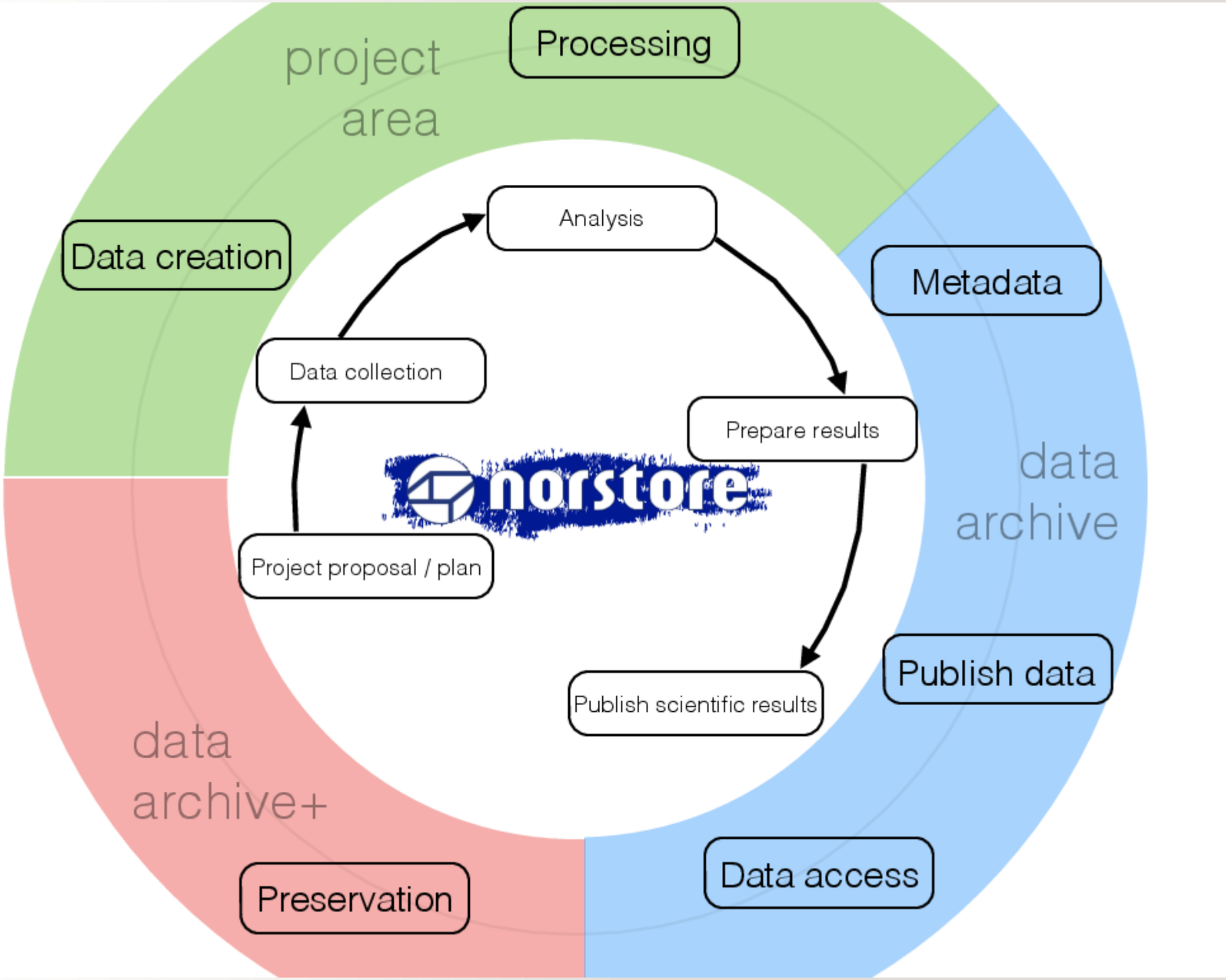
Archive I

7. Jan 2015

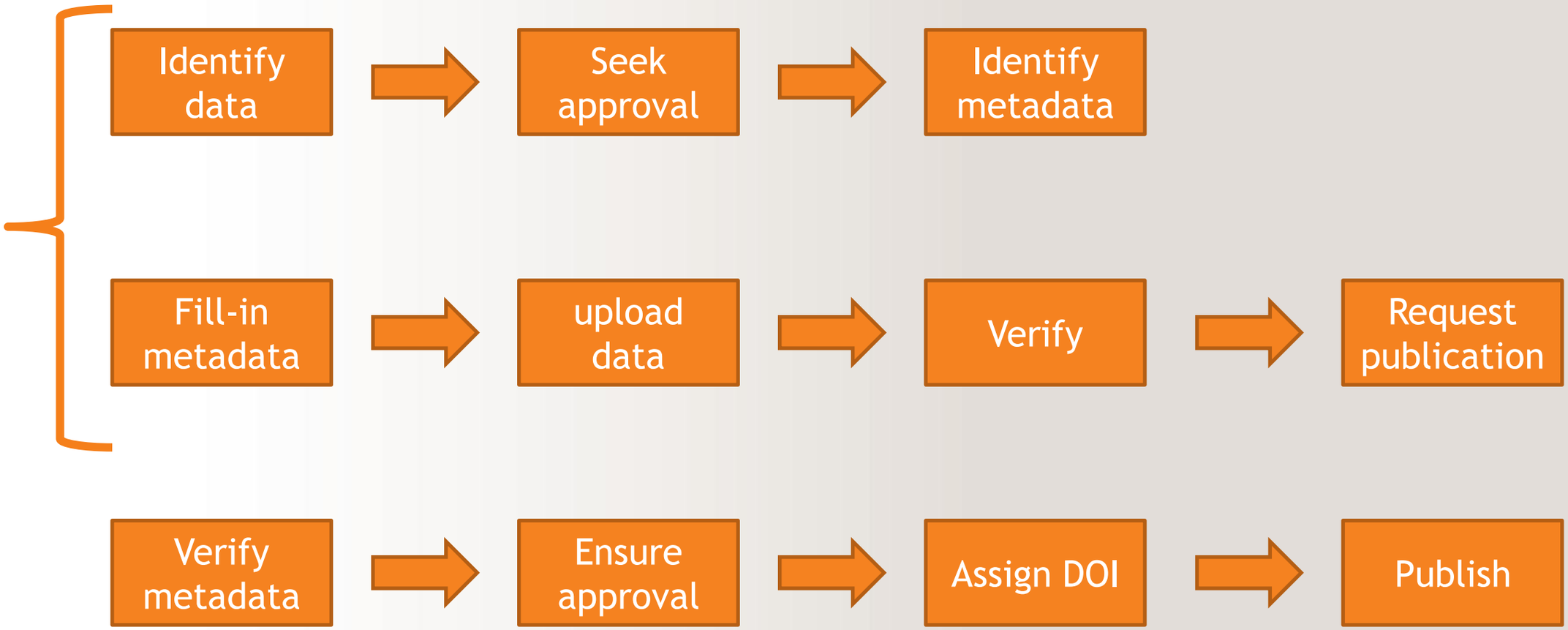
UNINETT


project
area
Processing
Data creation
Analysis
Metadata
Data collection
Prepare results
norstore
data
archive
Project proposal / plan
Publish data
Publish scientific results
data
archive+
Preservation
Data access

How to archive?

- We need to collect all the information that makes it possible to reuse our data at a later point in time.
  - Think what you would need in order to re-analyse or reuse the data yourself (applications, libraries, OS, services etc).
- Data – needs ideally to be in a standard or open format that makes it possible to migrate in case of obsolescence.
- Licensing – who can use the data, under what restrictions (see K Lyng's talk).
- Contact persons – in case of questions concerning the data.
- Metadata – description of the data.

