Free Software In Biology Using Debian-Med: A Resource For Information Agents and Computational Grids *

Andreas Tille
Robert Koch-Institut
Burgstraÿe 37
38855 Wernigerode
Germany
tillea@rki.de

Steffen Möller
University of Rostock
Institute of Immunology
Schillingallee 70
18059 Rostock
Germany
moeller@pzr.uni-rostock.de

ABSTRACT

The development of Free Software has much in common with scientific research: the sharing of knowledge and to make progress.

Software in science co-evolves with data that is available to feed it. In the data driven molecular sciences, the information technology is particularly concerned to ease the data flow between applications. This is of particular importance because of the biological data's heterogeneity while individual entries are strongly semantically interdependent. Bioinformatics has developed technologies to communicate between data and tools.

With agent and grid technologies, Computer Science has developed means to operate across multiple databases, hereto also connecting otherwise independent institutes across the world. For the agent community for the access of standard technologies and databases, and for the grid technologies in particular, a common problem is the accessibility of information sources in respective local installations. These may differ in version, location or access permissions. Because of these differences, a common infrastructure still requires considerable local maintenance, particularly for the incorporation of novel data sources.

For computational grids, VOs determine installation paths and inter-institutional access permissions. The actual work for implementing such a agreements is imposed on the respective site's maintainer. A further unresolved issue is the heterogeneity of underlying hardware. This paper describes Debian-Med, a special interest group within the Debian Linux organisation, aiming to provide a hardware-independent common view on Free software and databases for medical and biological research, and stresses its possible impact on the community as a backbone of grids and information agents in computational biology.

*This paper is available at the authors talk page

Multi-Agent Systems for Medicine, Computational Biology, and Bioinformatics MAS BIOMED'05, Utrecht, The Netherlands

CDD	Custom Debian Distribution
BOINC	Berkeley Open Infrastructure for Network Computing
BTS	Bug Tracking System
DFSG	Debian Free Software Guidelines
EBI	European Bioinformatics Institute
LSB	Linux Standard Base
RE	Runtime Environment
SETI	Search for Extraterrestrial Intelligence
SRS	Sequence Retrieval System
VO	Virtual Organisation

1. INTRODUCTION

1.1 Information integration in bioinformatics

The gold standard for the integration of information in bioinformatics today still is the SRS platform of Lion Bioscience as publicly accessible at the European Bioinformatics Institute (EBI) [28]. The developers take a particular pride in the flexibility of the tool, facilitating a straight-forward integration of local databases, a feature signing responsible for the commercial success of this technology far beyond its application in bioinformatics. While still available to the academic community for no money, the number of installations of SRS is surprisingly low. This speaks for the quality of the EBI installation, but to our understanding also for the extra burden such a system imposes in terms of maintenance. A companion tool, Prisma, addressing automated updates, is not free.

The problem of maintenance and coherence across sites is particularly obvious in the myGrid effort [21]. The development of workflows for the repository of data and their analysis across multiple tools and sites is mostly static. It is not possible for a machine already contributing to myGrid that has idle time, to help out other machines that are busy by a dynamic addition of services to those it already offers. Such would at least require the installation and deinstallation of the service's respective runtime environment in a fully automated manner. Further problems would require to be addressed for the notification of other sites in order to become aware of such a change and to subsequently react to to it. Such issues are addressed in agent research and standardised middleware like CORBA. The here presented Debian Linux distribution has implemented such dynamics.

The sharing of workloads in homogeneous environments is addressed by grid initiatives, like by the Globus-based [9] NorduGrid initiative[8]. For each site, the workload is addressed by one or multiple clusters of homogeneous machines, the coherence of in-

stallations across institutions is coordinated and supervised by so called virtual organisations (VOs). Those define Runtime Environments (REs) and clusters adhering to such specify such in their description that is utilised in the selection of clusters feasible for a job's execution.

To deploy one's algorithms and data, the programs are submitted as source code and (unless programmed in interpreted scripts or other hardware-independent languages) compiled prior to execution. VOs may ease this burden and require the sites to install a minimal set of programs for every machine by the inclusion on an RE. While this has been proven to be very functional for the grids' roots in particle physics, for bioinformatics, with its vast heterogeneity of small applications and comparatively tiny databases, even if an agreement could possibly be reached in a VO, the effect could hardly be maintained by a site's maintainer.

1.2 Status of Free Software in Biology

Common programs like a web server, or a mail user agent are installed on most computers and have a very large user base. Knowing this, many gifted programmers feel obliged for this kind of Free Software - they just need it for their own and they know to make a difference to the world. So one finds a fast, growing community around Free Software packages that have a widespread use. For specialised software in general and particularly for biological software, one needs to first explain what a particular software does. Often the developers get their satisfaction not from the software but from the excitement of extra insights in biological processes - the beauty of which is likely not to be accessible to regular software enthusiasts.

