CHAPTER 3

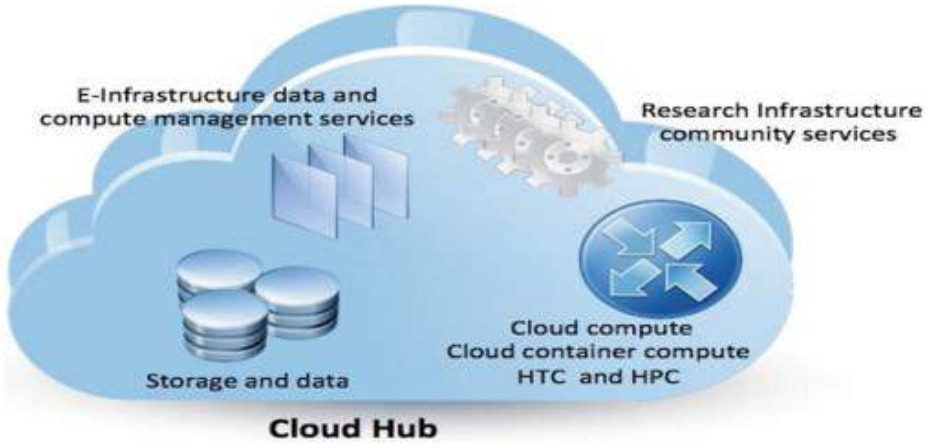
INFRASTRUCTURE ON OPEN SCIENCE IN PUBLIC FUNDED RESEARCH

3.1. BACKGROUND

Open Science can be implemented through online collaborative platforms that will link geographically-dispersed researchers or other stakeholders or interested parties. These platforms will enable them to cooperate seamlessly on their research, sharing research objects as well and ideas and experiences. Generally, in the form of online services, collaborative platforms provide a virtual environment that concurrently links to multiple people and allows them to work on the same task. These can range from extensive virtual research environments (VREs) that facilitate sharing and collaboration via web forums and wikis and, collaborative document hosting. The infrastructure of the European Open Science Cloud (EOSC) can be seen in Box 3.1. The infrastructure comprises research data, processing services, virtual laboratories and tools, that relies on a federated system of data and storage facilities.

In Malaysia, it is common for higher learning institutes to own a platform to deposit and share scholarly publications such as journal papers and students' theses. Based on the Landscape on Open Science in Malaysia, there are a few institutions that has advanced to develop their own research data repository such as Universiti Putra Malaysia (UPM) who has established an in-built research data repository, which at the moment is still at a project-based level, but will transition to the institutional level in the near future. Universiti Teknologi Malaysia (UTM) has set up the prototype and has identified potential users for a pilot test. Universiti Malaya (UM) also developed a prototype for a research data repository.

BOX 3.1. INFRASTRUCTURE ON OPEN SCIENCE



Source: ISSI Scientific Report Series Vol.15

[9] Virtual research environments have been defined as "innovative, dynamic, and ubiquitous research supporting environments where scattered scientists can seamlessly access data, software, and processing resources managed by diverse systems in separate administration domains through their browser" (Candela, Castelli and Pagano, 2013).

3.2. GUIDELINES FOR DEVELOPMENT OF INSTITUTIONAL REPOSITORY

When it comes to open data sharing through Open Science in research, researchers fear that their data will be misused for unethical purposes or they might get scooped or misinterpreted. Another concern relates to commercial entities who make use of freely available data with no strings attached. Therefore, a **robust policy is a must to secure and establish a trustworthy platform**. It is important to identify if an institution has established its own data sharing policy. For example, the University of Malaya is developing the UM research data management policy and Universiti Putra Malaysia (UPM) has also drafted a data sharing policy and is now waiting for endorsement from the top management.

In addition, an Open Science platform must have **a process that controls and oversees data usage** (e.g. who uses the data, for what purpose the data is used, and approvals from relevant authorities to access the data), and must ensure that data sharing practices **meet the FAIR (Findable, Accessible, Interoperable, Reusable) principle** and all data contributors and data users must be clear on their **responsibilities** and understand **ethical rules** when using the platform. Hence, it is of importance that fundamental technical elements are embedded in the platform to secure the deposited data.