Archive workflow

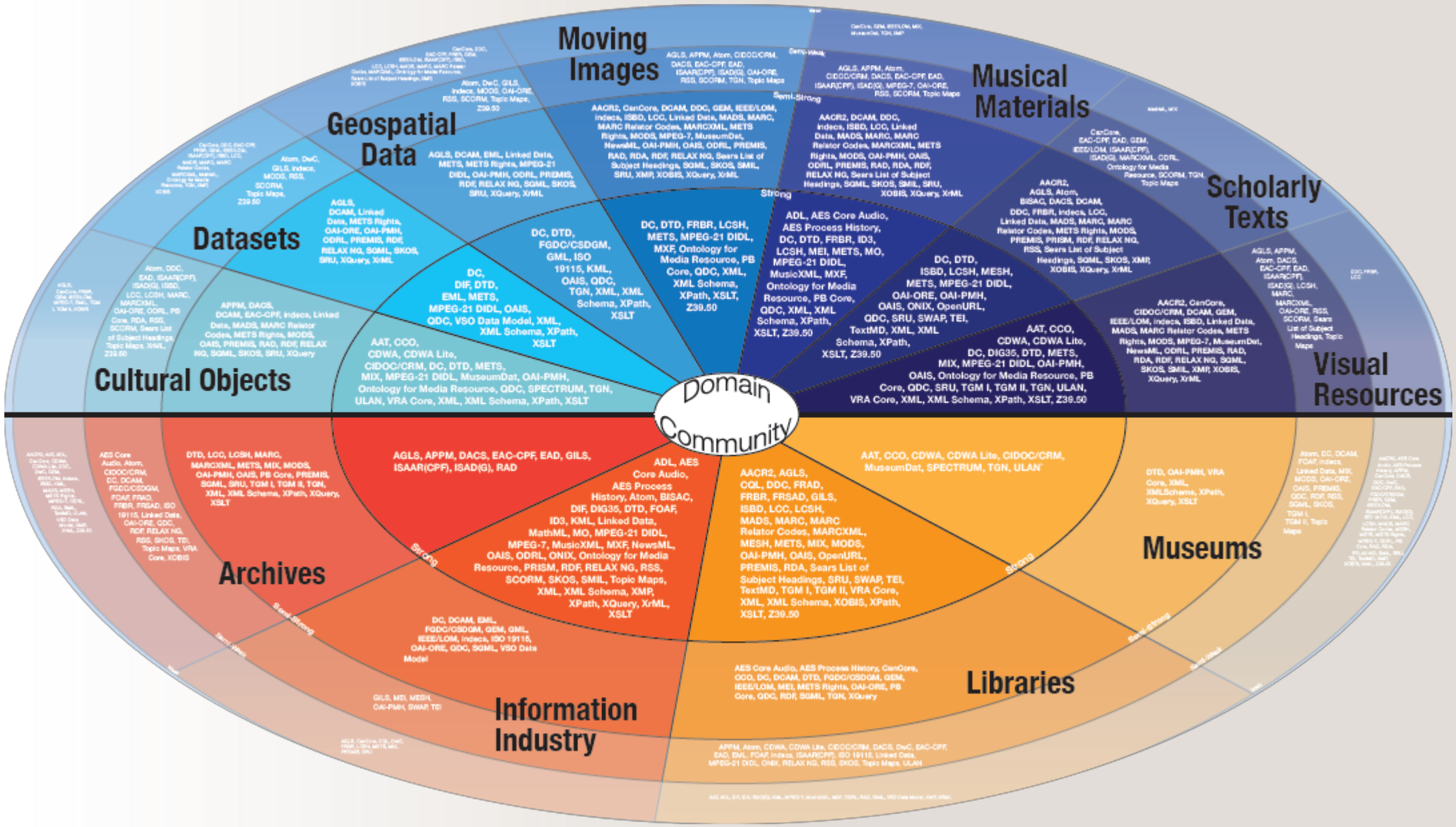


Metadata

- Essentially information about data.
- Describes how to successfully manage and reuse the data.
- Can be divided into 3 types:

  - **Descriptive** – what the data is, how to use it, special features in the data, workflows, etc.
  - **Structural** – how the data is arranged (e.g. files in directory X are configuration files, in directory Y are documentation and in directory Z are the raw data), also includes the formats descriptions.
  - **Administrative** – covers information to manage the data such as checksums, rights information, significant properties etc.

