BE INSPIRED
THE UNIVERSITY LIBRARY

UNIVERSITY OF LEEDS

Research Data - Systems integration

John Beaman

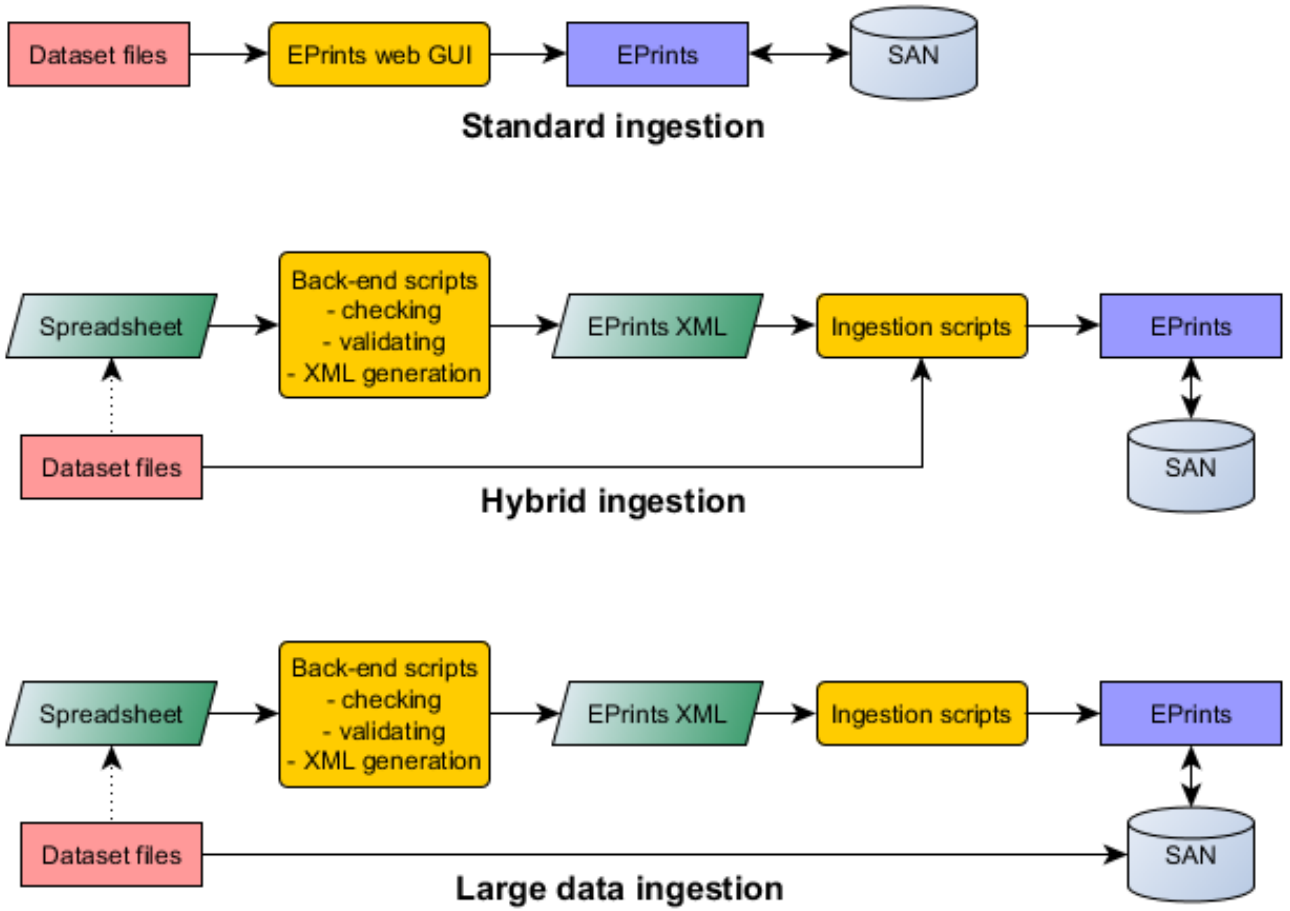
Some of my roles

- Library Systems Team member
- EPrints developer / maintainer / server administrator
- Primary repository administrator (DL, DUAL)
- Secondary repository administrator (WRRO, WREO)
- Research Data Leeds (RDL) Team
- RDL repository administrator
- RDL (and DL) data ingestion (huge area!)
- Data security, preparation, ingestion and preservation

