Scientific Grid Computing via Community-Controlled Autobuilding of Software Packages Across Architectures

Steffen Möller[1] Daniel Bayer[1] David Vernazobres[2]
Albrecht Gebhardt[3] Dirk Eddelbüttel[4]

[1]University of Lübeck, Institute for Neuro- and Bioinformatics, [2]Westphalian Wilhelms University of Münster, Institute for Evolution and Biodiversity, Division of Bioinformatics, [3]University of Klagenfurt, Institute for Statistics, [4]Debian Project, Chicago

NETTAB, Pisa
June 2007

Outline

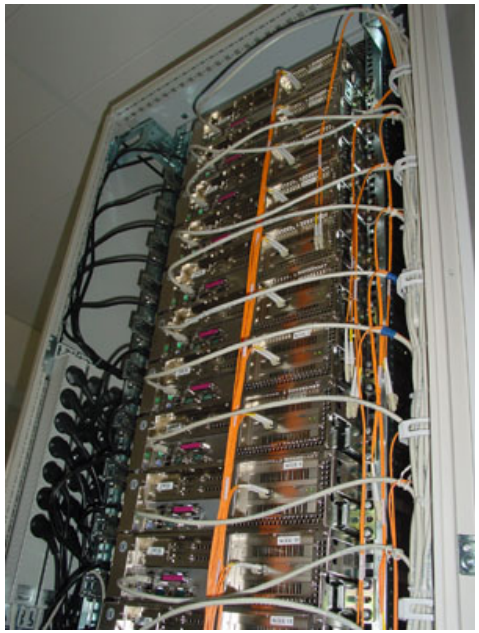
Motivation for Grid Computing in Bioinformatics Research

Large number of data parallel problems:

- Image analysis
- Sequence analysis
- Statistical genetics

Long-lasting jobs

- Ligand screening, Protein docking
- Monte-Carlo Simulation



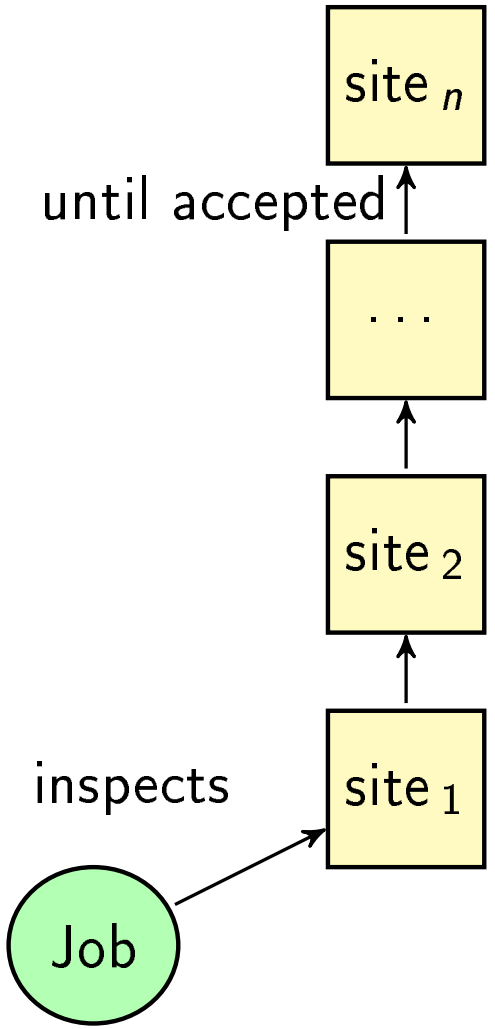
All data from biological high-throughput efforts

Principles of Grid Computing

"Integration of local batch systems"

- Users specifies a job
  - required software packages
  - cpu time
  - ...
- Site is selected that matches demands
- Job is executed on worker node of that site

half a working day to set up as server, 5 min as client ... once certificates are available.