Metadata



Seeing Standards: A Visualisation of the Metadata Universe.
J. Riley, D. Becker

Norstore Archive Metadata

- Large number of metadata standards.
- Many are tailored to specific communities.
- Many have Dublin Core either as a basis or have a strong overlap with Dublin Core.

- Dublin Core is an ISO standard.
  - The standard has 15 terms, extended Dublin Core has more terms.
- The Norstore Archive uses Dublin Core as a basis.
  - Additional metadata terms added that are not covered by DC, but are generic enough for all communities.
  - OAI-PMH based on DC so automatically compliant.

Norstore Archive Metadata

Descriptive Information

Category
Description
Identifier
Internal Identifier
Journal Article
Language
Phase
State
Subject
Title

Administrative Information

Access Rights
Contributor
Created
Creator
Data Manager
License
Lifetime
Preservation Level
Published on
Publisher
Rights
Rights Holder
Submitted
Terms and Conditions for Deposit

Structural Information

File Checksum
File Name
File Size
File Type

Descriptive Information (optional)

Bibliographic Citation
Conforms to
Comment
Geolocation
Label
Project
Provenance
Source
Temporal Coverage

- Bold terms are Dublin Core recommended terms.
- Top 3 boxes contain mandatory metadata.
- Terms in italics are automatically filled in by archive.