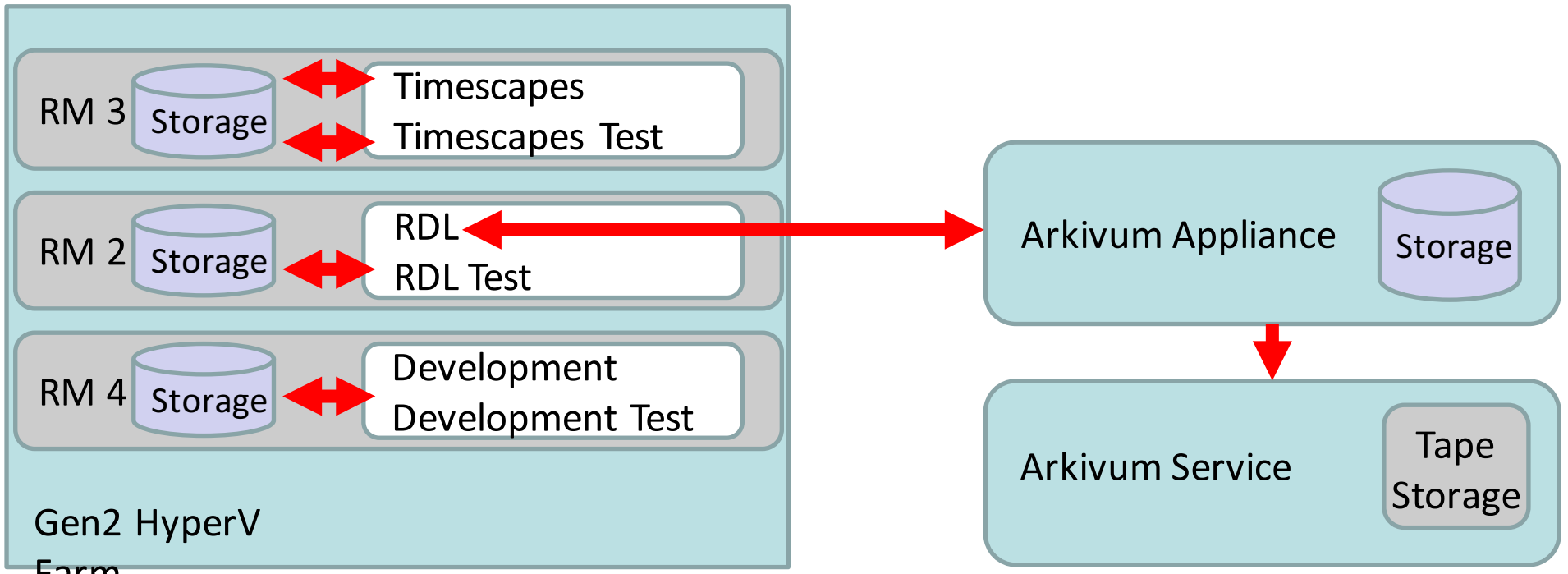
Research Data

- Huge variety of…
  - File types (e.g. docs, image, audio, video, equipment-generated)
  - File formats (for each type of file)
  - File sizes
  - Dataset structures, sizes, file naming considerations
  - Mixtures of classes of data / information
  - DPA, commercial, legal, ethical restrictions, embargos
- Presents many challenges
  - preparation, classification, ingestion, storage, management, preservation, access
- If it's going to happen, it probably will with research data!

