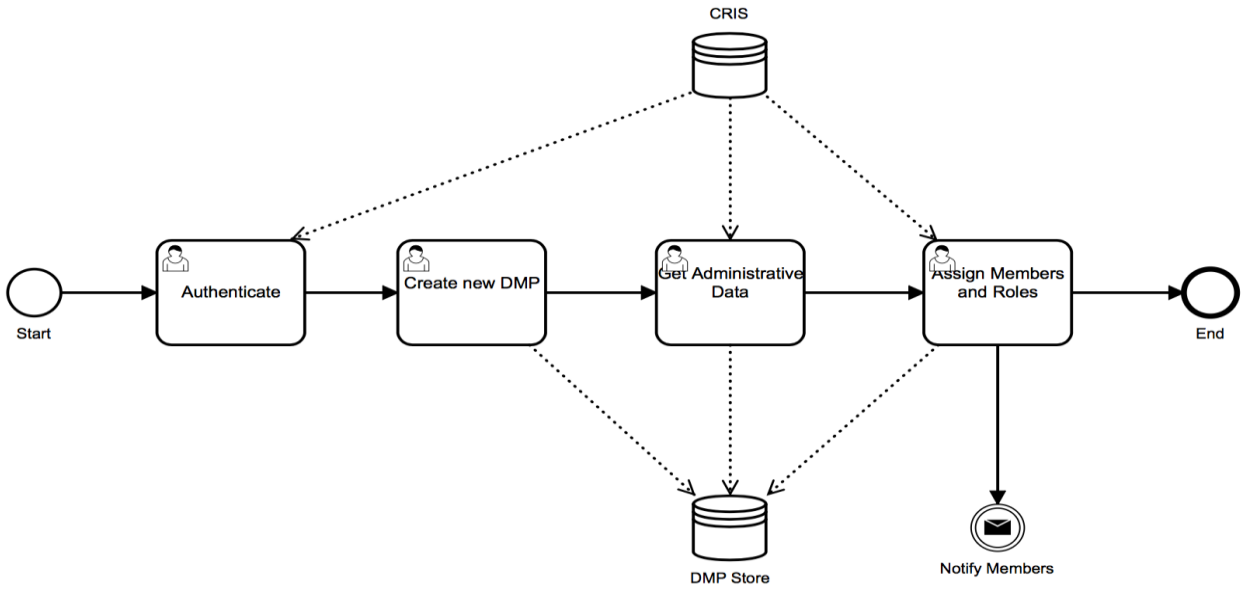
# BPMN Processes for machine-actionable DMPs

Simon Oblasser & Tomasz Miksa

## Contents

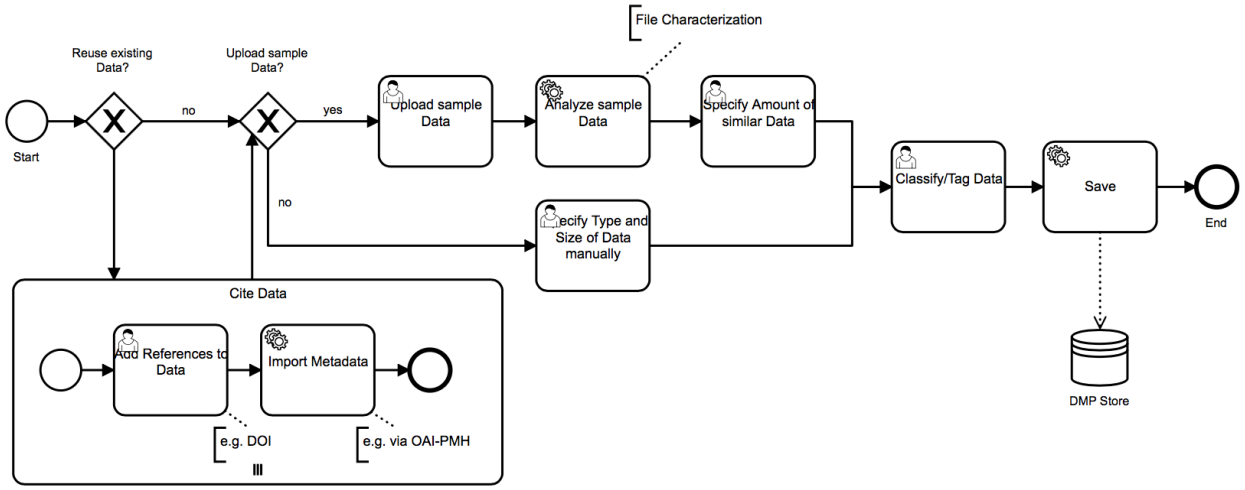
# Start DMP




The Start DMP workflow is all about creating an initial DMP.