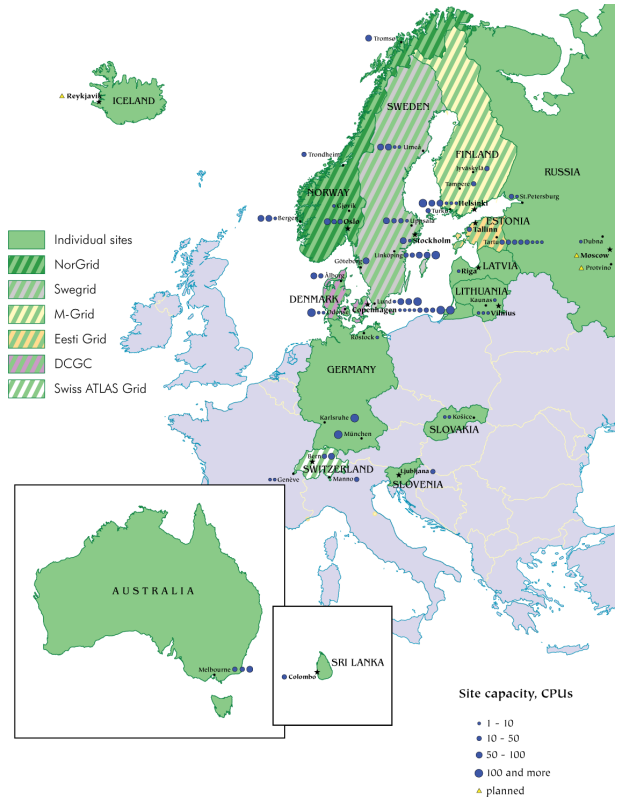
NorduGrid and ARC

- Compute and data sharing grid
- Launched in 2001
- $>$ 7500 active hosts

Special features:

- Integrates regular batch systems
- Distributed data handling
- Minimally-invasive – single machine config

Details on www.nordugrid.org



Map of NorduGrid sites

Bringing huge software repositories to the Grid

Software installations are traditionally performed by site administrators:

- Restricted availability of resources
- Serious validation of error-prone installation

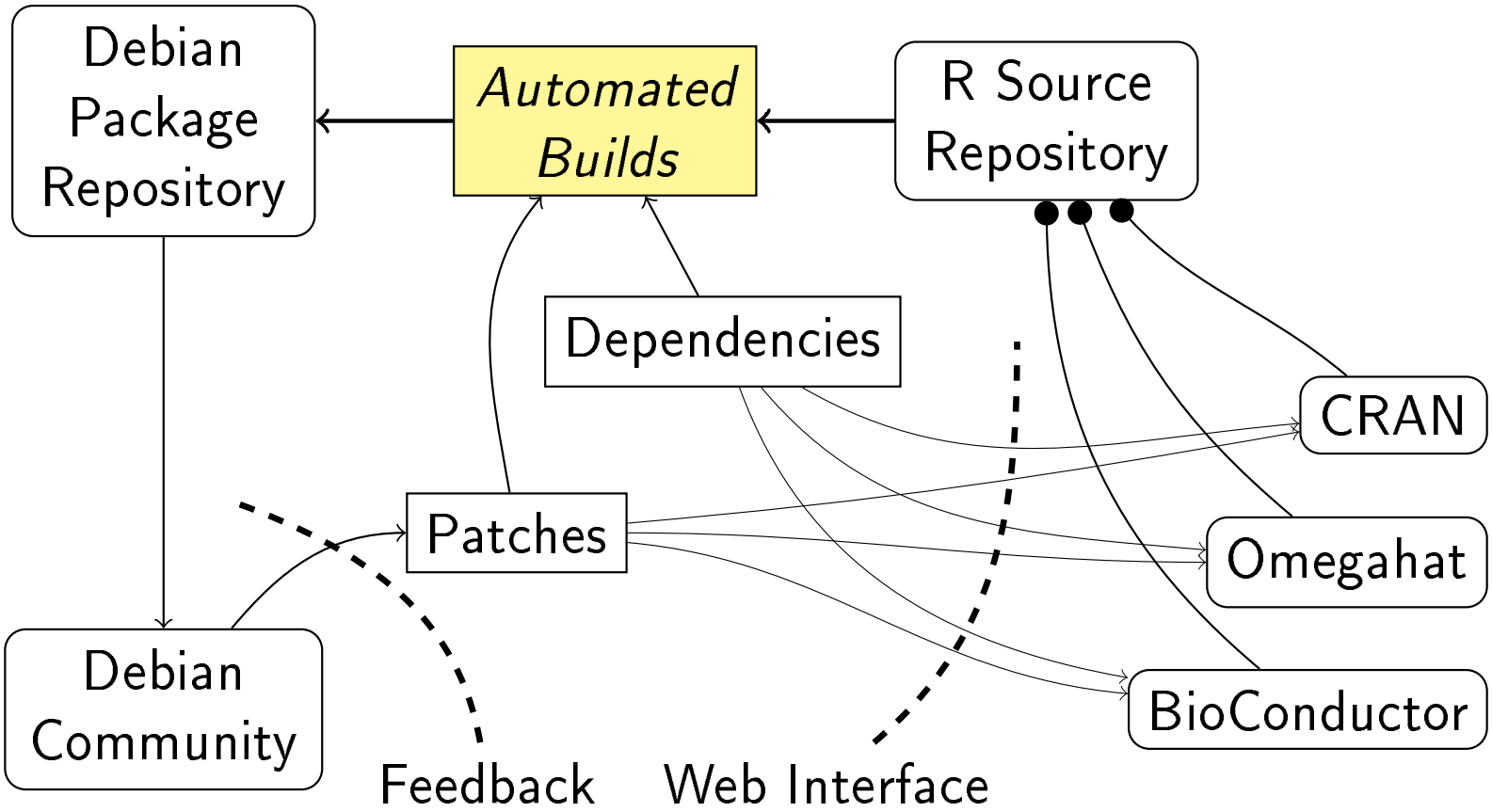
Heterogeneous communities do not know each other's software and research aims

