Preferred skills in people for semi-automated mark-up of legacy taxonomic works

ver. 1.3

Thomas Hamann

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

Introduction

This document aims at providing some guidelines for the type of skills that are useful during the various phases of legacy taxonomic work mark-up using Perl scripts and FlorML XML schema development, and also discusses how tasks could be distributed within a team of users with different skills to increase the efficiency of the mark-up process, including a risk assessment.

Table of Contents

Suggested skills

The following skills are suggested for users who want to perform all tasks by themselves:

Personal

- Excellent analytical skills:
  - Able to spot useful (and less useful) parts of text for scripting.
  - Able to extract generally applicable methods to process large parts of text consistently.
  - Able to determine proper processing order.
  - Able to identify problems and make improvements when required.
- Precise.
- Critical, also of self.
- Long attention span, undeterred by sometimes very large documents (plain text files several Mb large).
- Capable of unsupervised learning.
- Capable of working together usefully with both IT developers and taxonomists, speaking IT-language and yet being able to discuss things clearly with users with little computing experience.
- Pragmatic, with an attitude aiming at excellent practical results that both satisfy IT developers and taxonomists.
  - Be able to understand that what might seem a perfect IT solution may not be in line with the use taxonomists may desire, and be able to adapt such a solution in a manner that *does* work out properly for both sides, i.e. flexibility of mind.

Biology-related

- Must master main language used in taxonomic work (native speaker or university-level).
  - Basic knowledge of Latin required.
- Must have knowledge of taxonomic nomenclature and be willing to learn more.
- Knowledge of the morphology of the targeted taxonomic group(s), or willingness and ability to learn this.
- Understanding of background concepts and procedures leading to taxonomic work creation, or willingness and ability to learn this prior to starting employment (i.e. take a course).
- Understanding of the specific terminology and statistical concepts (e.g. categorical, ranges, etc.) used in character descriptions.
- Background understanding of the basic concepts of how taxonomic characters are defined and the associated pitfalls.
  - Ability to dissociate such understanding from the actual mark-up[1].

[1] Meaning that one should avoid applying their personal interpretation of a character to the mark-up itself. Definition of character-related terms is not the task of the person doing the mark-up.

- Understanding of the other topics often encountered in taxonomic works, such as geography, specimen collection activities, species concept, etc., and how these may change over time.

IT-related

- Basic knowledge of HTML/XML.
- Knowledge of XML schema design and associated programs (e.g. XML Spy), or willing to learn this.
- Capable of writing advanced regular expressions in Perl or willing to learn this.
- Knowledge of Command Line tools.
- Knowledge of Photoshop and Microsoft Office.
- Willing to learn and extend his or her knowledge of the programs above.
- Advanced knowledge of Windows may be helpful, for example to change advanced settings.

This skillset is also suggested for users who want to develop the FlorML XML schema further and/or who want to create Perl scripts.

Task suitability for different users

Table 1 shows which tasks are suitable for which type of users, based on the simplicity and skills required by the task.

Table 1: Suitability of tasks for different types of user.

Type of user	XML schema maintenance	Text preparation	Script creation	Script use	XML file proofreading	Image processing
Untrained non-technical user		X (with manual)				
Trained non-technical user		X		X		
Untrained technical user		X (with manual)		X (with manual)	X (with manual)	
Trained technical user (basic)		X		X	X	X
Trained technical user (advanced)	X	X	X	X	X	X

Non-technical users or technical users who cannot match the entire skillset can however do certain less complex tasks, and with appropriate training they can gain the ability to do more tasks.

Possibilities for distributing tasks

The different tasks that need to be completed for the mark up of legacy taxonomic works using Perl scripts can be distributed amongst a team of differently-skilled co-workers. This can speed up the work and increase efficiency. However, good communication between the people in the team is crucial for a qualitatively good result.