1. **Authentication**: The researcher authenticates via Current Research Information System (CRIS) of the institution or uses an external authentication mechanism like provided by ORCID. The institutional account can also be linked to the ORCID record.
2. **Create New DMP**: The researcher creates a new DMP which gets persisted to the DMP Store.
3. **Get Administrative Data**: Administrative data which is linked to the signed-in CRIS account (e.g. ORCID) is retrieved via API and populated on the DMP. Besides personal details of the researcher like ORCID, email and affiliation also information about works and projects could be imported.
4. **Assign Members and Roles**: The researcher can add members, e.g. other researchers, who are involved in the project to the DMP, again by retrieving personal details from the CRIS. Members can be assigned with roles and responsibilities for the data management planning. Once the research project team is set-up, a notification informing each member about the DMP and their role/responsibility is sent. Each member uniquely identified by the ORCID or institutional email will be granted access to the DMP.

# Specify Size and Type




The Specify Size and Type workflow covers the specification of research data which will be used and generated during the project.

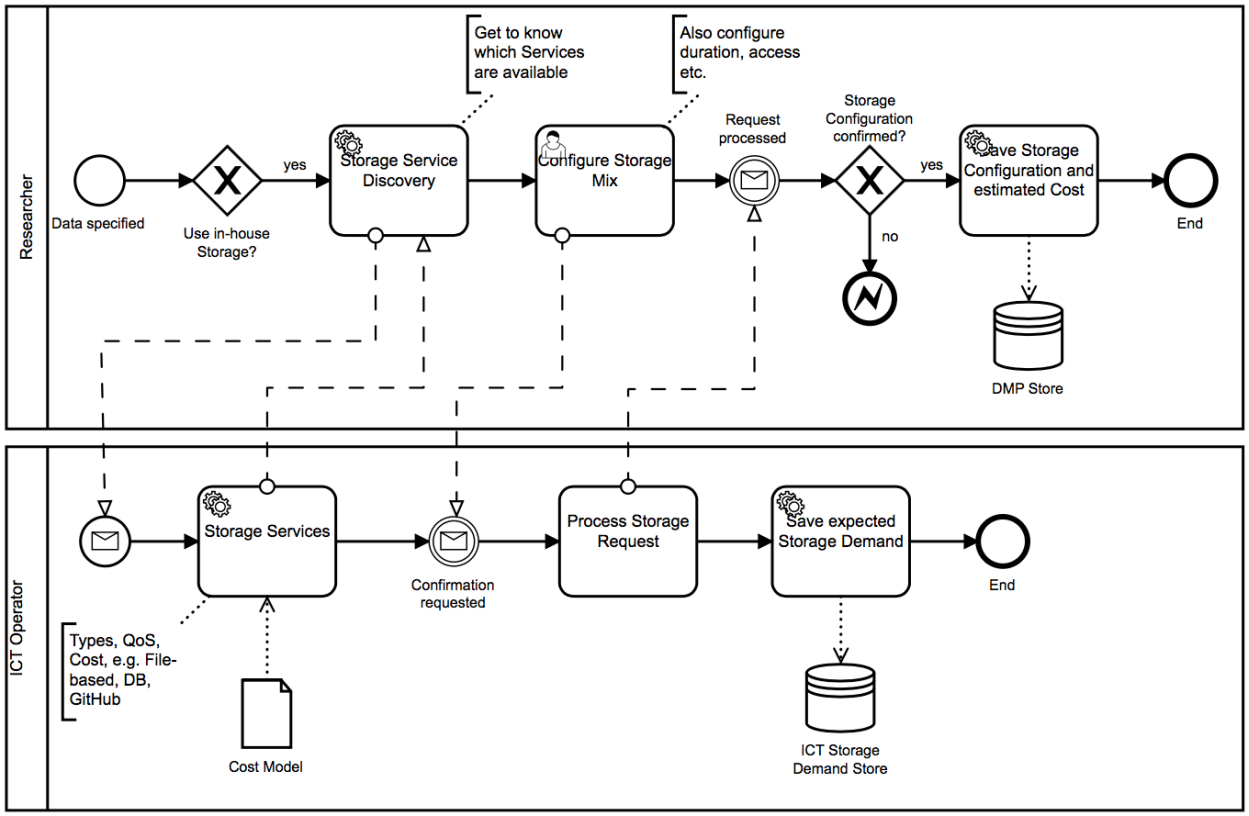
1. **Cite Data**: If existing data is being reused, ideally deposited in a citable repository, a unique and resolvable identifier like DOI can be entered and associated metadata be retrieved and added to the DMP. Many repositories support the Open Archives Initiative Protocol for Metadata Harvesting[1] (OAI-PMH) standard, other repositories like GitHub provide a REST-API[2] to collect metadata.
2. **Specify Output Data**: The researcher can specify the data which will be generated throughout the course of the project. The researcher can manually specify the expected types, file formats and size of the research data or get semi-automated support by uploading file samples which are analyzed by a file characterization tool. After the file samples got characterized, the researcher can specify the amount of expected similar files and a size estimation is being calculated.
3. **Classify/Tag Data**: The researcher can classify/tag the data with labels, e.g. input, output etc.
4. **Save**: The workflow concludes with storing the details about the research (input/output) data to the DMP.