- Limited motivation
- Homogenize descriptions of packages

Independence of human factors seeked

Provisioning of a Homogeneous Grid Environment

Bringing Science, Linux and Grid Communities Together



R Statistics Environments

Repositories: CRAN, BioConductor, Omegahat

R

- free software environment for *statistical computing* and graphics.
- available on UNIX , Windows and MacOS

CRAN

- $>$ 1000 R packages provided by the community
- central repository

R Statistics Environments

Repositories: CRAN, BioConductor, Omegahat

BioConductor

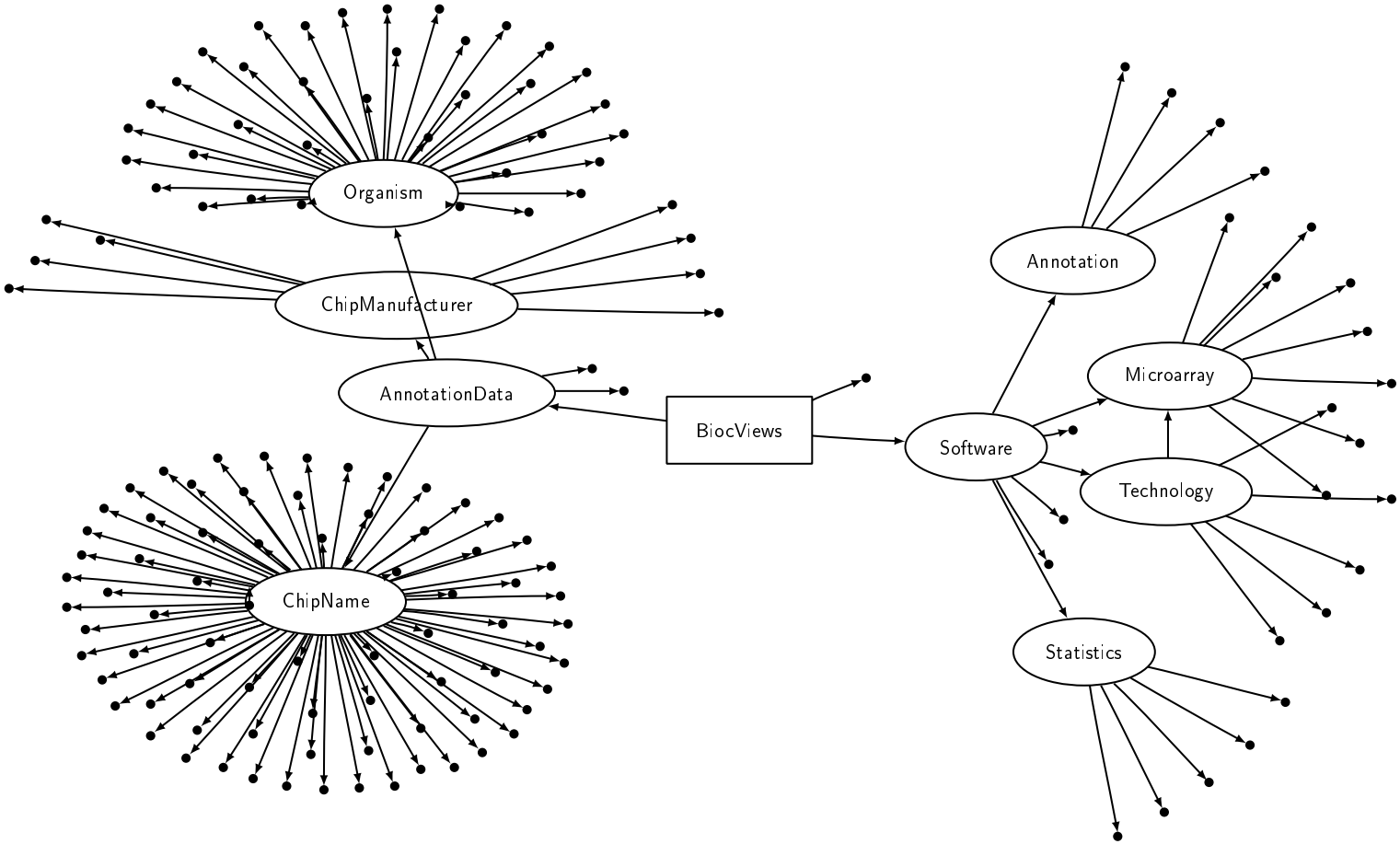
- $>$ 1000 R packages provided by the community
- central repository
- Methods supporting reserach in Bioinformatics (Microarrays, Proteomics, ...)
- Access to biological data and its visualisation

