RWTH AACHEN UNIVERSITY

UNICORE SUMMIT

JÜLICH FORSCHUNGSZENTRUM

MMF: A Flexible Framework for Metadata Management in ***UNICORE***

Waqas Noor[1], Bernd Schuller[2]

[1] RWTH Aachen Technical University

[2] Jülich Supercomputer Centre (JSC)

UNICORE Summit 2010, May 18th, 2010, Jülich, Germany

Outline

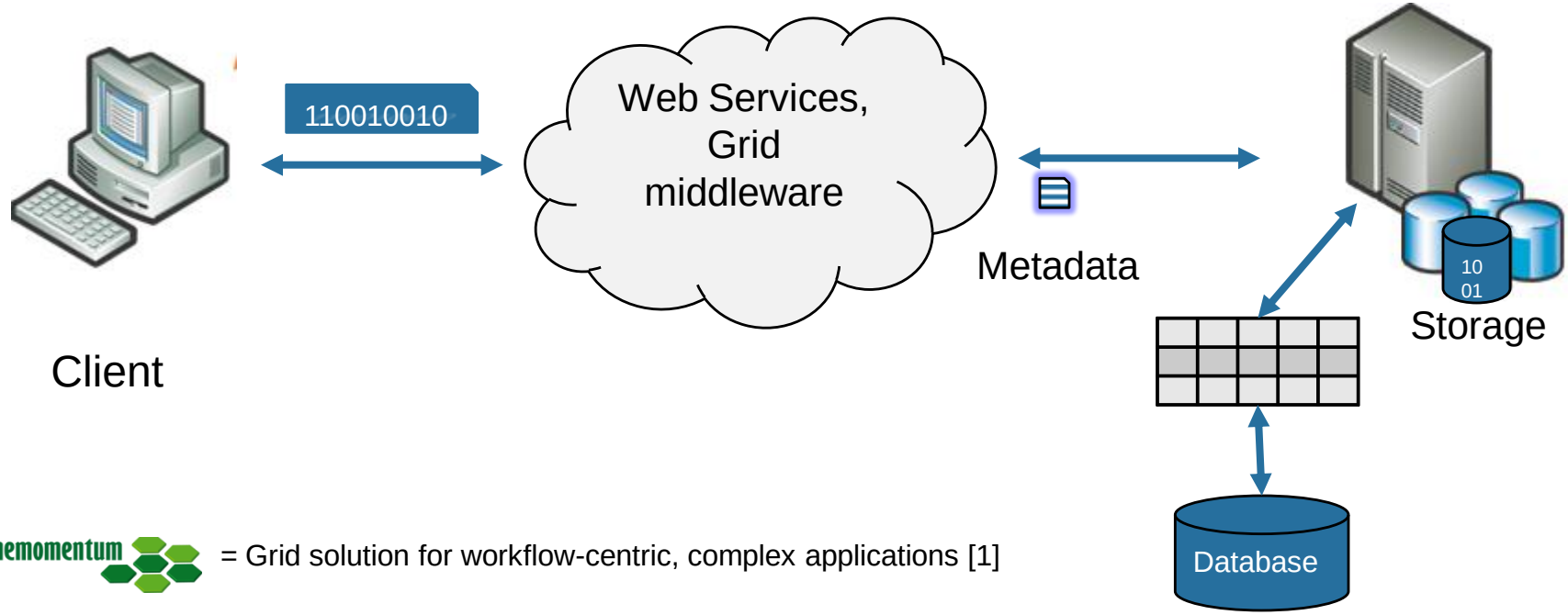
- Motivation
- Objectives
- Metadata Management
  - Architecture
  - Implementation
  - Performance & Evaluation
- Usage Scenarios
- Future Work

Motivation

- Metadata
  - Classification of data
  - Makes data highly searchable
- Metadata & Grid
  - No full text search
  - No flexible support for user defined metadata
  - Data and metadata stored separately
  - Data migration problem

Motivation

- Centralized repositories for metadata



Chemomentum = Grid solution for workflow-centric, complex applications [1]