[1] Open Archives Initiative
Protocol for Metadata Harvesting, URL: https://www.openarchives.org/pmh/
[2] GitHub Developer. REST API, URL: https://developer.github.com/v3/

# Get Cost and Storage

The Get Cost and Storage workflow deals with the selection, configuration, cost estimation and provisioning of various kinds of storage used for managing research data during the project (active data). The workflow is divided into two phases. In the first phase (Figure 10) the researcher configures the required storage offered by the ICT operator and gets a cost estimation. In the second phase (Figure 11), when the funder approved the DMP / project, the storage is actually booked and provisioned. By splitting up the storage provisioning process into two phases, only projects with backed-up funding get provided with computational resources to avoid stranded cost.

## Storage Configuration and Cost Estimation



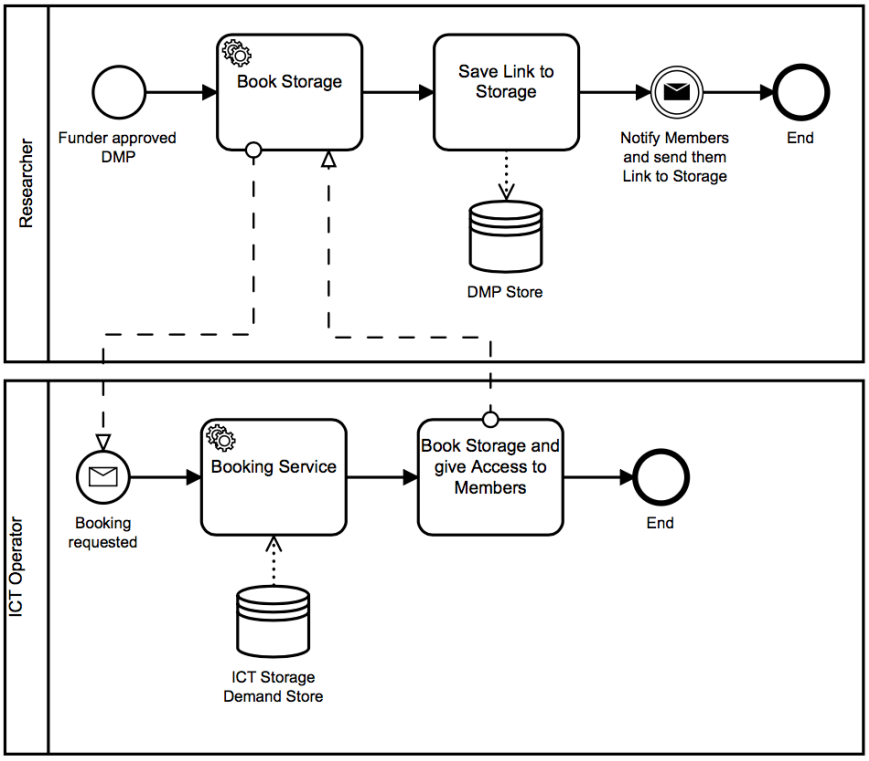

The workflow assumes that the type and size of the research data was specified (see Specify Size and Type).