Omeghat

- Additions for using Java, Perl, SOAP, ...

Applications in Biological Research

biocView controlled vocabulary of BioConductor



The Debian Linux Distribution

Debian package maintainers

- Automate compilation of software packages
- Completion (man pages, description)
- Dissemination to 11 architectures by autobuilders

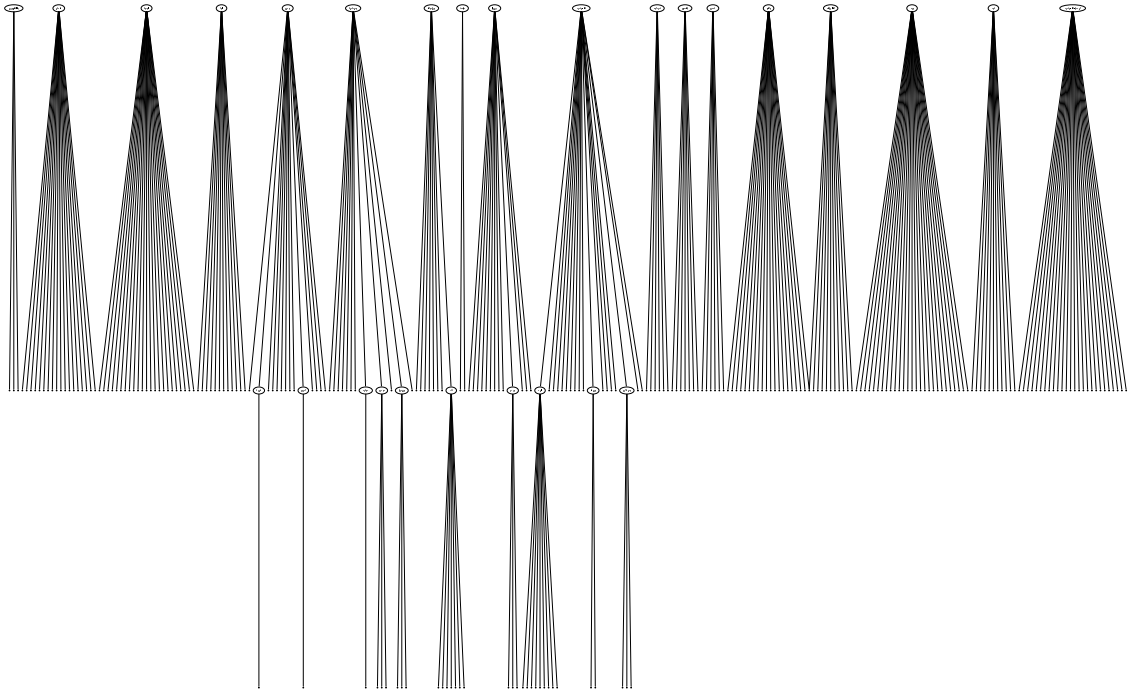
Community

- Package maintainers come directly from the users community
- Authentication as decentralised *chain of trust*
- QA by *homogeneity of platforms* and reporting system

Classification by Debtags

Facets based:

- accessibility
- admin
- devel
- field
- game
- hardware
- interface
- junior
- legacy
- mail
- network



