Stichting NIOC en de NIOC kennisbank

Stichting NIOC (www.nioc.nl) stelt zich conform zijn statuten tot doel: het realiseren van congressen over informatica onderwijs en voorts al hetgeen met een en ander rechtstreeks of zijdelings verband houdt of daartoe bevorderlijk kan zijn, alles in de ruimste zin des woords.

De stichting NIOC neemt de archivering van de resultaten van de congressen voor zijn rekening. De website www.nioc.nl ontsluit onder "Eerdere congressen" de gearchiveerde websites van eerdere congressen. De vele afzonderlijke congresbijdragen zijn opgenomen in een kennisbank die via dezelfde website onder "NIOC kennisbank" ontsloten wordt.

Op dit moment bevat de NIOC kennisbank alle bijdragen, incl. die van het laatste congres (NIOC2023, gehouden op donderdag 30 maart 2023 jl. en georganiseerd door NHL Stenden Hogeschool). Bij elkaar bijna 1500 bijdragen!

We roepen je op, na het lezen van het document dat door jou is gedownload, de auteur(s) feedback te geven. Dit kan door je te registreren als gebruiker van de NIOC kennisbank. Na registratie krijg je bericht hoe in te loggen op de NIOC kennisbank.

Het eerstvolgende NIOC vindt plaats op donderdag 27 maart 2025 in Zwolle en wordt dan georganiseerd door Hogeschool Windesheim. Kijk op www.nioc2025.nl voor meer informatie.

Wil je op de hoogte blijven van de ontwikkeling rond Stichting NIOC en de NIOC kennisbank, schrijf je dan in op de nieuwsbrief via

www.nioc.nl/nioc-kennisbank/aanmelden nieuwsbrief

Reacties over de NIOC kennisbank en de inhoud daarvan kun je richten aan de beheerder:

R. Smedinga kennisbank@nioc.nl.

Vermeld bij reacties jouw naam en telefoonnummer voor nader contact.

Software for Science NIOC Leeuwarden

Dr. Marten Teitsma

6 maart 2018

Who is Marten Teitsma

Some aspects of my identity:

- Lecturer at the Amsterdam University of Applied Sciences
- Coordinator Technical Computing, founder Software for Science
- PhD in Ontology Engineering
- Living in Utrecht, born in Fryslân
- Daughter 12, son nearly 8

What is the AUAS?

Amsterdam University of Applied Sciences:

- 46.000 students, 4000 employees
- 3% international students
- HBO-ICT (Department of Information Technology), 3000 students, 160 members of staff. Specialisations:
  - Software Engineering
  - Cyber Security
  - Game Development
  - Business Information and Management
  - Technical Computing

Software for Science

Software is changing the way we conduct science, in terms of the sophistication of the analyses we perform, and the volume of data we can process. [4]

Software for Science

Software is changing the way we conduct science, in terms of the sophistication of the analyses we perform, and the volume of data we can process. [4]
When do scientist use software or more general, computer technology?

Software for Science

Software is changing the way we conduct science, in terms of the sophistication of the analyses we perform, and the volume of data we can process. [4]
When do scientist use software or more general, computer technology?

- Developing hypotheses and finding literature.
- Gathering data.
- Analysis of data.
- Making models.
- Publication of results.

Software for Science

Software is changing the way we conduct science, in terms of the sophistication of the analyses we perform, and the volume of data we can process. [4]

When do scientist use software or more general, computer technology?

- Developing hypotheses and finding literature.
- Gathering data.
- Analysis of data.
- Making models.
- Publication of results.

Software pervades every domain of science [5]

Examples of software for science

- climate change research
- bio-informatics (dna sequences)
- use of social media
- word use
- deep-learning
- general use of statistics

Software is important for Science

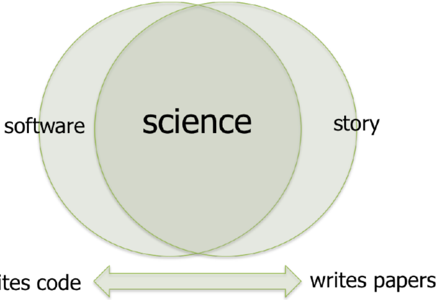


Figure: In many areas of research, science is now produced at the intersection of the 'software' and the 'story'. From: [3]

When software is so important, you should expect the software and its developers are mentioned. But no.