As a view shared by many in Open Source Bioinformatics, Ewan Birney of the EnsEMBL [11] project stresses Open Source to "ensure scientific progress". He also laid out that the Open Source of programs is a comparatively trivial issue when compared with the openness of biological data. The latter is often far more expensive to produce. While giving the data away might possibly diminish one's competitive advantage, the sharing of a program with others and respective citations is well accepted as fostering one's career.

1.3 Debian, distributed computing, and bioinformatics

This paper presents a development from the Debian Linux Free Software community that offers solutions to the prior mentioned problems. The basic idea is to automate updates, to share maintenance load under governance of a peer controlled strict policy and to use asymmetric cryptography for the packages' integrity and authenticity. The underlying technology may be applied to any program and be run independently from Debian. The technology has been ported to 11 different platforms, hence providing a considerable freedom in the selection of hardware.

A subproject of Debian is Debian-Med[23], a Custom Debian Distribution [24], supporting users with a special interest in medical and biological problems. The common goal of all CDDs is to make installation and administration of computers for their target users as easy as possible, and to serve in the role as the missing link between software developers and users well. The importance of this technology for computational grids and information agents has not been presented before.

2. METHODS

2.1 Unique technology supporting Debian's principles in packaging

2.1.1 `.deb` *packages*

Most distributors ship their distribution in binary packages. Two package formats are widely used[10]:

RPM (RedHat Package Manager) which is supported by RedHat, SuSE, Mandrake and others.

DEB (Debian Package) used by Debian and derived distributions.

It is this *adherence to policy* that causes a distribution to remain consistent within its own bounds. At the same time, this is the reason why packages can not always be safely installed across distribution boundaries. A SuSE `package.rpm` might not play well with a RedHat `package.rpm`, although the packages work perfectly well within their own distributions. A similar compatibility problem could also apply to packages from the same distributor, but from a different version or generation of the distribution.

For Debian, all files of a program are packed as the `ar` archiver, otherwise known for the creation of programming libraries. Every such DEB package has a companion source package from which the binary package may be built automatically. The distributions of SuSE, RedHat and derivatives use the RPM format. The program `alien` by Joey Hess (not to be mistaken for the Grid environment at CERN) can translate between these. Though scripts to be run upon installation may get lost in the translation process.

2.1.2 *Build daemons*

Sites external to the Debian main distribution may offer packages only for a subset of architectures. The Debian main distribution, however, automatically compiles software for all 11 architectures that are supported by the Debian effort. To get a package into Debian, be it novel or an update of an existing package, the maintainer of a package submits the source code of the program together with his changes on the code to create the package. The build daemons (or autobuilders) compile the packages for each of the supported systems and make the resulting package publicly available for download. Logs of the build platforms are available online for everybody's inspection.

2.1.3 *Bug tracking system*

Users should give immediate feedback about problems arising in using a package. They always have the choice of reporting these to the upstream developer, usually per email. A particular strength of SourceForget.net is to bring users of a particular software together. Earlier than this effort was the Debian Bug Tracking System (BTS). The maintainer of a software can decide if the bug should be forwarded to the upstream developers of the package or if it is fixed by himself. All problems are made public and hence the whole community may contribute to solving a particular issue.

2.2 Divide and conquer of package maintenance

The Debian Project is an association of individuals who share the intention to create the best possible free operating system. This operating system that which is created is called Debian GNU/Linux, or simply Debian for short. Everybody in the internet may initiate a site and offer packages for the installation in Debian. A local administrator has to decide, if this public source may be trusted.

For Free Software development to work it requires a critical mass of supporters. Development without feedback prior to the submission of the final product is disadvantageous. The development of programs is not the main concern of a regular Linux Distribution. However, with the focus on Free Software and smooth local compilation, Debian considerably eases the contribution of comments and concise feedback of the technically skilled early adopters. Debian such helps to bring developers and users of applications together.

All members of the Debian project are connected in a web of trust, which is woven by signing GnuPG (www.gnupg.org) keys1. A central requirement to become a member of the Debian project is to have one's GPG key signed by an already accepted member of the Debian community. When Debian developers first meet in person, they sign each other's keys. Thus, the web of trust is woven.

Debian does its best to have every member profit of somebody else's work as quickly as possible.

2.2.1 Debian Policy

All GNU/Linux distributions have a certain amount of common ground, and the Linux Standard Base (LSB)[27] is attempting to develop and promote a set of standards that will increase compatibility among Linux distributions, hereby enabling software applications to run on any compliant system. The very essence of any distribution, (whether delivered as RPMs, DEBs, Source tarballs or ports) is the choice of *policy statements*.

Policy statements in Debian[12] specify configuration files to reside in `/etc/\$package/\$package.conf`, logfiles go to `/var/log/\$package/\$package.log` and the documentation files to be located in `/usr/share/doc/\$package`. CGI-scripts are installed in `/usr/lib/cgi-bin`.

