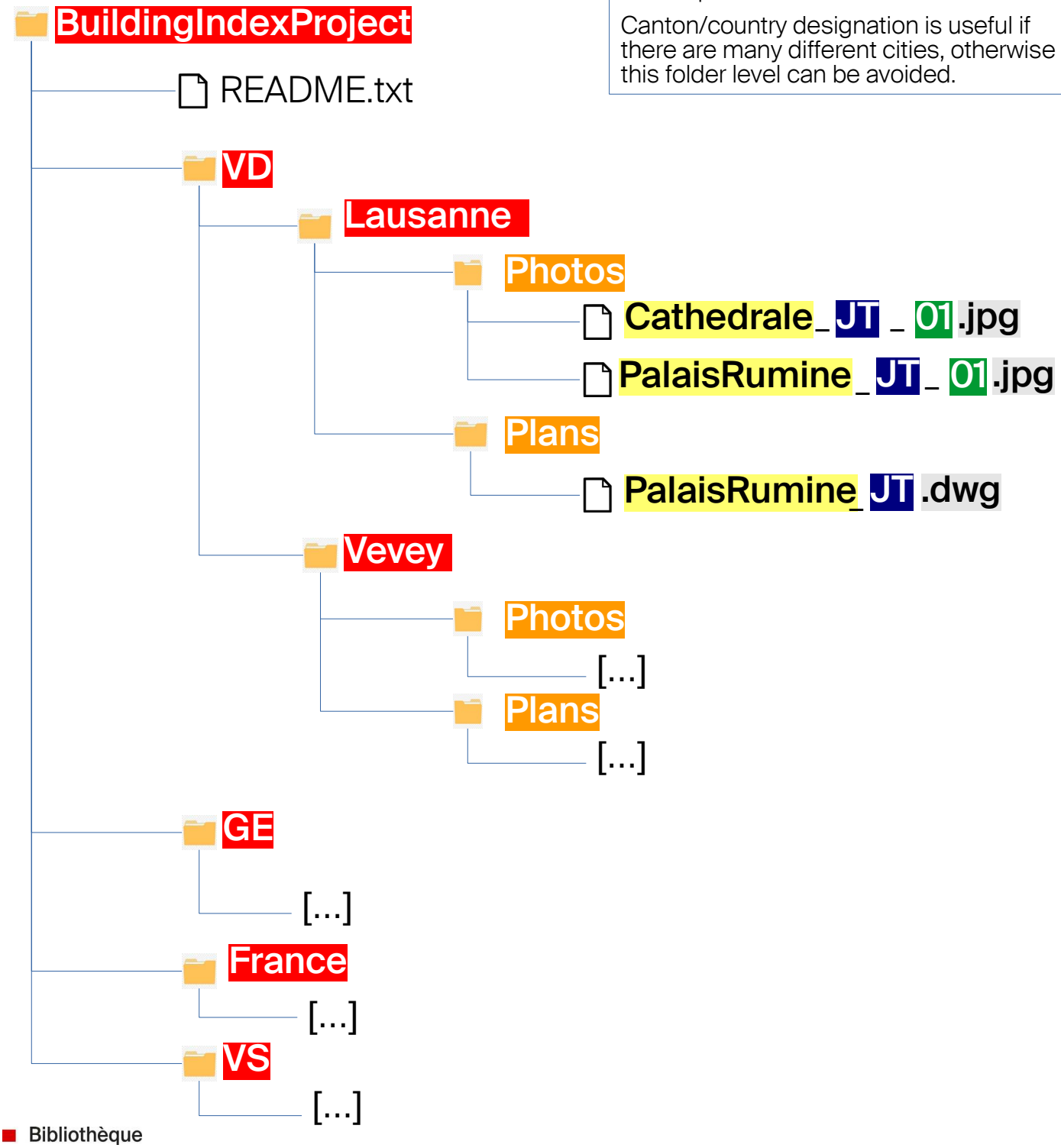
Here the objects (WHAT) are basically sub-projects, thus they take priority over the methods (HOW).

