

Elettra Big Data and Open Data Challenges

R. Pugliese

F. Billè, R. Borghes, F. Brun, V. Chenda, A. Curri,
D. Favretto, G. Kourousias, M. Lonza, M. Prica, M. Turcinovich
Elettra – Sincrotrone Trieste S.C.p.A, Trieste, Italy

Big Data and Open Data Workshop, Brussels, May 7-8 2014

Outline

- Elettra and FERMI@Elettra light sources
- CERIC-ERIC and the role of Elettra Sincrotrone Trieste
- A set of use cases and their specific big data issues
- Elettra Sincrotrone Trieste scientific raw data production capacity
- Elettra Scientific Data policy
- Elettra – CINECA distributed data storage infrastructure
- Relevant aspects of Big and Open scientific data: standards, data gravity, compression, economics
- Elettra Virtual Laboratory and future plans towards a sustainable solution to the Elettra Big Data Challenges

Elettra Sincrotrone Trieste in numbers



400 employees

100000 m²

5000 light hours / year of synchrotron radiation (Operation H24)

10 laboratories user support (biology, material preparation, micrometrics, ...)

26+ Beamlines / ELETTRA storage ring and 3+ / FERMI@Elettra (each line is an independent analytical laboratory)

More than 1000 users per year, from more than 25 countries



CERIC-ERIC



NEW

CERIC-ERIC Zero Call for proposals is open.

[Read more](#)

CERIC-ERIC is evaluating the applications received for up to 10 (Ten) Fellowships in the start-up and development activities of the consortium.

CERIC will be a distributed research facility, set up as an ERIC, by a group of proposing countries (**Austria, Croatia, Czech Republic, Hungary, Italy, Poland, Romania, Serbia, Slovenia**) and open to other interested countries.

The specific scope of this ERIC will concern the offer as an integrated service to external researchers of the access to *synchrotron light and other microscopic probes for analytical and modification techniques* notably for materials preparation and characterization, structural investigations and imaging in Life Sciences, Nanoscience and Nanotechnology, Cultural Heritage, Environment and Materials Sciences and to their various technological and industrial outcomes ranging from energy to biomedical and of interest to most manufacturing industries.

CERIC's mission will be to bring the integrated service to world-level quality thus contributing to the attractiveness of the European Research Area and stimulating a beneficial impact on the scientific and economic development of the entire region, also helping to introduce a strong interchange between scientists and technicians and attraction of scientists from other regions.

CERIC-ERIC

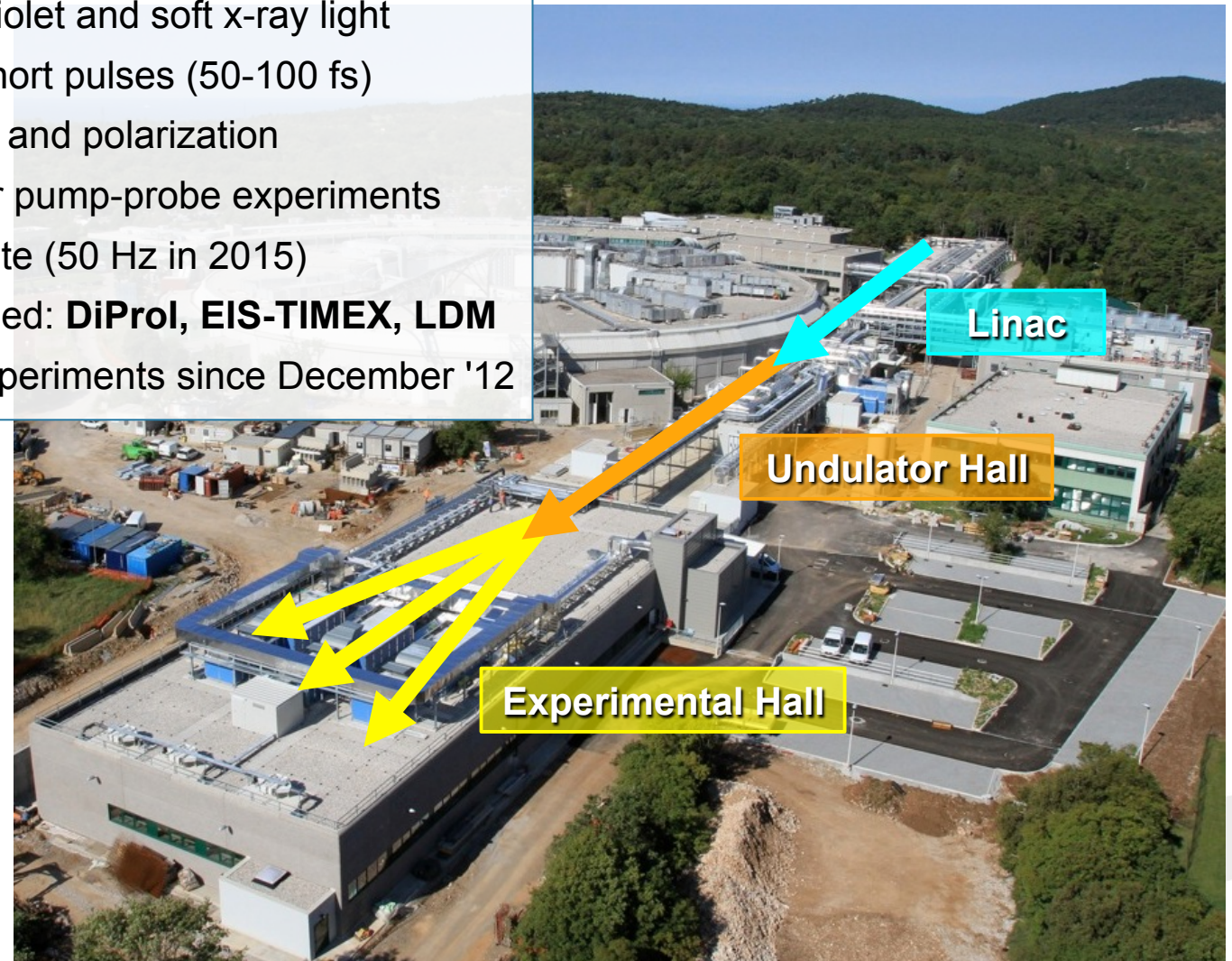
Elettra is the statutory seat of CERIC-ERIC, a distributed Research Infrastructure with:

