The DUTCH e-infrastructure

LIFE SCIENCE COMMUNITY

Establishing life science ICT infrastructure SURFsara ELIXIR – NL: big data context

ELIXIR VT, January 16, 2013



Life Science ICT support – NL

SURF:

- SURFnet : network, collaboration
- SURFmarket : software
- SURFshare : innovation
- SURFsara : computing, storage



From January 2013

JRF

BiG Grid: computing infrastructure

Netherlands Bioinformatics Centre

- BioAssist and Bioinformatics Research Support
- BioRange: science collaboration

Netherlands e-Science Centre: software, interdisciplinary



BiG Grid





ELIXIR VT – January 16, 2013





e-BioGrid : BiG Grid e-infrastructure gateway for life science research

When: 1 September 2010 – 31 december 2012 Funding: 2 m€

* **BiG Grid computing resources** Enabling life science compute and data intensive research Selected technology area main projects (2 years), e.g. data a Small ad-hoc support projects (2-3 months)

* e-BioGrid human resources (UvA, UL, SARA)

Multidisciplinary e-core coordination team e-Science infrastructure support team Decentralised scientific programmers (located at selected bioscience labs) An evolving e-bioscience community to build and use the infrastructure





Technology exchange

NBIC BioAssist Taskforces



SURF SARA

ELIXIR VT – January 16, 2013

e-BioGrid: results

 Community building, e.g. regular workshops and user meetings, newsletters

٠

- Growing number of life science users (universities, medical centres)
- Life scientists trained to use computing resources
- Guiding, monitoring and managing life
 science e-infrastructure needs
- Awareness of needs to plan and budget
 e-infrastructure

- Technology specfic infrastructure, e.g.
 - Human genome analysis pipeline
 - analyse phenotyping and genotyping (SNP) data from large-scale plant breeding experiments
 - Integration of computationally intensive software in a metabolomics data processing tool chain
- Generic infrastructure, e.g.
 - Life Science Grid Portal
 - NBIC Galaxy on Cloud

www.ebiogrid.nl/results

٠



e-BioGrid: example projects



NeCEN

EURO-BIOIMAGING

- Analysis of 750 human genomes
 - Genome analysis pipeline running on Grid
 - Using Compute: <u>http://www.molgenis.org/wiki/ComputeStart</u>

- Infrastructure image analysis
 - Authentication
 - Cloud computing





NBIC Galaxy Server on Cloud

유 nbic Galaxy 🖥 Galaxy / NBIC Analyze Data Workflow Data Libraries Admin Help User Tools Options V History Options V Map with Bowtie for Illumina **Wavelet Analysis** 0-Will you select a reference genome from your history or use a built-in index?: Graph/Display Data **Unnamed history Regional Variation** Use a built-in index 0 Multiple regression Built-ins were indexed using default options 38: GAPSS - FASTQ to @ 0 💥 **Multivariate Analysis** FASTA on data 22 Select a reference genome: Evolution Human_UCSC_hg19_complete \$ 37: GAPSS - FASTQ to @ 0 💥 Metagenomic analyses if your genome of interest is not listed - contact Galaxy team FASTA on data 22 **FASTA** manipulation Is this library mate-paired?: NGS: QC and manipulation 36: Map with Bowtie 🛛 👁 🖉 💥 Single-end 0 NGS: Mapping for Illumina on data 22 NGS: Indel Analysis FASTQ file: 35: Map with Bowtie 👁 🖉 💥 **NGS: Expression Analysis** 22: FASTQ Groomer on data 2 🗘 for Illumina on data 22 NGS: SAM Tools Must have Sanger-scaled quality values with ASCII offset 33 NGS: Peak Calling 34: Map with Bowtie 💿 🖉 💥 Bowtie settings to use: for Illumina on data 22 SNP/WGA: Data; Filters Commonly used 2 SNP/WGA: QC; LD; Plots 31: VarScan -· 1 × SNP/WGA: Statistical Models For most mapping needs use Commonly used settings. If you want full control use Full parameter list pileup2snp on data 30 NGS: Snip Detection Suppress the header in the output SAM file: NGS: Tools LUMC \checkmark 30: Generate pileup • / X GAPSS on data 29 Bowtie produces SAM with several lines of header information by default Map with Bowtie for Illumina output in SAM format: 29: SAM-to-BAM on · 1 × \checkmark 2 GAPSS - FASTA to FASTQ data 28 The output file will be in SAM format GAPSS - FASTQ to FASTA 28: Map with Bowtie 🛛 🖉 💥 Execute for Illumina on data 22 GAPSS - SCARF to FASTO 308 lines, format: sam, database: - CADCC . javascript:void(0); **NBIC** Tools **Control panel** History panel



e-BioGrid: lessons learned

- Local infrastructures are sometimes favored
- Abundant communication concerning user needs, available expertise and solutions, and progress is essential.
- Training is best set up build on usecases and on-site
- Monitoring of resources usage is not straightforward and may be improved
- Different user groups require different solutions and support: from pilot to workflow frameworks