The policy statements are followed by the tool-chains and libraries used to build the software, and the lists of dependencies, which dictate the prerequisites and order in which the software has to be built and installed.

Policies in Debian are developed within the community. Commonly with a single person or a small team drafting it, with further refinements being discussed in respective mailing lists.

While every single maintainer of a Debian package has to build the package in compliance with the policy he has the ability and the right to decide which software is worth packaging. Normally maintainers choose the software which is used in their own work and they are free to move the development of Debian in a certain direction (as long as they follow the rules of the policy). This is referred to as *Do-o-cracy* in Debian which means: The doer decides what is done.

2.3 Selection of packages

Debian contains nearly 10000 binary packages, and this number is constantly increasing. There is no single user who needs all these packages. The regular user is interested in a subset of these packages. To specify packages of one's particular interest, several options are provided by Debian:

tasksel Provision of a reasonable selection of rather general tasks that can be accomplished using a set of packages installed on a Debian GNU/Linux system. However, these are not yet covering scientific applications. The CDD toolkit which is currently developed will also support `tasksel` to enable selecting for instance Debian-Med right after a fresh installation of a general Debian system.

standard package management `dpkg` and `apt` provide means to search for packages of particular interest by its name or words in the package's description. Every package also indicates, as set by its maintainer, references to other packages of potential interest.

In its current development, an ontology of applications of software, upon which semantical queries could be performed, analogous to Moby-S and BioMoby [17, 26] effort, is not available.

A package management system is a very strong tool to manage software packages on your computer. A large amount of the work of a distributor is building these software packages. The Debian package management tools have been ported to MacOS X[22] and other Linux Distributions[14].

Debian officially maintains 11 different architectures with many more not officially supported ports to other operating systems, which includes some that run another flavour of UNIX. Its technology for package management has been adopted for other operating systems, i.e. Fink on MacOSX (fink.sourceforge.net).

A *distribution* is a collection of software packages around the GNU Linux operating system that satisfies the needs of the target user group. There are general distributions, which try to support all users, and there are several specialised distributions, which each target a special group of users.

Distributors are those companies that are building these collections of software around the GNU Linux operating system. Since the software is Free, the user who buys a distribution pays for the service that the distributor is providing. These services might be:

- Preparing a useful collection of software around GNU Linux.
- Caring for smooth installation that the target user is able to manage.
- Providing software updates and security fixes.
- Writing documentation and translations to enable the user to use the distribution with maximum effect.
- Selling Boxes with ready to install CDs and printed documentation.
- Offering training and qualification.

The best established Distributors of GNU/Linux systems are Mandrake, RedHat, SuSE (now owned by Novell) and Debian. Linspire, Xandros, MEPIS and Ubuntu are well known derivates of Debian.

3. RESULTS

3.1 Bioinformatics and Debian-associated repositories

Bio-Linux Bioinformatics package repository The Bio-Linux Bioinformatics package repository contains the Bio-Linux