Research Data – three types of ingestion



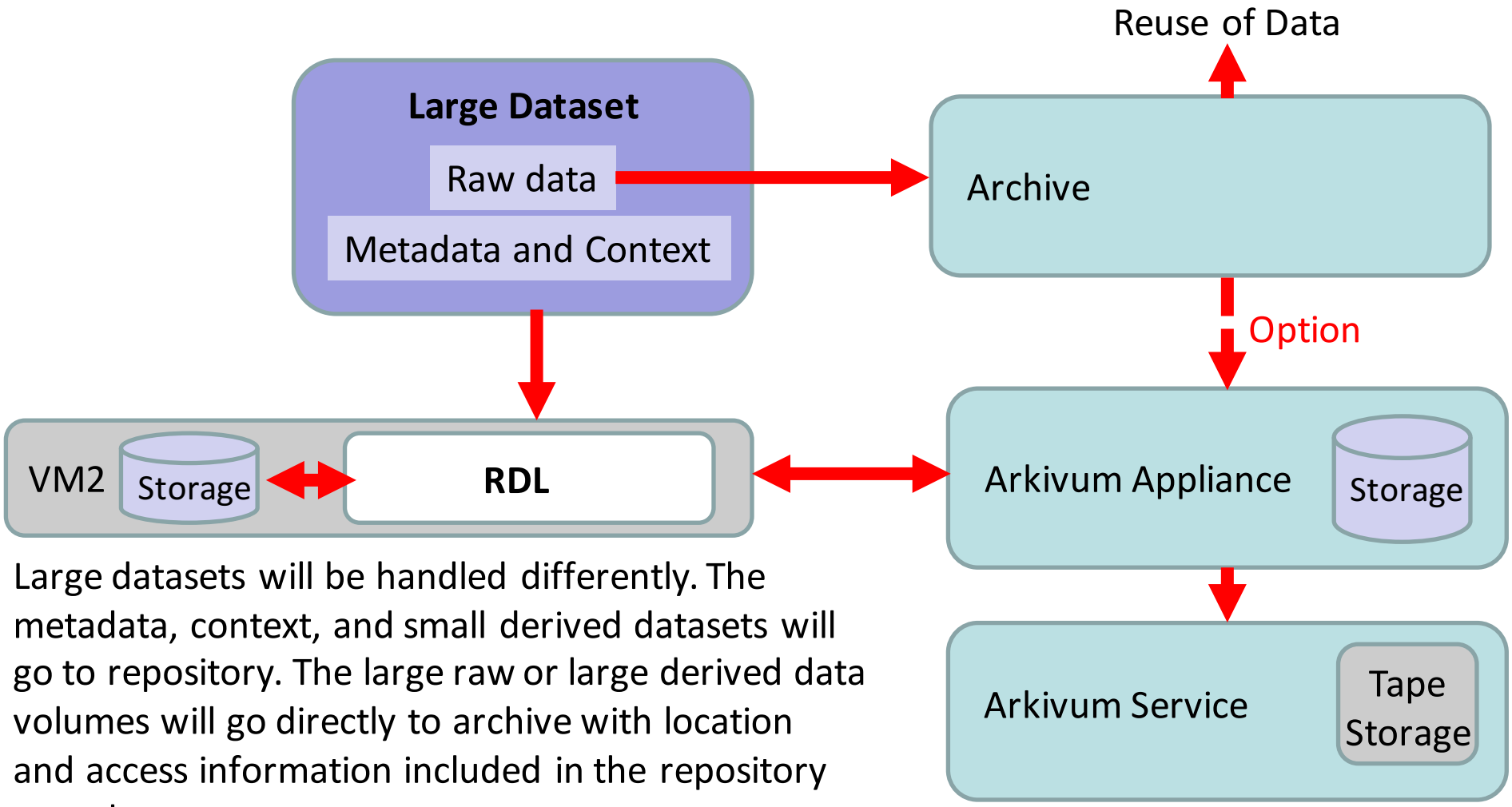
RDL – Phase 1



Each of the 3 Virtual Machines runs an instance of EPrints. An instance of EPrints can have one or more data repositories. The repositories each use local storage associated with the VM.

