Cremlin Connecting Russian and European Measures for Large-scale Research Infrastructures

Document information

Deliverable no.	D2.3
Deliverable title	Recommendations on e-infrastructure initiatives
Deliverable responsible	DESY
Related Work-Package/Task	WP2/ Tasks 2.4 & 2.1
Type (e.g. Report; other)	Report
Author(s)	M. Sandhop, V. Gülzow, P. Fuhrmann, DESY
Dissemination level	Public
Submission date	31.08.2018
Download page	https://www.cremlin.eu/deliverables/

Project full title	Connecting Russian and European Measures for Large-scale Research Infrastructures
Project acronym	CREMLIN
Grant agreement no.	654166
Instrument	Coordination and Support Action (CSA)
Duration	01/09/2015 – 30/08/2018
Website	www.cremlin.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654166.

Recommendations on e-infrastructure initiatives

CREMLIN Deliverable D2.3

WP2: Exchange platform

Task 2.4: Link of Russian megascience projects to global and European e-infrastructure initiatives.

Lead partner: DESY; NRC KI

Introduction:

DESY has together with the NRC KI organised a European-Russian "Big data workshop" in February 2017 in Moscow. This workshop has been carried out in the framework of CREMLIN WP2, and addresses the Task 2.4, dedicated to establishing "Links of Russian megascience projects to global and European e-infrastructure initiatives" and to working out specifi8c Recommendations in this field. The key aim of this event is to address Task 2.4 and to support the integration of the Russian megascience projects into the European and global e-infrastructure initiatives. The five European and global big data initiatives involved are: EOSC, EGI, RDA, PRACE and GÉANT.

Workshop on Big data Management

The workshop has been realised at the premises of the NRC KI in Moscow, during 15-16/02/2017.

Summary of the Workshop and its results:

About:

This CREMLIN WP2 Workshop on "Big data management" brought together around 70 experts both from European and Russian CREMLIN partners, as well as from external European and global platforms and initiatives, and Russian universities.

First discussion level:

- Which is the status of collaboration?
- Which are typical best practice examples in the collaboration?
- Which are the key challenges for the future collaboration?
- Which are the next steps in addressing these challenges?
- Collaboration in data management with respect to:
    - research with neutron sources around the upcoming PIK facility in Gatchina (WP4)

- Particle physics (WP7) and research with ion sources including the NICA facility in Dubna (WP3)
- Photon science (WP5) and research with High-power lasers (WP6)

Second discussion level:

A second level of discussion refers to the five European and global big data initiatives and networks that were represented at the WS by high-level stakeholders. The WS functioned as a platform to present these five initiatives, to sound out the mutual collaboration interests and possibilities and potential next steps to enter into closer collaboration, where mutually intended:

1. EOSC - European Open Science Cloud) (*Barend Mons; former chair of HLEG)*
2. GÉANT - Pan European Infrastructure and services for research and education *(Vincenzo Capone)*
3. EGI (*Yannick Legré)*
4. PRACE aisbl - Partnership for Advanced Computing in Europe *(Florian Berberich)*
5. RDA - Research Data Alliance *(Mark Parsons)*

Outcome & Conclusions:

The workshop clearly pointed out that there is an urgent need to increase the European-Russian connectivity, to work out common data policies, and to use also CREMLIN for joint work on these tasks. It was agreed, highlighted and recommended:

- Among the key challenges: joint EU-Russian software development and joint work on meta data; Long term data preservation
- NRC KI serves as a Tier1/Tier2 facility for CERN experiments ATLAS, ALICE, LHCb
- JINR hosts Tier1for CMS; Tier2 for all LHC experiments
- European XFEL upcoming operational phase 2017: connectivity to Russia via high speed data links needs to be enlarged in order to federate compute and storage resources for a seamless analysis environment
- For EU-Russian collaboration along PIK: urgent need for data analysis services in order to provide a consistent ecosystem for EU-Russian users of neutron sources; follow-up meeting agreed for September 2017
- For lepton collider SCT in Novosibirsk: proposed to increase the data link between Novosibirsk and central Europe and to increase the size of the existing data centre; CERN and BINP to launch a web-based platform in April 2017 to share experience in software and computing for lepton collider projects

- Russia invited to participate in the GO FAIR initiative (FAIR: “Findable; Accessible; Interoperable; Reusable”)
- Russia invited to increase collaboration with GÉANT, e.g. via eduROAM, eduGAIN
- Suggested to Russian partners to propose to host 2019 RDA plenary.