An ideal feature of a trusted data sharing platform is having a **good data request handling process** and that the re-used data must be properly cited and acknowledge by the data originator. The cost and infrastructure setup involved should be planned to ensure **sufficient storage capacity** to deposit raw research data. The future expansion for the increasing capacity of stored data should be considered.

3.3. INFRASTRUCTURE DEVELOPMENT

3.3.1. Types of storage options or solutions

There are various storage solutions available but these solutions must be compared against two criteria:

a. The value of the data and its potential for reuse.

b. The types of components which give value to data, such as its discoverability, curation and whether the storage is reliable, large and sustainable.

Researchers incline towards using individual or project data storage (e.g. USB, hard drive on individual laptop, local drives etc.) as it is a simple, convenient and quick solution to store data. However, this option reduces the potential for data reuse, as well as the discoverability, reliability and sustainability of the data, thus reducing the value of the data itself. Institutions should encourage researchers to share their raw research data on institutional repositories to ensure that the stored data is reliable, well-curated and identifiable with appropriate metadata. The repository should be able to support the submission of raw research data at any stages of the research cycle, either the initial data, working data or final data stages. Researchers should plan at the start of a project how they will store data, and to outline its budget. Such planning must be documented in a data management plan, as described in Section 5.4.1 of Chapter 5.

[10] https://www.ands.org.au/guides/data-storage

3.3.2. Interaction between storage solutions and with metadata stores

An example of Malaysia Open Science Platform (MOSP) as indicated in **Figure 3.1**. enables researchers to easily store, discover, access and share their data for better research impacts. Storage of raw research data are held in institutional repositories and are made discoverable to users using the MOSP Portal. The MOSP portal is the central getaway to Malaysia's research data. The MOSP Portal forms a registry or a catalogue that harvests metadata from multiple institutional, agency-based and domain-specific repositories. The harvested metadata are archived and stored in the MOSP Central Portal. Using interoperable metadata standards and Application Programming Interface (API), the MOSP Portal can be integrated with multiple types of repositories. This includes institutional repositories and domain-specific repositories. The portal will be an extendable and flexible platform for data sharing. Integration of the MOSP Portal with existing repository systems must be complemented with applicable risk management mechanisms to identify potential risks associated with integration and interoperability issues during implementation.