- a common access point describing and offering the available services;
- a common entry point for users proposals and allocate access time to the integrated services;
- free and open access based on quality selection only;
- support and logistic services as required;
- a common legal form, allowing a single and effective governance;
- a single management board in charge of its integrated operation;
- a common scientific Big Data infrastructure?



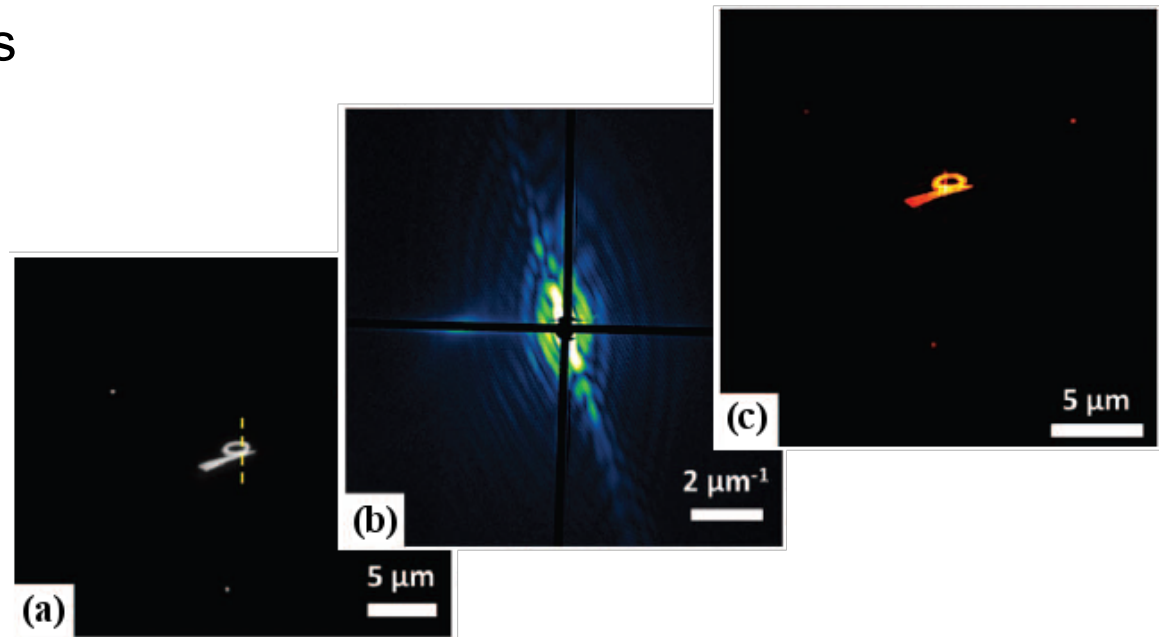
FERMI use cases

- **FERMI** is the sole seeded Free Electron Laser facility
- providing extreme ultraviolet and soft x-ray light
- intense (up to 300 μJ) short pulses (50-100 fs)
- tunable light wavelength and polarization
- a user laser available for pump-probe experiments
- **10 Hz** pulse repetition rate (50 Hz in 2015)
- three end-stations installed: **DiProl, EIS-TIMEX, LDM**
- open to external user experiments since December '12



