WISE SBOD-WG

Definition of Big and Open Data

In order to reduce the scope of the white paper on Security issues that SBOD wants to produce, it is necessary as first step to provide a clear definition of what the WG considers Big and Open data.

First of all the data SBOD will focus on is Research data. Even though it would be interesting and beneficial to discuss the issues arising when dealing with data in general, the WG feels that it would be too broad and out of scope. Research data is the data that the e-infrastructures and projects mainly involved in WISE deal with in their daily job.

From now on when referring to data in all documents of the WG it will be implicitly assumed that it is research data.

Big Data vs Open data

In general Big Data can be defined as petabytes and exabytes of data, being generated daily in large experiments as starting from medical scans by magnetic resonance tomographs, weather and climate databases, up to data sets generated and distributed by the Large Hadron Collider at CERN[1] and the Square Kilometre Array (SKA) project[2] starting up in Australia and South Africa.

When we talk about security with this Big Data, we do not concentrate on the generation and first time storage. This is mostly done with specialized programs and protocols structuring and compressing data, generating meta data catalogs describing the data itself, its structure, producer, and access privileges for future access.

Most of the issues to look at appear when the user comes into account. Who is allowed to access the data? Which protocols, communications as well as authentication and authorization protocols, are used?

Some security, like integrity, also comes into account when transferring the data across infrastructures, even if no user interaction is involved.

We know there is a definition of Big Data that uses 5Vs but for our purposes we find the original definition more appropriate.

The Gartner analyst Doug Laney defines Big Data with 3Vs:

- Volume - Refers to the quantity of generated and stored data. A vast amounts of data that is generated every second.
- Variety - Different forms of data are collected and stored and in different formats, such as structured and unstructured.

[1] http://home.cern/topics/large-hadron-collider
[2] https://www.skatelescope.org/

- Velocity - Refers to the speed at which the data is generated and processed.

These three main characteristics of Big Data are the aspects SBOD-WG is mainly interested in.

As the name already says, Open data is available to everybody. Normally when data is first produced, it is private data. Private not in the sense of data containing personal information, but data, the producing user is the only one which is allowed to access. In a second step the owner of the data might decide to let her data be private, restricted or open. Think about a user in a multi user operating system. The user may allow access to other users of the same system, i.e. the group she is belonging to in a Linux system. She may also allow access to all users of the system. This will go in the direction of restricted data.
So anything that is fully shared and doesn't have any type of restriction can be considered Open data.

What will the WG focus on?

The main focus will be on Big and Open Research data. In particular for big research data (true also for open data) a special focus will be kept on handling distributed data and the exchange of data among e-infrastructures.
We consider medical data as a special case with too many controversial aspects. As long as medical data is anonymized, it is in our focus. At this point "personalized" medical data is considered out of scope.
Regarding data transfer, very often big volumes implicate to make as less transfers as possible: a user from one infrastructure might need to execute a program on another one where data is stored. What the user gets back is the result of that program.
But when talking about big data, data transfer is still not so easy. The issue resides in where the process of extracting data should be located. Usually e-infrastructures don't deal with this: "We provide the storage but are not responsible for extracting data". Often though a user is interested in small parts of the data only and needs procedures to extract it. How can this be done in a secure way? Installing procedures on the remote systems, where the data is located, should be allowed to local admins only. But those don't know about internals of the data (structure, privacy, …) Both approaches are mutually exclusive. How to proceed?
It is important to have a common language for data localization and access rights descriptions - (should be based on some metadata), so that exchange of data between e-infrastructures can be implemented.The SBOD-WG doesn't feel that it is possible to fully answer this question at this stage. The WG believes there is the need of analyzing possible use-cases in order to clarify the scope. After this, available reliable and safe solutions can be proposed or missing solutions be highlighted.