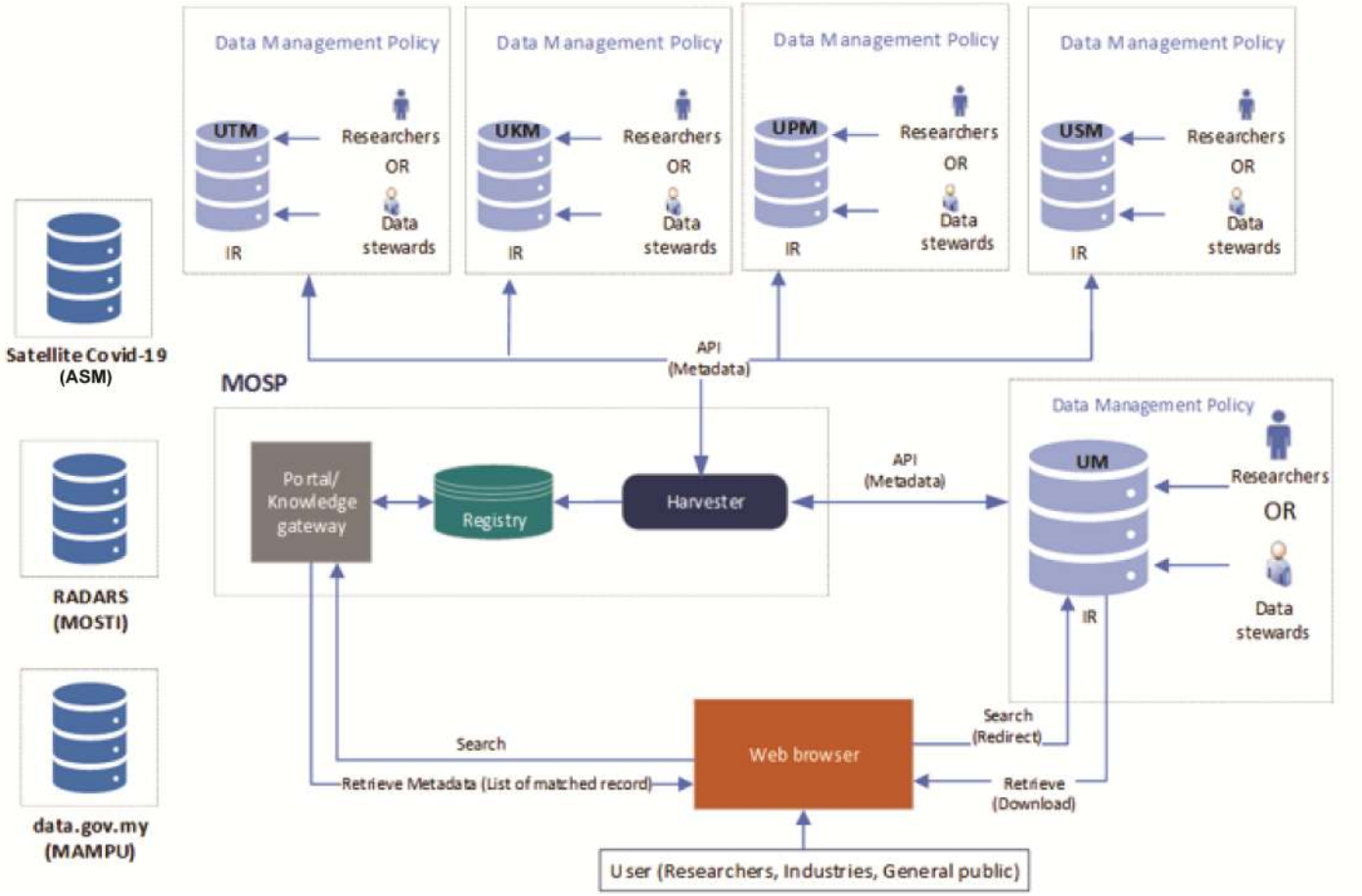


Figure 3.1: Architecture of the Malaysia Open Science Platform (MOSP) pilot project
Source: MOSP, 2020

3.3.3. Identifier[11]

A persistent identifier is any label that is used uniquely to guarantee that the deposited datasets can be managed and kept up to date over a defined time. An identifier will be assigned to each entry of research datasets by respective repositories. The IT team is responsible for keeping the system running while the data originator is responsible for providing up-to-date information about the entry that is being identified. The identifier serves as one of the metadata elements and will be captured by the MOSP Portal for each entry registered in the MOSP Portal.

3.3.4. Publishing and sharing sensitive data

Sensitive data is data that must be protected against unwanted disclosure. Access to sensitive data should be safeguarded. Protection of sensitive data may be required for legal or ethical reasons, for issues pertaining to personal privacy, or for proprietary considerations.

Malaysia has strong regulations regarding personal (for example, Personal Data Protection Act 2010) and non-personal data. Examples of sensitive data are:

1. Personal data - Name, photographs, Identification Card (IC), bank details, medical records, bank details.

2. Confidential data - Physical or mental health or condition of a data subject, his political opinions, his religious beliefs, interview transcripts containing identifiable individuals' sensitive personal data such as drug dependence, research data/information/IP with significant commercial value/obligations.

3. Biological data – endangered or threatened species whose survival depends upon protection of their habitat location.

When handling and dealing with sensitive data, it is important that careful measures must be undertaken when collecting, processing, handling, and storing data throughout the research process. As such, appropriate permits, and informed consent must be sought before initiating the research process. Anonymisation of personal data should be taken into account to ensure that these data are non-identifiable when being deposited. The sensitivity of datasets must be identified and appropriate ways of handling these data must be written in a Data Management Plan.

[11] https://www.ands.org.au/guides/persistent-identifiers-expert
[12] https://www.openaire.eu/sensitive-data-guide and https://www.ands.org.au/guides/sensitivedata