1. **Storage Discovery**: If in-house storage is being used a Storage Discovery Service aids the researcher in finding out which services are available and seem fit for the kind of data at hand. The Storage Discovery Service communicates with a Storage Service offered by the ICT operator. The Storage Service provides information about the types of available storage services (e.g. file-based, database, code-repository) and Quality of Service (QoS) metrics, e.g. access speed,

availability, backup, etc. Based on an internal cost model the Storage Service can provide information about the associated costs.

2. **Configure Storage Mix**: Once the researcher knows about the available services and costs, he or she can configure a storage mix and provide further information, like who shall have access and the duration the storage is needed. A request containing the desired storage mix is sent to the ICT operator and processed.
3. **Save Storage Configuration and estimated Cost**: If the requested storage mix is confirmed, the details and associated costs are stored to the DMP. On the ICT operator side, the storage request is stored and the expected demand in the near future is known, but no booking/provisioning is done yet.

## Storage Provisioning




The storage booking workflow is triggered when the funder approves the DMP / project.

1. **Book Storage**: A Book Storage task requests the booking of the previously configured storage mix from the ICT operator. A Booking Service retrieves the details of the storage mix and the actual booking and further required processes (e.g. set-up of code-repository) are executed.
2. **Give Access**: Access to the members is given and access details (e.g. link) are fed back to the DMP.
3. **Notification**: Once the storage is ready, all the project members are notified about the availability of the storage mix, containing respective details for access.

