Research Ideas and Outcomes 3: e12095
doi: 10.3897/rio.3.e12095

OPEN ACCESS

Software Management Plan

Gentoo Linux for Neuroscience - a replicable, flexible, scalable, rolling-release environment that provides direct access to development software

Horea-Ioan Ioanas[‡], Bechara John Saab[§], Markus Rudin[‡]

‡ Institute for Biomedical Engineering, ETH and University of Zürich, Zurich, Switzerland
§ Preclinical Laboratory for Translational Research into Affective Disorders, DPPP, Psychiatric Hospital, University of Zurich, Zurich, Switzerland

Corresponding author: Horea-Ioan Ioanas (ioanas@biomed.ee.ethz.ch)

Reviewable v1

Received: 03 Feb 2017 | Published: 07 Feb 2017

Citation: Ioanas H, Saab B, Rudin M (2017) Gentoo Linux for Neuroscience - a replicable, flexible, scalable, rolling-release environment that provides direct access to development software. Research Ideas and Outcomes 3: e12095. https://doi.org/10.3897/rio.3.e12095

Abstract

Background

Gentoo is a GNU/Linux metadistribution designed to maximize and simplify user control of the software environment. All determinants of a Gentoo environment are recorded in a small number of plain-text configuration files, from which the software make-up of the system can be reconstructed entirely. As such, Gentoo constitutes a replicable and transparent software infrastructure - as mandated by research valuing reproducibility. Of equal scientific interest is the flexibility of Gentoo's package management. All software is distributed in a rolling-release fashion, giving the user full control over which versions (including live versions and branches/tags from version control) of which programs to install, and with which compilation options. All of the above is accompanied by automatic, version-aware dependency resolution, which also tracks static library linking and prompts for rebuilds as necessary.

We believe Gentoo is excellently suited to address many of the challenges in neuroscience software management; including: system replicability, system documentation, data analysis reproducibility, fine-grained dependency management, easy control over compilation options, and seamless access to cutting-edge software releases.

New information

We have made a substantial set of neuroimaging and data analysis packages - including their entire dependency stacks - available for any system using Gentoo's Package Management Standard. Neuroscientific software now usable under Gentoo includes but is not limited to:

- Dipy (Garyfallidis et al. 2014)
- FSL (Jenkinson et al. 2012)
- Nipype (Gorgolewski et al. 2016)
- Nilearn (Abraham et al. 2014)
- PsychoPy (Peirce 2008)

Herein we describe the implementation and current capabilities of this environment, as well as its ability to accelerate and improve research.

Keywords

gentoo, portage, linux, software management, repository, dependency management, dependency resolution, neuroscience, flexible, scalable, rolling-release, live software, versioning, source-based, gnu, gnu/linux

Introduction

Neuroscientific Software Management

Neuroscientific data analysis commonly relies on a multitude of software packages, which many scientists still resort to managing manually. Across the scientific community, manual software management is a major cause of effort duplication, resource waste (Hanke and Halchenko 2011), and lack of portability and reproducibility of data analysis. NeuroDebian (Halchenko and Hanke 2012) was until recently the only notable framework for automatic software managemnet under Linux. It has met with considerable success, thus providing an incentive to make even more feature-rich scientific software management systems available to even more users.

Gentoo

The Gentoo Linux metadistribution[*1] is chiefly characterized by its FreeBSD-ports-style package manager, Portage, which conforms to the feature-rich Package Management Specification (Bennett et al. 2015). As such, Portage packages can be used by at least two other package managers: Paludis (Paludis Contributors 2016) and pkgcore (pkgcore Contributors 2016). These "packages" - called ebuilds - are short bash-like syntax text files, and contain metadata such as source code location, a brief description, dependencies, and - in case the package structure cannot be automatically understood - additional instructions for installation. This package management style makes Portage repositories extremely lightweight, and removes reposited binaries as an obligatory intermediary between users and developers. Bugs or enhancements can thus be resolved directly between the user and the source, with users able to patch local sources if they so wish, and “live” packages enabling seamless installation of the very newest upstream bug fixes or enhancements. Portage thus eases the workflow of user-developers (which increasingly many researchers are), and encourages a more active involvement of users in the testing and development process. As a consequence, using Gentoo for neuroscience eases and improves not only the management and distribution of software, but also its development.