Problems with scientific software I

Specific problems with respect to software for science [7, 2, 3, 5]:

- designing software
  - quickly changing requirements,
  - bespoke software,
  - how do you test bespoke software,
  - lack of metrics,
- research is (big) time constrained
  - time is the constraint instead of functionality (say the same with other words),
  - competition between maintainable and performance code,
  - pressure to rapidly produce scientific publications (is cause)
- academic culture
  - no need to make software sustainable,
  - many scientist adopt and use software critical to their research on non-scientific reasons,
  - shame for being called not smart (creating a bug or fail to understand an experiment)

Problems with scientific software II

- misunderstanding of
  - domain
  - software engineering process
  - computation
  - translation theory to code and the inherent social skills needed for this,

Replication

Replication is the ultimate standard to judge scientific claims.

Replication

Replication is the ultimate standard to judge scientific claims.
Standard is not realistic because of:

- costs of research,
- time needed to do research,
- problems software for science as mentioned.

When replication is not possible, what about reproducibility?

Reproducibility

In a Nature survey of 1,576 researchers 70% of researchers have tried and failed to reproduce experiments. About 80% blamed this, partially, on the unavailability of methods and code. [1]

Reproducibility

In a Nature survey of 1,576 researchers 70% of researchers have tried and failed to reproduce experiments. About 80% blamed this, partially, on the unavailability of methods and code. [1]

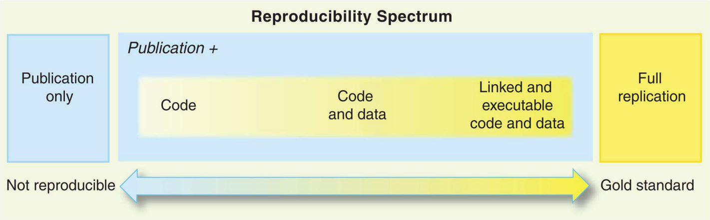


Figure: The spectrum of reproducibility. From: [6]

Reproduction of research is better than no replication. We should create a culture of reproducibility [6]

Intrinsic sustainability

Sustainable software is software you use today and which will be available in the future (Software Sustainability Institute).

Intrinsic sustainability

Sustainable software is software you use today and which will be available in the future (Software Sustainability Institute). Intrinsic sustainability concerns aspects of the software itself[2]. Software should be (in decreasing order of importance):

- well documented
- tested
- easy to read
- modular
- using third party libraries (with a large user base)
- usefulness
- scalable

Extrinsic sustainability

Extrinsic sustainability concerns the environment of the software [2]. Software should be (in decreasing order of importance):

- open available
- resourced (e.g. funded)
- actively maintained
- supported
- shared or co-owned
- independent from infrastructure

Recommendations

Ways to enhance sustainability[2]:

- actively maintain software
- use third party libraries and embrace standards
- make software available and discoverable
- rethink resourcing
- build a community around the software
- build a research software engineering community

FAIR

The FAIR guiding principles for data are [8], to be:

- Findable.
- Accessible.
- Interoperable.
- Reusable.

By use of metadata stating all sorts of information about data. FAIRness should be applied to both human-driven and machine-driven activities.

Findable

To be Findable:

1. (meta)data are assigned a globally unique and persistent identifier
2. data are described with rich metadata (defined by R1 below)
3. metadata clearly and explicitly include the identifier of the data it describes
4. (meta)data are registered or indexed in a searchable resource

Accessible

To be Accessible:

1. (meta)data are retrievable by their identifier using a standardized communications protocol
   1. the protocol is open, free, and universally implementable
   2. the protocol allows for an authentication and authorization procedure, where necessary
2. metadata are accessible, even when the data are no longer available

Interoperable

To be Interoperable:

1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
2. (meta)data use vocabularies that follow FAIR principles
3. (meta)data include qualified references to other (meta)data

Reusable

To be Reusable:

1. meta(data) are richly described with a plurality of accurate and relevant attributes
    1. (meta)data are released with a clear and accessible data usage license
    2. (meta)data are associated with detailed provenance
    3. (meta)data meet domain-relevant community standards

Repositories

Some repository frameworks which have been developed:

- Dataverse: open-source data repository software. Makes Digital Object Identifier (DOI) public when dataset is published. A landing page giving access to metadata, data files, dataset terms, etc.
- FAIRDOM: platform to produce a FAIR data and model management facility for Systems Biology.
- ISA: a community-driven metadata tracking framework for life science datasets.

Research Software Engineer

Research Software Engineers (RSEs) are the people that closely collaborate with researchers to understand the challenges they face, and then develop research software to provide the answers. Various RSE organisations:

- RSE United Kingdom
- RSE Germany
- RSE Netherlands

Other interesting links:

- Software Sustainability Institute
- Open data CERN

The Software for Science initiative

Students develop software to make scientific research possible. Research is being done by:

- interns (midterm and final),
- teams of students during a minor and summer school,
- lecturers/researchers,
- teams at other universities.