# Get License

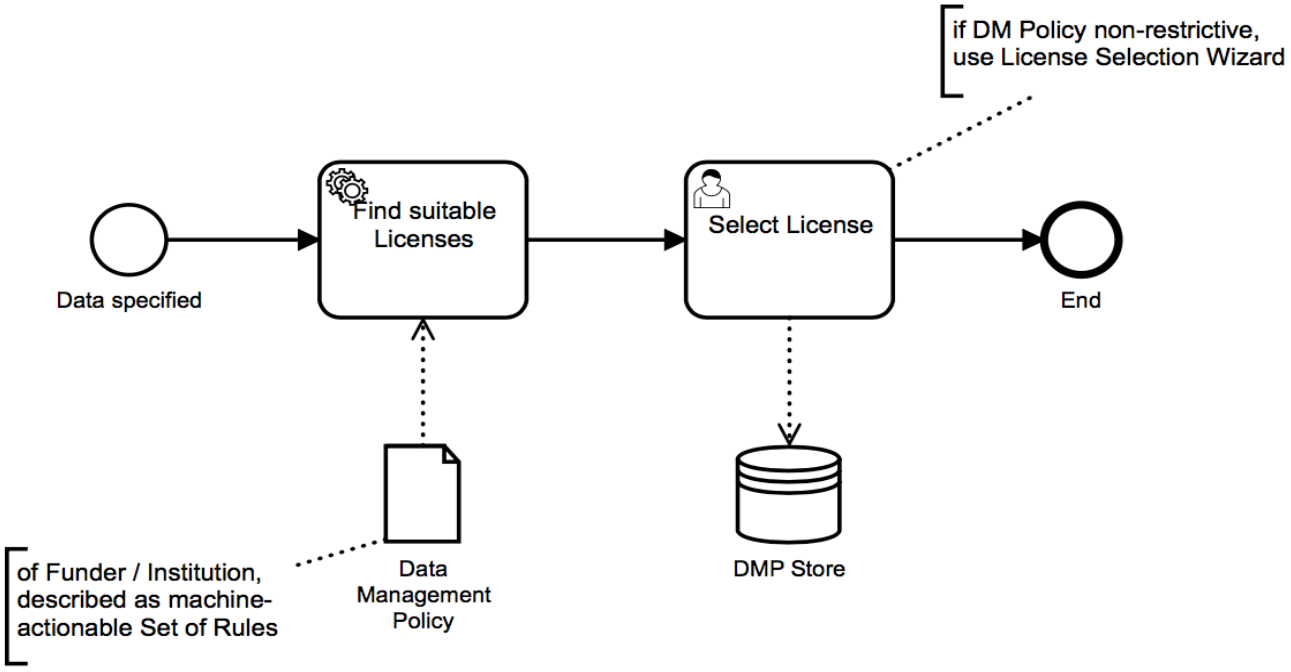
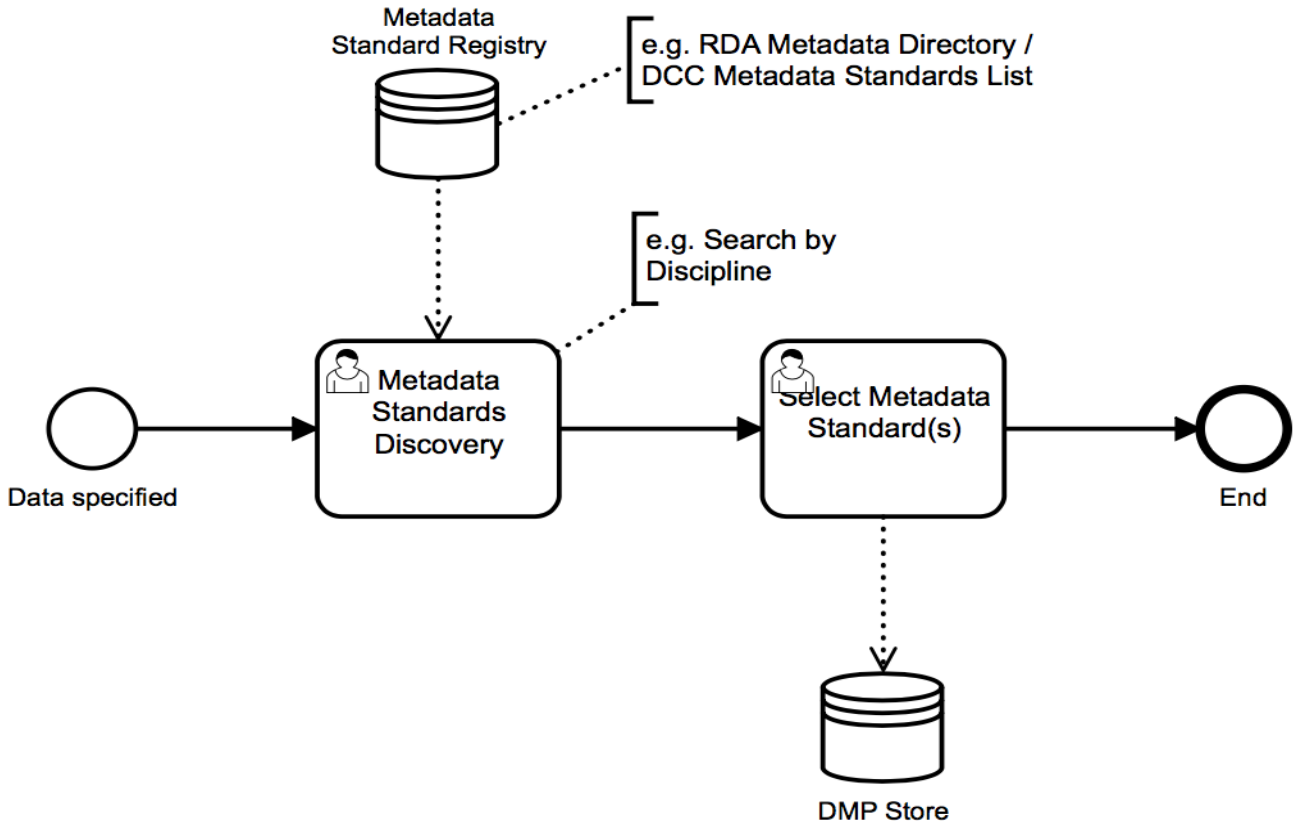



Figure shows the workflow for research data license selection.