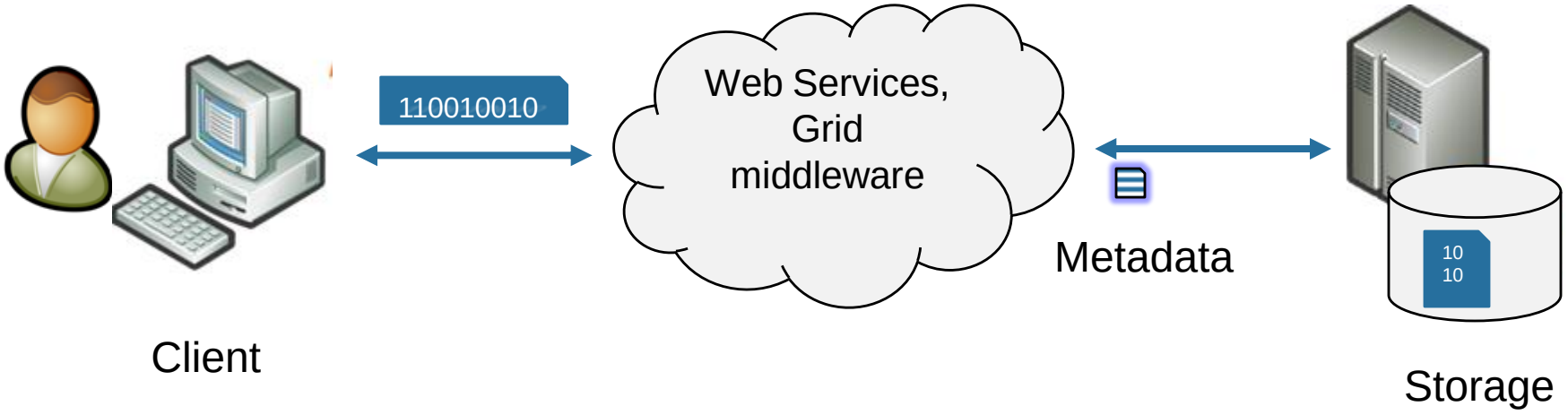
ARDA = Metadata catalogues for gLite grid middleware [2]

Approach

- No centralized metadata management
- Schema free metadata model
- User defined metadata as tags
- Full text search
- Extensible & scalable

Metadata Management Framework (MMF)

Metadata Management Framework (MMF)



