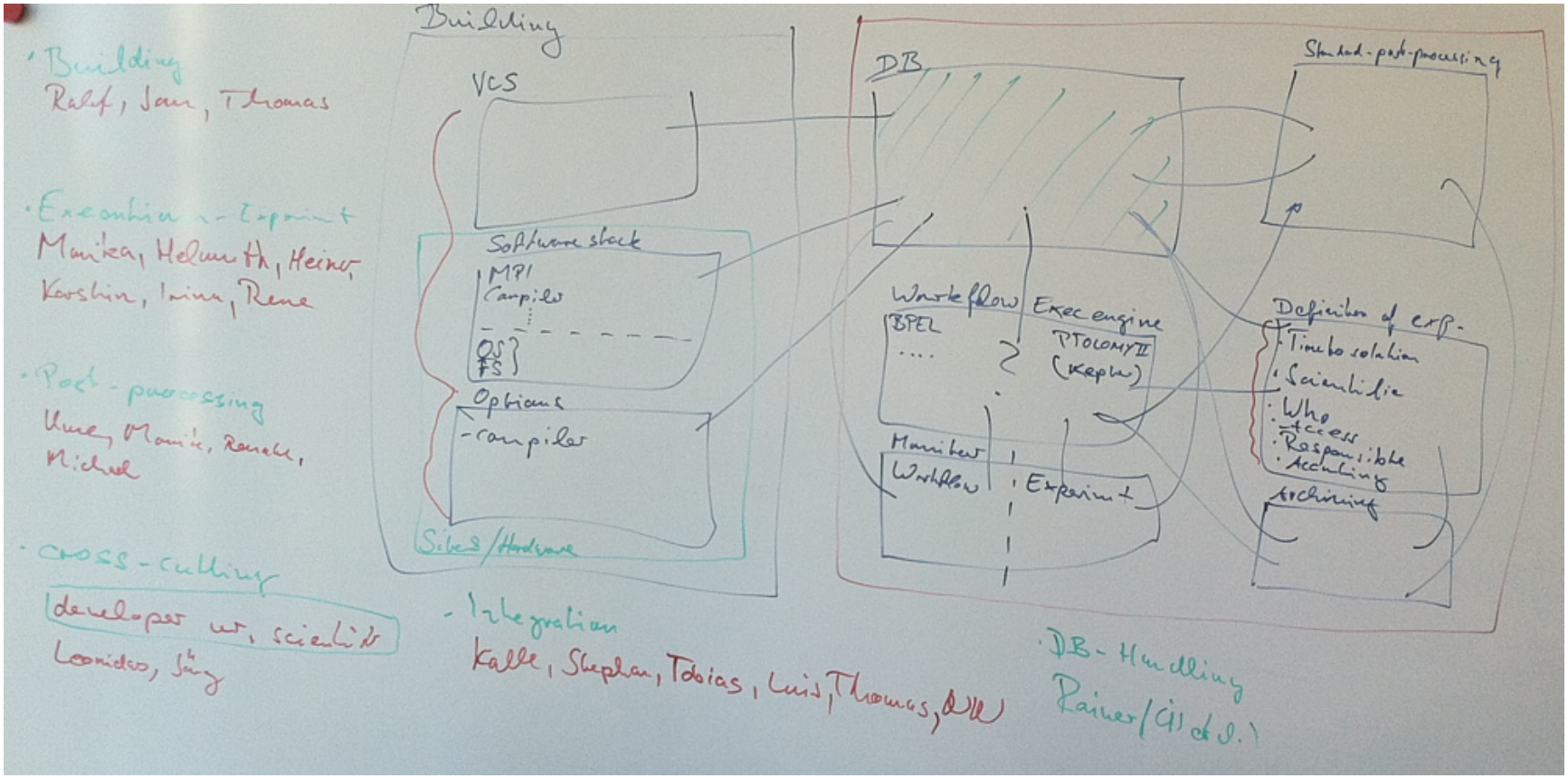
Provenance data collection

Miklip and HD(CP)^2
Runtime Environment Development Projects

Luis Kornblueh, Deike Kleberg, Kalle Wieners,
Uwe Schulzweida, Kameshwar Rao, Ketan Kulkarni,
and
Kerstin Fieg, Pavan Siligam, Kerstin Ronneberger,
Mathis Rosenhauer, Tobias Weigel

