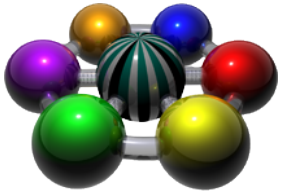
Data Services Integration Team

How To: KIT Data Manager supporting Metadata

Volker Hartmann, Thomas Jejkal

LSDMA

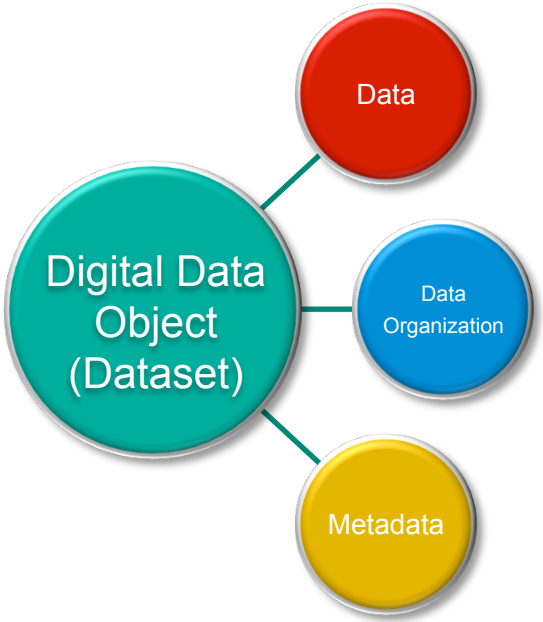
KIT Data Manager



Repository

Managed location/destination/directory/bucket where **digital data objects** are

- Registered
- Permanently stored
- Made accessible and retrievable
- Curated



Digital data object (dataset):

Consists of

- Data
- Description for re-use

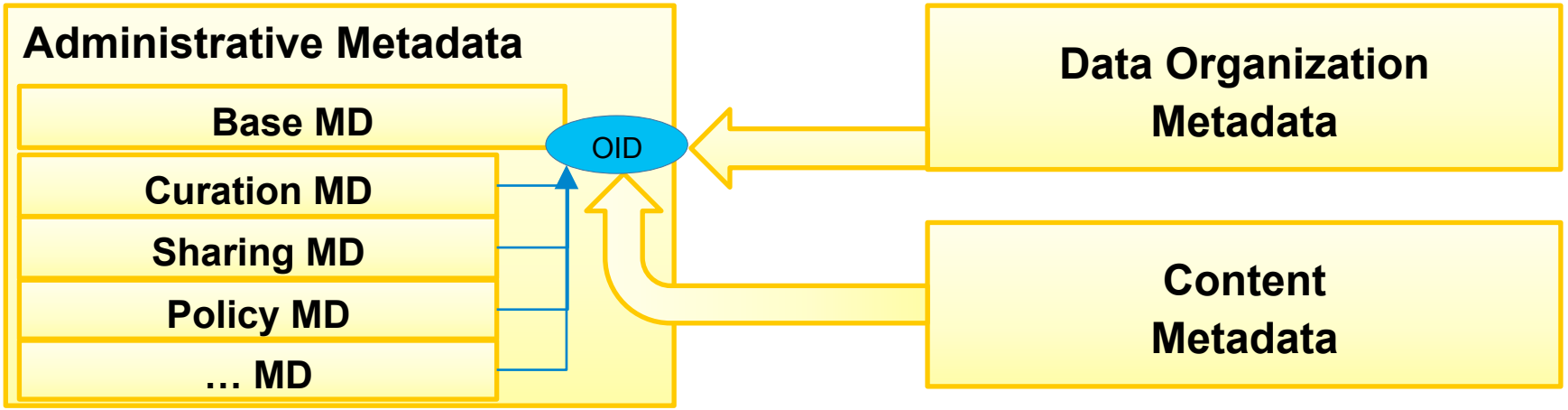
Scientific Metadata

- Many standards for scientific metadata
  - http://www.dcc.ac.uk/resources/metadata-standards
- Mostly very domain specific and hard to map
- No common standard for scientific metadata
  - Dublin Core not sufficient!