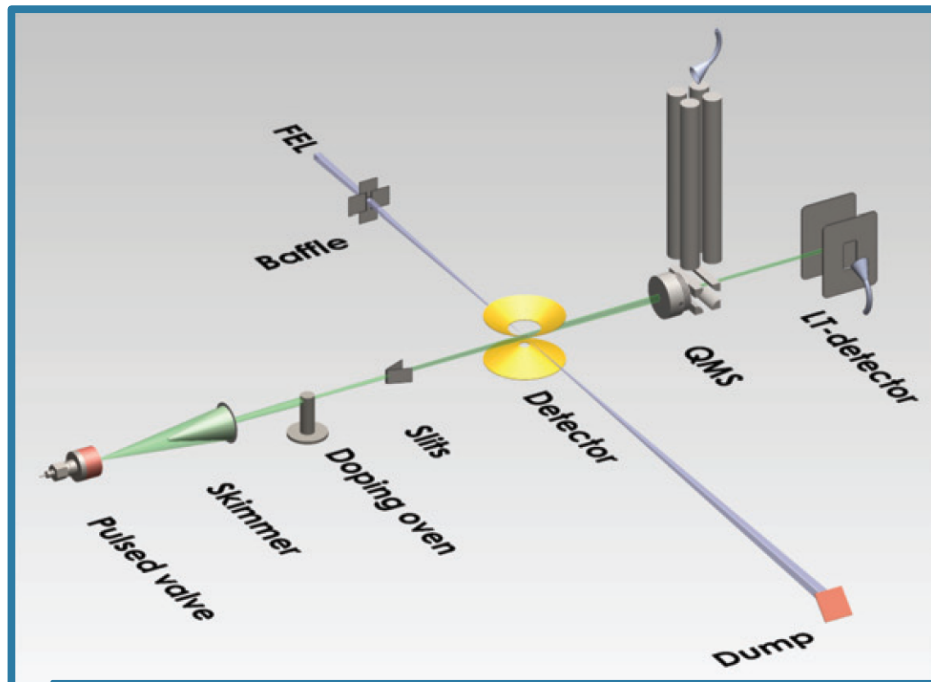
DIPROI end-station

- single shot Coherent Diffraction Imaging experiments
- patterns measured by a Princeton CCD (2048x2048 pixels, 16 bit depth)
- wavelength and polarization tuning
- pump & probe experiments
- complex data analysis



* Images from
F. Capotondi et al., "Coherent imaging using seeded free-electron laser pulses with variable polarization: First results and research opportunities", Review of Scientific Instruments, Vol. 84 - 5 (2013)

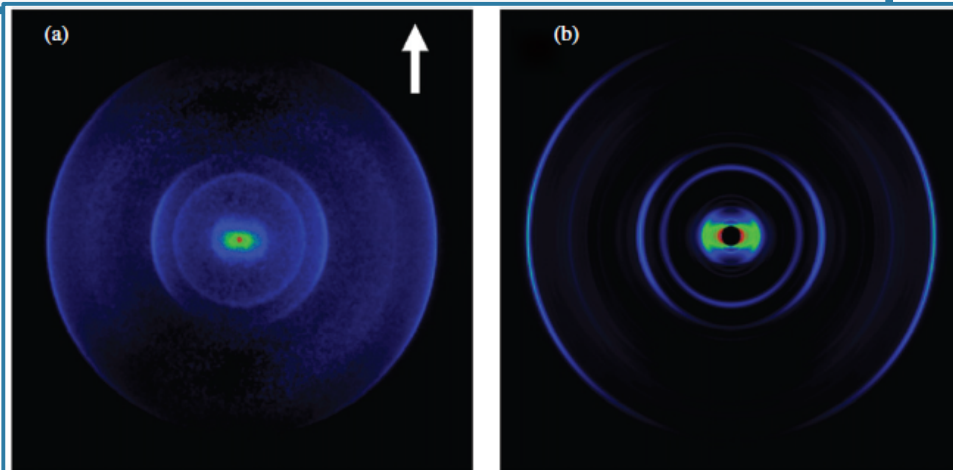
LDM end-station



- Low Density Matter investigations
- a pulsed valve provides a target jet
- atomic, molecular and cluster targets
- pump & probe experiments
- wavelength and polarization tuning
- Velocity Map Imaging spectrometer
- Time Of Flight mass spectrometer

sCMOS Andor R Neo camera 2560x2160 pixels, 16 bit depth; CAEN digitizer (VX1751, 1 GS/s, 10 bits, 8 channels)

