Init3dbw2.pm21 - Provenance metadata to support Semantic Workflows

Contents of this Page

The following are links to some useful external materials

- Requirements for caBIG Infrastructure to Support Semantic Workflows

```
historical link: https://wiki.nci.nih.gov/download/attachments/20285848
/Requs_caBIG_InfrastructureSupportSWorkflow.ppt
```

- UNLINK Recommendations for caBIG to Support Semantic Workflows

```
historical link: https://wiki.nci.nih.gov/download/attachments/20285848
/Recommendations_caBIG_SupportSemWorkf.ppt
```

- Use Cases for Semantic Workflows in caBIG

```
historical link: https://wiki.nci.nih.gov/download/attachments/20285848
/UseCasesForSemanticWorklowIncaBIG.doc
```

The following are high level use case statements related to these requirements

- Semantic Metadata
  - Define semantic metadata for analytical service
  - Define semantic metadata for scientific data
  - Define semantic metadata for translation services
- Dynamic Workflows
  - Define workflow constraints
    - Desired output
    - Desired input
    - Data query parameters
    - Analytical parameters
    - Desired operations
    - Computational constraints/requirements
    - Time constraints/requirements
    - Storage constraints/requirements
  - Generate workflow
  - Validate workflow
  - Run workflow
  - Track workflow
  - Share workflow
  - Share dynamic workflow (template/constraints)
  - Version workflow (design, creation, evolution)
- Provenance Tracking
  - Create intermediate data
  - Fetch intermediate data
  - Link data (process)
  - Establish data ownership and security (attribution)
  - Version data (republishing/updates)

The following are non-functional requirements that do not result in actor-oriented use cases

- Define a semantic workflow standard encoding (e.g. OWL-S, WSMO, SWSL, SWSF)
- Define a provenance standard encoding

The following are some basic discovery related use cases that pertain to the requirements

Discover data of interest

Use Case Number	Init3dbw2.pm21.1
Brief Description	Discover data of interest: A researcher wants to find data that has already been collected for use with caArray. They are able to find the data and to inspect the system to learn about what type of cells are in the database, what type of pathology is available for the data, etc.
Actor(s) for this particular use case	Cancer Researcher
Pre-condition **The state of the system before the user interacts with it**	Data services exist and are accessible.
Post condition **The state of the system after the user interacts with it**	Data of interest is discovered.
Steps to take **The step-by-step description of how users will interact with the system to achieve a specific business goal or function**	1. The Cancer Researcher identifies the characteristics of the data that he would like to discover (e.g. type, size, specific data fields, etc.) 2. The Cancer Researcher performs a discovery query and gets back a number of datasets. 3. The Cancer Researcher interrogates those datasets to determine if they are of interest.
Alternate Flow **Things which would prevent the normal flow of the use case**	None.
Priority **The priority of implementing the use case: High, Medium or Low**	High.
Associated Links **The brief user stories, each describing the user interacts with the system for the one function only of the use case. There would potentially be a number of user stories that make up the use case.**	• Init3dbw2 - Provenance metadata to support Semantic Workflows
Fit criterion/Acceptance Criterion **How would actor describe the acceptable usage scenarios for the software or service that meets the actor's requirement?**	None.

Discover related data

Use Case Number	Init3dbw2.pm21.2
Brief Description	In some cases, two semantically equivalent data element can be annotated with different semantic concepts that may or may not themselves be related. In these cases, there needs to be a mechanism to define semantic equivalence between the data elements, the concepts, or expand/contract the scope of the semantic query in the case of related concepts. An example of this use case is that there needs to be a way to discover data elements both with StartDate and Begin+Date, e.g. through a semantic equivalence of the two or through a widening/narrowing query.
Actor(s) for this particular use case	Metadata Specialist, Cancer Researcher
Pre-condition **The state of the system before the user interacts with it**	Two data element exist and are individually discoverable
Post condition **The state of the system after the user interacts with it**	The two data elements are discovered as semantically equivalent

Steps to take **The step-by-step description of how users will interact with the system to achieve a specific business goal or function**	1. A Metadata Specialist individually discovers the two data elements 2. The Metadata Specialist determines (manually) that these two data elements are semantically equivalent 3. The Metadata Specialist defines a rule that the data elements are semantically equivalent 4. A Cancer Researcher performs a discovery query that would normally (if there were no rules defined) return one of the data elements 5. Both of the data elements are returned to the Cancer Researcher
Alternate Flow **Things which would prevent the normal flow of the use case**	If the two data elements are annotated with related concepts, the following alternate flow is possible: 1. A Cancer Researcher discovers one of the data elements through a semantic query 2. The Cancer Researcher widens the semantic query to include additional related concepts (up the tree for less specific, down the tree for more specific) 3. Both of the data elements are returned to the Cancer Researcher
Priority **The priority of implementing the use case: High, Medium or Low**	High.
Associated Links **The brief user stories, each describing the user interacts with the system for the one function only of the use case. There would potentially be a number of user stories that make up the use case.**	None.
Fit criterion/Acceptance Criterion **How would actor describe the acceptable usage scenarios for the software or service that meets the actor's requirement?**	None.

Aggregate data

Use Case Number	Init3dbw2.pm21.3
Brief Description	Aggregate data of interest: A researcher is able to query the system to find data that can be combined with their data. It is able to compare the characteristics of the dataset to ensure that the data are combinable, for example .
Actor(s) for this particular use case	Cancer Researcher
Pre-condition **The state of the system before the user interacts with it**	A number of datasets have been identified for aggregation.
Post condition **The state of the system after the user interacts with it**	Combinable data has been aggregated.
Steps to take **The step-by-step description of how users will interact with the system to achieve a specific business goal or function**	1. The Cancer Researcher performs a discovery query to find available datasets that can be aggregated with his dataset 2. The Cancer Researcher selects datasets to be aggregated 3. The Cancer Researcher selects aggregation parameters (e.g. data elements to combine) 4. The Cancer Researcher performs the aggregation
Alternate Flow **Things which would prevent the normal flow of the use case**	None.
Priority **The priority of implementing the use case: High, Medium or Low**	Low.
Associated Links **The brief user stories, each describing the user interacts with the system for the one function only of the use case. There would potentially be a number of user stories that make up the use case.**	• Init3dbw2 - Provenance metadata to support Semantic Workflows

Fit criterion/Acceptance Criterion **How would actor describe the acceptable usage scenarios for the software or service that meets the actor's requirement?**	None.

The following use cases have direct overlap with these requirements but have been captured under Init1dbw6.pm8.U0 - Support caB2B to integrate services on caGrid

- Init1dbw6.pm8.U0 - Support caB2B to integrate services on caGrid