1. **Find License base on policy**: Based on the funder or institutional policy a license can be selected automatically, e.g. the funder requires to publish all research data under CC-BY license. In order to make this approach work, the data management policy needs to be transformed into a machine-actionable format, which can be consumed by the license selection service.
2. **Select License**: If the policy is less restrictive, a license selection wizard such as the EUDAT License Selector[3] can be integrated to guide the research through the decision process of selecting a suitable license. After the license is selected, the decision is stored to the DMP.

[3] EUDAT License Selector, URL: https://www.eudat.eu/services/userdoc/license-selector

# Get Metadata Standard

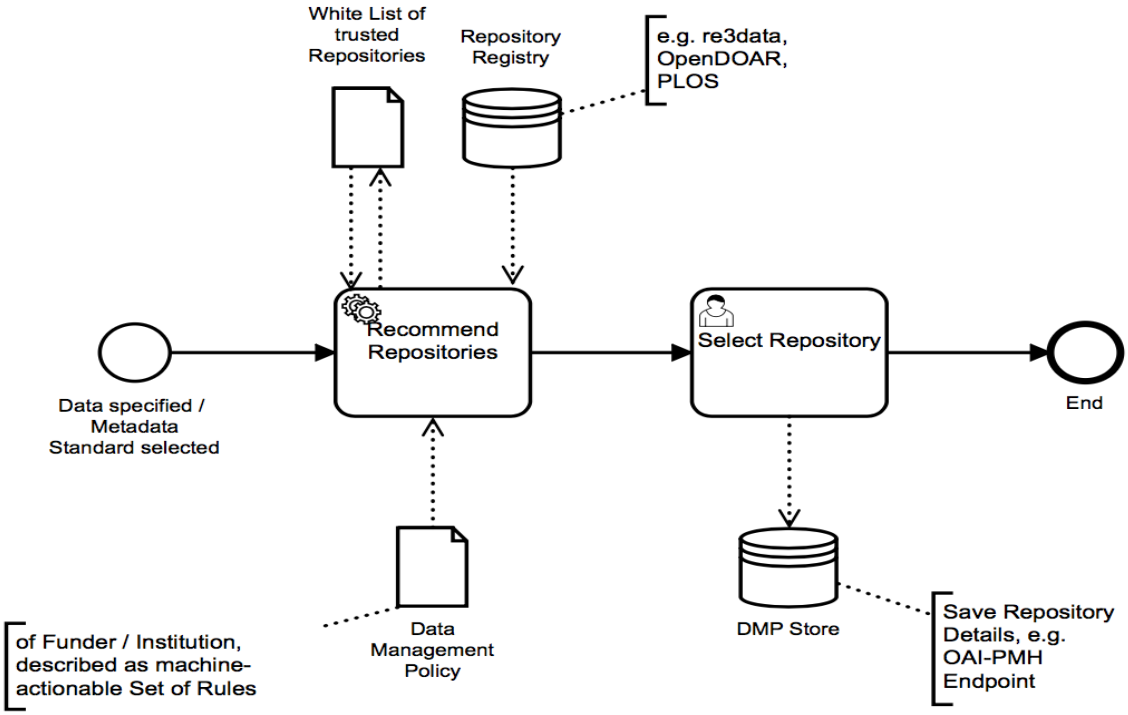



The Get Metadata Standard workflow aids the researcher with selecting suitable metadata formats to describe the research data.

1. **Metadata Standards Discovery**: The researcher consults a Metadata Standard Discovery service, which allows to browse through lists of standards and apply filters, e.g. by discipline, type of data etc. The service collects the metadata standards from a metadata standard registry, which could be backed-up by trusted sources like the RDA Metadata Directory[4] (recommended by H2020[5]) or the DCC List of Metadata Standards[6].
2. **Metadata Standards Selection**: Once suitable metadata standards are found, the selection gets persisted to the DMP.