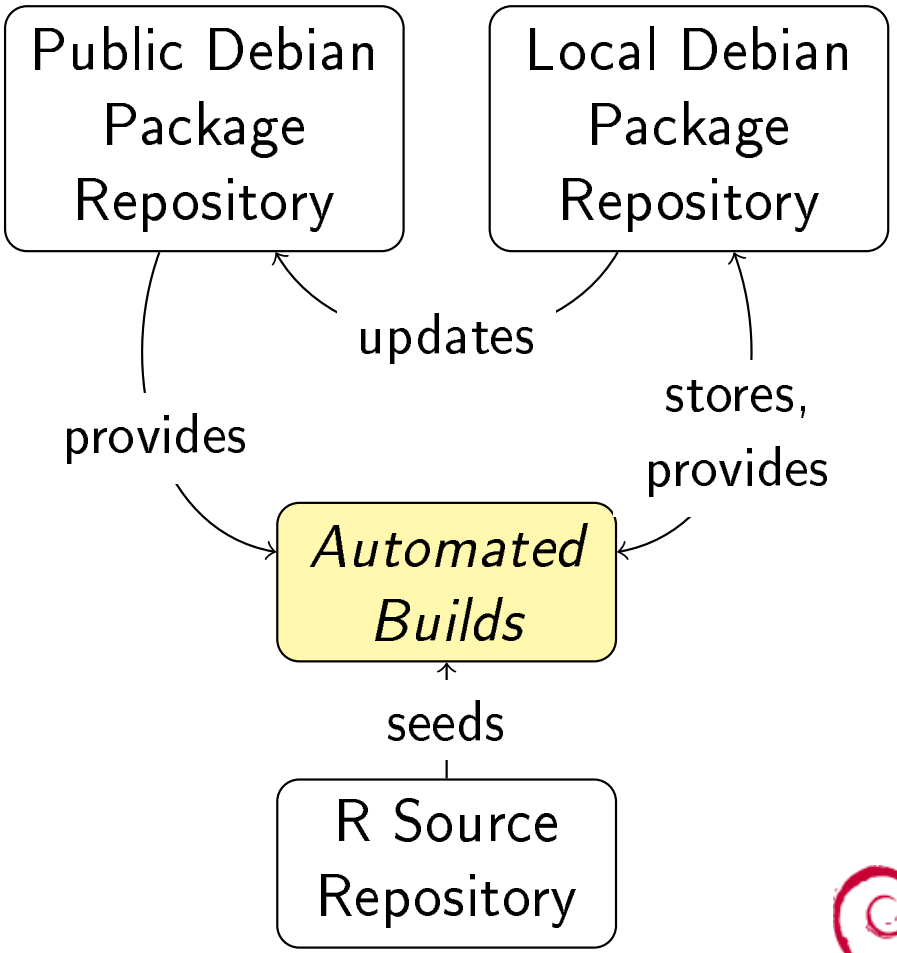
Automated Builds of Debian Packages

Problems

- Not all packages installable (not yet existent, disk space)
- Order of packaging (inter-dependencies)

Solution: Debian's pbuilder

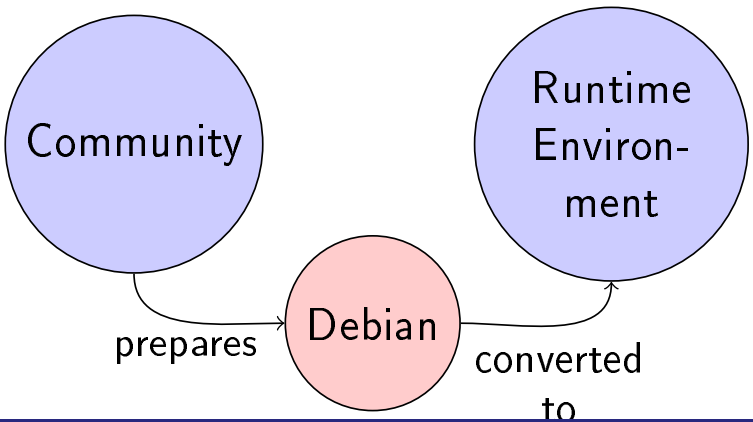
- Deps resolved dynamically
- Planning build order



Debian as *Lingua Franca*

- Automated provisioning of packages for 11 Platforms
- Automated installations
- Detailed descriptions