In addition to a package-version-aware dependency graph, Portage provides USE flags - parameters which can be used to specify how packages should be built. This fine-grained control is useful for reducing disk space footprint and memory usage, but can also - among many other things - allow administrators to select whether a package is built with static libraries or not. As package version differences and library linking are a leading factor impeding data analysis reproducibility (Glatard et al. 2015), Portage is excellently suited to support not only software environment replicability, but also data analysis reproducibility efforts. Conversely, when optimization has a higher priority than exact result reproducibility (e.g. when developing new embedded systems) the ability to fine-tune compile-time options can be used to create very heterogeneous and specialized systems on a multitude of architectures.

Furthermore, the Gentoo Prefix project allows users of any GNU/Linux distribution - and even of some non-Linux operating systems - to set up a Portage software environment in userspace. This is especially relevant for researchers who use high-performance computing environments where they are not awarded administration rights (Amadio and Xu 2016). In many cases Gentoo Prefix may also be used as a more lightweight alternative to containers.

Approach

We tackle the advanced software management needs faced by neuroscience by leveraging the manifold capabilities of the Gentoo metadistribution and the Portage package manager. This task materializes chiefly in writing ebuilds for the most popular neuroscientifc packages and their dependencies, integrating these into the Gentoo ebuild repositing model, and testing the resulting environment in present research scenarios.

Ebuilds are reposited in directory trees called *repositories*, which can be enabled by the addition of a simple text file defining a small number of parameters (such as name, location, and priority) to the package manager configuration directory. In addition to the main Gentoo repository, containing just under 20.000 packages, a number of other repositories enjoy official status, and their users can rely on support from the entire Gentoo community. Of these we distribute neuroscience ebuilds via the Gentoo Science overlay (Lecher et al. 2015), which now provides just under 1000 highly specialized scientific packages. Thus we ensure our ebuilds receive support from the broader Gentoo community, and attain a higher flexibility and responsiveness compared to the main Gentoo ebuild repository.

Results

We have contributed and are maintaining ebuilds for about 40 neuroscientifically relevant software packages to the Gentoo Science overlay. This set encompasses highly specialized software, as well as a few more general scientific packages, and a number of dependencies not previously available for Gentoo. The ebuilds for dependencies not directly related to scientific applications are scheduled for migration to the main Gentoo repository.

Our *contributed ebuild set* (Table 1) incorporates the top-level packages of a full-fledged neuroscientific software environment, and is decisevly broader than what a neuroscientist would commonly require for one particular research project. We use this set to seed dependency graphs in Fig. 1. These graphs depic *all* of the software packages a Gentoo system would contain after the package manager is prompted to install the full *contributed ebuild set*. Compared to other GNU/Linux distributions, these graphs are notably small, showcasing Gentoo's capacity to create powerful, fully featured, but lightweight systems.

Table 1.

The list of packages written in order to facilitate automated neuroscientific software management on Gentoo platforms. It should be noted that very many packages with only incidental use for neuroscience (e.g. scikit-learn (Pedregosa et al. 2011)) have long been available for Gentoo and are not showcased in our *contributed ebuild set*.

dev-python/imageio
dev-python/tqdm
dev-python/moviepy
dev-python/matrix2latex
dev-python/prov
dev-python/pydotplus
dev-python/pymvpa
dev-vcs/datalad
dev-tex/pythontex
media-libs/avbin-bin

sci-biology/afni
sci-biology/ants
sci-biology/bru2nii
sci-biology/dipy
sci-biology/fsl
sci-biology/dcmstack
sci-biology/mne-python
sci-biology/nilearn
sci-biology/nistats
sci-biology/nitime
sci-biology/nireg
sci-biology/psychopy
sci-biology/pybrain
sci-biology/pysurfer
sci-biology/spm
sci-libs/itk
sci-libs/nibabel
sci-libs/nipype
sci-libs/nipy
sci-libs/nipy-data
sci-libs/nipy-templates
sci-libs/pydicom
sci-libs/scikits_image
sci-libs/vxl
sci-mathematics/mdp
sci-visualization/mricrogl
sci-visualization/mricron
sci-visualization/surf-ice