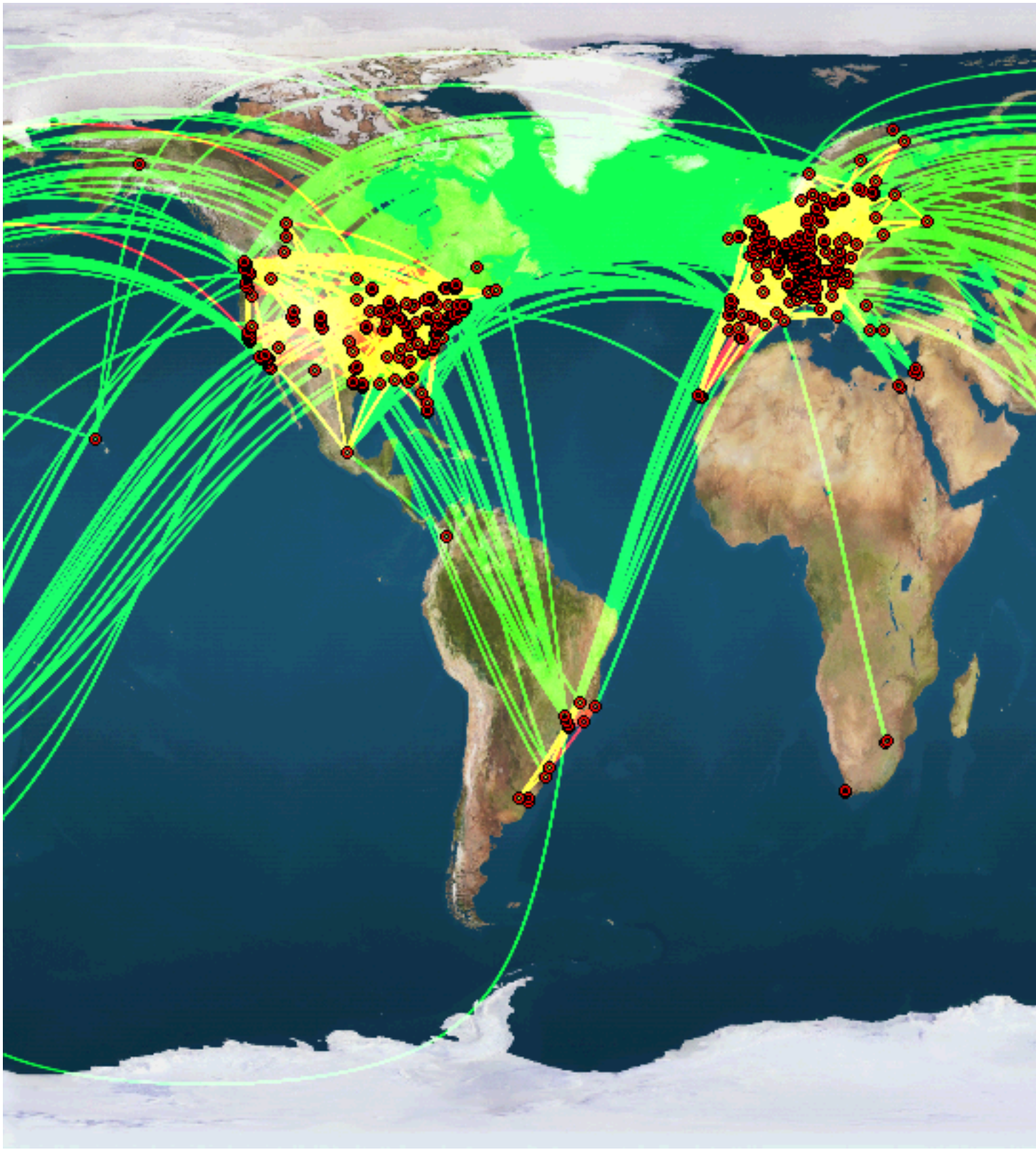

Figure 1: Debian web of trust

4 bioinformatics software and can be installed from a centralised repository located on the EGTDC server. The packages available from this site have been created by the EGTDC specifically for the Bio-Linux project and are in deb format.

The projects claim to be compatible with any installation of Debian or Debian variants. This might be a weak sentence because Debian variants are quite different. It is better to say exactly Debian "stable" or Debian "testing" and draw a line between native Debian and perhaps commercial derivatives which might differ in certain aspects.

The packages are not, however, core Debian packages. Please read the notes below for further details about Bio-Linux packages. They add the following notes:

1. *All* of the packages have a dependency on the package `bio-linux-base-directories`, as this is the package that sets up the directory structure and dependencies of several of the Bio-Linux system packages. This creates a number of directories and configuration files for Bio-Linux but will NOT overwrite anything on your system. Therefore you may wish to install this package first, or force the install of other packages if you do not want to install this package.
2. *Warning*: The `bio-linux-bldp-files` package contains files which install to a folder called `/var/www/boinformatics`. Thus, please ensure that you do not already have a directory on your system of this name, or that, at the very least, you are not concerned with overwriting any index.html file in such an existing directory.
3. All packages install to `/usr/local/bioinf` and create symlinks in `/usr/local/bin` with the exception of the bio-linux-genespring software which installs to `/home/db/SiliconGenetics` (for more information regarding installation of GeneSpring, see our GeneSpring Web Page.)
4. `Bio-linux-base-directories` installs two files into `/usr/local/bioinf/config\_files` (`bioenvrc` and `aliasrc`). The file *`bioenvrc` should be sourced* by your shell if you want to pick up the correct environmental variables for the software. We recommend the use of the z-shell for Bio-Linux. You will also want to `sourcealiasrc` if you wish to run some of the Java based software - for example `maxdView`, `maxdLoad2`, `mesquite`, `forester`, etc. You will *also need to edit* the `aliasrc` file to reflect the location of the Java binaries on your system.

The authors did a great job in collecting a certain amount of very useful software for biologists and the Debian-Med project might try to include some packages which are not yet available in Debian officially.

BioLinux-BR Project A similar project is the BioLinux-BR Project which is a project directed to the scientific community. Their goal is to create a Linux distribution for people with little familiarity with the installation of the operational system and mainly for people who do not know to proceed unpacking a program, compile and install it correctly.

For these reasons, they want to give contribution, developing a Linux system that aims to be easy to use and still offering packages that will be part of the BioLinux-BR. Giving this option, we will not be compelling anybody to change its Linux distribution, since there are packages developed for the greater Linux distributions.

"BR" here stands for "Brasil" which might lead to the assumption that some internationalisation effort was done, but according to the authors this was not their main concern but rather a side effect.

