Task: LaMachine: Maintenance, support and continued development

ID: T098

Author: Maarten van Gompel

Introduction:

LaMachine is a unified software distribution for Natural Language Processing. We integrate numerous open-source NLP tools, programming libraries, web-services, and web-applications in a single Virtual Research Environment that can be installed in various forms on a wide variety of machines.

At the core, LaMachine provides provisioning scripts (powered by Ansible) for the installation and configuration of various software (either developed in CLARIAH or 3rd party). It supports multiple common Linux distributions and to a limited degree also macOS. LaMachine can build various targets, which we call *flavours*: a Docker container, an LXC container, a Virtual Machine (using vagrant and virtualbox), local installation in a user-space (virtualenv), global installation.

LaMachine serves as a distribution both for individual researchers and hosters such as CLARIN centres.

What exists:

LaMachine exists and currently distributes a wide variety of software developed in the scope of CLARIAH/CLARIN:

- Frog, ucto and all dependencies
- All FoLiA tools and libraries, including also FLAT
- CLAM; a variety of CLAM-based webservices to various software is offered
- PICCL
- gecco (valkuil)
- Various Automatic Speech Recognition projects developed at CLST, Radboud University Nijmegen, often powered by kaldi. In collaboration also with Stichting Open Spraaktechnologie.

Additionaly, we ship a large number of well-known Python libraries/tools (spaCy,pytorch,transformers), 3rd party NLP tools (CoreNLP, Freeling, Moses, Kaldi). Which software will be installed is configurable on a per-instance basis.

LaMachine provides a portal page to installed webapplications/webservices. This is automatically generated on the basis of software metadata that LaMachine collects automatically for all installed software. LaMachine uses codemeta as a sofware metadata standard, which is a collective international effort for software metadata. For researchers/developers, LaMachine ships a Jupyter Lab installation providing Jupyter Notebooks.

What must be adapted / extended / created anew:

LaMachine by definition needs to be constantly maintained as it lives in an ever-moving ecosystem of software. The main focus is on supporting actual users (again these can be CLARIN centres or individual institutions/users), e.g. people come to us with questions. We add new software where necessary and ensure LaMachine is up to date with the latest versions and runs on a wide variety of platforms.

We are very much open to inclusion of tools by other CLARIAH partners in LaMachine (provided they comply with certain software quality & sustainability guidelines). Hitherto, however, it has proven difficult to bring other partners to contribute in this project.

Why important for CLARIAH:

LaMachine ensures that a subset of CLARIAH software is practically installable and usable both on local user machines (which can be preferable in many cases) as well as for CLARIN centres or other hosters.

LaMachine also offers a counterweight to relying on the services of others, as by definition it can be installed by anyone. This also makes it a means to "bring the tools to the data", and run in restricted (even non-networked) environments. There have been two such use cases already in a medical setting.

Targeted/Actual users: developers, researchers, students, CLARIN centres, other hosters

Actual use (quantify!):

LaMachine is widely used by researchers/developers, there are a few known installations I'm aware of at universities/institutions/CLARIN centres.

Current use of LaMachine can be tracked in our statistical report: https://applejack.science.ru.nl/lamastats/lamachinestats.html

Lead: Maarten van Gompel

Estimated PM: 9PM, spread over the entire duration of CLARIAH-PLUS

Participants: Maarten van Gompel (DI, KNAW)

Actually allocated PMs: 3PM **(This is not enough for proper support and maintenance and has already been used up!!)**

Deliverables

1. (T098D1) Software: LaMachine
   - (T098D1.1) Documentation
   - (T098D1.2) Website
   - (T098D1.3) Software Image: Docker image with a default subset of software: LaMachine docker
   - (T098D1.4) Software image: Vagrant/Virtualbox image with a default subset of software: LaMachine VM image
2. (T098D2) Software: Labirinto - Powers the portal website inside LaMachine
3. (T098D3) Software: Recipes for Frog and dependencies for Homebrew: LaMachine Homebrew recipes
4. (T098D4) Software: CodemetaPy - A tool to convert software metadata between different formats, used by LaMachine. This is our contribution to the codemeta project
5. (T098D5) Service: LaMachine deployment at CLST (Radboud University, Nijmegen), hosting a wide variety of services.

Milestones

Due to the ongoing nature of this task, milestones are often defined as the task is ongoing and completed in short development cycles. At this stage, the following milestones are open:

1. (T098M1) Testing Milestone - LaMachine#182 - Integration of webservice/webapp testing/monitoring facilities (Late 2020, begin 2021)
2. (T098M2) Authentication Milestone - LaMachine#171 - Integration of a CLARIN-compatible authentication solution (aimed at Spring 2021)

If other CLARIAH partners are interested, integration of their software would make another milestone.

Changes

- Increased estimated PM from 6 to 9 (the officially allocated 3PM is very insufficient!)
- Much more elaborate description
- More explicit deliverables and milestones
- Added various dependencies of LaMachine as deliverables
- Since 14 July 2020 - This task has moved from CLST, Radboud University Nijmegen to Digital Infrastructure, Humanities A comprehensive LaMachine intallation hosting various webservices is retained and supported at CLST.

References

- M. van Gompel and I. Hendrickx (2019). LaMAchine: A meta-distribution for NLP software. Selected papers from the CLARIN Annual Conference 2018. pp 214-226 - https://ep.liu.se/ecp/159/022/ecp18159022.pdf