

# **WISE WG**

## **Security in Big and Open Data**

Ralph Niederberger  
Jülich Supercomputing Center (FZJ)  
[r.niederberger@fz-juelich.de](mailto:r.niederberger@fz-juelich.de)

Wise Workshop, Amsterdam, The Netherlands  
March. 27th 2017

# Setting the scene

Nowadays, 'Big data' and 'Open data' are often-heard buzzwords.

Want to make your work interesting -> Work on BIG Data

Want to get financing -> Produce OPEN Data

But open data does not mean you don't need to take care of it.

There are issues we have to take into account, like

- » confidentiality,
- » integrity, and
- » availability

# Some definitions

- Big data refers to large datasets.
- Public or available restricted only by a number of people, an organisation or a community.
- Open data refers to data available to everyone, republishable without restrictions. Those may be not always large or “big”.
- There are big datasets which have to be
  - accessible worldwide by distinct people or working groups only.
  - replicable for security reasons (damage) or
  - accessible with high-speed at different sites to spread download capacity ...
- A clear example of overlap between big and open data are large datasets from scientific research sources.

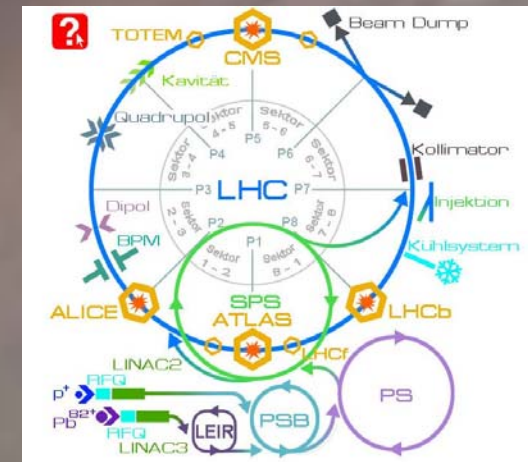


## Some examples (1): LHC

Output from Large Hadron Collider:

Data volume from all experiments:  
150 Mio. sensors provide data 40 Mio. times / sec.

- ATLAS: 320 Megabyte per second
- CMS: 220 Megabyte per second
- LHCb: 50 Megabyte per second
- ALICE: 100 Megabyte per second



That's about 15 PetaByte/year accessible to thousands of physicists around the world.



# Some examples (2):



Output from Square Kilometre Array (SKA):

Simulating a huge radio telescope (> 1000 antennas)

Around 1 square kilometer area

10.000 times faster than current telescopes

Assumed daily data volume: 960 Peta Byte

Analysis of data worldwide on HPC systems

Accessibility of data to huge community worldwide.





# Some examples (3):



Data handling: EUDAT project

Collaborative Pan-European infrastructure which provides research data services, training and consultancy



## Services:

- B2DROP
- B2SHARE
- B2SAFE
- B2STAGE
- B2FIND

- Sync and Exchange Research Data
- Store and Share Research Data
- Replicate Research Data Safely
- Get Data to Computation
- Find Research Data



EU project, but when in production similar problems and risks arise with data accessible to huge communities worldwide.



# Some examples (4):



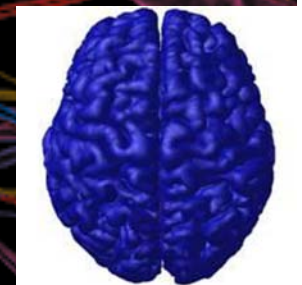
The HBP is a EC Future and Emerging Technologies Flagship.

- aims to put in place a cutting-edge,
- ICT-based scientific RI for brain research,
- cognitive neuroscience and brain-inspired computing.

It promotes collaboration across the globe, and is committed to driving forward European industry.

Similar projects are ongoing worldwide.

Huge amounts of data for simulating the brain and real brain images are anticipated to be stored, cataloged, and analyzed. Assuming, data are anonymized, usable as open data too, at least for all medical scientists.



# The need of data reduction

Transmitting data from storage to computation, which is not handled at all, does not make sense.

Data reduction on storage side should be done.

But sometimes, administration of data and competence/knowledge of data structure differ.

Data archives contain huge amounts of data for different communities. How can this be handled.

Applications may need access to the whole database (root privileges), but how can system administrators trust software developers from communities.

Those use cases, SBOD is looking for.

We will ask: Is there a solution available, already? Can it be generalized?



# The scope of SBOD-WG

- SBOD-WG focuses on security issues that arise when dealing with big and open data especially within the e-infrastructures.
- Security issues in this context concentrate on (as stated above):
  - Confidentiality regulates access to the information,
  - Integrity assures that the information is trustworthy, i.e. has not been changed without authorisation
  - Availability guarantees access to the information by authorised people at any time.
- SBOD intends to focus on high level security issues only.
- CSIRT issues are out of scope.

# How do we work

- Emails are the means of communication of the working group.
- SBOD will mainly meet via teleconferences, but if needed face-to-face meetings will also be considered and organised.
- GET INVOLVED  
<https://wise-community.org/security-in-big-and-open-data/>
- Subscribe to our mailing lists  
[sbod-wg@lists.wise-community.org](mailto:sbod-wg@lists.wise-community.org)



# What has been done so far...

Published on the SBOD Wiki

- Case Statement
- Definition of Big and Open Data

We are currently working on identifying possible use cases.

If you have any, get in touch with the chair(s):

[r.niederberger@fz-juelich.de](mailto:r.niederberger@fz-juelich.de)

n.n.

# Questions