http://www.benjamintseng.com/wp-content/uploads/2011/08/standards.png

Scientific Metadata in KIT Data Manager

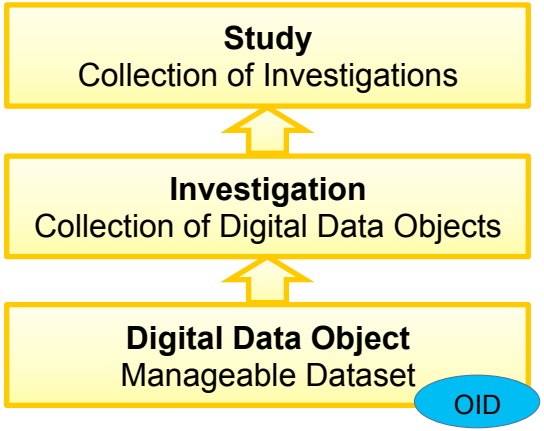
- Metadata model organized in three main categories:



- Allows to handle metadata according to their usage patterns
  - Different access patterns
  - Different user interfaces/access permissions
  - Different technologies

Administrative Metadata

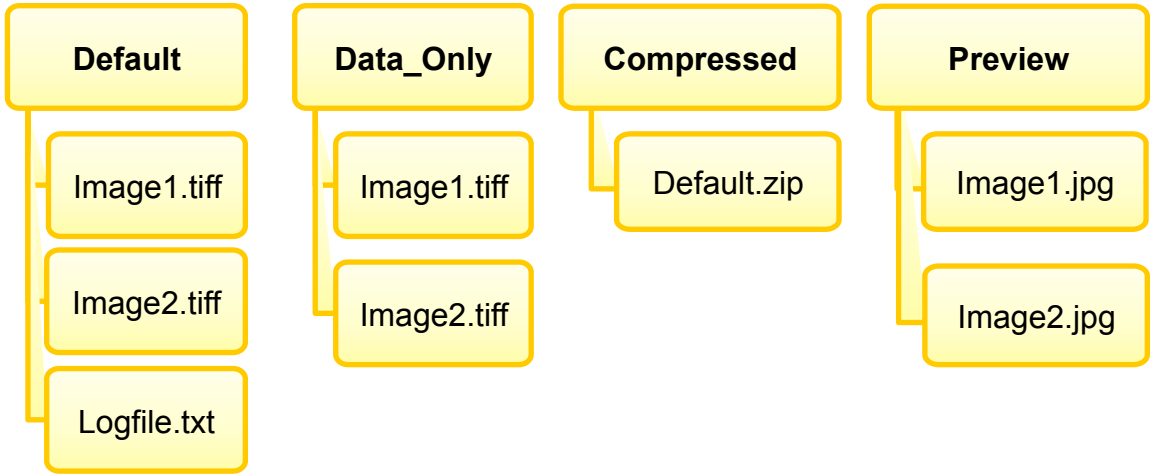
- Provides common set of metadata with well-known schema
  - Stored in RDBMs
- Mainly for internal/system use
- Contains Base Metadata
  - Defines basic metadata available for all datasets
  - Oriented towards Core Scientific Metadata Model
  - Might be provided by user or mapped from existing properties
- Other administrative metadata linked via OID to Base Metadata