Max-Planck-Institut für Meteorologie and DKRZ

Joint DKRZ/MPIM initial brainstorming



Courtesy of Joachim Biercamp, DKRZ

An experimentation howto

An experiments lab report

- Title, statement of problem, and hypothesis

An experimentation howto

An experiments lab report

- Title, statement of problem, and hypothesis
- Site, hardware, software stack, used programs (versioned), sources, and documented input and boundary data

An experimentation howto

An experiments lab report

- Title, statement of problem, and hypothesis
- Site, hardware, software stack, used programs (versioned), sources, and documented input and boundary data
- Step-by-step description in such a way that the experiment can be repeated

An experimentation howto

An experiments lab report

- Title, statement of problem, and hypothesis
- Site, hardware, software stack, used programs (versioned), sources, and documented input and boundary data
- Step-by-step description in such a way that the experiment can be repeated
- Observations made during and analysis of the results

An experimentation howto

An experiments lab report

- Title, statement of problem, and hypothesis
- Site, hardware, software stack, used programs (versioned), sources, and documented input and boundary data
- Step-by-step description in such a way that the experiment can be repeated
- Observations made during and analysis of the results
- Discuss possible errors that could have occurred in the collection of the data (experimental errors)

An experimentation howto

An experiments lab report

- Title, statement of problem, and hypothesis
- Site, hardware, software stack, used programs (versioned), sources, and documented input and boundary data
- Step-by-step description in such a way that the experiment can be repeated
- Observations made during and analysis of the results
- Discuss possible errors that could have occurred in the collection of the data (experimental errors)
- Publication of the results

The context

- What is our context?
  - digital-produced ESM output data
  - various processed derivatives
  - eventually observational data, e.g. remote sensing imagery
- What is its characteristics?
  - complex, non-standardized toolchain
  - various processing steps by various actors
  - no single infrastructure

Quality of scientific data

- The processing history of a data object forms an important part of its scientific context.
- Users who did not create a data product must be able to understand the implications that went into its creation.
- Data may be reused many years after creation.

Reproducibility

- If processing steps are recorded in detail, a future user may reproduce them to get the exact same results
- May be impossible for ESM output data in all its depth
- We cannot archive the supercomputer itself
- Yet, try to capture as much as possible

Attribution

- Give credit to the original data producer
- Citing a data cite DOI may not be enough
- Who is using data that is generated with which resources?

Provenance can enable anyone to trace back to the original source and producer.