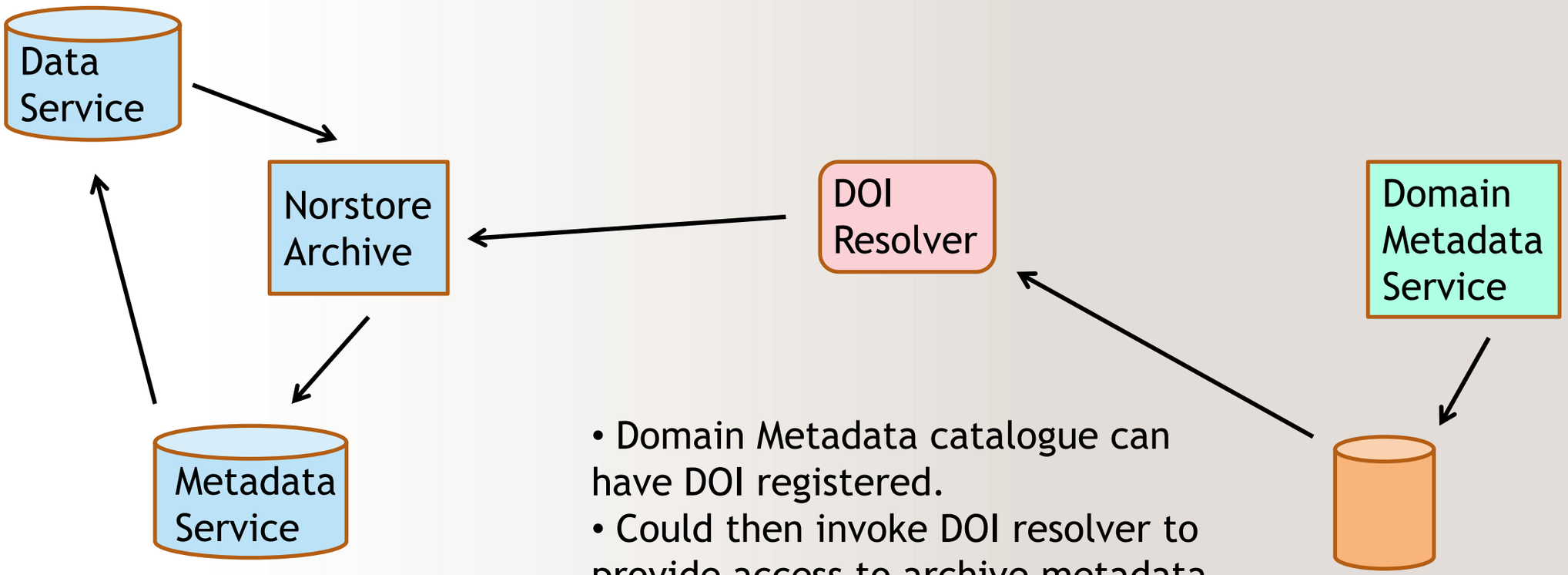
Norstore Archive Metadata

- You may find many of the optional terms very useful:
  - Geolocation – used to define geographic locations (WGS84 and UTM supported).
  - Temporal Coverage – used to define durations.
  - Bibliographic Citation – how the dataset should be cited in follow-up research that uses this dataset.
    - E.g. Authors/Rights Holders, Title, Year Published, DOI
  - Provenance – the history of the dataset (ie how it was created).
  - Source – the source dataset from which this dataset was derived.
  - Conforms to – which standards does the dataset follow.

Norstore Archive Metadata

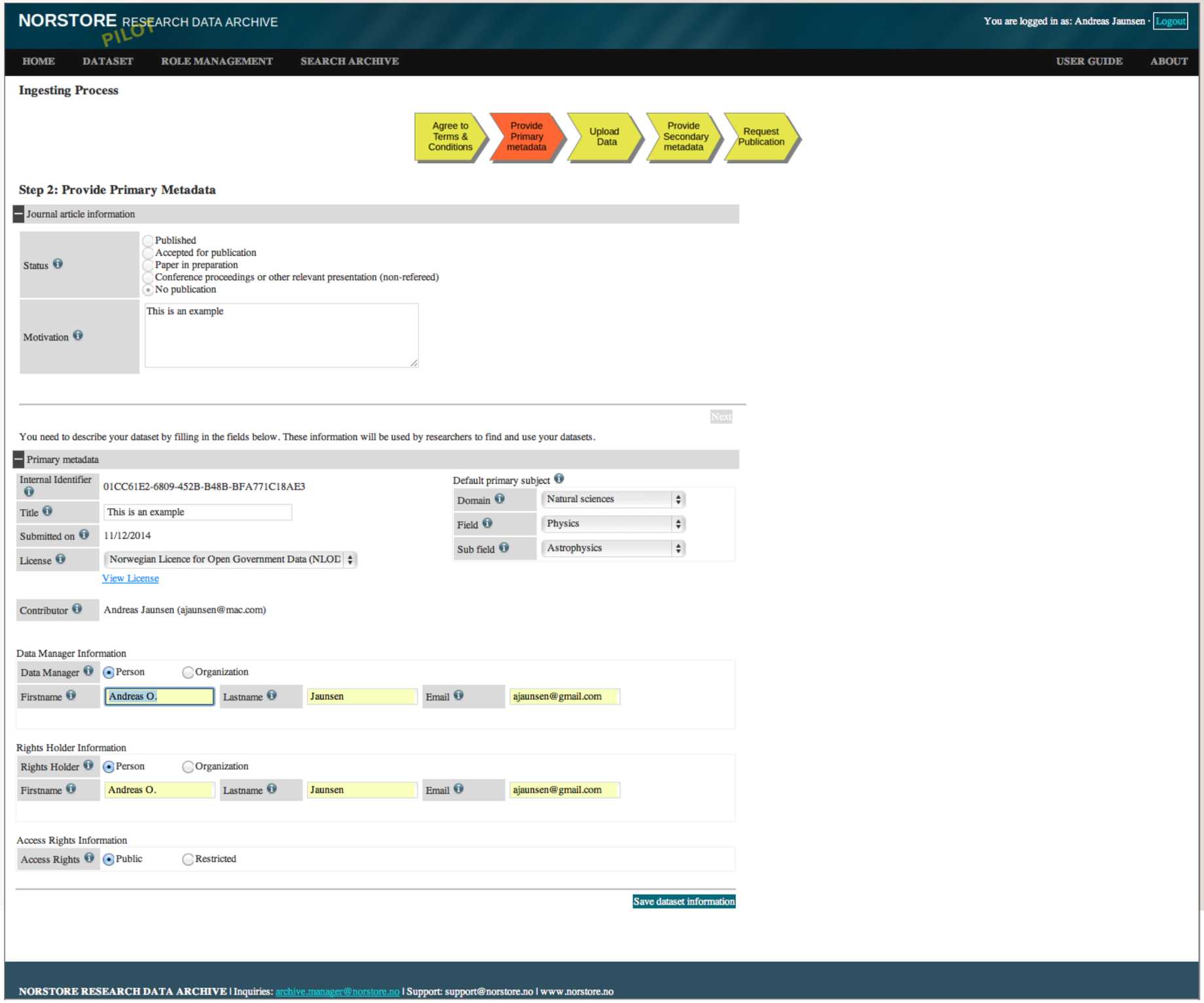
- Norstore metadata designed to be as generic as possible.
- Sufficient to locate data and understand how to use the data.
- More detailed information can be contained in domain-specific catalogues.
- Can have domain specific catalogues use the DOI as a handle to the data (resolving the link will provide access to the data).
- Could think about the archive referencing an external domain specific catalogue.