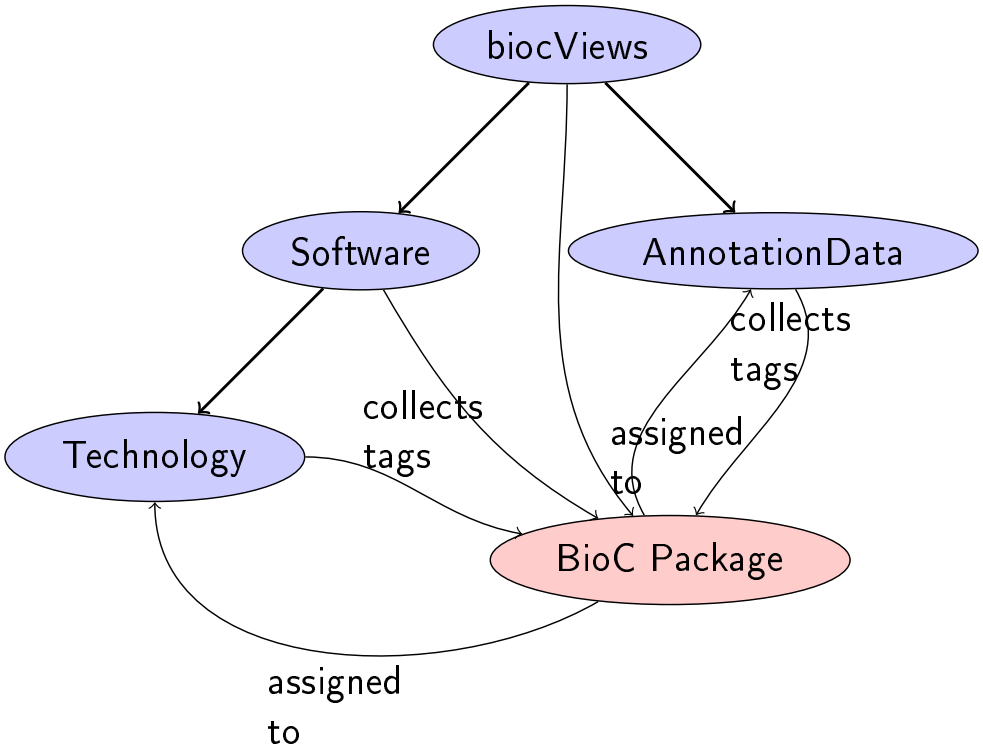
  Formal: Debtags
  Verbose: Package descriptions




```
Package: bash
Priority: required
Section: shells
Installed-Size: 1848
Architecture: i386
Version: 3.1dfsg-8
Depends: base-files (>= 2.1.12),
Suggests: bash-doc
Size: 872884
Description: The GNU Bourne Again
 Bash is an sh-compatible command
 commands read from the standard
 incorporates useful features fro
Tag: implemented-in::c, interface
     scope::utility, uitoolkit::n
```

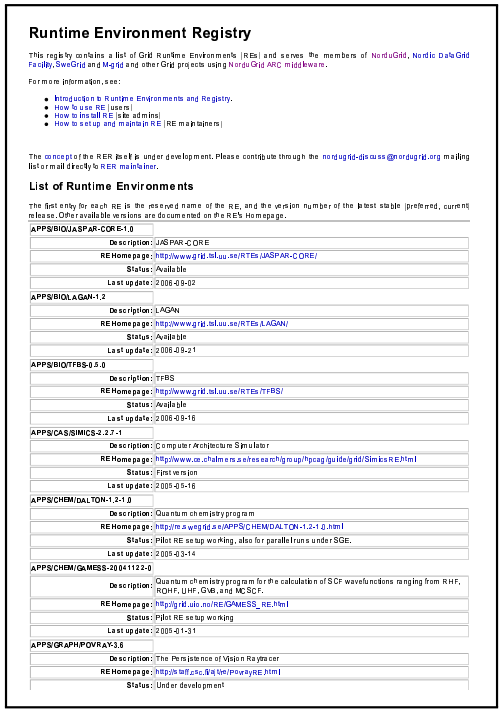
Conversion of biocView vocabulary to Debtags

1. Selected nodes in biocViews tree are annotated with Debtags
2. Packages receive all Debtags associated with referenced biocView entries



Traditional Grid *Runtime Environments*

- A much respected special feature of the ARC grid middleware:
  1. Site-maintainers install a particular softare for all worker nodes
  2. Software installation is promoted via Grid Information System
- Job descriptions explicitly mention required runtime environments
- Runtime environments are organised via a web site: `http://gridrer.csc.fi/`



Runtime Environment Registry

This registry contains a list of Grid Runtime Environments (REs) and serves the members of NorduGrid, Nordic DataGrid Facility, SweGrid and M-grid and other Grid projects using NorduGrid ARC middleware.

For more information, see:

- Introduction to Runtime Environments and Registry.
- How to use RE (users)
- How to install RE (site admins)
- How to setup and maintain RE (RE maintainers)

The concept of the RER itself is under development. Please contribute through the nordugrid-discuss@nordugrid.org mailing list or mail directly to RER maintainer.

List of Runtime Environments

The first entry for each RE is the reserved name of the RE, and the version number of the latest stable (preferred, current) release. Other available versions are documented on the RE's Homepage.

