The eScience technology platform

A coherent set of technologies to tackle the grand challenges in eScience

Rob van Nieuwpoort
R.vanNieuwpoort@esciencecenter.nl
The eScience landscape

Research Disciplines
- Environment & Sustainability
- Life Sciences & eHealth
- Humanities & Social Sciences
- Physics & Beyond

Enhanced Science
- (optimized research)

eScience Instruments
- (software, workflows)

Data & Computational Sciences
- Optimized Data-Handling
- Big-Data Analytics
- Efficient Computing

e-Infrastructure
- Computing, Networking, Storage & Visualization

eScience Services
- (@SURF, @NFU etc.)
NLeSC eScience competences applied in research

1. Optimized data handling
Data integration, data base optimization, structured & unstructured data, real time data

2. Big data analytics
Statistics, machine learning, visualization, text mining

3. Efficient computing
Distributed & accelerated computing, efficient algorithms
- Key expertises are used in many projects

- Projects often use quite a number of different competences and technologies
Cross-cutting basic skills

- Code quality and best practices
- Integration of software
- Scaling of software
- Analytics and statistics
- Visualization
eStep

The eScience technology platform

A coherent set of technologies to tackle the grand challenges in eScience
**Software**

- algorithms, interfaces, libraries, tools, workflows, applications, ...

**Knowledgebase**

- best practices, tutorials, training material, white papers, ...

**eScience research**

- Journal and conference publications, book chapters, posters, demos, ...

---

**eStep**
eStep Goals

- Prevent fragmentation and duplication
- Promote the exchange and re-use of best practices
- Represent our expertise and knowledge base
- Improve the eScience state of the art with a fundamental eScience research line
NLeSC projects

Adopt

eStep

Tailor

Develop

Generalize

NLeSC projects
Main criteria for integrating technology in eStep:

- State-of-the-art / best-of-breed?
- Generic and overarching?
- Match with our expertise areas?
- Includes externally developed software

Open platform!
Our sustainability approach

- Prevent duplication, fragmentation

- Build something that is worth sustaining!
  - Sufficiently generic
  - Modular
  - High quality
  - Must be taken into account from the start

- Enforce software engineering guidelines and best practices

- Educate partners with [software carpentry](#) and [data carpentry](#)

- Open source / open access, open standards, unless…
- Community coding
- Standardization for software and data formats
- eStep is an open platform
• Make researchers more productive by teaching them basic lab skills for scientific computing

• All lessons are freely available
• Workshops, teacher trainings
• Example lessons
  – Version Control and Unit Testing for Scientific Software
  – Shell, Git, Scientific Python
  – Testing and Continuous Integration with Python
  – From Excel to a Database
  – Data Management in the Ocean, Weather and Climate Sciences
  – Visualizing Your Data on the Web Using D3
  – Working With Data on the Web
  – Intermediate/Advanced R Lessons
  – Programming with GAP
Develop and teach workshops on the fundamental data skills for research in all domains
Covering the full lifecycle of data-driven research
Introductory computational skills for data management and analysis
Domain-specific lessons, from life and physical sciences to social sciences
Build on existing knowledge, enabling quick application of new skills to own research
Examples:

- Ecology
  - Data Organization in Spreadsheets, Data Cleaning with OpenRefine, Data Management with SQL, Data Analysis and Visualization in R, Data Analysis and Visualization in Python

- Genomics
  - Introduction to cloud computing for genomics, Introduction to the command line, Data wrangling and processing, Data analysis in R, Data visualization in R

- Social sciences
  - Social sciences text mining

- Biology

- Geospatial data
Coding Style

- Nicholas C. Zakas: Why Coding Style Matters

- Use is mandatory
- We provide editor configuration
  - http://editorconfig.org/

EditorConfig
Conventions & Guidelines

- **Web development**
  - General frontend guidelines: [https://github.com/bendc/frontend-guidelines](https://github.com/bendc/frontend-guidelines)
  - AngularJS: [https://github.com/johnpapa/angular-styleguide](https://github.com/johnpapa/angular-styleguide)
  - Airbnb JavaScript Style Guide: [https://github.com/airbnb/javascript](https://github.com/airbnb/javascript)

- **Python**
  - PEP8: [https://www.python.org/dev/peps/pep-0008/](https://www.python.org/dev/peps/pep-0008/)

- **Java**
  - Code Conventions for the Java™ Programming Language (Oracle)

- **Google Style Guides**: [https://github.com/google/styleguide](https://github.com/google/styleguide)

Quality Improvement Tools

Article about good development practices:
The Joel Test: 12 Steps to Better Code.