Data Organization Metadata

- Contains information about dataset organization and location

- Multiple organizations possible for each dataset



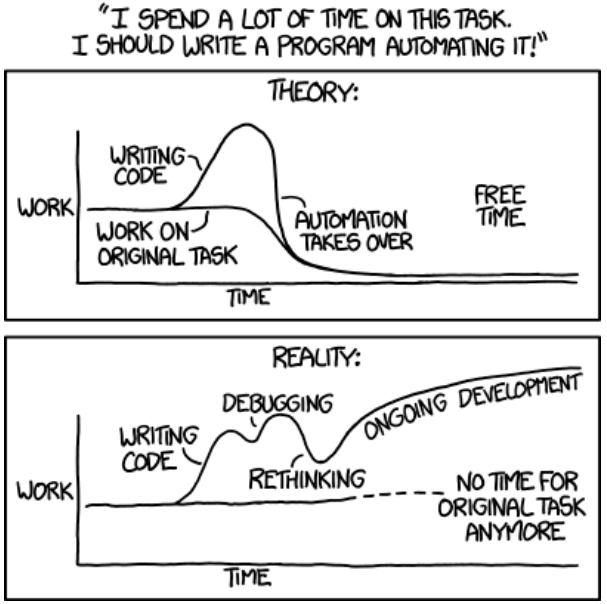
- Currently file based Data Organization supported

Content Metadata

- Depends on (sub-)community and use case
  - Some communities have a common standard (e.g. OME in Microscopy)
  - Some communities have many standards (e.g. DARIAH, Climatology)
  - Others have no established standards

- Retrieval can be difficult and hard/impossible to automate

An example…



http://imgs.xkcd.com/comics/automation.png

Content Metadata

FF D8 FF E0 ... What's that?

Automatically extracted

Filename: IMG_1405.jpg
Type: JPG
Date: 2013-02-14
Width: 509px
Height: 506px
Size: 107kB

Manually added

Persons: M. Hausmann, J. Hesser, N. Kepper, R. Stotzka, V. Hartmann, R. Grunzke
Location: IKP/Heidelberg
Note: F2F meeting DLCL Key

Use Case: BESS (1)

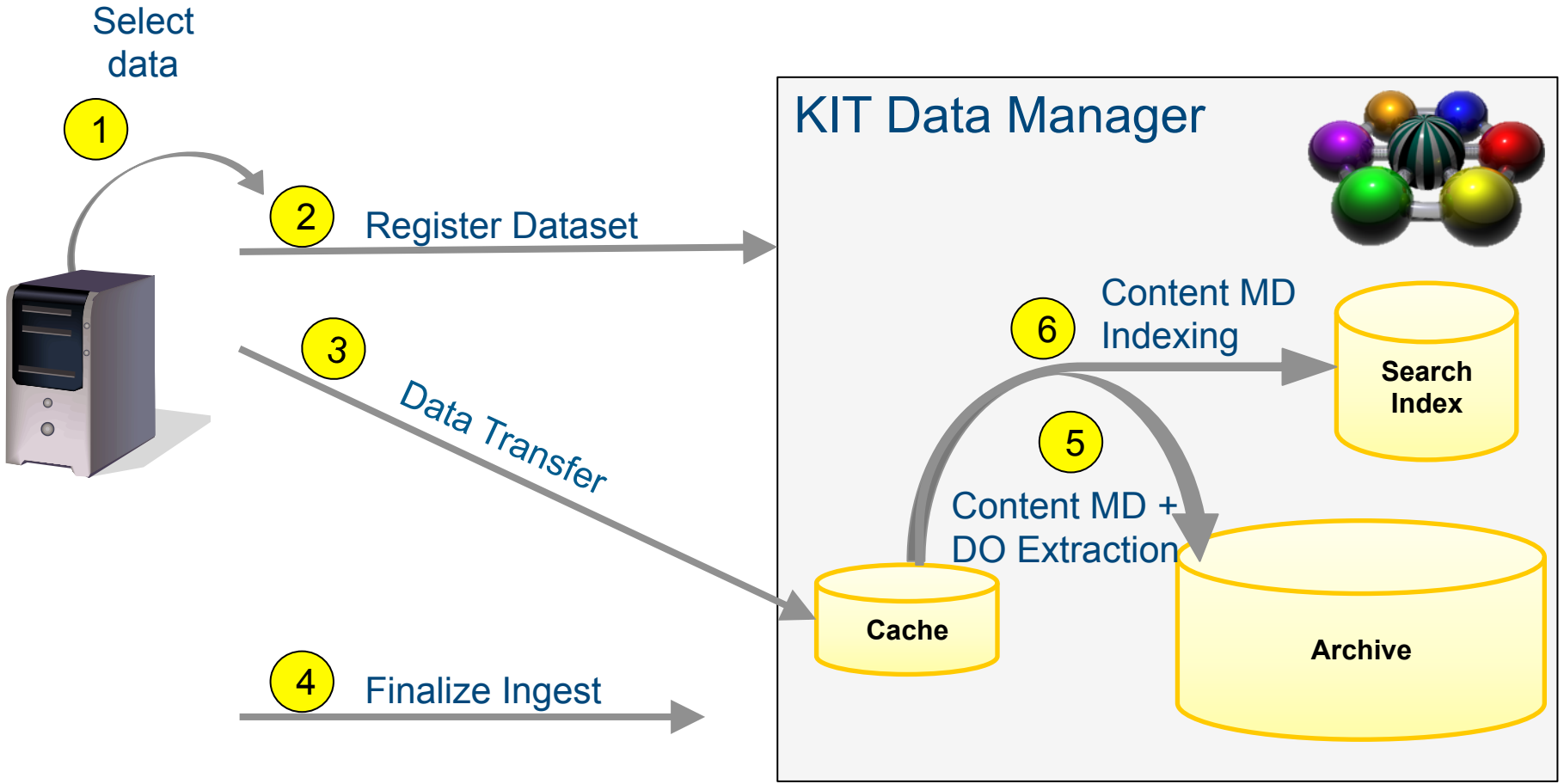
- Archiving of measurement database dumps in repository