data throughput ~120 MB/s



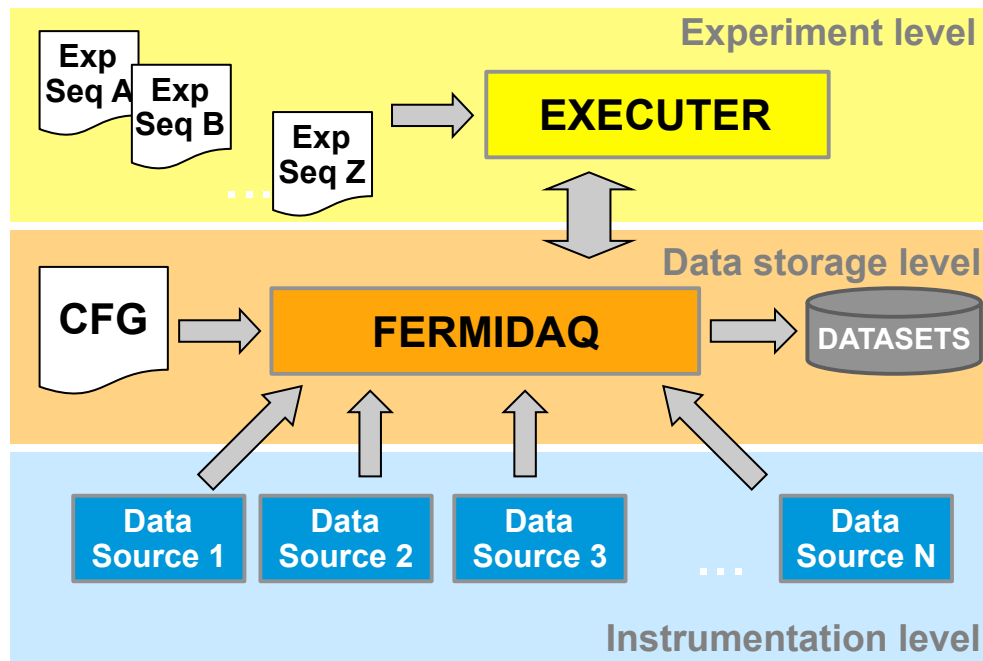
* Images from:
V. Lyamayev et al., "A modular end-station for atomic, molecular, and cluster science at the low density matter beamline of FERMI", J. Phys. B: At. Mol. Opt. Phys. Vol. 46 - 16 (2013)

Data Acquisition Requirements

- Data have to be **acquired and tagged** with the corresponding FEL pulse identification number (bunch number)
- Number and type of data sources continuously change, the acquisition framework **should be easily configurable**
- In order to meet the users' experimental requirements the framework **should allow for easy adaptation** and implementation of new experimental procedures and sequences
- Keep it **simple and reusable**
- Development based on TANGO



Data Acquisition System Overview



The development has been broken up into three logical levels:

- at **EXPERIMENT level** there is the script engine **EXECUTER**, capable of implementing different experimental sequences in the form of Python scripts
- at **DATA STORAGE level** there is a single centralized, configurable software device, named **FERMIDAQ**, that organizes and stores data coming from multiple sources
- At **INSTRUMENTATION level** there are multiple shot-by-shot data acquisition devices, capable of buffering and exporting data tagged with the bunchnumber

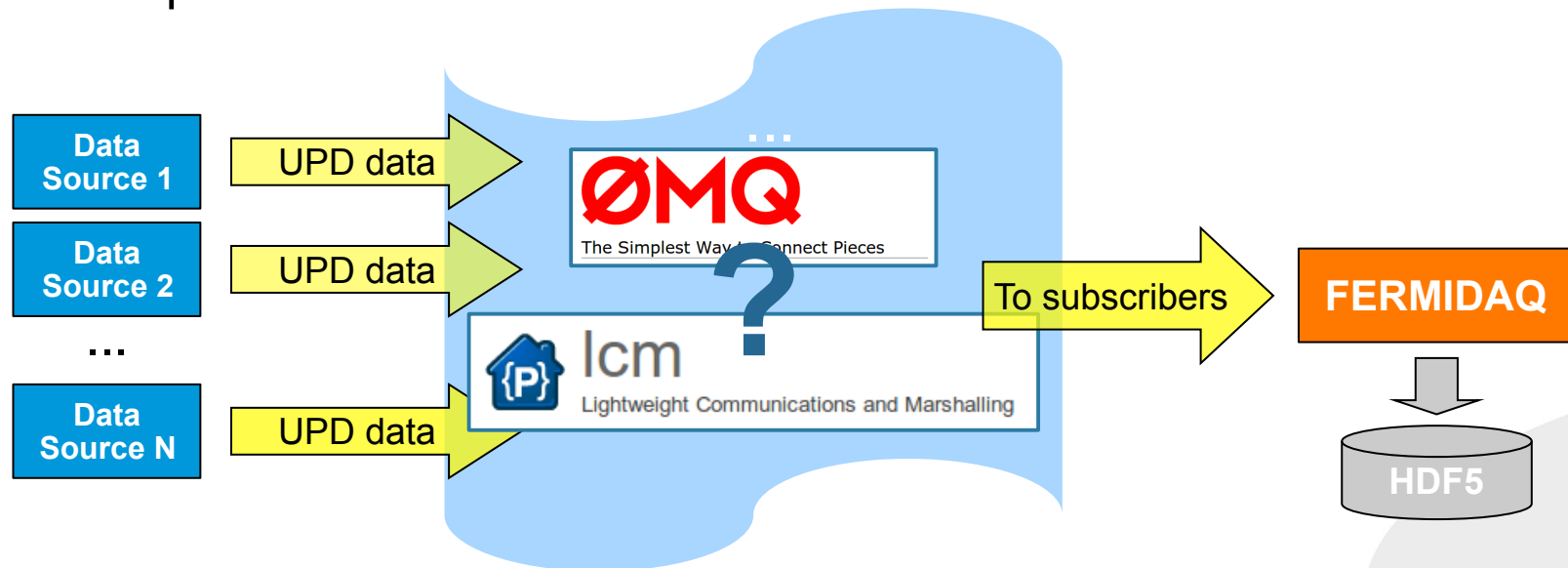
Ongoing developments

3 new end-stations under construction:

- . EIS-TIMER
- . MAGNEDYN
- . TERAFERMI

upgrade of the repetition rate of the facility to 50 Hz scheduled for 2015:

We need to minimize the communication overhead, by transforming the data acquisition devices to *active UDP sources*



Elettra Big Data Issues

- We have estimated an yearly raw data production capability of 1PB (PetaByte)
- In order to obtain this number we have considered bottom-up, analogical and historical grow rate approaches
- The annual growth rate of storage requirements of Elettra in the last 10 years has been 100% (i.e., we double our storage requirements every year).
- Currently the Elettra external network link capability is 1Gbit/s and the storage capacity is 0.3PB.
- Data sets are so big they cannot be transported by the user in a removable disk or transferred via network in a reasonable time. For example, the LDM 10MB detector acquiring at 50Hz is producing data at 500MB/s, i.e., 1.8TB per hour.
- With the current link capabilities it will take over 10 hours to transfer this dataset somewhere else.

Approaching our Big Data Issues

- Elettra is part of the Pan-Data initiative since the very beginning and has been one of the partners of Pan-Data-Europe and PanDataODI FP7 EC projects.



Elettra Scientific Data Policy

- Defined together with the experimental scientists in a 2 year-long process as a deliverable of PaNdata project; refined during the PaNdata-ODI project

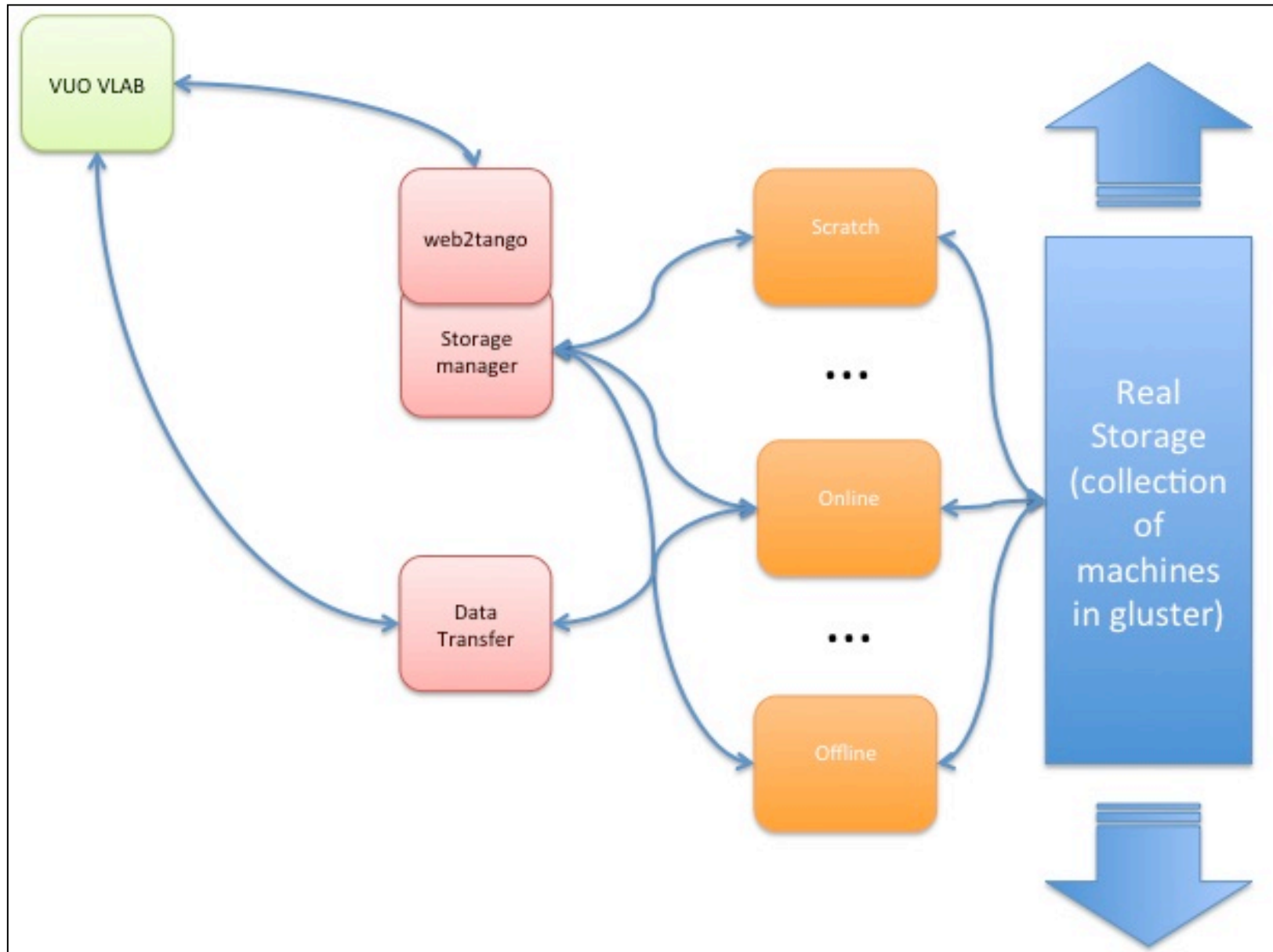
When	Access	Note
0 - 3 y	Restricted to PI and his/her team	The PI can set it "open access" whenever he/she wants
3 - 10	Restricted to PI and his/her team Unless the PI decides to let it open	The system reminds the PI to set the dataset "open access"
10 -	-	Life cycle complete