[4] Research Data Alliance. Metadata Standards Directory WG, URL: http://rd-alliance.github.io/metadata-directory/
[5] Guidelines on FAIR Data Management in Horizon 2020. URL: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
[6] DCC. List of Metadata Standards, URL: http://www.dcc.ac.uk/resources/metadata-standards/list

# Get Repository

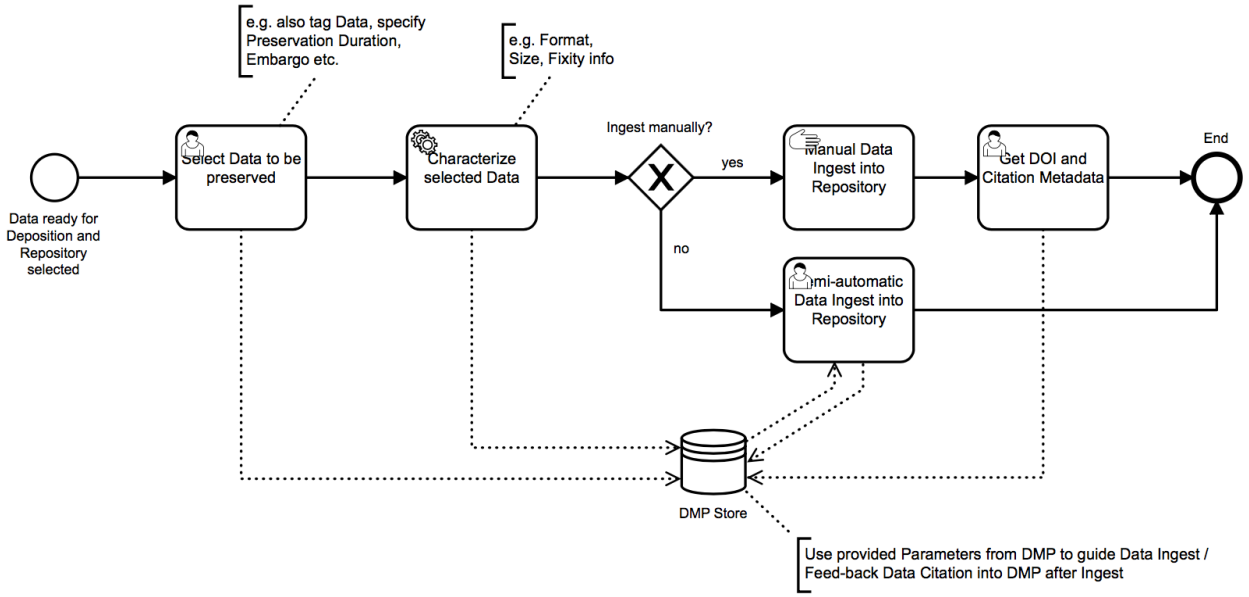



The Get Repository workflow is concerned with a repository selection for long-term preservation of the research data.

1. **Recommend Repository**: Based on parameters such as the type of research data, field of research, selected metadata standard, preferred geographic location of the repository and funder policies, a repository recommender service can be consulted to find a suitable repository. Similar as described in the Get License workflow, funder policies need to be transformed to a machine-actionable format in order to be consumed by this service. As an example a funder policy could prescribe that the research data need to be deposited in an open access repository located in Europe. Internally the repository recommender service could be backed-up by data from well-known research data repository registries like re3data[7], OpendDOAR[8] or a list of trusted repositories issued from PLOS[9]. These registries index repositories and provide metadata which could be used for filtering. The repository recommender could black-list undesired repositories and rank repositories based on institutional needs and policies (whitelist), e.g. the policy could state that the institutional repository shall be used, therefore it gets ranked #1 in the list of recommended repositories.
2. **Repository Selection**: Once a suitable repository is found, the selection including useful metadata, e.g. OAI-PMH endpoint of the repository gets persisted to the DMP.

[7] Registry of Research Data Repositories, URL: https://www.re3data.org/
[8] Directory of Open Access Repositories, URL: http://www.opendoar.org/
[9] PLOS. Recommended Repositories, URL: http://journals.plos.org/plosone/s/data-availability#loc-recommended-repositories

# Deposit Data




The Deposit Data workflow models the process when a researcher wants to publish his or her research data in a repository.