In fact, this project has assembled a huge amount of packages, probably the most complete collection of Free Software in biology. Packages for multiple distributions are provided, which includes Debian, and a live CD.

bioinformatics.debian.net Debian developer Matt Hope has created a small set of packages for Debian that he has not yet uploaded to the main distribution. Most are in a very nice state he just should submit.

3.2 Bioinformatics Live CDs

The concept of a live CD allows to create a CD or DVD that boots a computer, starts a defined set of application without a user's intervention and has all tools in place that suits a particular community. Such provide fully featured Linux workstations without additional installations of access to local disk space, alternatively booting via the network is supported by Debian, which particularly appeals to Blades or large clusters. The most successful such LiveCD is the Debian-derived Knoppix[13].

The descriptions below only summarise a few efforts but give insights into the vast realm of possibilities also for distributed computing. It should be pointed out that all Knoppix-derived[13] bioinformatics Live CDs may become members of the NorduGrid with the effort of Niels E. Larsen[15], who provides a respective script to create a LiveCD that lets one immediately join a grid with 4000+ CPUs[16].

The Quantian Scientific Computing Environment Quantian is a remastering of a well established effort (Knoppix). Recent versions of Quantian [7] are based on ClusterKnoppix [25] and add support for OpenMosix [6], including remote booting of light clients in an openMosix terminal server context. Earlier releases are still available; see below for URLs for downloads as well as ordering information. More detailed information are available at the Quantian homepage.

The interesting part for biologists is that Quantian contains in addition all interesting packages of Debian-Med. The author Dirk Eddelbuettel, who is a Debian developer himself, just used the simply to install biological software feature we provide and thus made a great profit from Debian-Med.

Vigyaan - the biochemical software workbench Vigyaan is an electronic workbench for bioinformatics, computational biology and computational chemistry. It has been designed to meet the needs of both beginners and experts. VigyaanCD is a live Linux CD containing all the required software to boot the computer with ready to use modelling software. VigyaanCD v0.1 is based on Knoppix v3.3.

Vigyann contains some programs which are not yet contained in Debian. It might be mutually beneficial to include these provided that the license fits the DFSG.

BioKnoppix BioKnoppix is a customised distribution of Knoppix Linux Live CD. It is a very similar project to the previous

which specialises Knoppix for computational biology and chemistry.

VLinux Bioinformatics Workbench Also VLinux is at the time of writing a Live CD based on the same outdated Knoppix version 3.3 as Vigyann and includes a slightly changed software selection and surely a different background layout.

These are too many different initiatives that could all well do much more in order to share the burden of maintenance and updates. With Debian they have he right basic infrastructure. The time will show, whose packages will gain most momentum.

3.3 Bioinformatics and Debian-Med

Debian-Med is a Custom Debian Distribution with the aim to develop Debian into an operating system that is particularly well fit for the requirements for medical practice and research. The goal of Debian-Med is a complete system for all tasks in medical care which is build completely on free software.

On the technical side Debian-Med contains a set of meta packages that declare dependencies on other Debian packages, and that way the complete system is prepared for solving particular tasks. A special user menu will be created to enhance usability for the user working in the field of medicine.

On the organisational side the project tries to attract people working in the field of Free Software in medicine to share the effort of building a common platform which reduces the amount of work for developers and users. Sharing the effort in publishing free medical software to a large user base is a main advantage of Debian-Med.

Currently inside Debian-Med applications are provided in certain categories: medical practice and patient management, medical research, hospital information systems, medical imaging, documentation, molecular biology and medical genetics and others. The last part seems to be the most interesting and will be introduced in more detail.

There are two so called meta packages which are named `med-bio` and `med-bio-dev`. The sense of a meta package is that you have to install only one single package using a package management software inside Debian to get all interesting packages which are necessary for a single task. For instance if a user types in:

```
apt-get install med-bio
```

all applications inside Debian which are related to the field of molecular biology and medical genetics will be installed. Moreover system users will get an extra menu which contains all these applications. The `med-bio-dev` package just installs programming libraries and tools which are interesting for users who want to develop biological applications (for instance the NCBI library[4, 1, 3]).

The strength of Debian is the huge number of developers (more than 1000) all over the world working in different fields1. Some of them are working in the field of biology or medicine and thus have a natural interest in developing a rock solid system they can relay on for their own work (not only commercial interest to sell service per accident). That is the reason why Debian is often the platform of choice for researcher in the field of biology: They just find what their colleagues all over the world are using. The more the Debian user in the field of biology report back about problems or wishes the more Debian maintainers are able to enhance their system for their own and their users profit.