XML schema maintenance

The XML schema development and maintenance can be the task of a person not involved with the other tasks, as long as good feedback loops are available between all people involved. The XML schema developer should be kept up to date with regards to incompatibilities between the XML schema and the documents being marked up, while the script developer and the person doing the proofreading should be able to request the help of the XML schema developer when required. The XML schema developer should also consult expert taxonomists whenever a legacy taxonomic work contains peculiarities that complicate mark-up or that require additional functionality. Therefore they need to have the conversational skills to communicate with non-technical users, understand their suggestions or concerns, and act upon them.

Text preparation

Based on Table 1, the less technical task of text-preparation can be performed by non-technical users, who then provide the person developing the scripts with these texts and notes on potential problems, useful formats and a suggested processing order.

Script creation

Script creation requires the attention of a person with good analytical and technical skills. They should be capable of writing suitable regular expressions in Perl. More importantly, they should be able to determine the proper processing order for the scripts, both on the level of the whole taxonomic text and at the level of a single task. It also requires that this person is well-informed of the problems that are present in the actual taxonomic texts and those that remain after automated mark-up, so they can make the required improvements to the Perl scripts. This requires being able to obtain useful feedback from the other team members.

Script running

After scripts have been developed by an experienced developer, the non-technical users who prepared the text could then actually run the scripts, too. However, this present a pitfall, as script creation is initially intricately linked to running the scripts. Furthermore, technical skills are required to understand the mark-up after a script has been run. Therefore, although the task of running scripts is suited for a non-technical user, a technical user will need to assist them to actually interpret the resulting files. It also means a good feedback loop back to the script developer needs to be in place. This signifies that completely dissociating script creation and script running will likely cause problems.

An alternative option is to have the non-technical person run only the scripts that require the least technical attention, while the more complex scripts are left to the script developer. However, this does not remove the requirement for a good feedback loop, should any issues arise.

XML proofreading

The XML proofreading can be performed by a separate person who is sufficiently skilled to properly read XML and manually edit it. A good knowledge of the FlorML XML schema is also required. Again, a feedback loop towards the script developer is required. Furthermore, a feedback loop with the XML schema maintainer is required. The person proofreading the XML documents should also be able to consult expert taxonomists whenever required by peculiarities in the marked up text.

Image processing

The task of image processing involves work that anyone with basic Photoshop skills can perform, but the process of adding metadata is more complex and technically involved and requires the expert eye of a trained technical user who is comfortable with working with Windows batch files.

However, it is possible to have this work performed by a closely-knit team of two or three co-workers. This consists of having one or two persons perform the Photoshop work and basic metadata entry in Microsoft Excel, while another, more technical person then adds the metadata to the image files. In this case, the most important task is actually performed by the non-technical users, because if they make errors with the filenames or metadata the technical user will not be able to fix this.

A feedback loop should be in place between the people doing the image processing and those doing the XML proofreading to ensure the image file IDs are properly communicated.

Supporting personnel

Irrespective of whether the digitalisation of legacy taxonomic works is performed by one person or a team, supporting personnel is required to assist the people performing the digitalisation task.

Expert taxonomists should be available for any questions regarding taxonomic contents in legacy taxonomic works, especially regarding nomenclatural or specimen data. It is preferable if the experts involved are actually acquainted with the taxonomic work(s) in question, for example because they were major contributors or the editor of the work.

Likewise, other people involved in taxonomic work creation, such as editors of the taxonomic works involved, are useful as a resource. They often can help out with obvious errors that were missed during the editorial process or providing files of older computer-produced taxonomic works.

Risk assessment

Distributing the various tasks involved in the semi-automatic mark-up of legacy taxonomic texts has several risks associated with it. Below is a case-by-case discussion of these risks. Most of these are also applicable when only one person is performing all of the work.

Improper communication

The larger the team, the more communication issues will arise. This can be alleviated by having clear communication protocols in place; however such protocols may also lead to tunnel-vision side effects where problems are ignored or downplayed, damaging the quality of the results. To avoid such things, the team members can best have an open, constructive, yet critical attitude to each other's work.

Lack of expert consultation

The consultation of taxonomic experts and other supporting personnel is essential towards a satisfactory result. Ignoring this may lead to a result that is acceptable from an IT point of view, but not acceptable for the main target group: biologists[2].