- SonarQube: http://www.sonarqube.org
- Code climate: https://codeclimate.com
- Codacy: https://www.codacy.com
- Scrutinizer: https://scrutinizer-ci.com
- Landscape: https://landscape.io
- Coveralls: https://coveralls.io
- See also
  - https://github.com/ripienaar/free-for-dev#code-quality
  - http://shields.io/
Unit & Integration Testing

- **Guide: Writing Testable Code**
- Continuous integration testing with **Travis-CI** and **Jenkins-CI**
- We require at least 70% code coverage

- Java: **junit**
- Javascript
  - **Jasmine**, a behavior-driven development framework for testing JavaScript code.
  - **Karma**, Runs tests in web browser with code coverage.
  - **PhantomJS**, headless web browser on CI-servers.
- Python
  - **Unittest, nose** and **pytest**.
- R
  - **testthat**
- Web development
  - To interact with web-browsers use **Selenium**.
  - **Sauce Labs** hosts a matrix of web-browsers and Operating Systems for testing.
  - AngularJS applications can be tested with **Protractor**.
Documentation

- Document at multiple levels
  - Source code comments
  - API documentation
  - Installation and usage documentation
- Comments at each level should take into account the different target audiences
- Use Markdown, a readable lightweight markup language that can be converted to many formats
Version Control

• Git and GitHub

• A successful and simple Git branching model: GitHub Flow
  • https://guides.github.com/introduction/flow/

• Commit messages are formatted and formulated in a readable way
  • http://tbaggery.com/2008/04/19/a-note-about-git-commit-messages.html
  • http://who-t.blogspot.nl/2009/12/on-commit-messages.html
Releases and packaging

• Tag versions, use github releases
• Semantic versioning
• Keep changelogs

• Packaging is important
  – Use packaging that is well known and appropriate for user community: pypi, npm, maven, docker

• Make your code and data citable: get a DOI (Zenodo)
A Common Workflow @ NLeSC

GitHub

Test and Deploy with Confidence. Easily sync your GitHub projects with Travis CI and you'll be testing your code in minutes!

Jenkins

We run a Jenkins CI instance locally. Used for private repositories and repositories requiring HPC middleware.

Travis CI

Test and Deploy with Confidence. Easily sync your GitHub projects with Travis CI and you'll be testing your code in minutes!

Open platform for building, shipping and running distributed applications.
• technology.esciencecenter.nl

• Non-technical, targets general audience
estep.esciencecenter.nl

All eScience software and knowledge you need, in one place

Technical, targets developers

The eScience Technology Platform

An important aspect of eScience is the development of new software technologies to enable scientists to make new scientific breakthroughs. The route from data to information to insight should take optimal advantage of modern ICT facilities and e-infrastructures. This often requires specialist knowledge which is beyond the level of expertise of domain scientists. Our aim is to let researchers be engaged with scientific challenges rather than with the idiosyncrasies of ICT.