1. **Data Selection**: The researcher must select which files shall be preserved. For this interactive approach to work, the researcher must be presented with a structured list of the files located in the storage. The researcher can tag files to provide additional metadata. The preservation duration and embargo time for the selected files shall also be specified. Once the list of files to be preserved is ready, a file characterization tool analyzes the files and provides metadata for each file (e.g. size, format, fixity).
2. **Data Ingest**: The researcher can choose to manually ingest the research data into the repository or choose a semi-automatic approach. This of course depends on the repository supporting a standard way of automatic data deposition, like Simple Web-service Offering Repository Deposit[10] (SWORD). Common repositories supporting SWORD are DSpace[11], Fedora or EPrints.
3. **Get DOI and Citation Metadata**: After the ingest into the repository completed and a DOI was issued by the repository, citation metadata is collected from the issuer and stored along the DOI in the DMP.

[10] SWORD, URL: http://swordapp.org/
[11] DURASPACE. SWORDv2 Server, URL: https://wiki.duraspace.org/display/DSDOC6x/SWORDv2+Server

# Get Help

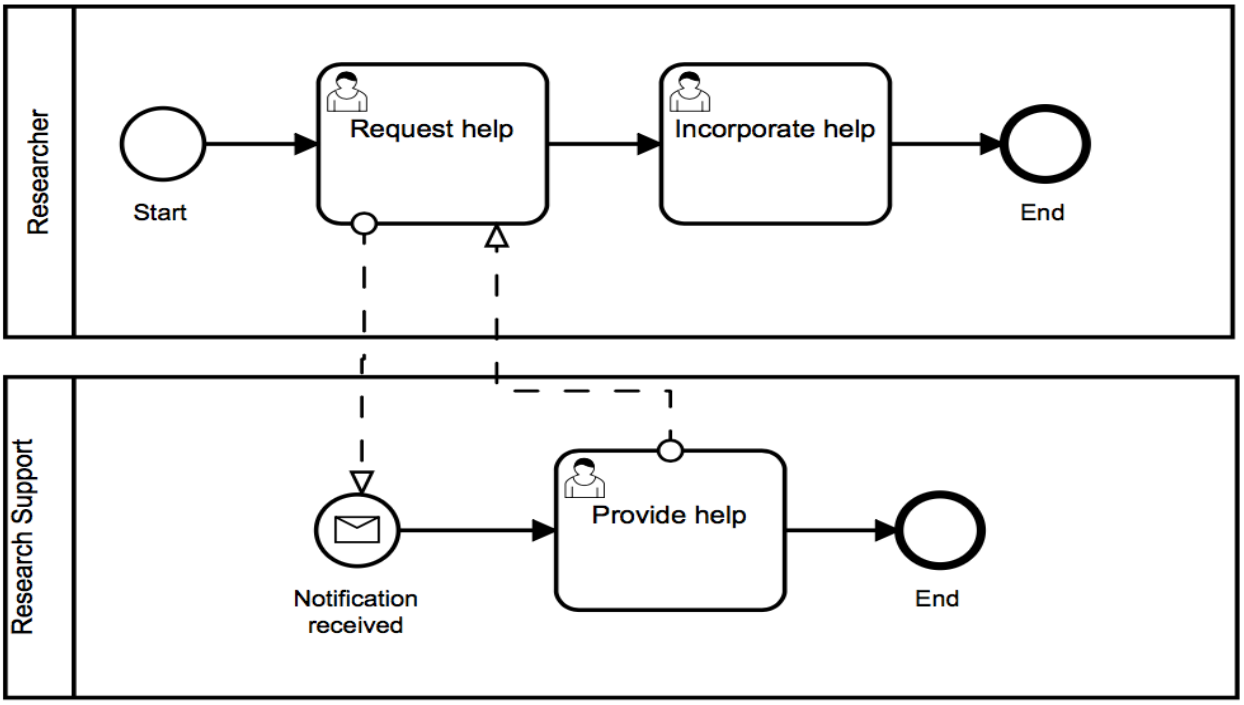



Figure depicts the Get Help workflow which illustrates the integration of research support into data management planning.

1. **Request help**: If a researcher needs help/advice with data management planning, he or she can create a help request which will be sent to a research support helpdesk.
2. **Notification**: Employees of the helpdesk get notified, can process the help request and provide feedback. The helpdesk could use some kind of ticketing system and may get access to the DMP for the time of assist.