LaMachine v2

Maarten van Gompel

Centre for Language and Speech Technology, Radboud University Nijmegen

What is LaMachine?

- A **software (meta) distribution** for **open-source NLP software**
  - installation and configuration recipes for software
  - especially useful in case of *complex inter-dependent* software setups
  - facilitates installation on a variety of platforms
- A kind of **Virtual Research Environment** in its own right
  - initially geared towards more tech savvy researchers, aka "the 20%"
    - command line tools and programming libraries
  - **But:** also includes webservices and web applications
    - webserver with simple **portal** application
    - software configured out of the box

Different “flavours”

- A local user environment (virtualenv)
- Globally on a system
- A **Virtual Machine**
- A **Docker** container
- On a remote system (production)

Features

- A single bootstrap script and simple interactive installer: *one command starts all.* (see `https://proycon.github.io/lamachine`)
- **Modular:** you can build LaMachine with various software combinations of your choice.
- **Updatable:** once you have a LaMachine installation you can easily update it to get the latest version of sofware.
- **Three versions:** stable, development and custom
- **Test** framework

Target and audience

- For **data scientists, developers, hosting providers** (e.g. CLARIAH centres)
- Supports several major **Linux** distributions (Debian/Ubuntu, RedHat/CentOS, Arch)
- Also support for Mac OS X (to a more limited degree)
- Windows users can use the VM or the Windows Linux Subsystem

Architecture

- LaMachine consists of installation recipes (using ansible)
- It does not copy/fork any software!
- Software is obtained from the upstream providers
- Software **MUST** be in proper industry-standard **repositories**
  - *e.g. Github, Python Package Index, Maven Central, CPAN, CRAN*
  - Lamachine simply uses those (harvesting available metadata where possible)
  - in doing so; enforces/encourages some of the CLARIAH software sustainability guidelines
  - LaMachine is just a convenience or courtesy and not a requirement or substitute

Technologies

- **Provisioning (all flavours):** Ansible
- **Virtualisation:** Vagrant and Virtualbox
- **Containerisation:** Docker
  - No longer a need to write your own Dockerfile
  - Maybe Singularity support later

What is LaMachine *NOT*?

- *NOT* an NLP pipeline/workflow system; rather it may install such systems or components required by such systems.
  - *e.g PICCL (powered by Nextflow), Frog, perhaps Newsreader in the future?*
- *NOT* a system for archiving/preserving legacy software
  - software **MUST** be maintained
- *NOT* only for Nijmegen software
- *NOT* a portal to search/access data collections
  - with LaMachine you can bring the tools to the data

Next Steps

- Collaborate with other partners
  - Help other partners include their software if they are interested
  - Enhance LaMachine for demands of production deployment (with e.g. INT)
  - Offer an authentication solution
- Make a decent portal inside LaMachine
  - Working towards accommodating "the 80%" of non-technical researchers
  - Provides access to **webservices** and **web applications**.
  - This *might* be powered by the **CLARIN LR Switchboard** (by Claus Zinn)
  - Offer simple data upload and sharing facilities
    - LaMachine already takes care of a shared data mount by default
  - Provide a **Jupyter** Notebook scripting environment (Python, R, ...)
  - Harmonize software metadata (automatically harvested at run time from upstream repositories where possible!)
- Coordinate with other solutions emerging from WP3 VRE plan

Links

- **Website:** `https://proycon.github.io/LaMachine`
- **Source repository:**
  `https://github.com/proycon/LaMachine`
- **Documentation:**
  - `README.md` - *Main documentation*
  - `CONTRIBUTING.md` - *Contributor guidelines and technical specification*