Contributions:

Mikhail Popov (NRC KI) mentioned in his welcome address the Scientific and Technological Development Strategy of the Russian Federation adopted in December 2016. The development of Big Data Management Systems and large-scale scientific infrastructure in Russia are identified in the Strategy among its main priority tasks.

He also expressed his confidence that the Strategy opened a new opportunity to broaden the cooperation in the Big data area especially within megascience projects.

It was especially emphasized the success of the current collaboration between NRC KI, JINR Dubna (both host WCLG Tier 1 centers) and CERN.

Frank Lehner (DESY) introduced the objectives of the WS, that are both related to the CREMLIN thematic WPs (addressing the collaboration in the specific communities), as well as related to the European and international initiatives and their collaboration with Russian partners.

In his welcome note, **Richard Burger** (Science Counsellor at the European Union Delegation to Russia) highlighted the Cremlin project as a “flagship project” and “key initiative” in the European-Russian collaboration with a high visibility in this multilateral collaboration context. The European Commission will be looking forward to the recommendations that will be produced in this project.

As to the European Open Science Cloud, **Wainer Lusoli** (European Commission) gave the keynote lecture “From Vision to Action: The European open science Cloud (EOSC)”. He introduced the European Cloud Initiative, that will lead to the Open Science Cloud by 2020, as part of the Digital Single Market (DSM) Strategy. Mr Lusoli highlighted the policy actions that are foreseen by the EC, and introduced the Commission High Level Expert Group European Open Science Cloud (EOSC), and its First Report (published October 2016). The H2020 EOSC Pilot project has recently started.

Links to EOSC, and EOSC Pilot:

http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

http://eoscpilot.eu/

Overview talks:

Volker Gülzow (DESY) presented “Big data for large-scale facilities” and developed a possible model for large scientific computing facilities. In this context he highlighted significant contributions by previous and ongoing EU Projects, as there are the PANDATA FP7-project on Photon and Neutron

data infrastructures and the H2020 EOSC Pilot. The latter is addressing typical challenges like “Data is not always openly accessible (FAIR Principles)”; “the lack of incentives and rewards for scientists for sharing their data”; “the lack of interoperability required for data sharing” and “the fragmentation between data infrastructures that are split by scientific and economic domains, countries and governance models”. Furthermore, he reported on the success of the EOSC Pilot in providing a prototype of a governance framework for the European Open Science Cloud, being officially inaugurated in Nov, 2018. The EOSC Pilot developed a number of science demonstrators and engaged with a broad range of stakeholders in that domain. However, key challenges remain: Open and interoperable software is needed and consistent and reusable Meta Data models is required to guarantee future usability of scientific data.

Vasily Velikhov (NRC KI Supercomputing Centre) gave an overview on “Big data in Russian context”. KI Big Data providers: HEP; Materials Science (nano-bio) – SR, Neutron sources, E-Microscopy; Genetics; Brain research.

- International collaboration:
    - in terms of big data at large-scale facilities: LHC; European XFEL; ITER; FAIR; ESRF
    - with e-infrastructures: WLCG, EGI (including CSIRT), operate National certification authority (X.509)
- NRC KI serves as a Tier1/Tier2 facility for ATLAS, ALICE, LHCb; and operate HPC center (1+ PFlops peak), interconnected with Tier1 at 240 GBit/sec., providing HPC/HTC integration. NRC KI is connected to LHC OPN and LHC ONE and provides connectivity for branch institutes IHEP (Protvino), ITEP (Moscow) and PNPI (Gatchina) -60 GBit/sec.
- Technologies we use/extend: CERN EOS and dCache: both as parts of a production in Tier-1 and R&D activity for federated cloud + WLCG/XFEL demonstrators, also CERNbox/EOS as KI infrastructure project
- Job management/scheduling: Torque/Maui, Slurm, CREAM CE, ARC CE Storage: Lustre, UFS/ZFS-based NFS, CERN VM FS, HTTP/Rsync/SSH-based access
- Management: HP CMU, CFEngine, Puppet, own deployment engine Analysis : ANN & ML algorithms
- Proposing:
    - extend LHC ONE to all CREMLIN mega science facilities?
    - support off-line data processing and simulations for mega science facilities on selected large HPC centers (as FZJ, KI)?

Patrick Fuhrmann (DESY) reported about Challenges in Storage. He introduced initiatives within recent EU projects like the INDIGO-DataCloud and the eXtreme Data-Cloud to automatically manage Quality of Service in Storage and to orchestrate scientific “Data Life Cycles”, along examples from the

SKA Square Kilometer Array and the European XFEL. In both cases, the expected data rates make the use of “Smart algorithms” inevitable, possibly using machine learning methods. Furthermore, he elaborated on typical issues of “Long Term Data Preservation” like bit file preservation including long term encryption mechanisms, content preservation and last but not least: governance and legal issues, e.g. the long term ownership of scientific data.