APPS/BIO/JASPAR-CORE-1.0	
Description:	JASPAR-CORE
RE Homepage:	http://www.grid.tsl.uu.se/RTEs/JASPAR-CORE/
Status:	Available
Last update:	2006-09-02
APPS/BIO/LAGAN-1.2	
Description:	LAGAN
RE Homepage:	http://www.grid.tsl.uu.se/RTEs/LAGAN/
Status:	Available
Last update:	2006-09-21
APPS/BIO/TFBS-0.5.0	
Description:	TFBS
RE Homepage:	http://www.grid.tsl.uu.se/RTEs/TFBS/
Status:	Available
Last update:	2006-09-16
APPS/CAS/SIMICS-2.2.7-1	
Description:	Computer Architecture Simulator
RE Homepage:	http://www.ce.chalmers.se/research/group/hpcag/guide/grid/SimicsRE.html
Status:	First version
Last update:	2005-05-16
APPS/CHEM/DALTON-1.2-1.0	
Description:	Quantum chemistry program
RE Homepage:	http://ra.wegrid.se/APPS/CHEM/DALTON-1.2-1.0.html
Status:	Pilot RE setup working, also for parallel runs under SGE.
Last update:	2005-03-14
APPS/CHEM/GAMESS-20041122-0	
Description:	Quantum chemistry program for the calculation of SCF wavefunctions ranging from RHF, ROHF, UHF, GVB, and MCSCF.
RE Homepage:	http://grid.uio.no/RE/GAMESS_RE.html
Status:	Pilot RE setup working
Last update:	2005-01-31
APPS/GRAPH/POVRAY-3.6	
Description:	The Persistence of Vision Raytracer
RE Homepage:	http://staff.csc.fi/ulfw/povrayRE.html
Status:	Under development

Traditional Grid *Runtime Environments*

List of Runtime Environments

The first entry for each RE is the reserved name of the RE, and the version number of the lates release. Other available versions are documented on the RE's Homepage.

APPS/BIO/JASPAR-CORE-1.0	
Description:	JASPAR-CORE
RE Homepage:	http://www.grid.tsl.uu.se/RTEs/JASPAR-CORE/
Status:	Available
Last update:	2006-09-02
APPS/BIO/LAGAN-1.2	
Description:	LAGAN
RE Homepage:	http://www.grid.tsl.uu.se/RTEs/LAGAN/
Status:	Available
Last update:	2006-09-21
APPS/BIO/TFBS-0.5.0	

Automated Grid Runtime Environments

Additional development seeded in Lübeck

1. Software packages are organised in Catalogs
   - RDF description (architecture, debtags, dependencies)
   - automated location-independent installation
   - Reference to binary for download

2. Service at sites
   - install requested runtime environments on demand
   - purge legacy installations

Conversion from Debian to Runtime Environments

Current Implementation: Debian → Tar files

- Script retrieves files to be repacked as tar file
- Automatically prepared install script to set environment variables for R
- No support for dependencies to *non-R* Debian packages

Alternative: Virtualisation

- Preparation of image for virtualisation
- Directly functional for all Debian packages
- ETA: 6 months

Deciding on the Eligibility of Packages

- Users positively select single packages for their computation
  - the selection of an R package is perceived as an integral part of the scientific application and
  - not specific to Grids
  - the selection is always a positive selection
- Site administrators
  - select classes of applications/libraries
  - both *positively* (ok to install) and *negatively* (not of interest) using
    - regular expressions or
    - SPARQL queries on Catalogs

