Research Data Management Platform

A system approach to the management of research data

Donald Scobbie

Jim Galloway

RDMP

- University of Dundee Health Informatics Centre is a data supplier
  - Serves a diverse research community
  - 30 core clinical data sets (470K patients contributing data since 1952)
  - 20 research specific data sets, biobanks and longitudinal studies (10k to 20k patients)
  - 100 linked data releases annually
  - Cohort prospecting and feasibility

RDMP

- NHS clinical data sources are unconsented
- Data must be anonymised on a project basis to prevent cross-linkage
- Reproduce cohorts and extracts
- On-demand data refresh with new events
- Governance process is per project and integrated into data release and access
- Approved researcher status for access

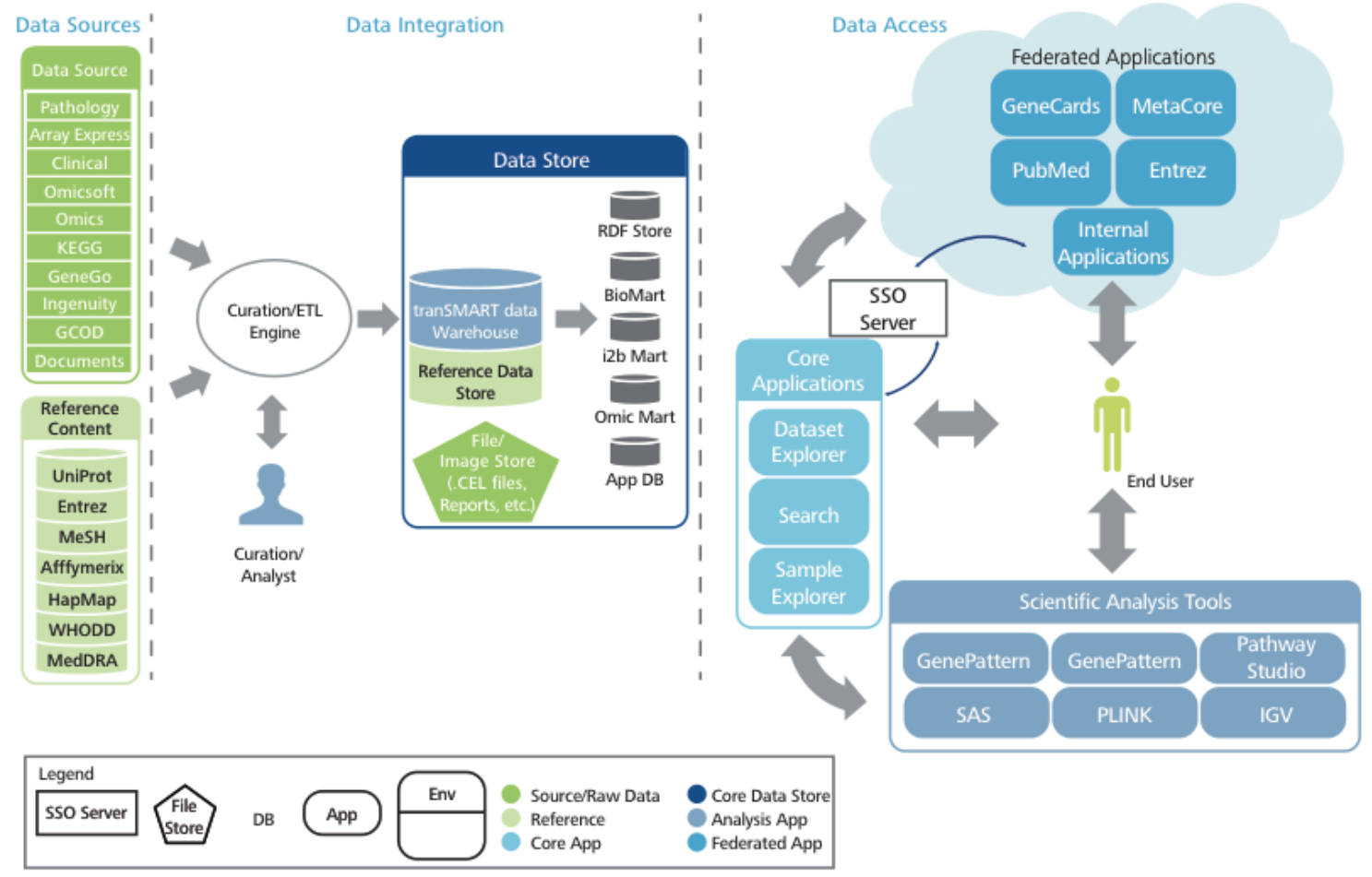
RDMP

- Biochemistry - 28 years, 800k patients and 180m events, updated weekly
- Prescribing – 14 years, 600k patients and 160m events, updated daily
- Acute Admissions – 30 years, updated monthly
- Genetic data studies for cohorts of 5k to 17k
- Image data studies for cohorts of 2k

DATA LINKAGE SERVICE (DLS)

health informatics centre services

tranSMART



http://www.recomdata.com/solutions/open-source/

RDMP

- Two related, but distinct activities
  - Data Repository Management (Data)
    - Preservation
    - Metadata generation (feature extraction)
    - Data profiling and quality control
    - Cohort discovery, data linkage and extraction
  - Research Data Management (Project)
    - Data quality assessment and control
    - Data transformation
    - Data analysis

hic health informatics centre services

Research Data Management

- Central thesis
  - *Research repository data management is best done through data preservation in original form and metadata generation*
  - Metadata allows dynamic processes for feature extraction and attribute discovery to be captured
  - Metadata facilitates integration of disparate data types, especially when the original artifact is of little concern but its derived attributes are of interest

