Research Data Architectures in Research Institutions IG

Healthy architectures for healthy data - sharing approaches for sensitive data architectures

14th RDA Plenary Meeting, Helsinki, 24 Oct 2019, 11:00 - 12:30

Agenda

Introduction

Research Data Management Life Cycle. Research data processes can be different, it is not always a circle. All phases shall be discussed in this IG, concentrating on technology and system architecture. Also uses cases of institutions.

Lightning Talks

Sensitive data architectures and services @ University of Helsinki (Ville Tenhunen)

Researchers have a need to capture, store, share and distribute sensitive data. Architectures are made by IT and user need is often neglected. GDPR protects personal data. For sensitive personal data there are more requirements. Sensitive data can be classified information by law, economical reasons, etc.

Data protection != Data Security
Fundamental right vs security based risk assessments

Sensitive data could be FAIR.

Architecture Framework shall be tested at University of Helsinki. http://strategia.helsinki.fi/en

Drivers:

- Changes in business environment
- Strategic goals
- Development areas
- Actions
- Abilities

AAI (Access, authorisation), logs, user tracking, encryption are important questions.

Some possible solutiox ns:

- Umpio, NAS solution
    - Encryption, no sharing
- Ceph + Nextcloud

Questions & Answers

Q: Do you plan to use commercial cloud services like Azure, AWS at University of Helsinki?
A: Cost is an issue

- C: Authorization & access. We register the person who has proffered the invitation. This person is responsible for the invitee's actions.
- Q: Have you considered how to integrate public cloud services into your infrastructure?
  - A: we have agreements w AWS, AZURE, etc. They are expensive. We are creating local services, they are much less expensive. Also, not all researchers have Azure ID and can access Azure, so it's not really an open environment.
  - C: many institutions are engaging with these integrations
- Q: Virtual organizations may be an option?

Infrastructure for Sensitive Data at UCL (James Wilson)

SLIMS deals with medical data and life sciences. Data storage and services are mostly developed in-house, for other services in the research data lifecycle external tools are used (e.g. DMP online, ePrints, figshare@UCL etc.)

Data Safe Haven: technical solution for storing, handling and analysing identifiable data. Environment where researchers can work safely, comes with some inconvenience because it is designed for security. More compute power is needed in the future, rapid provisioning, Metadata shall be standardised in the data repository.

Q: How are you dealing with backing up archives?
A: Backups are located in and around London.

Q: What is the concern with sharing data on dropbox?
A: Visibility and accidentally sharing is a risk.

Q: How do you handle with data from collaborators?

- Q: Replication?
  - A: have secure backup. Not great geographic security.
- Q: Have you considered Data Lakes?
  - A: Under consideration. We use this for business information for university. Safe Haven distinct from this.

- Q: Policies for retention & disposal
  - A: Not for safe data haven. But significant policies for other services. They are not necessarily complied with.
- Q: What about multi-party computing solutions?
  - A: not sure if these have been considered. Perhaps for associated hospitals?
- Q: Why do you object to Dropbox for sensitive data?
  - A: It's because of the governance and training and not having visibility to what is happening. It is very easy to accidentally share something. This lack of oversight at the uni level is the risk. Not so much the technology o f dropbox
- Q: How do you handle data from collaborators, from other countries.
  - A: There are practices for these, something with encryption, but not sure
- Q: Any pressure from agencies to give up walled garden approach, because of perceived benefits to linking with other data, e.g. geographical data.
  - A: Mainly security first. Not sure where this might go in future

Research Data Management in Munich (Stephan Hachinger)

rdmuc - WG: Bavarian state library, LMU, Leibnitz Supercomputing Center, UB, TUM
Topics: Metadata, rights management, electronic lab notebooks, PID, …
RDM services: new tools for better RDM, e.g. Datacite Metadata generator
LMU RDM services: metadata
TUM: archival of scientific output & research data, archiving + workbench (RDM tool that also contains provenance data)
LRZ: Make Big Data public (e.g. HPC output) - findable and accessible. Connect data silos to metadata catalog (OAI-PMH), use of DataCite (DOI) - make data citable
Sensitive data was kept in an isolated network of hospital servers. GDPR prohibits processing
Making the computing center processing sensitive data, two pilot biomedical research projects (bare metal setup):

- DigiMed Bayern
- Bavarian Genomes

Q: How to deal with sensitive data storing in a repository, when there are no protection services in place?
A: Encryption could be a first step (e.g. BitLocker). Inform people. Not use Dropbox, hard drives, USB sticks etc. Store Metadata separate from the Data so it is accessible. Jurisdiction and fines for data breaches might encourage development of services at the institution. Dataverse example: metadata on dataverse, actual data stored somewhere else with QoS agreements.

Governance Survey

Benchmarking Research Data Services at Research Institutions.
https://www.rd-alliance.org/group/research-data-architectures-research-institutions-ig/wiki/rdari-survey-distribution

Complete before Nov. 15th https://opinio.ucl.ac.uk/s?s=63105

Objective: Sharing case studies from different institutions

Intermediate results:
24 complete responses mainly from universities, but also other institutions like NGO

Q: BoF session from morning
A: driven by CODATA, benchmarking of RDM services at institutions

Seeking for WG topics

RIDA: Repository Interface for Data Analytics. Interface jungle with different access methods from repository to repository. How could a standard interface look like?

Comments

- The European Genome-phenome Archive (EGA) is a service for permanent archiving and sharing of all types of personally identifiable genetic and phenotypic data resulting from biomedical research projects. It is now being developed to a federated system. Presentation of this work at the session Access to Human Data Makes a Difference Friday before lunch.
-