Summer School

- creating software for international scientific experiments
- 4 ects
- two weeks
- writing a technical report

Teams at other universities

Invitation to other universities for participating in this initiative.
NHL University for Applied Science.

The minor

Characteristics of the minor:

- 30 ects (half a year),
- theoretical and practical elements,
- international.

Educational elements

Elements of the minor are:

- personalised educational route (students write a motivational letter and curriculum vitae),
- boot camp for 3 or 4 weeks, 4 ects
- two courses for 10 weeks, 2 * 6 ects
- developing a product for 15 weeks, 14 ects

Bootcamp

During the bootcamp students learn to handle the tools for working on their project. Several tools are studied:

- Domain knowledge
- LaTeX
- GitHub
- Ansible
- Travis CI

Several other tools specific for the project:

- C++
- JavaScript
- MySQL
- etc

Courses

Scientific Method

- statistics (R)
- research methods
  - for own research
  - as context
- writing skills

Scientific Programming

- distributed computing
- design patterns
- comments
- FAIR
- testing
- other best practices

Project

Various aspects of the project:

- teams of 3 or 4 students
- agile method (scrum)
- lecturer/researcher is product owner
- regular meetings with ultimate stakeholder
- writing and research skills:
  - technical report (team product),
  - research paper (individual product)
- assessment
- presentation during a (small scale) symposium

The eScience Center is the Dutch national hub for the development and application of domain overarching software and methods for the scientific community. Their business model is subsidizing experiments by lending scientific programmers. The eScience Center is active in the fields of:

- life sciences
- physics
- astronomy
- humanities
- social sciences

Via Appia

Figure: Tomb of Hilarus Fuscus

Development of a 4D geographic information system for archaeological purposes.

Pulsar detection

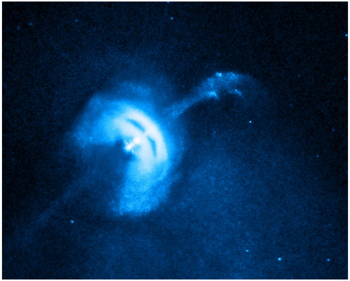


Figure: The Vela pulsar

Re-engineering an application made by an astronomer to detect pulsars by filtering data.

Figure: Westerbork Telescopes

ASTRON is involved in the:

- Westerbork Synthesis Radio Telescope
- Low-Frequency Array (LOFAR),
- Square Kilometre Array (SKA).

Monitoring on energy consumption

Large scientific experiments such as the LHC or SKA are in need of an enormous amounts of energy. This energy is needed for the experiments themselves and for computer facilities. Monitoring the energy consumption of computer technology is not a default function of monitoring frameworks.
The AUAS does research on exposing detailed energy consumption metrics to general purpose system monitoring tools for trend analysis.

Visualising high dimensional scientific data in Virtual Reality

In Virtual Reality more dimensions can be shown than on traditional displays. Goal is to improve scientific analysis of high-dimensional data sets with affordable virtual reality gear.

CERN (European Organization for Nuclear Research, Conseil Européen pour la Recherche Nucléaire)

- 1953 convention signed by 12 founding members (Belgium, Denmark, France, the Federal Republic of Germany, Greece, Italy, the Netherlands, Norway, Sweden, Switzerland, the United Kingdom and Yugoslavia)
- today 22 members states
- 2500 employees
- 600 institutes and universities use CERNs facilities
- 12.000 visiting researchers

Latest discovery in 2012 of the Higgs boson (explaining mass in the Standard Model).

LHC

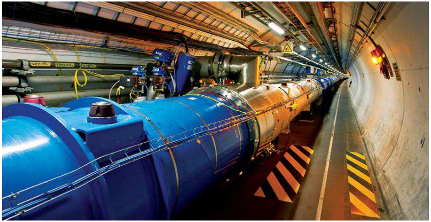
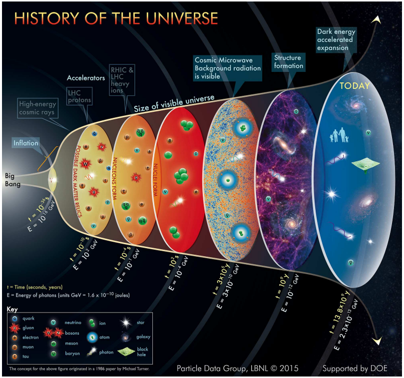


Figure: The Large Hadron Collider is the world's largest and most powerful particle accelerator (Image: CERN)