RDMP

- Minimal set of *processes* required to support the system requirements for an IDR
  - *Load* (import selected data items into repository)
  - *Transform* (validate, clean, impute)
  - *Quality Management* (profiling and reporting)
  - *Summarise* (create discovery metadata)
  - *Extract* (link and merge)
  - *Export* (anonymise and format)

hic health informatics centre services

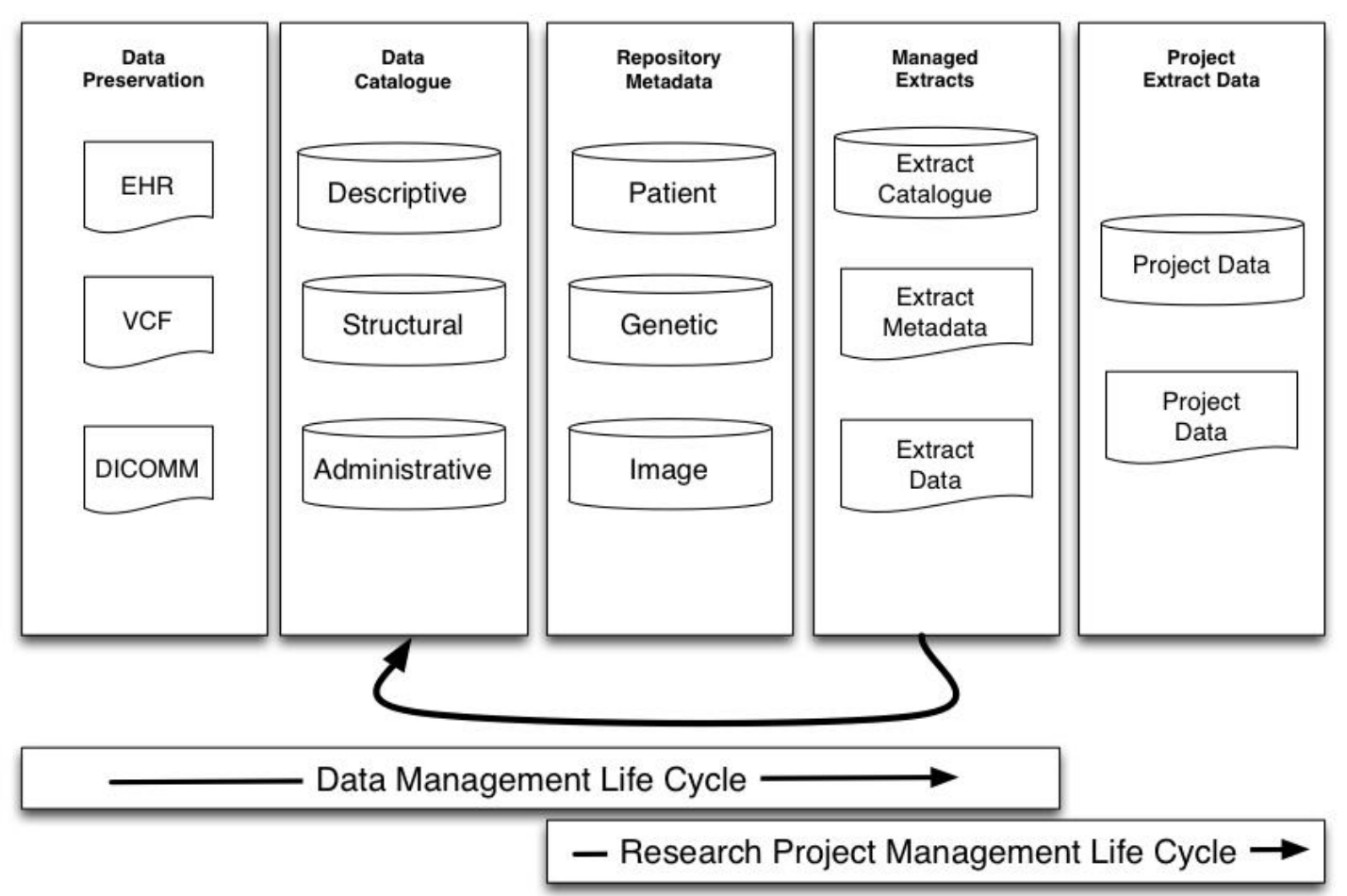
Research Data Management

- Data Catalogue is central in the architecture for process co-ordination
- Metadata process capture
  - Data validation and cleaning
  - Data transformation (code mapping)
  - Data summarisation (feature extraction)
  - Imputation
  - Data set assessment (baseline and stability)
- Accessibility, Reproducibility, Transparency

Data Catalogue

- Principle mechanism for integration
  - Descriptive
  - Structural
  - Administrative (dynamic process)
- RDMP catalogue a synthesis of elements
  - Dublin Core
  - DDI Alliance
  - World Bank

Research Data Management



hic health informatics centre services

Research Data Management

- Metadata integration
  - Semantically different data may remain so
  - Lazy evaluation and JIT transformation
  - Data may be selected and linked (integrated) *after* mapping processes have been applied
  - The metadata mapping processes may resolve semantic mismatches in a transparent way at a late stage in the data life cycle
  - Metadata allows multiple different views of a repository to exist simultaneously

hic health informatics centre services

Research Data Management

- Data should only be converted at the final stage of the research process when the required data is narrowly defined and well understood
- i2b2, OMOP and openEHR schema conversion are not preliminary stages in data preparation, they are final stage *output formats*