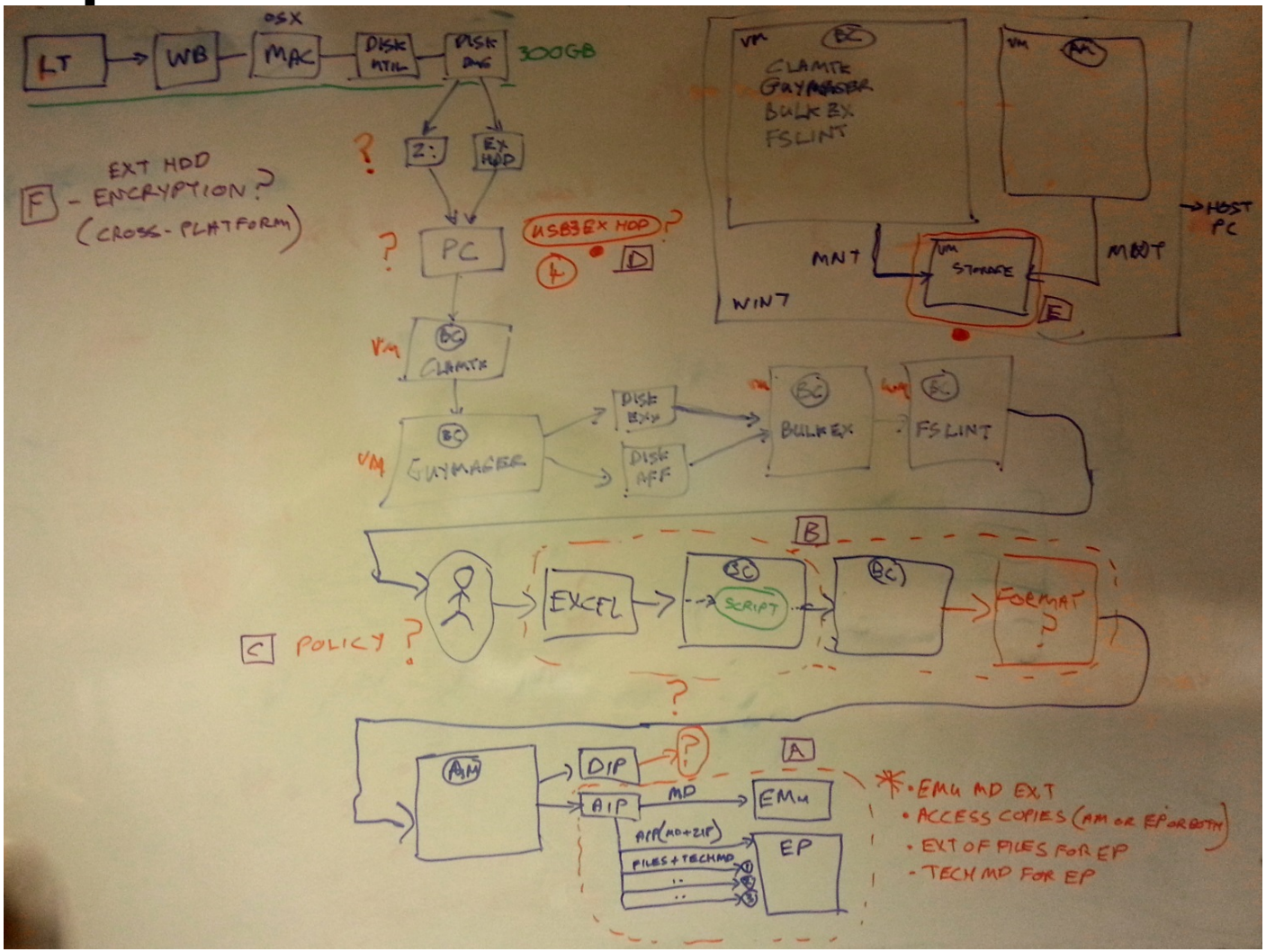
The RDL data repository uses storage on the Arkivum appliance which in turn transfers a copy of the data to the remote Arkivum Service

RDL – Phase 2 (Large Datasets)