- Dynamic field of life science research
 - requires both structural support for e-infrastructure developments, as well as ad-hoc support for short implementation projects.
 - Short-cycle implementation projects often favor cloud over grid computing.



e-BioGrid website

info@ebiogrid.nl



e-BioGrid

Building the Dutch e-science infrastructure for life science research



 $\widehat{}$

www.e-biogrid.nl

About | Projects | Results | Get involved | News & events | Contact

Biobanking

Mass spectrometry

Microarray technology

Nanoscopy and imaging

Next generation sequencing

NMR Spectroscopy

Generic

e-BioGrid: enabling Life Science Research and Technology

Life sciences today deals with data-intensive and high-throughput experimentation. e-BioGrid supports life-science research through the development and exploitation of a computing infrastructure. High-performance computing hardware as well as infrastructure software is build and made available to enable more effective research in technology areas such as next-generation sequencing, microarray technology, mass spectrometry, nanoscopy and biobanking.





Examples include the availibility of grid computing for computational intensive DNA sequence

Subscribe to our newsletter here

2012-05-15 -Mass spectrometry

2012-05-15 The project "Implementing a proteomics Taverna workflow onto Grid and Cloud" has been added/updated.read more



ELIXIR VT - January 16, 2013



SURFsara

Dutch e-infrastructure

High-performance computing and storage resources,

- The Life Science Grid
- National supercomputer
- National compute cluster, GPU's
- High performance compute Cloud
- Hadoop
 Had
- Network and light paths
- Petabytes of disk and tape storage

© Tailored support, training and advice



SURFsara



ELIXIR VT – January 16, 2013

SURF SARA

Life Science Grid



WHAT IT IS

•Several clusters at academic medical centers and universities.

•Open to all life scientists based in the Netherlands.

•Clusters can be used separately or together by using Grid middleware.

•Currently more than 3700 job slots for concurrent processing.

•Distributive storage capabilities and automatic replication of data.

•Maintained and supported by SARA.



Life Science Grid sites



LUMC – Leiden WUR – Wageningen UU – Utrecht NKI – Amsterdam AMC – Amsterdam Radboud Universiteit – Nijmegen ErasmusMC – Rotterdam Technische Universiteit - Delft Rijksuniversiteit Groningen Keygene - Wageningen

Each of these clusters have either 16 or 32 cores and 18 Tb of storage space.



SURFsara HPC Cloud



SURF SARA

ELIXIR VT – January 16, 2013

SURFsara

Continues

- Life science research e-infrastructure support (e-BioGrid)
- Life science community building
- Development of e-infrastructure based on life science research requirements

Ongoing:

WS-Pgrade install (with AMC)

We would like to hear from users how they experience WS-Pgrade

- Upgrade of Life Science Grid
- trainings



technology exchange, e.g. Workshop with CSC Finland





SURFnet : network and lightpaths





ELIXIR - NL



Compute Training Tools Standards



DUTCH TECHCENTRE FOR LIFE SCIENCES

DTL: Dutch techcentre for life sciences DISC: Data Integration and Stewardship Centre

Services: SURF, NLeSC, and others

Netherlands Proteomics Centre (NPC) Netherlands Metabolomics Center (NMC) Center for Translational Molecular Medicine (CTMM) Wageningen University & Research Centre Radboud University Medical Centre UMC Utrecht Hubrecht Institute TNO Quality of Life UMC Groningen Netherlands Center of Systems Biology (NCSB) Leiden University Medical Centre VU University Amsterdam



LIFE SCIENCE IS THE NEXT BIG SCIENCE

HIGH TECH - 'BIG' SCIENCE

discovery & innovation depend on advanced technologies and on cross tech collaboration

easy and cost-effective access to key technologies is essential

HIGH VOLUME - 'BIG' DATA

data volumes double every 9 months data integration needed to understand complex biological systems progress needs data sharing & open access of public data strategy needed for an integrated technology & data infrastructure

DTL



DISC Services



DTL >>>

SET UP PPP CONSORTIUM BASED UPON EXISTING NETWORK



ELIXIR - NL



- Continuing life science ICT infrastructure support: SURF, NLeSC, and others
- Establising DTL-DISC
- Providing affiliated, and possibly commissioned services
- Head node workshop, March 21-22, Amsterdam

• April 2013, final recommendation of the ELIXIR Interim Board on Node Application.



Contact: irene.nooren@surfsara.nl (i.m.a.nooren@uva.nl)

www.surfsara.nl

See you at EGI community forum, April 8-12?