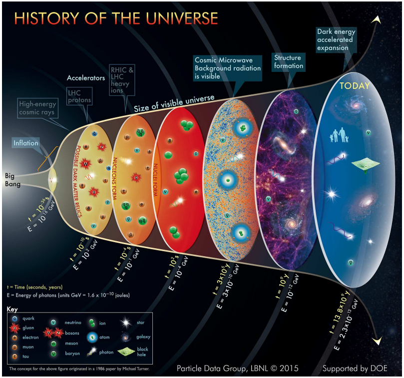
- 7 detectors: ATLAS, CMS, LHCb, ALICE, TOTEM, LHCf, MoEDAL
- 27 kilometer, about 100 meter underground, France and Switzerland (near Geneva)

ALICE



- Quark-Gluon Plasma
- 18 detectors

ALICE



- Quark-Gluon Plasma
- 18 detectors

Amsterdam University of Applied Sciences is associate member of the ALICE collaboration.

Load balancing

In Run 3 (2021-2023) ALICE produces 1.1 TB/s. Research is done to distribute these data from 250 computers to 1500 computers.

- interns doing research already
- experiments on cluster at Nikhef
- cooperation with Frankfurt Institute for Advanced Studies

Monitoring

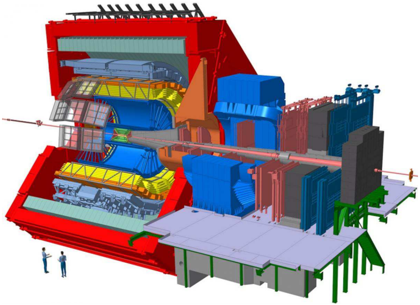


Figure: A Large Ion Collider Experiment

To monitor the machine and all its elements research is done.

Bookkeeping

To do research on the data produced by ALICE one needs to know how this data is produced. For this is bookkeeping needed. Data is worked on for various reasons:

- calibration
- interpretation
- specification
- production of Monte Carlo simulations
- research itself (popularity of data)

system control and data acquisition, i.e. a logging system.

Questions?

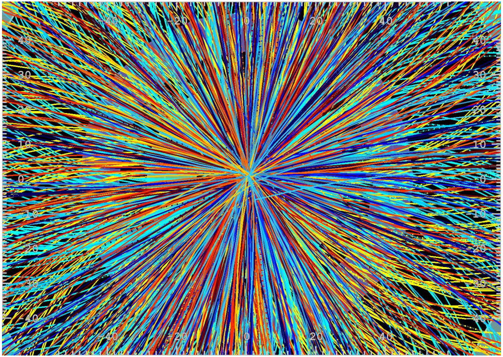


Figure: Events recorded by the ALICE experiment from the first lead ion collisions, at a centre-of-mass energy of 2.76 TeV per nucleon pair. From: cdsweb.cern.ch/record/1305399

Contact me: m.teitsma@hva.nl

Monya Baker.
Is there a reproducibility crisis? a nature survey lifts the lid on how researchers view the'crisis rocking science and what they think will help.
Nature, 533(7604):452–455, 2016.

Mario Rosado de Souza, Robert Haines, and Caroline Jay.
Defining sustainability through developers eyes: Recommendations from an interview study.
In *Proceedings of the 2nd Workshop on Sustainable Software for Science: Practice and Experiences (WSSSPE 2014)*, 2014.

Caroline Jay, Robert Haines, Markel Vigo, Nicolas Matentzoglu, Robert Stevens, Jonathan Boyle, Alan Davies, Chiara Del Vescovo, Nicolas Gruel, Anja Le Blanc, David Mawdsley, Dale Mellor, Eleni Mikroyannidi, Richard PeterRollins, Andrew Rowley, and Julio Vega.
Identifying the challenges of code/theory translation: report from the code/theory 2017 workshop.
Research Ideas and Outcomes, 3:e13236, 2017.

Caroline Jay, Rawan Sanyour, and Robert Haines.
Not everyone can use git: Research software engineers recommendations for scientist-centred software support (and what researchers really think of them).
Journal of Open Research Software, 2017.

Lucas N Joppa, Greg McInerny, Richard Harper, Lara Salido, Kenji Takeda, Kenton O'hara, David Gavaghan, and Stephen Emmott.
Troubling trends in scientific software use.
Science, 340(6134):814–815, 2013.

Roger D Peng.
Reproducible research in computational science.
Science, 334(6060):1226–1227, 2011.

Francisco Queiroz, Raniere Silva, Jonah Miller, Sandor Brockhauser, and Hans Fangohr.

Good usability practices in scientific software development.
arXiv preprint arXiv:1709.00111, 2017.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al.
The fair guiding principles for scientific data management and stewardship.
Scientific data, 3:160018, 2016.