Programming libraries:	BioPerl, BioPython
Sequence similarity:	BLAST2, Cluster3 *, ClustalW, e-PCR
Pretty printing:	Boxshade, TeXshade *, Textopo *
Phylogeny:	Molphy, Phylip, Treeview
Tool collection:	EMBOSS **, EMBASSY **, Bioconductor *, ncbi-tools library and programs, ARB, Primer3 *
Molecular modeling:	Garlic
Gene detection:	Glimmer, Artemis **
Genetics:	R/qtl
Viewers:	Rasmol, Treetool
Pattern discovery:	SMILE *, HMMer

Table 1: Overview on packages in Debian-Med. * marks packages of collaborators of Debian Med that are not yet part of the Debian main distribution, ** marks Debian packages made available through third parties.

In order to achieve a cooperation between virtual organisations, which seems essential since a single Grid has many virtual organisations and the virtual organisations may work across Grids, an infrastructure and policies for their cooperation is required. We regards this as equivalent to the collaboration between individuals, which is exactly what Debian was developed for.

Debian harbours the most known and well accepted tools in bioinformatics. An incomplete overview of these is given in table 1. A more detailed overview is available at the Debian-Med website. Debian has means to inform the community of programs that should be packed and the Debian-Med mailing list serves for an additional information transfer.

4. DISCUSSION

Technically the Debian community has implemented functionalities for package management that have not been seen elsewhere. Originally motivated to achieve platform independence, all packages in the Debian Main distribution are required to be completely compilable and installable by a standardised set of commands. For all platforms, build daemons fully automated build binary packages upon the package maintainer's submission of a new source package. The principle to dynamically add all required packages (i.e. specialised libraries) and the subsequent removal of these packages, should be adopted by the agent and grid communities, fostering increased flexibility and the better utilisation of human and machine resources.

The Debian-Med project serves as a common platform for all Free Software that may be utilised in medical care. Tools developed in computational biology is just a part of it because it is an important brick in medical science. With Debian-Med's ambition to become the platform of choice for biological work, conform with the principles of the Debian Policy [12], by the means of the distribution of development within the Debian Society, a well established reference platform for bioinformatics research and its medical applications has evolved and will continue to improve. The organisation is open, both to new members and to external sites offering packages for installations. Initiatives for agent research and computational grids should strongly consider to utilise the prepared packages.

From the perspective of data security for Grids or Agent Environments, it should be stressed that Debian GNU/Linux has the unique feature of the automated creation of `chroot` environments. In a dedicated directory, the minimal set of programs is stored to run a a Debian GNU/Linux system. A process started in such will not have file access to the remainder of the system with otherwise complete functionality. The process is separated, even when started as root. Hence, arbitrary packages can be added and computations performed, be it for the build daemon, for the grid or for agent environment, without accidentally or intentionally impeding the functioning of the underlying operating system.

The concept of integrating common tools in Bioinformatics to form a grid is recently exploited in the Lattice project in the group of Michael Cummings [19]. It is based on the libraries of the BOINC platform [5], the successor technology of the internet-distributed search for extraterrestrial intelligence from Berkeley, for which libraries for Debian exist.

4.1 Differences from other distributions

The Debian GNU/Linux distribution differs from others in several ways. Firstly, Debian is a non-commercial organisation of volunteers, that does not sell anything. The second and most appealing difference is the peer review and continuous pressure among the members to provide a high quality of packages. The Debian society has a constitution, elects its leader, and transparently describes policies for the creation of packages utilising specific technologies.

With these principles, Debian achieved the largest collection of ready-to-install Free Software on the Internet.

4.2 Remaining issues

Programs not submitted to the main Debian distribution, i.e. for the size of the binary, because of being of interest only to a very small user base or because of license restrictions, will require to be maintained externally. A Grid's VO might decide to create such a shared repository also for an improved accessibility of a repository for non-debian in-house developers, hereby improving the communication with the Grid's user base and its developers.

4.2.1 Licensing issues

Several existing programs that might be useful for specialists are not free in the sense of the Debian Free Software Guidelines [20]. Programs that are incompatible with the DFSG cannot be included in the Debian main distribution. One famous example of this group is PhyLip and the same hold for ClustalW. Both programs' licenses contain a clause like

> Permission is granted to copy and use this program provided no fee is charged for it and provided that this copyright notice is not removed.

As a consequence no reseller of Debian would be allowed to sell Debian because one CD contains PhyLip. Hence, this program may not be redistributed per default with the main and essential tools of Debian. Problems are, starting with those for the user:

- Need to obtain PhyLip from a different source
- If the user is not using a "common" architecture like `i386` he might run into problems in compiling the latest version as only the packages of the main distribution are submitted to the build daemons to save resources.
- The user might not even notice that something like PhyLip exists at all. Debian-Med cares for pointing users to relevant software and thus the user will be pointed to each single program package *inside* Debian which of interest for biological research.

From the developers point of view we face also drawbacks:

- Possibly smaller user base (see last point above)
- Fewer bug reports and thus lower chance to increase the quality of their software.
- Porting problems to different architectures might not be revealed early.