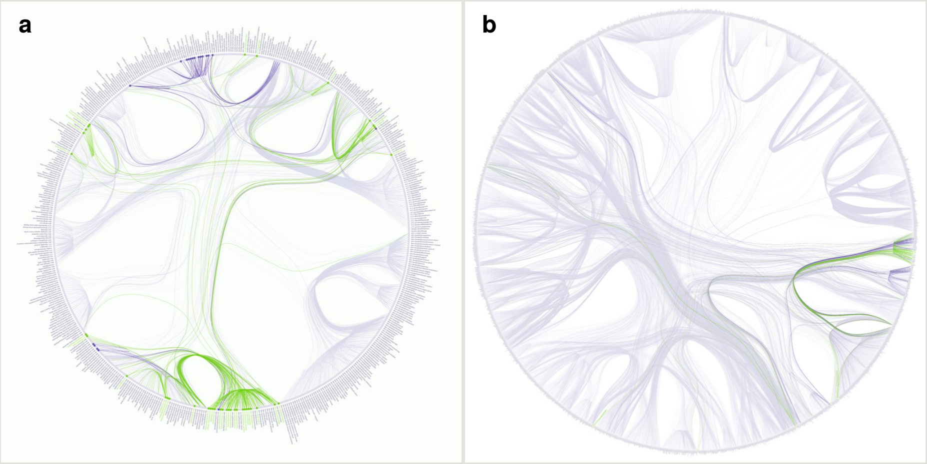

Figure 1.

Dependency graphs with hierarchical edge bundling, depicting packages as vertices and dependency relationships as edges. The graphs are seeded by the ~40 packages which we maintain and have contributed to the Portage environment primarily for neuroscience use. Graph **(a)** covers the set's entire non-optional dependency stack, and totals ~550 packages. Graph **(b)** covers the set's entire dependency stack, including all optional dependencies, and totals ~3500 packages. The seed packages and their dependency relationships are highlighted in green. Dependencies provided by the Gentoo Science repository and their dependency relationships are colored purple. Dependencies provided by the main Gentoo repository and their dependency relationships are colored purple-tinted gray. The graph shows a tight clustering of neuroscientific Python packages, indicating the infrastructure cohesiveness and application diversity of scientific Python. The graph shows that Portage neuroscience packages make use of ~20 lower-level packages from Gentoo Science - illustrating the benefit of integrating scientific software management across disciplines. It is also notable that this graph includes deep Haskell and TeX dependency stacks - which are pulled in by DataLad (Halchenko et al. 2016) and PythonTeX respectively. Both of these packages are very optional; PythonTeX in particular would only be required if the system were designed to support re-executable publications (Poore 2015). This is a theoretical system make-up, and in practice a Gentoo neuroscience data analysis system may be even more lightweight. The figures were generated with DeGraVi (Christian 2017), which makes considerable use of the graph-tool module (Peixoto 2014).

a: Minimal (excluding all optional features) dependency graph of the contributed neuroscience package set.
b: Maximal (including all optional features) dependency graph of the contributed neuroscience package set.

Neuroscientific research and teaching was performed on Gentoo platforms using our ebuilds on at least 4 physical machines and over 100 virtual machines by at least 40 students and researchers at least at 4 academic institutions. The testing process demonstrated the usability of our software management solution, and illustrated areas which could most benefit from improvement, notably the ease of distribution for base Gentoo systems.

Conclusion

We have made a comprehensive set of neuroscientific software packages available for the wide family of Gentoo distributions and derivatives. Via Gentoo-prefix, these neuroscientific software packages are, in fact, also accessible to users of many other operating systems.

Having demonstrated the feasibility of Gentoo for neuroscientific research we seek to further improve the system, by augmenting packaging with outstanding issues, and compiling a detailed overview of the easiest ways to obtain a base Gentoo distribution - tailored to popular research usage scenarios.

Acknowledgements

We thank the Gentoo community and developers, in particular Ted Rodgers, Andrew Savchenko, and Benda Xu. We thank the Gentoo Science Project members, in particular Justin Lercher and François Bissey. We thank the brainhack organizers and attendees, and the NeuroDebian and DataLad developer, Yaroslav Halchenko.

