Security in Big and Open Data - WG

Alessandra Scicchitano

GÉANT - Project Development Officer

Ralph Niederberger

Jülich Supercomputing Center (FZJ)

SBOD-WG

- 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 your data

- 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 org or a community.

- Open data refers to data that is available to everyone and can be republished 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 of BIG and OPEN Data issues (1)

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 Mio Gigabyte / year accessible to thousands of physicists around the world.

Some examples of BIG and OPEN Data issues (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

Accessible to huge community worldwide.

Some examples of BIG and OPEN Data issues (3)

Data handling: EUDAT project

the collaborative Pan-European infrastructure
providing research data services, training and consultancy

Services:

B2DROP	–	Sync and Exchange Research Data
B2SHARE	-	Store and Share Research Data
B2SAFE	-	Replicate Research Data Safely
B2STAGE	-	Get Data to Computation
B2FIND	-	Find Research Data

Currently an EU project, but when in production similar problems and risks arise.

Again, data should be accessible to huge communities worldwide.

Some examples of BIG and OPEN Data issues (4)

Data handling: The Human Brain Flagship project

The Human Brain Project (HBP) is a European Commission Future and Emerging Technologies Flagship. The HBP aims to put in place a cutting-edge, ICT-based scientific Research Infrastructure for brain research, cognitive neuroscience and brain-inspired computing. The Project promotes collaboration across the globe, and is committed to driving forward European industry.

Similar activies are ongoing at least in the US, also.

Here again huge amounts of data for simulating the brain are produced and analyzed. Also images of the brain are stored and analyzed or cataloged. Assuming that the data are anonymized, those data can be used as open data too at least for all medical scientists

Some examples of BIG and OPEN Data issues (5)

Data handling: the problem of data reduction

It does not make sense to transmit data from storage to computation, which is not handled at all.
Data reduction on storage side should be done.
But sometimes, administration of data and competence/knowledge of data structure differ.

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

Questions:
Is this something SBOD should cope with? What about transfer protocols? Should we cope on those?
Generally: Is there an issue or has anything already be solved, but not everyone knows about?
Could this be an issue for SBOD? Making solutions available/ known?

The scope

- The 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 we work

Emails are the means of communication of the working group.

SBOD mainly meets via teleconferences, but if needed face-to-face meetings will also be considered and organised.

GET INVOLVED

Subscribe to our mailing list

So far…

Published on the SBOD Wiki

- Case Statement
- Definition of Big and Open Data

The WG is working right now on identifying possible use cases.
If you have any, get in touch with the Chairs:

alessandra.scicchitano@geant.org
r.niederberger@fz-juelich.de