Results of the Sessions:

Session I: Best practice examples from ongoing EU-Russian collaborations (chair: Vasily Velikhov)

Massimo Lamanna (CERN) presented updated figures of the systems CERN has developed and uses to match the data management requirements from the LHC Run 2. 2016 was a record year with a peak of 10 PB collected in a single month.

He described three strategic systems for CERN: EOS, CERNBox and SWAN. EOS is an open software storage solution developed at CERN to deliver affordable commodity storage to scale to the growing requirements of CERN experiments. CERNBox is the cloud storage service to allow users to synchronise and share their data in a seamless way. SWAN (Service for Web based ANalysis) is the platform to perform interactive data analysis in the cloud (via web browser).

Alexei Klimentov (NRC KI) was talking about Big data handling challenges in HENP. The largest scientific instrument in the world – the Large Hadron Collider (LHC) – operates at the CERN Laboratory in Geneva, Switzerland. Experiments at the LHC explore the fundamental nature of matter and the basic forces that shape our universe. To address an unprecedented multi-petabyte data processing challenge, experiments are relying on the deployed computational infrastructure of the Worldwide LHC Computing Grid (WLCG). More than 6,000 scientists from 200 universities and laboratories in 45 countries analyze the LHC data in search of new discoveries. Since the start of LHC data taking, experiments operate under conditions in which contention for computing resources among high-priority physics activities happen routinely. The experiments will collect a factor of 10-100 more data during the next 3-5 years.

The Big data technology Lab at NRC KI has started several R&D projects in collaboration with CERN, DESY, EU and US Universities to integrate High Throughput Computing (Grid) with High Performance and cloud computing, to address High-Luminosity (2022-2025) LHC challenges. The major projects are Federated data storage, workload management system (megaPanDA) and Data Knowledge Base (DKB). The status of the projects and the first results has been reported at the workshop. It was explicitly stressed how HENP WMS can be used for bioinformatics and other data intensive sciences.

Krzysztof Wrona (European XFEL): reported on the data management challenges to be faced by the European X-Ray Free Electron Laser Facility. Using the X-ray flashes of the European XFEL, scientists will be able to map the atomic details of viruses, decipher the molecular composition of cells, take three-dimensional images of the nanoworld, film chemical reactions, and study processes such as

those occurring deep inside planets. In order to do so, scientists will be performing experiments which generate huge amount of data at very high rates. A single instance of a 2D detector would be able to produce up to 10GB/sec of raw data. This data needs to be recorded and inspected in the semi-real time allowing fast feedback to be provided to the next experiment stage.

The detailed data analysis will be performed offline using compute and storage resources which are located at the data center on site. A possibility to enlarge the computing model of the European XFEL by including external computing centers, located in partner countries like Russia and connected via high speed data links was presented. A computing facility in Russia would help increasing the Russian user base for the European XFEL and would serve as a focus point in Russia. A proposed pilot project would allow both sites to test whether this approach is useful for the likely increased needs in the operation phase of the European XFEL facility.

Anton Teslyuk (NRC KI): gave a short progress report about research activities in Kurchatov Institute in the domain of XFEL diffraction patterns classification and sorting. Also preliminary results in software development for diffraction simulation, CXI storage database and structure reconstruction were presented.

Thorsten Kollegger (GSI) reported about the computing challenges at the Facility for Antiproton- and Ion Research (FAIR). The diverse nature of research programs at FAIR result in a large variety of requirements: the online reconstruction of the larger experiments will have to handle up to 1 TByte/s input rates while some of the smaller are in the MByte/s range, the experimental collaborations are quite different in size ranging from few people up to several hundred with different proficiency in data management and analysis. Common to all is the data management challenge of sharing the data and its processing between the international researchers. The common software development work and tools with the NICA project at JINR were also presented.

Victor Braguta (ITEP NRC KI) pointed out his project "Study of quark-gluon matter within lattice simulation of QCD". This project is supported by the FAIR-Russia Research Center (FRRC). The aim of the project is to study properties of quark-gluon plasma at nonzero baryon density. The quark-gluon plasma is strongly correlated system and the only way to carry out first principle study is lattice simulation of theory of strong interactions. In order to do the lattice simulation, Victor uses the NRC KI HPC facilities and the FAIR-ITEP supercomputer.