While every author is perfectly free to choose this kind of license, Free Software experts agree that this kind of restrictions is possibly a drawback for those programs because they do not fully use the spinning power of Free Software development.

4.2.2 Importance of community support

That strong support within the community of users is essential for the development of software, for quality assurance, feedback on features, and not at least for the motivation of staff, all commercial distributors are well aware of. E.g., RedHat has initiated Fedora as a free supplement to their commercial distribution. It is this reason why Debian-Med is part of Debian and why groups external to the Debian society, like BioLinux, are also keen on close collaborations with the community.

4.2.3 Road map to come closer to Debian-Med

1. Join the Debian-Med mailing list.
2. Check what projects are missing and ask Debian maintainers for official inclusion. There is a sponsoring program by which even non Debian developers can provide packages which are checked and uploaded by official maintainers. There is no point in keeping good quality softwares outside of Debian.
3. Verify whether one needs special configuration for your project. If yes, verify which possibilities are given in the Custom Debian Distribution effort. It is more than collecting software but bringing the software to your target users while taking the burden from any configuration issues from his back.
4. The only reason to keep things outside of Debian are licenses which are not compatible with DFSG. All other parts of your projects can be included and your time for everyday package building tasks can be saved and the workload shared with other people following the same road.

There are two ways to obtain Debian GNU/Linux:

1. Installation from a CD that may be borrowed from a friend, or bought from a commercial vendor. This may be together with a computer magazine on a newsstand or from a redistributor on the Internet.

2. Download Debian from the web for free from a local mirror.

The latter is the common way, a net install will only download the required packages.

4.2.4 Biological databases

The Debian community has yet not addressed the problem of incorporating biological databases into Debian. It seems not likely that this will happen in the Debian main distribution. The extra burden to maintain copies of e.g. the EMBL DNA sequence database on multiple mirrors of Debian puts too much of a burden to the mirrors with only little gain. And not everybody requires regular updates with the associated network traffic and the induced instability when exchanging the files while scripts might still be running reading the data. The provision of a set of tools that provides updates on demand seems a more likely scenario. This would then also need to manage the update of indices of e.g. sequence similarity tools.

For now, Debian offers libraries like BioPerl with its facility to access online repositories, circumventing problems with the updating of local data. Debian is well suited though to address the issue because of its means of introspection, the programmer can tell which databases are installed and what files are available.

5. CONCLUSIONS

We have shown that there is a considerable heterogeneous shape of Free Software in biology. The continuous updates of data and the addition of novel important tools for a general bioinformatics environment cannot be performed by a single maintainer. The adherence to a policy and the sharing of maintenance are basic technologies to allow inter-institutional software projects as in computational grids and mobile agent technology.

Debian and its special dedication to software in computational biology in Debian-Med, but also the technical infrastructure behind this community project renders a comfortable solution. The volunteers behind Debian-Med strive to support everybody's specific projects as best as they possibly can. It is the particular challenge of users of Free Software, to determine together with the community the available packages that already serve their needs or may be adapted respectively.

For Debian to become the race-horse for agents and computational grids, the respective Debian packages and respective policies need to be created. The authors have created a Debian-based installation and experimental packages of the NorduGrid in Rostock. In cooperation with Emanuela Merelli and Ezio Bartocci from Camerino the installation of the BioAgents [18] environment is addressed under Debian. The development of a Debian package of the JADE agent environment [2] is currently being addressed.

Explicit links between service descriptions in bioinformatics by MOBY-S and the Debian package management have not been implemented. Similarly, no efforts are known to combine it with the site selection in computational grids. Both seems technically feasible.

Whatever choice for an infrastructure is made, with Debian it is available to all collaborating sites almost instantly.

Acknowledgements

The authors thank the Debian Developers for all their work and particularly those who have contributed to Debian-Med and the Custom Debian Distributions effort. Debian is supported by donations through Software in the Public Interest, a non-profit umbrella organisation for free software projects. Balazs Konya is thanked for his comments on the manuscript. This work was supported by the BMBF NBL3 program (FKZ 01ZZ0108) and BMBF project "Emerging Foodborne Pathogens in Germany" (FKZ 01KL9901). Special thanks go to the Robert Koch-Institut for its supports of A.Tille and his work on Debian-Med.

6. REFERENCES

[1] The ncbi. publicly available tools and resources on the web. *Methods Mol. Biol.*, 132:301–312, 2000.

[2] Developing multi-agent systems with a fipa-compliant agent framework. 31:103–128, 2001.

[3] The ncbi c++ toolkit book. 2004. URL `http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=toolkit.chapter.ch_datam%od`.

[4] Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, 33(1 (Database issue)): D39–D45, 2005.

[5] D. P. Anderson. Boinc: A system for public-resource computing and storage. In *5th IEEE/ACM International Workshop on Grid Computing*. Pittsburgh, USA, 2004. URL `http://boinc.berkeley.edu`.