Hosting institution

Institute for Biomedical Engineering, ETH and University of Zürich

References

- Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, Gramfort A, Thirion B, Varoquaux G (2014) Machine learning for neuroimaging with scikit-learn. Frontiers in Neuroinformatics 8 https://doi.org/10.3389/fninf.2014.00014
- Amadio G, Xu B (2016) Portage: Bringing Hackers' Wisdom to Science. arXiv URL: https://arxiv.org/abs/1610.02742
- Bennett S, Faulhammer C, McCreesh C, Müller U (2015) Package Manager Specification. https://web.archive.org/web/20151006123755/http://dev.gentoo.org/~ulm/pms/head/pms.html. Accession date: 2017 1 25.
- Christian H (2017) DeGraVi -First Alpha Release. Zenodo https://doi.org/10.5281/zenodo.259741
- Garyfallidis E, Brett M, Amirbekian B, Rokem A, der Walt Sv, Descoteaux M, Nimmo-Smith I, Contributors D (2014) Dipy, a library for the analysis of diffusion MRI data. Frontiers in Neuroinformatics 8 https://doi.org/10.3389/fninf.2014.00008
- Glatard T, Lewis L, da Silva RF, Adalat R, Beck N, Lepage C, Rioux P, Rousseau M, Sherif T, Deelman E, Khalili-Mahani N, Evans A (2015) Reproducibility of neuroimaging analyses across operating systems. Frontiers in Neuroinformatics 9 https://doi.org/10.3389/fninf.2015.00012

- Gorgolewski K, Esteban O, Burns C, Ziegler E, Pinsard B, Madison C, Waskom M, Ellis DG, Clark D, Dayan M, Manhães-Savio A, Notter MP, Johnson H, Dewey B, Halchenko Y, Hamalainen C, Keshavan A, Clark D, Huntenburg J, Hanke M, Nichols BN, Wassermann D, Eshaghi A, Markiewicz C, Varoquaux G, Acland B, Forbes J, Rokem A, Kong X, Gramfort A, Kleesiek J, Schaefer A, Sikka S (2016) Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python version 0.12.0-rc1. Zenodo https://doi.org/10.5281/zenodo.50186
- Halchenko Y, Hanke M (2012) Open is Not Enough. Let's Take the Next Step: An Integrated, Community-Driven Computing Platform for Neuroscience. Frontiers in Neuroinformatics 6 https://doi.org/10.3389/fninf.2012.00022
- Halchenko YO, Poldrack B, Hanke M (2016) DataLad – decentralized data distribution for consumption and sharing of scientific datasets. In: Organization of Human Brain Mapping Poster. Organization of Human Brain Mapping Annual Meeting, Geneva, Switzerland, 2016.
- Hanke M, Halchenko Y (2011) Neuroscience Runs on GNU/Linux. Frontiers in Neuroinformatics 5 https://doi.org/10.3389/fninf.2011.00008
- Jenkinson M, Beckmann C, Behrens TJ, Woolrich M, Smith S (2012) FSL. NeuroImage 62 (2): 782-790. https://doi.org/10.1016/j.neuroimage.2011.09.015
- Lecher J, Bronder J, Amadio G, Evans B, Xu B, Fabbro S, Shvetsov A, Savchenko A, Junghans C, Szuba M, Bock N, Seifert D, Maier M, Kahle T, Kowalik K (2015) Gentoo Science Overlay Project. https://web.archive.org/web/20150919074807/https://wiki.gentoo.org/wiki/Project:Science/Overlay. Accession date: 2017 1 24.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12: 2825-2830. URL: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html
- Peirce JW (2008) Generating stimuli for neuroscience using PsychoPy. Frontiers in Neuroinformatics 2 https://doi.org/10.3389/neuro.11.010.2008
- Peixoto T (2014) graph-tool. Figshare https://doi.org/10.6084/M9.FIGSHARE.1164194
- Poore GM (2015) PythonTeX: reproducible documents with LaTeX, Python, and more. Computational Science & Discovery 8 (1): 014010. https://doi.org/10.1088/1749-4699/8/1/014010

Endnotes

[*1] As a metadistribution, Gentoo consists of a collection of tools allowing users to create their own distributions - of which many have emerged and some have gained significant popularity in their own right: Sabayon, Calculate Linux, and Kogaion - just to name a few. These are distinct from Gentoo derivatives, for which Funtoo and ChromeOS would be better examples. (All of these platforms, however, can benefit from neuroscience software management solutions designed for Gentoo.)