Norstore and Domain Metadata



- Domain Metadata catalogue can have DOI registered.
- Could then invoke DOI resolver to provide access to archive metadata and data.

How to fill in Norstore Metadata



Norstore Metadata

- Currently metadata must be supplied via the web interface.
- Metadata divided into 2 phases:
  - Before uploading the data – consists of mandatory metadata that is needed by the archive for managing the data (eg contact information, title of the dataset etc)
  - After uploading the data – consists of remaining mandatory descriptive information and optional information.
- There is a 3 month time limit to fill in metadata
  - User will be reminded during this period of need to complete metadata.
  - At the discretion of the Archive Manager the dataset may be deleted if the metadata remains incomplete after this limit.
  - Typically metadata is completed within 2 weeks.

Tips for Norstore Metadata

- Try to avoid duplicating information.
- Think what information you would need in order to re-analyse your data.
  - What applications, libraries, workflows, manuals would you need (versions etc)? Any other data that would be needed?
  - Any other information (such as auxiliary files or databases) required?
- Are there any features or peculiarities of the data that are worth noting?
- Is the environment the data was collected in described?
  - Date, location the data was collected. If using additional instruments the configuration of those instruments.

Tips for Norstore Metadata

- Try to use the description field to describe what the dataset is and how to use it.
  - If you have a lot of documentation that covers this you could think about including it as part of the dataset and describe where to find it in the description.
- Try to make use of the Journal metadata to provide a reference to an article that describes the dataset.
  - In this case your description can be more succinct.
- If your dataset has temporal or spatial information consider using the optional metadata to capture that information.

Example

- Project to model deposition of Saharan sand in country X
  - Project develops a model for the summer aerial transport of sand from Sahara to country X.
  - The model consists of a C++ computer program that makes use of GNU Scientific Library and a series of input parameter files. The parameter files are structured ASCII files. The model output are structured ASCII files.
  - The project also contains recordings of the actual amount of sand deposited in 20 regions in country X. The recordings are stored in structured ASCII files.

Example Metadata

- Description metadata should contain:
  - What the dataset is, what project it is associated to, when it ran, assumptions in the model any notable features in the data.
  - Layout of the data (ie directory structure what's where).
  - "This dataset consists of a model for the summer aerial deposition of Saharan sand in country X. The model has been developed for Summer 2014 deposition. The dataset is arranged in the following manner: src – contains the model as C++ programs; input –contains the sample input configuration files needed by the model; output – contains the sample model output; data – contains the sand deposition in soil samples taken in regions X1, X2, … X20; doc – contains detailed documentation on building, configuring and running the model as well as required libraries and documentation on interpreting the results."
- Use the Journal metadata to reference the article based on the dataset.
- Use the geospatial coverage to identify the country studied.
  - Optionally use the temporal coverage if the studies span a range in time.