Elettra Scientific Storage 3 stage Logical Organization

DATA CATALOGUE	NO (Without metadata)		YES (With metadata)	
	STANDARD PANDATA STATION	1	2	3
AUTOMATIZED PANDATA STATION		→	1 ↔ 2	
STORAGE SYSTEM	SCRATCH Can be organized Can be processed	ONLINE Can be organized Can be processed Can be transferred	OFFLINE Can be transferred	



Elettra Scientific Storage Architecture



The VUO Virtual Laboratory

Virtual Laboratory *

[My investigations](#)

[All investigations](#)

[My tags](#)

[All tags](#)

[My tunnels](#)

[All tunnels](#)

[My applications](#)

[All applications](#)



VUO - Virtual Unified Office

VUO - Welcome to the Virtual Unified Office

Login

Username: [\[Login\]](#)
 Password:

Indicate as username your identification code (USER ID) or **your e-mail** and the password (for Sincrotrone Trieste users it is valid also the password used for the e-mail system [Marconi](#)).

Umbrella System: [login](#)

Lost password

If you are already registered but **you don't remember your identification code (USER ID) and/or password** please don't try to register again but click [here](#) to retrieve the lost information via e-mail.

Registration

If you are a [new user select this link](#) and go on with your registration. You will receive as soon as possible an identification code and a password.

Visits to the Elettra laboratory

If you are planning a visit to our laboratory just click [here](#) and fill the form. You will be contacted by our visitors office.

Se intendi pianificare una visita al nostro laboratorio seleziona [questo link](#) ed inserisci i dati della tua richiesta. Sarai contattato dal nostro "Ufficio visite" quanto prima.

Our visits [statistics](#)

Strategic committee agenda *

Show here the [year planning](#) of the Strategic Committee (*Restricted access*)

Resource booking *

Show here a [calendar](#) of the usage of the meeting rooms of the Elettra site. To book an event you must login in the VUO using username and password as indicated in the «Login» section.

Calendar

For details on Beamtime Allocation Calendar have a look to [Elettra](#) or [FERMI](#) Calendars.

Seminars

Forthcoming [seminars@Elettra](#)
 Forthcoming [seminars@TASC](#)

Elettra Library

Search [here](#) for Books & CDs in the Elettra Library.

Publication Search & Submittal

Please note that all publications resulting from measurement runs or research done at Elettra must be entered into the Elettra Publication Database.

Authors are invited to complete the [Elettra Publication Search and Submittal Form](#) online for each contribution, i.e., journal article, conference presentation, book or book chapter, thesis, contributed news articles, internal staff reports, general Elettra publications, brochures, etc. Only published contributions should be submitted using this form.



The VUO Virtual Laboratory

VUO - Investigations

Logged as: **Roberto PUGLIESE** (738) [[Sudo](#)] - [[Logout](#)]

[Create Help](#)

[Home/](#) [My investigations/](#) [All investigations/](#) [My tags/](#) [All tags/](#) [My tunnels/](#) [All tunnels/](#) [My applications/](#) [All applications/](#) [Unix users/](#)

Enter your query [Clear](#)

[\[Search\]](#)

Investigations				
sorted by "Tag" ascending then by "Code" ascending				
	▼ Tag ▲	▼ Code ▲	Principal Investigator ▲	▼ Proposal ▲
[Select]	DIPROI	140220_Test	Emanuele PEDERSOLI	
[Select]	DIPROI	140224Test	Emanuele PEDERSOLI	
[Select]	DIPROI	140225_Test	Emanuele PEDERSOLI	
[Select]	DIPROI	140226_Magnetism	Emanuele PEDERSOLI	
[Select]	DIPROI	140302_Mg	Emanuele PEDERSOLI	
[Select]	DIPROI	20129031	Iwao MATSUDA	[Proposal]
[Select]	DIPROI	First_test	Flavio CAPOTONDI	
[Select]	DIPROI	Stracazzi	Flavio CAPOTONDI	
[Select]	DIPROI	TEST_INVESTIGATION	Flavio CAPOTONDI	
[Select]	EIS-TIMEX	20124039	Manuel IZQUIERDO	[Proposal]
[Select]	EIS-TIMEX	20129002	Luca POLETTI	[Proposal]
[Select]	EIS-TIMEX	20129014	Emiliano PRINCIPI	[Proposal]
[Select]	EIS-TIMEX	20129024	Alexander FOHLISCH	[Proposal]
[Select]	EIS-TIMEX	20129030	Martin BEYE	[Proposal]
[Select]	EIS-TIMEX	20131208	Riccardo MINCIGRUCCI	
[Select]	EIS-TIMEX	Manuel_test	Manuel IZQUIERDO	
[Select]	EIS-TIMEX	Manuel_test_2	Manuel IZQUIERDO	
[Select]	EIS-TIMEX	TEST_INVESTIGATION	Emiliano PRINCIPI	
[Select]	EIS-TIMEX	bkg_acquisition	Riccardo MINCIGRUCCI	
[Select]	EIS-TIMEX	inhouse_Mar2014	Emiliano PRINCIPI	