Harald Reichert (ESRF) described challenges in data management for Photon Science. The Photon Science community is a very large and heterogeneous user community exploiting a well-developed European network of accelerator-based photon sources (storage rings, XFELs). Most of the photon sources are set up as national sources, albeit with a major fraction of user time being allocated to users from across Europe (transnational access). The Photon Science community has been organising itself since 2006 in order to deal with the data challenge, mostly fostered by European projects. Basic

points for an open data policy for all facilities have been established (PaNData-ODI project) and are in the process of being adopted and implemented at various facilities.

Using the ESRF as an example, the challenges of data management at the next generation of facilities, that are already in planning or under construction, were introduced with data production rates reaching the level of PBytes/week. Currently, the facilities are trying to organise a European network for the development of shared data analysis tools which are mandatory to exploit the next generation of facilities. This will require support on the European level since (i) it goes beyond the level of support available from national funding agencies and (ii) the strong transnational use of the facilities fully justifies the use of common resources.

Session II: European and international Big-data related platforms and networks (chair: Volker Gülzow)

Barend Mons (former EOSC HLEG Chair): Introducing both the EOSC and the report of the HLEG (until December 2016); key element GO FAIR, which is: *Global Open FAIR*; where FAIR stands for: making fragmented and unlinked research data Findable, Accessible, Interoperable and thus Reusable. The EOSCpilot: to support the implementation of the first phase of the EOSC. We need an "Internet of FAIR Data & Services (IFDS)". FAIR Data Stewardship.

Vincenzo Capone (GÉANT – the Pan European Infrastructure and services for research and education), gave an introduction into the GÉANT mission and services. GÉANT serves around 50 million users in 10,000 institutions in 40 countries in Europe (geographical Europe), and interconnecting more than 100 countries globally. Services are e.g. Connectivity & network management; End to end Performance; or Trust, Identity and Security (eduGAIN – Secure access, single sign-on; eduroam – Seamless Wi-Fi access for research and education around the world). GÉANT is providing services to all relevant LHC experiments. Also, GÉANT´s Support for International Users was mentioned (Dedicated User Support Team; Single point-of-contact for international collaborations and organisations; Providing a one-stop-shop; ..). There are some Russian connections to GÉANT existing, e.g. to the NRC KI to JINR. GÉANT intends to increase the collaboration between European and Russian user communities, e.g. via eduROAM, or eduGAIN".

Yannick Legré, EGI: the EGI is a European and worldwide federation of more than 300 computing and data centres, and is providing a wide range of offers to e.g. Research Infrastructures, multinational projects. For instance cloud computing, Storage and data, Training, Data management, Security. The WLCG Grid is the "largest resource provider and service consumer of the EGI Federation". EGI is open for further collaborations.

Florian Berberich, PRACE aisbl - Partnership for Advanced Computing in Europe, explained that PRACE is a pan European HPC e-Infrastructure, with 25 members. Access to PRACE is provided via peer review, free of charge, main criterion is scientific excellence. Also, Training is an important pillar of PRACE, for instance the Seasonal Schools (Upcoming Seasonal School: 10 – 12 April 2017, PRACE Spring School 2017, Sweden - HPC in the Life Sciences). Training and event homepage: http://www.training.prace-ri.eu/nc/training_courses/index.html . PRACE Summer of HPC (SoHPC).

Possible Cooperation with Russian RI / Some ideas for future cooperations in HPC: Exchange of Experience; Network; Operation; Training; Applications Enabling.

Mark Parsons, RDA - Research Data Alliance: The RDA is an alliance of 4,908 individual members, and 46 organisational members, with the mission to "build the social and technical bridges that enable open sharing of data". Aims at "Accelerating data sharing and interoperability across cultures, communities, scales, and technologies". RDA members form Working and Interest Groups that address targeted issues around data sharing and reuse. Members come from 118 countries and RDA conducts Plenary meetings every six months. Invitations are open for proposals where to host the RDA Plenary Meeting in 2019 and beyond, perhaps Russia. CREMLIN may also want to consider forming or joining an RDA Working and Interest Group or conducting a side event around the dates of a future Plenary.

Sessions III-IV: Reports from Preparatory discussions in three parallel sessions on community-specific challenges in EU-Russian collaboration, including session with five EU/ international initiatives experts:

Volker Gülzow (DESY), **Vasily Velikhov** (NRC KI):

Big data and data management in research with neutron sources (WP4)

Chair: *Jean-François Perrin, ILL;* co-chairs: *S. Grigoriev, A. Kiryanov PNPI NRC KI*

Jean-François Perrin, ILL: This session was the first meeting on the subject of Data Management between the different partners. We had to slightly extend the scope of the discussions to user communities, instruments, data reduction and analysis, in order to obtain a comprehensive overview of data management for neutron facilities.