[6] M. Bar and B. Knox. Openmosix: The other kind of hpc cluster. *ClusterWorld*, 2(12), 2004. URL `http://www.openmosix.org`.

[7] D. Eddelbuettel. Quantian: A scientific computing environment. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, 12(2): 291–301, March 2003. URL `http://dirk.eddelbuettel.com/quantian`. Vienna, Austria.

[8] P. Eerola, T. Ekelöf, M. Ellert, J. Hansen, A. Konstantinov, B. Konya, J. Nielsen, F. Ould-Saada, O. Smirnova, and A. Wäänänen. Science on nordugrid. In *ECCOMAS 2004*, 2004. URL `http://www.nordugrid.org/documents/eccomas04.pdf`.

[9] T. S. Foster I., Kesselman C. The anatomy of the grid: Enabling scalable virtual organizations. *International J. Supercomputer Applications*, 15(3), 2001.

[10] J. Hess. Comparing linux/unix binary package formats. 2003. URL `http://www.kitenet.net/%7Ejoey/pkg-comp`.

[11] T. Hubbard, D. Andrews, M. Caccamo, G. Cameron, Y. Chen, M. Clamp, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, T. Down, R. Durbin, X. Fernandez-Suarez, J. Gilbert, M. Hammond, J. Herrero, H. Hotz, K. Howe, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, F. Kokocinsci, D. London, I. Longden, G. McVicker, C. Melsopp, P. Meidl, S. Potter, G. Proctor, M. Rae, D. Rios, M. Schuster, S. Searle, J. Severin, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, S. Trevanion,

A. Ureta-Vidal, J. Vogel, S. White, C. Woodwark, and E. Birney. Ensembl 2005. *Nucleic Acids Res.*, 33(1): D447–D453, 2005.

[12] I. Jackson, C. Schwarz, and Debian. Debian policy manual. 2005. URL http://www.debian.org/doc/debian-policy.

[13] K. Knopper. Knoppix live cd. 2005. URL http://www.knopper.net.

[14] A. K. Kojima. An rpm port of apt. 2000. URL http://freshmeat.net/articles/view/192, http://apt4rpm.sourceforge.net.

[15] N. E. Larsen. Arc knoppix. 2005. URL http://cvs.nordugrid.org/knx/arcknoppix.iso.

[16] N. E. Larsen. Script to add the nordugrid to the basic knoppix distribution. 2005. URL http://cvs.nordugrid.org/mkknx.

[17] P. Lord, S. Bechhofer, M. Wilkinson, G. Schiltz, D. Gessler, D. Hull, C. Goble, and L. Stein. Applying semantic web services to bioinformatics: Experiences gained, lessons learnt. In *ISWC*, pages 350–364. Springer-Verlag Berlin Heidelberg, 2004.

[18] E. Merelli, L. Culmone, and L. Mariani. Bioagent: A mobile agent system for bioscientists. In *NETTAB02 Agents in Bioinformatics, Bologna*, 2002. URL http://www.bioagent.net.

[19] D. S. Myers and M. P. Cummings. Necessity is the mother of invention: a simple grid computing system using commodity tools. *Journal of Parallel and Distributed Computing*, 63(5): 578–589, 2003. URL http://lattice.umiacs.umd.edu.

[20] B. Perens, E. Schuessler, and Debian. Debian free software guidelines. June 1997. URL http://www.debian.org/social_contract#guidelines.

[21] R. Stevens, H. Tipney, C. Wroe, T. Oinn, M. Senger, P. Lord, C. G. C.A., A. Brass, and M. Tassabehji. Exploring williams-beuren syndrome using mygrid. *Bioinformatics*, 20 (Suppl. 1):i303–i310, 2004. URL http://www.mygrid.org.uk.

[22] F. D. Team. Fink. 2005. URL http://fink.sourceforge.net.

[23] A. Tille. Freie software im gesundheitswesen. In *Proceedings LinuxTag 2003*. LinuxTag, July 2003. URL http://people.debian.org/%7Etille/talks/200307_ltk/paper-305-de.html.

[24] A. Tille. Custom debian distributions. June 2004. URL http://people.debian.org/%7Etille/cdd.

[25] W. Vandersmissen. Clusterknoppix. 2005. URL http://bofh.be/clusterknoppix.

[26] M. Wilkinson, H. Schoof, R. Ernst, and D. Haase. Biomoby successfully integrates disributed heterogenous bioinformatics web services. the planet exemplar case. *Plant Physol.*, 138:1–13, 2005.

[27] L. Workgroup. Linux standard base. 2005. URL http://www.linuxbase.org.

[28] E. Zdobnov, R. Lopez, R. Apweiler, and T. Etzold. The ebi srs server-new features. *Bioinformatics*, 18(8):1149–1150, 2002. URL http://srs.ebi.ac.uk.