Page: 1/11

Items: 1-20/206

[\[Next\]](#)



The VUO Virtual Laboratory

VUO - Investigation

Logged as: **Roberto PUGLIESE** (738) [[Sudo](#)] - [[Logout](#)]

[Create Help](#)

[Home/](#) [My investigations/](#) [All investigations/](#) [My tags/](#) [All tags/](#) [My tunnels/](#) [All tunnels/](#) [My applications/](#) [All applications/](#) [Unix users/](#)

LDM

20124009

[He_10_bar_1s5p/](#) [He_10_bar_50-76/](#) [He_Meta_and_Fluo/](#) [He_VMI/](#) [TEST/](#) [VMI_g](#)

Investigation details	
Name:	20124009
Description:	20124009
Max 400 characters	
Principal investigator:	(14637) ZITNIK Matjaz [JSI -
Proposal:	20124009

[\[Edit\]](#)

Other Investigators	
	Name
[Delete]	AVALDI Lorenzo
[Delete]	BUCAR Klemen
[Delete]	CORENO Marcello
[Delete]	JOURNAL Loic
[Delete]	MARCHENKO Tatiana
[Delete]	MIHELIC Andrej
[Delete]	O KEEFFE Patrick
[Delete]	PIANCASTELLI Maria Novella
[Delete]	PLEKAN Oksana Kudelich
[Delete]	PRINCE Kevin Charles
[Delete]	RICHTER Robert
[Delete]	RUBENSSON Jan Erik
[Delete]	SODERSTROM Johan

[\[Add a new investigator\]](#)

Experiments	
	Code
[View]	He_10_bar_1s5p
[View]	He_10_bar_50-76
[View]	He_Meta_and_Fluo
[View]	He_VMI

VUO - Experiment

Logged as: **Roberto PUGLIESE** (738) [[Sudo](#)] - [[Logout](#)]

[Create Help](#)

[Home/](#) [My investigations/](#) [All investigations/](#) [My tags/](#) [All tags/](#) [My tunnels/](#) [All tunnels/](#) [My applications/](#) [All applications/](#) [Unix users/](#)

LDM

20124009

[He_10_bar_1s5p/](#) [He_10_bar_50-76/](#) [He_Meta_and_Fluo/](#) [He_VMI/](#) [TEST/](#) [VMI_and_Meta_HEn/](#) [Xe/](#) [test](#)

↓

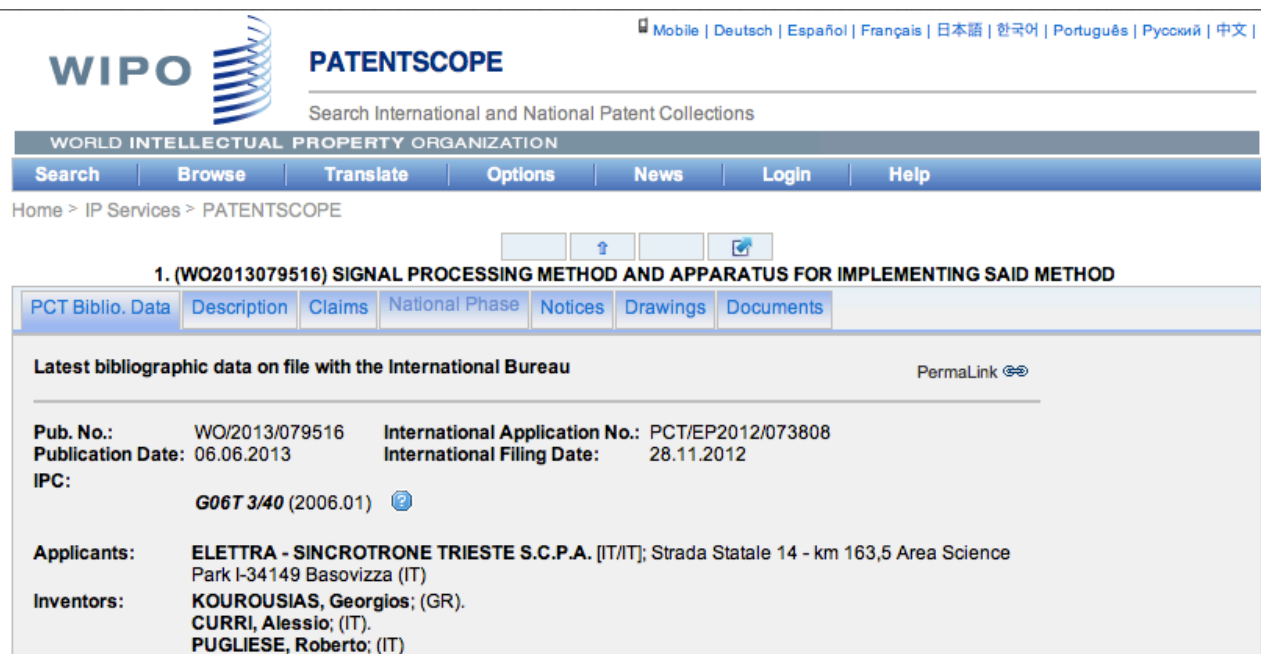
Experiment details	
Name:	He_10_bar_1s5p
Description:	He_10_bar_1s5p
Max 400 characters	