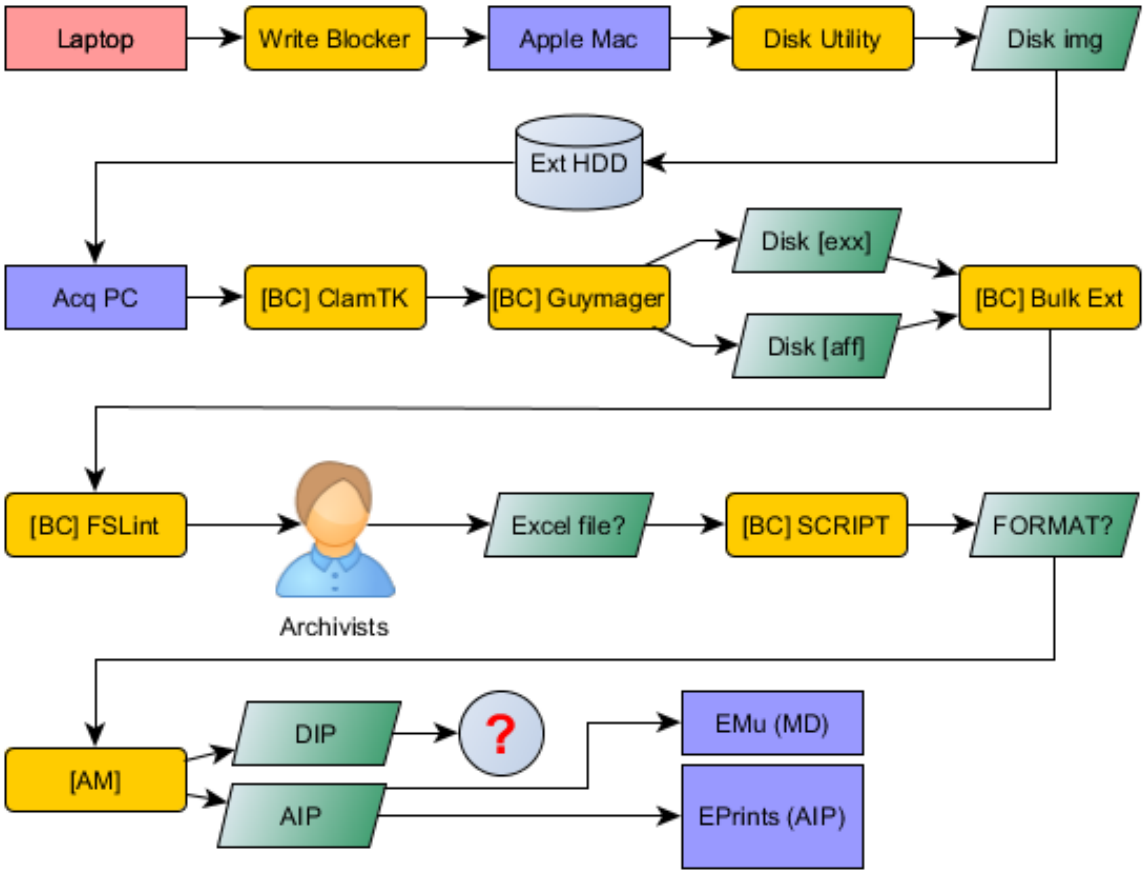


Large datasets will be handled differently. The metadata, context, and small derived datasets will go to repository. The large raw or large derived data volumes will go directly to archive with location and access information included in the repository metadata.

A sample workflow



A sample workflow



Some workflow questions

- Which processes could / should be integrated into an Archivematica-based workflow?
- How might this be done?
- What are the benefits?
- What are the challenges?

Some of the challenges

- Data preparation
  - file location, file selection, filenames, dataset structures (good researcher tools and utilities)
- Information security
  - Identification of potentially confidential data
  - Data classification / anonymisation / pseudonymisation
  - Data encryption (during processing / in final storage)
- Large datasets (processing, storage, delivery)
- Integration with other internal systems
  - Publications repository, CRIS, finance system etc.
- Integration with third-party services

Some options

- Identify common requirements, processes, issues, challenges, solutions
- Start within institution (remove silos)
- Spread the word, share the experience (e.g. RDMF)
- User groups (Archivematica, Symplectic etc.)
- A proposition…. Arkivum user group