The main aim of the eScience Technology Platform (eStep) is to scout, adopt, research, develop, integrate and make available an extensive and stable set of advanced scientific software technologies. These can take the form of compute kernels, interfaces, libraries, tools, and scientific workflows. Tools in eStep help to manage large collections of data, utilize advanced compute infrastructures, storage facilities, high-speed networks, high-resolution visualization equipment, and instruments.

eStep explicitly aims to promote the exchange and re-use of best practices and to prevent fragmentation, duplication. Therefore, eStep contains software that is developed in-house, but also externally developed software that we have expertise with. A key idea behind eStep is to have high-level, and sometimes domain-specific, solutions on top of generic low-level libraries, thus maximizing software re-use.
eStep Software used in Projects

### Software

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AHN2 point cloud viewer</td>
<td>WebGL point cloud visualization of AHN2</td>
</tr>
<tr>
<td>AMUSE</td>
<td>The Astrophysical Multipurpose Simulation Environment</td>
</tr>
<tr>
<td>CClusTer</td>
<td>3D web tool for interactive visualization of hierarchically clustered big data</td>
</tr>
<tr>
<td>Cesium-ncWMS</td>
<td>3D Globe Visualization of NetCDF data.</td>
</tr>
<tr>
<td>Common Sense</td>
<td>User-friendly web application for showing (GIS) data on a map.</td>
</tr>
<tr>
<td>Cross-perspective Topic Modeling</td>
<td>A Gibbs sampler that implements Cross-Perspective Topic Modeling</td>
</tr>
<tr>
<td>Datalax</td>
<td>Technology of Attachment to a DBMS of large file repositories.</td>
</tr>
<tr>
<td>Differential Evolution</td>
<td>Differential Evolution global optimization algorithm, with Metropolis for uncertainty estimation</td>
</tr>
<tr>
<td>eAstroViz</td>
<td>This tool can convert and visualize radio astronomy measurement sets, as well as most LOFAR intermediate data products. It also does RFI mitigation</td>
</tr>
<tr>
<td>eEcology Annotation Tool</td>
<td>Visualize &amp; annotate GPS measurements of bird movements</td>
</tr>
<tr>
<td>eEcology Tracker calendar</td>
<td>Calendar overview with daily statistics of GPS-tracker</td>
</tr>
<tr>
<td>eWaterLeaf</td>
<td>Web-based visualization for the eWaterCycle project</td>
</tr>
<tr>
<td>ExtJS-Date-Time</td>
<td>Date/Time input field for ExtJS</td>
</tr>
<tr>
<td>FAIR Data Point</td>
<td>FAIR Data Point Metadata Service</td>
</tr>
<tr>
<td>Google Earth Toolbox for MATLAB</td>
<td>Export data from MATLAB to GoogleEarth's KML format.</td>
</tr>
<tr>
<td>Historic Embodied Emotions Model (HEEM) dataset</td>
<td>279 17th and 18th century Dutch theater texts with HEEM labels</td>
</tr>
<tr>
<td>Kernel Tuner</td>
<td>A simple CUDA/OpenCL kernel tuner in Python.</td>
</tr>
</tbody>
</table>
Knowledge base

- knowledge.esciencecenter.nl
- training and education
- best practices
- tutorials
- white papers
- training resources

Software development Checklist available
More info on eStep

technology.esciencecenter.nl

estep.esciencecenter.nl

R.vanNieuwoort@esciencecenter.nl
Logo Bingo

Osmium

CommonSense

xtas

NLTK

EDAL

Semanticizer

gensim

rapidlasso

MATLAB

cesium

AHN2 viewer

PostgreSQL

CloudCompare

Open for Innovation

KNIME
Support levels

- **S0**: generic software or hibernating software that is currently not used in the NLeSC project portfolio
  - Fortran, Python, vBrowser, TwiNL, XNAT, ...
  - No support, not disseminated

- **S1**: software where NLeSC maintains expertise on, and that is used in projects, as well as external software that NLeSC extends and improves
  - Potree, OpenDA, ElasticSearch
  - Support for project partners only
  - Contribute improvements back to community

- **S2**: software developed in-house, where NLeSC is the specialist
  - Xenon, Magnesium, Osmium, xtas, esiBayes, Aether, ...
  - Full support for project partners, limited support for Dutch scientific community, best effort for international community
IP & Software Licenses

- NLeSC does not develop IP portfolio
- Ownership of research results is shared property of partners, including NLeSC
- IP protection is possible
- Software is open source / open access unless agreed otherwise
- Default software license: Apache 2.0
- Deviations possible, discuss with NLeSC MT