The analytical facilities present a different range of techniques and instruments for the scientists. In order to conduct their research, scientists use the different tools, complementary techniques and travel between the facilities; this is demonstrated by the regular PaNData user surveys (http://pan-data.eu/Users2014-Results). It is therefore extremely important to coordinate efforts in order to provide a consistent ecosystem for our users in terms of software, policies, infrastructures and services.

The current work to harmonize and support the development of the data treatment software (Mantid, MuhRec & KipTool, Born Again, SASView, NSXtool …) have been presented and discussed as well as the Data Management (i.e: Data and Metadata preservation, Policies, use and impact of DOIs for data). The urgent need for data analysis services has been highlighted by the evolution of experimental data in term of volume and complexity. The PaNDaaS collaboration, which tries to bring this data analysis service to the X-ray and neutron users' community as a whole, and the different approaches were explained.

In order to further develop the discussion between EU and Russian partners, a follow up meeting is foreseen for September 2017.

Big data in Particle physics and in research with ion sources (WP3; WP7)

Chair: *Eva Sicking, CERN;* co-chairs: *Vladimir Koren´kov, JINR; Yuriy Tikhonov, BINP*

Eva Sicking, CERN: presented a summary of the Big data challenges discussed in the WP3 and WP7 session. For WP3, the projects FAIR@GSI (presented by Thorsten Kollegger) and NICA@JINR (presented by Vladimir Korenkov) were discussed, and for WP7, CLIC@CERN (presented by Eva Sicking) and SCT@BINP (presented by Iouri Tikhonov) were discussed.

FAIR is currently under construction and it is planned to start operation in 2025. FAIR has an unprecedented data challenge with 1TByte/s to be fed into on-line farms and data to be stored in the order of 35 PByte/year to disk and 30 PByte/year to tape. To tackle this challenge, a data centre for FAIR and GSI called "Green Cube" was put into operation. For an efficient use of person power, a common software framework for all FAIR experiments and beyond (including for instance ALICE at the LHC) is currently being developed.

NICA's computing challenge can build on large, existing infrastructure at JINR and on a close collaboration with FAIR and LHC. At JINR, there are already numerous existing collaborations with other organisations and experiments. For instance, JINR hosts a Tier 1 for CMS and a Tier 2 for all LHC experiments.

The lepton collider CLIC that would start operation in 2035 has rather pure collision events which correspond to a data challenge far below the one of FAIR and NICA. The CLIC data of 10GByte/s rates would be 10 times the rates of today at the LHC. Due to the small linear collider community, many cooperations with other projects are already in place in computing and software development. Common software solutions are shared for example between CLIC, ILC, CALICE, FCC, neutrino experiments and the LHC experiments. For simulations= studies performed to identify the physics potential and to work on detector R&D, CLIC uses existing infrastructures such as the Grid and EOS developed for the LHC.

The lepton collider SCT - which currently finalizes the CDR - will have data output of 8GByte/s. To tackle this, it was proposed to increase the data link between Novosibirsk and central Europe and to increase the size of the existing local data centre.

To share experience in software and computing among the WP7 members, it was agreed to set up a web-based platform to share existing information and software tools between lepton collider projects. As a potential option for further collaboration between BINP and CERN it was discussed to integrate the Grid information system CRIC currently developed at BINP into the CLIC grid submission system.

Note: The lepton collider platform has been launched in April 2017:
http://leptoncolliderplatform.web.cern.ch/

Big data and data management in Photon science and in research with High-power lasers (WP5; WP6)

Chair: *Harald Reichert; Alexander Sergeev, IAP RAS*

Harald Reichert, ESRF: The group discussed common challenges in data management which are different in accelerator-based photon sources and at high-power laser facilities.

For Photon Science at accelerator-based sources, there is currently very little interaction on issues of big data and data management. This will become more pertinent with the start of user operation of the European XFEL where Russia is a major partner and issues such as data transfer and data analysis capacities have to be dealt with. The construction of a 4th generation photon science facility at the ESRF and in Russia will also elevate the level of big data issues (large data sets, limited data analysis capacity for individual user groups) for the Russian photon science community and is an docking point for common development efforts.

Challenges in high-power laser facilities are less in the absolute amount of data, but more on the controls side and peak data production rates. This is a common issue at all high-power laser facilities and can be tackled by sharing best practices and technology.

Workshop materials:

https://indico.desy.de/indico/event/16462/

CREMLIN news Workshop:

https://www.cremlin.eu/news/2017/cremlin_wp2_workshop_on_big_data_management/

Picture of the Workshop:

Picture: NRC KI. Group picture Big data WS, 16/02/2017, Moscow