Target: provenance of the whole data life cycle

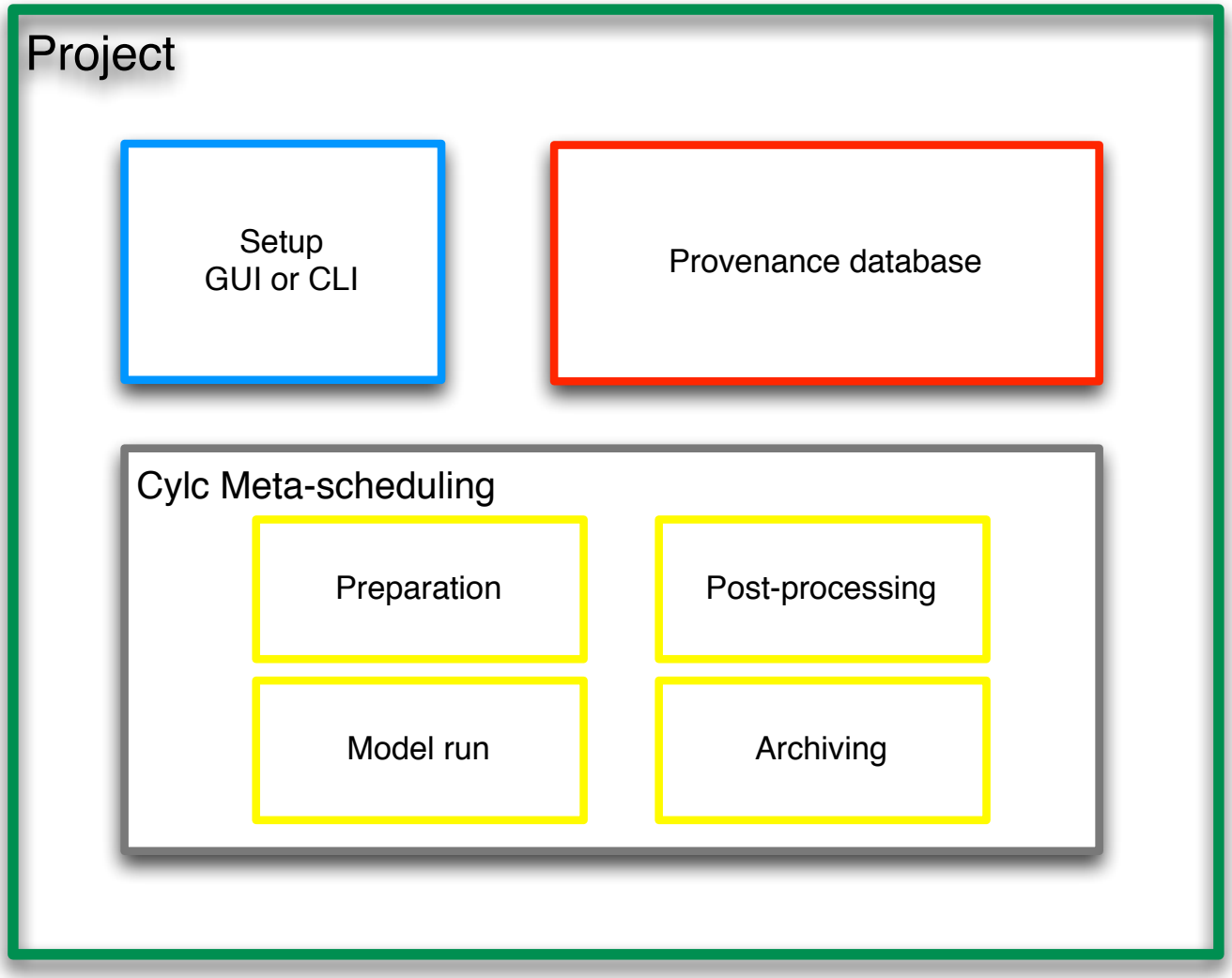
Focus on adding:

- data generation
- data processing

Already available to a large extent:

- data publishing
- data distribution

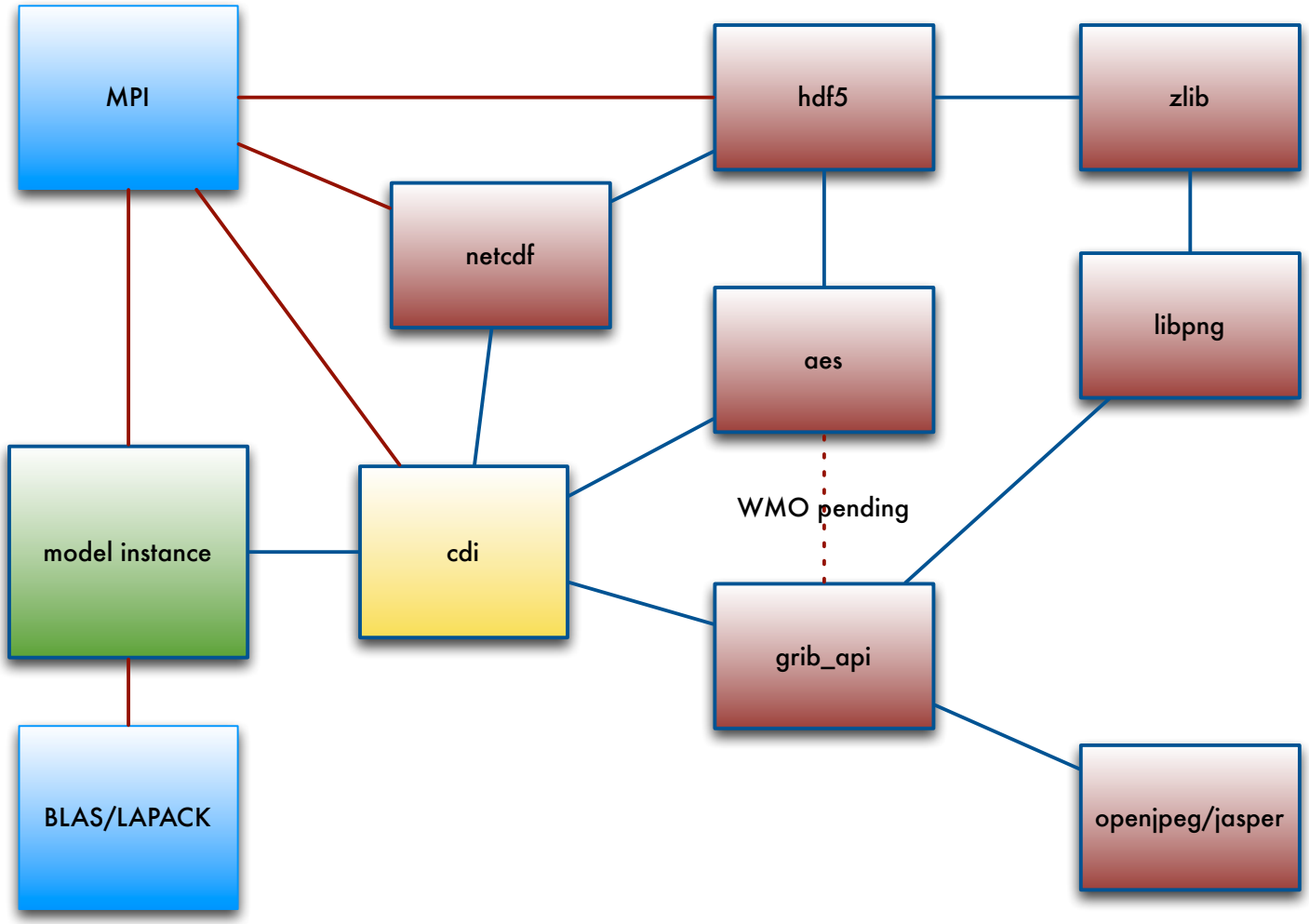
Components of the basic systems



Defining, selecting, and ...

- Define: who, title, statement of problem, and hypothesis
- Select/change model and initial and boundary conditions
- Select site
- Build

Libraries . . .


MPI
hdf5
zlib
netcdf
libpng
aes
model instance
cdi
WMO pending
grib_api
BLAS/LAPACK
openjpeg/jasper

Tools

Packages in use:

- python as scripting language
- postgres - provenance data collection
- subversion/git (migration to git for parts or all later, if model developers get convinced)
- cmake (migration from autotools and self-maintained makefile generator)
- Web interface for site and compiler dependencies; dependencies versioned in line with model code
- namelist migration to xml as model source code, user interfaces: a kind of namelist and a GUI.

Setup provenance data collection

User project/experiment definition

- Principle investigator and experiment description
- Source code (revision)
- Runtime parameter (namelists)
- Compiler and compiler flags
- Libraries
- System information (Site, OS, Hardware)

Experiment organization

cylc - the Meta-Scheduler

- design distributed suites of inter-dependent cycling tasks efficiently, modular and reusable
- control complex running suites
- diagnose failures easily
- simplify failure recovery
- benefit from expert experience with a specialized tool for meteorological forecasting systems
- validate and visualize workflows on the fly

Courtesy of Hilary Oliver, NIWA and some contributors

A task modeling framework

Cylc controled tasks and provenance collection

- high level programming language — python
- abstract task description
- embedded provenance data collection (database stored)
- tightly connected to cylc
- connect workflow to ESGF data distribution

Status

- Cylc management server provided and co-maintained by/with DKRZ
- Necessary network setups on compute nodes done by DKRZ
- Postgres database server provided by DKRZ
- Joint design of optimized ensemble workflow for decadel runs with Hilary
- Design for task modeling language done, implementation is work in progress
- Setup GUI work in progress
- Development of refactoring tool for replacing namelists by xml based setup finished - needs to be implemented in model
- Plenty of small bits-and-pieces for provenance data collection created

Todo next

- Model Miklip decadel hindcast system
- Model Miklip decadel ensemble prediction system
- Model HD(CP)^2 experiments
- Replace IMDI driven experiments
- Build templates for the MPIM standard experiments
- Do tutorials and training