Fictitious example 3

Inventory of historical buildings around Lake Geneva

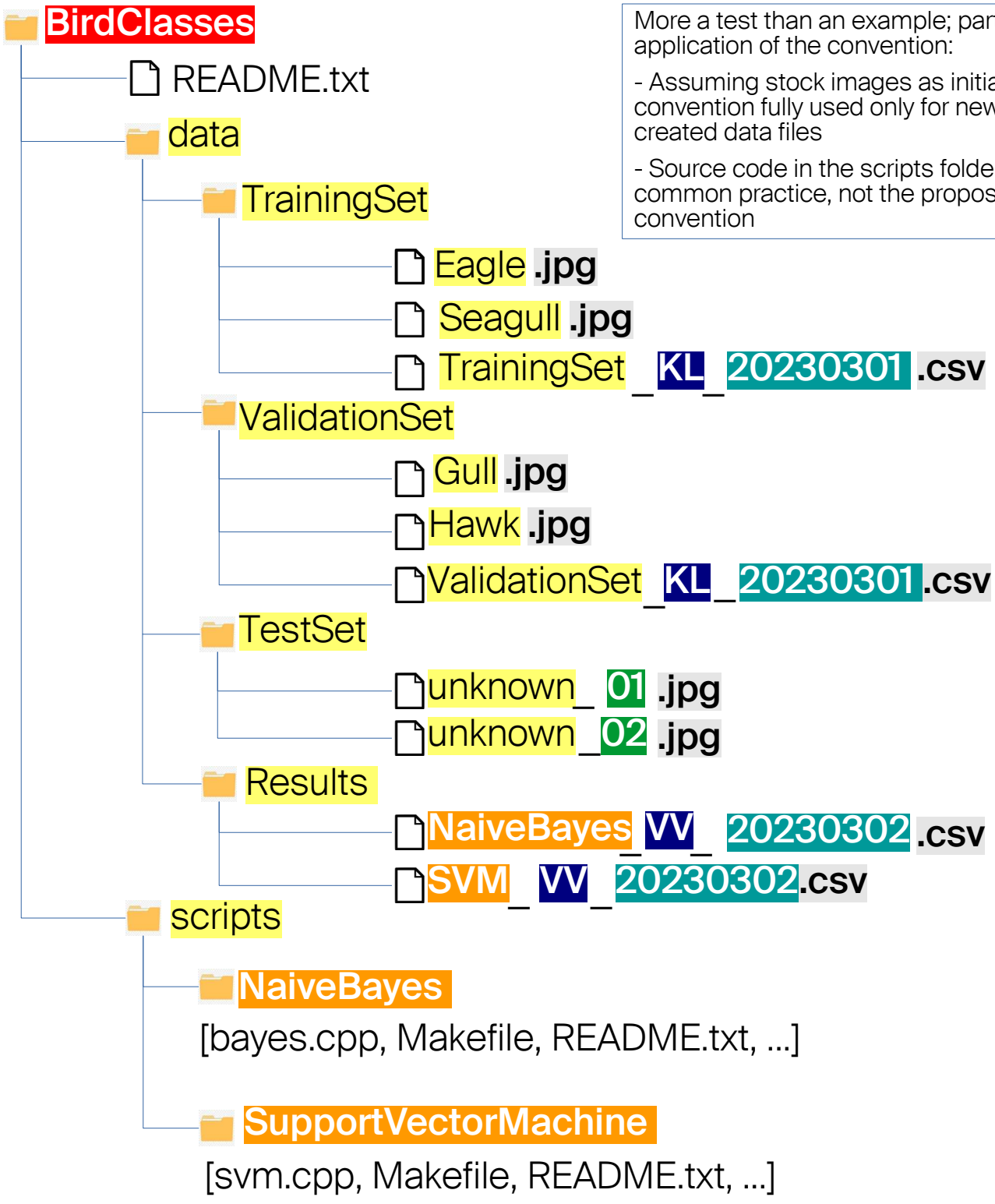
Pictures, plans, etc. for multiple buildings in multiple locations

Canton/country designation is useful if there are many different cities, otherwise this folder level can be avoided.



Fictitious example 4

Comparison of supervised learning algorithms for bird image classification



More a test than an example; partial application of the convention:

- Assuming stock images as initial data ; convention fully used only for newly created data files
- Source code in the scripts folder follows common practice, not the proposed convention