Therefore, for the skilled technical workers it is important to understand that consulting with supporting non-technical personnel prior to and during development of new features is crucial. Consider having the technical workers trained in usefully communicating with non-technical users. A condescending attitude towards non-experts in matters touching the technical should be avoided as it will lead to animosity in the workplace.

Taxonomic peculiarities vs. IT practices

If the expert opinion of the taxonomic expert does not stroke with IT practices, it should not be dismissed for this reason.

Furthermore, data in legacy taxonomic works may not match current practices. For example, that data in a description in a legacy taxonomic work is not normalized does not imply it is wrong. Likewise, older taxonomic works often include taxonomic names that do not adhere to current nomenclatural rules. This does not mean that part of the data can be safely left out; the solution chosen must include it. The reason for this is simple: such data is part of the scientific (and historic) function of taxonomic works.

Therefore: Work with taxonomic peculiarities, do not eliminate them. Come up with original solutions to deal with them.

[2] A good example of such unacceptable results can be found in online databases that provide images of herbarium sheets, which sometimes are too small and/or not sharp enough to be of any actual use in taxonomic work.

Boredom due to repetitive tasks

This is a risk that should not be underestimated. By subdividing the entire digitalisation process into parts, each person involved performs only a small part of the work. Although this theoretically means the efficiency is raised, in practice repetitive work leads to a serious diminishment of attention after some time, which in turn leads to errors and a qualitatively subpar result. Due to the attention to detail that some tasks require, this may have very negative consequences. Furthermore, repetitive tasks and associated boredom also lead to a diminished mental state of the users, which may lead to further reductions in quality.

Therefore, working on digitalization in too large a team may actually be counterproductive, such that the efficiency and quality of the work are actually reduced compared to a smaller team.

Time pressure

Although some time pressure is not necessarily bad, qualitatively good results cannot be obtained if the work has to be performed in too little time. It is better to work on further reductions in the time required for manual operations by further automatisation. However, it is suggested that enough time remains allocated for the proofreading stage (= quality control).

Noisy environments

Both the script development (programming) and the mark-up of legacy taxonomic works are tasks that require much concentration and careful thinking. Noisy environments, such as large co-working workspaces with no physical separation between desks, where multiple workers with vastly different tasks cohabitate, are harmful to a proper and productive working environment. Furthermore, the sensitivity of people to noise depends from person to person. As an absolute minimum, ensure that there are rules in place regarding the amount of noise people may produce, and enforce those rules. Also provide desks with soundproof separations (cubicles) and people with equipment designed to keep out sound. Better yet, provide smaller rooms for people performing critical tasks.

Lack of task synchronisation and coordination

As most of the tasks depend on the correct execution of one or more of the previous tasks, and the users performing each task require proper feedback from the other users, the time tables for all tasks should be synchronised with each other. If this does not happen, delays will occur in the production chain.

In a large(r) team, a coordinator should be appointed to keep overview and ensure that the abovementioned risks are minimized and taken care of in a professional manner should they occur. Due to the nature of the work, this coordinator can best have a good IT-oriented technical background.

Concluding remarks

Based on the required skills, possible task distributions and the risk assessment, the following concluding remarks are possible:

- Assigning each individual task to a single person may cause more problems than it resolves. As the tasks are rather intertwined, distributing them over too many people may cause communication and synchronisation issues to arise.
- The repetitive nature of many of the simpler tasks may negatively impact productivity if each person only has to perform a single task.
- Interaction with outside experts and a good balance of biological and IT knowledge is required. Not doing so will result in a product that is unsuitable for the primary target group.

It is therefore suggested that using one of the two following strategies is optimal:

1) Have users work in parallel (where each worker performs all of the tasks by themselves).
2) Use a team of no more than 2-3 co-workers, with the script developer doubling as the team leader, as this person is the one who needs to be able to oversee the whole process for their work. Assign the simpler tasks, if required including those involving development and use of the simpler scripts, to the other, but let the script developer handle the script use of the more complex scripts.

In both cases the XML schema maintainer can be a separate person if required.