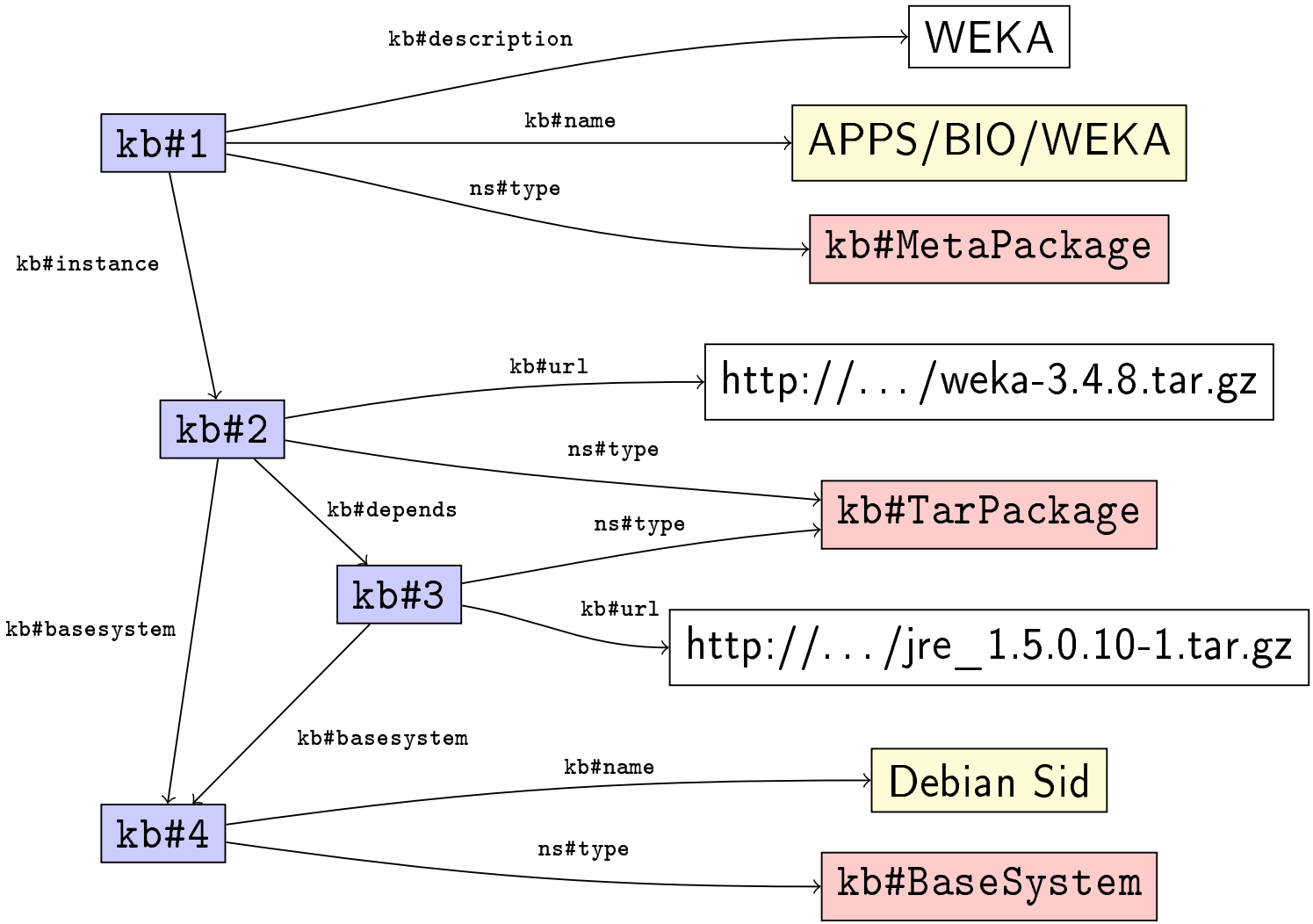
Catalog of R packages for the Grid

- 1700 Packages are made available as Grid Runtime Environments
- Complete automation of software updates
  1. from Community to Debian
  2. from Debian to Grid
- Presentation

  HTML to users

  RDF to machines

RDF triplets in the Catalog



SPARQL for the retrieval of packages I

SPARQL is an intrinsic component for the retrieval of information from RDF files:

- Retrieval of packages in catalog
- Request for a constraint to match

```
PREFIX kb: <http://knowarc.eu/kb#>
SELECT ?url
WHERE {
        ?id kb:name "APPS/BIO/WEKA" .
        ?id a kb:MetaPackage .
        ?id kb:instance ?package .
        ?package kb:basesystem ?basesystem .
        ?package kb#url ?url .
        ?basesystem kb#name "Debian Sid" .
}
```

SPARQL for the retrieval of packages II

Arbitrary constraints can be implemented:

- Maintainer of package (Virtual Organisations, . . .)
- Category of software
- Access to software
- . . .

Why are you Preparing for Heterogeneity and Complexity in the Grid

1. Allow for a heterogenous set of users
2. Allow for complex interactions
   - User-driven modelling of workflows (Taverna, . . .)
   - Automated Agents for cross-site communication

Strengths of RDF exploited

RDF is not essential for the current functionality, Debian provides core facilities today, but

1. it can be extended easily
   - more attributes
   - for more categories
2. database-like features
   - catalogs/ontologies are easily associated with entries
   - query language
3. it is a standard

Summary

- Integration of heterogeneous communities for Grid Computing
- Reference to software cathegorisation from within RDF Catalogs
  - No single system - allowing for multiple communities
  - Basis for decision of eligibility of packages for installation

Outlook

- Virtualisation: Mainstream Linux $\longrightarrow$ Grid
- Complex workflows on the Grid

Acknowledgements

Grid Computing

- The KnowARC EU project and
- the NorduGrid at large (www.nordugrid.org)

Debian

- `pkg-bioc` Community (pkg-bioc.alioth.debian.org)
- Maintainers of alioth.debian.org

R - CRAN - BioConductor - Omegahat

- all contributors