[\[Edit\]](#)

	Code	Datasets	Status
[View]	He_003		Filled
[View]	He_004		Filled
[View]	He_005		Filled
[View]	He_006		Filled
[View]	He_007		Filled
[View]	He_008		Filled
[View]	He_009		Filled
[View]	He_010		Filled
[View]	He_011		Filled
[View]	He_012		Filled
[View]	He_013		Filled
[View]	He_014		Filled
[View]	He_015		Filled
[View]	He_016		Filled
[View]	He_017		Filled
[View]	He_018		Filled
[View]	He_019		Filled
[View]	He_020		Filled

Reducing Big Data Costs

- In general we think it is really important to invest on the **study of the data format** as the computational cost is decreasing and even low compression rates can save a huge quantity of money
- A **first compression** can be obtained designing a suitable data format (HDF5) and a **second compression** can be applied on the data once it has been stored in a suitable data format
- We have estimated that we can save >50% of storage space using advanced compression algorithms



WIPO PATENTSCOPE
Search International and National Patent Collections

WORLD INTELLECTUAL PROPERTY ORGANIZATION

Search Browse Translate Options News Login Help

Home > IP Services > PATENTSCOPE

1. (WO2013079516) SIGNAL PROCESSING METHOD AND APPARATUS FOR IMPLEMENTING SAID METHOD

PCT Biblio. Data	Description	Claims	National Phase	Notices	Drawings	Documents
Latest bibliographic data on file with the International Bureau PermaLink						
Pub. No.:	WO/2013/079516	International Application No.:	PCT/EP2012/073808			
Publication Date:	06.06.2013	International Filing Date:	28.11.2012			
IPC:	G06T 3/40 (2006.01)					
Applicants:	ELETTRA - SINCROTRONE TRIESTE S.C.P.A. [IT/IT]; Strada Statale 14 - km 163,5 Area Science Park I-34149 Basovizza (IT)					
Inventors:	KOUROUSIAS, Georgios; (GR). CURRI, Alessio; (IT). PUGLIESE, Roberto; (IT)					

Data Gravity

- Data is something that continues to accumulate over time, and could be considered to become more dense, or have a greater mass. As density or mass accumulates, the data's gravitational pull increases.
- Services and applications have their own mass and; therefore, have their own gravity. **But data is much bigger and denser** than the two.
- And therefore, as data continues to build mass, **services and applications are more likely to be drawn to the data**, rather than vice versa ... (i.e., the data processing/analysis should data place where the data are stored and not on laptops)
(source Techopedia)

Project DIESEL

- Elettra has just launched the internal project DIESEL (Distributed Infrastructure for Effective Storage of data produced by Elettra & related Laboratories) with the sole purpose of managing the scientific data produced by Elettra, FERMI@Elettra, CERIC-ERIC, NFFFA, ...
- The project will increase the storage capacity of Elettra up 1PB to store scratch and online data
- Offline data will be moved to CINECA where Elettra will rent storage space thus leveraging the national storage infrastructure
- The project will also increase the transfer capacity of the network link between Elettra and CINECA and the computing power of Elettra VLAB (see data gravity)

Towards a sustainable solution to Elettra Big Data Issues

- The future is ... H2020 projects (PanDaaS, ...)
- First, we should be aware of the added value of storing big scientific ... **data micro-economics?**
- Once we are certain we save only data that has to be saved we must decide how to (who will) pay for the costs.
- Think of the whole value chain of scientific data production:
 - we should not design the next generation of super-detectors without examining how to reduce, process and store the data deluge generated by them.
- Should we introduce a Data Readiness Level (DRL) of a Research Infrastructure similar to the TRL?



Elettra
Sincrotrone
Trieste

Thanks!

Contacts:
Roberto.Pugliese@elettra.eu

www.elettra.eu