- Base metadata provided by user
  - E.g. dataset name, start date, end date
  - Collected during dataset registration

- File-based Data Organization as only database dumps are stored

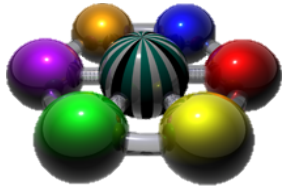
- Content metadata extracted from uploaded data
  - E.g. sensor names
  - Stored in XML format and published in elasticsearch index

Use Case: BESS (2)



Conclusions

- Metadata model based on three different categories of metadata
  - **Administrative Metadata** suitable for all communities
  - **Data Organization** supporting different representations
  - **Content metadata** for community-specific metadata



- Support for **automatic extraction** of metadata

- Model easily **extensible**

- Built-in support for transforming and publishing metadata
  - Harvesting via **OAI-PMH**
  - Access/Retrieval via **elasticsearch**

Backupfolien

Extract Metadata (community)

```
public class ExtractYourMetadata extends AbstractMetadataExtractor {

@Override
public String createMetadataDocument(TransferTaskContainer pContainer)
 throws TransferClientOPException {
    LOGGER.debug("createCommunitySpecificElement");
    String metadataAsXml = „<noMetadata />“;

    ICollectionNode root = pContainer.getFileTree().getRootNode();
    IFileNode metadataFile = (IFileNode) Util.getNodeByName(
                                        (ICollectionNode) root, METADATA_FILENAME);

    // Extract metadata

    return metadataAsXml;
  }
```

Example: BESS

sh bin/lsdma4bess [command] [command options]

Available commands:

init	Initialize ingest settings for KIT Data Manager. (hostname, user, …)
archive	Archive configured databases for given date to repository (e.g.: LSDF).
lookup	Lookup for possible databases and dates.
download	Download selected data from repository to local machine.

Example: BESS

sh bin/lsdma4bess lookup

Result:

Found Databases: name count

****

Database: database1 (3)

****

Start date:

Start date: 2014-08-01(1)

Start date: 2014-07-01(1)

Start date: 2014-06-01(1)

****

Database: database2 (2)

****

Start date:

Start date: 2014-07-01(1)

Start date: 2014-08-01(1)

****

Time period: 2013-09-01 - 2014-09-01