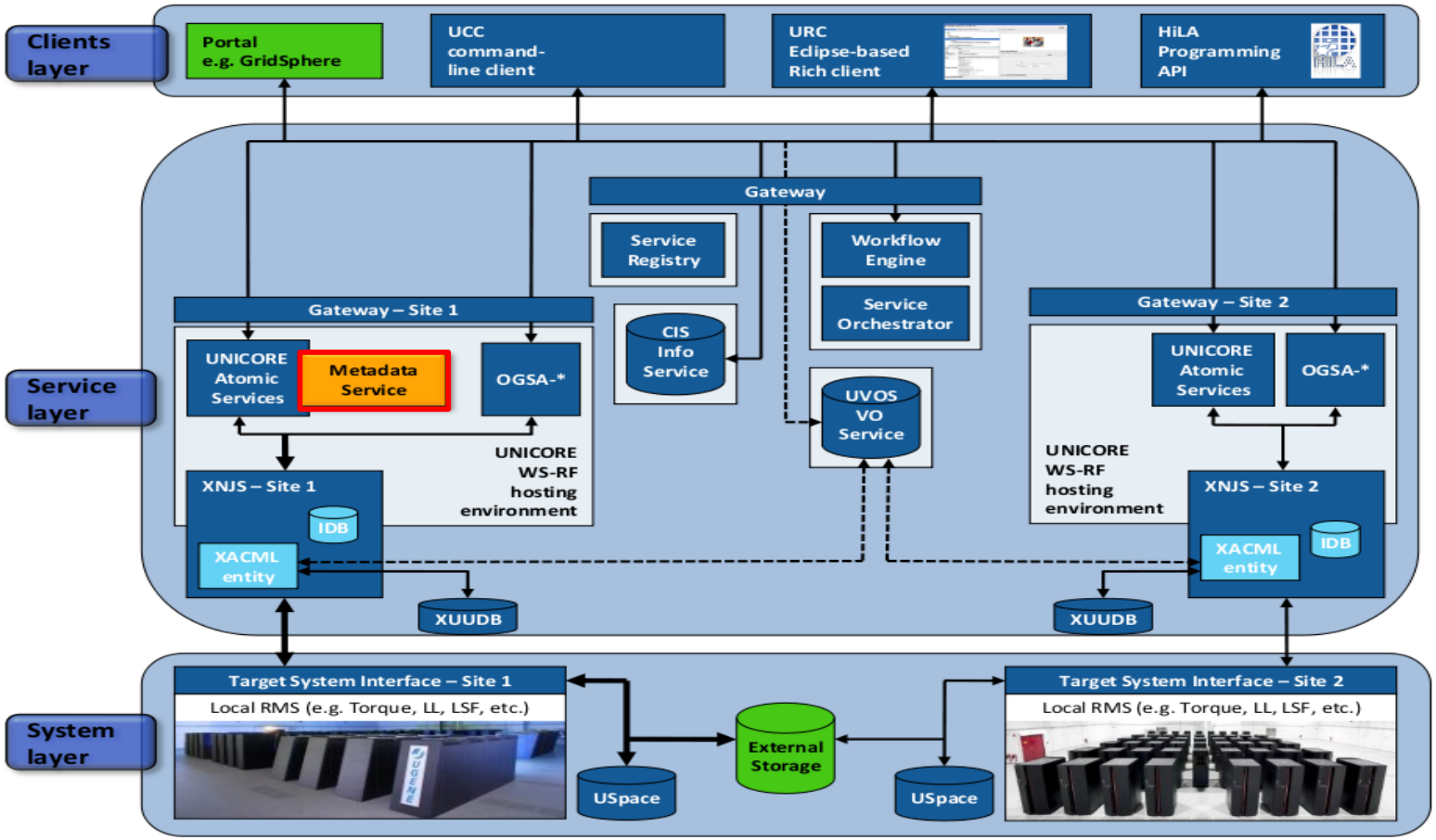
Architecture

UNICORE
(*Uni*form *I*nterface to *Co*mputing *Re*sources)

- Open source, ready to run grid middleware
- Provides seamless access to grid resources
- OGSA based architecture
- Workflow can be submitted to grid
- Involved in various grid R&D projects [3]
  - FIT4Green
  - WisNetGrid
  - SLA4D-Grid
  - DEISA2
  - And many more ...

SLA4D-GRID

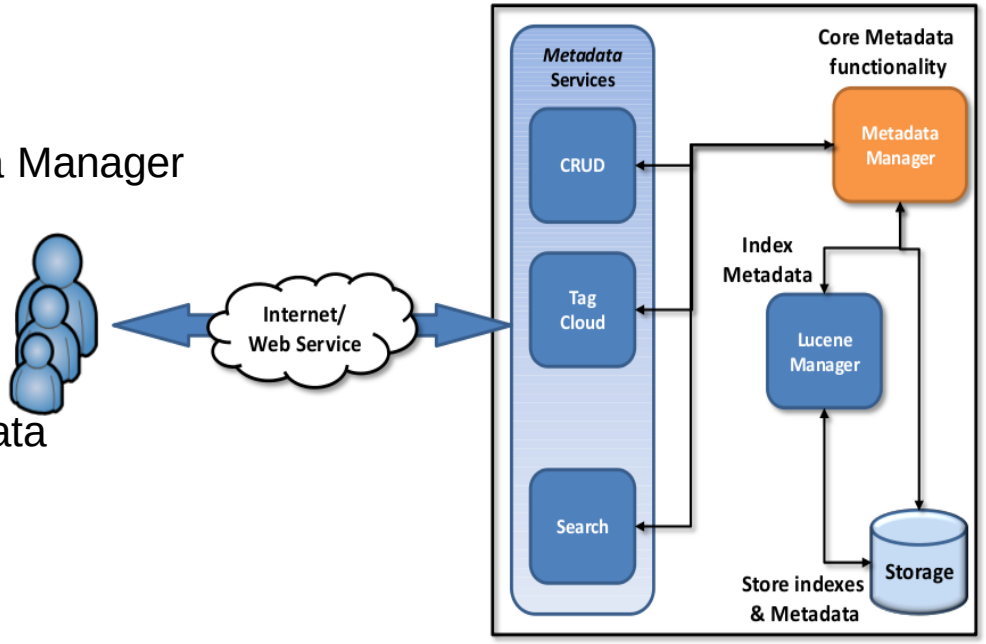
UNICORE Architecture



Metadata Management Architecture

Metadata Service

- OGSA based Web Service
- Functionality is backed by Metadata Manager



Metadata Manager

- Provides core functionality of metadata Management
- Communicate with Lucene Manager

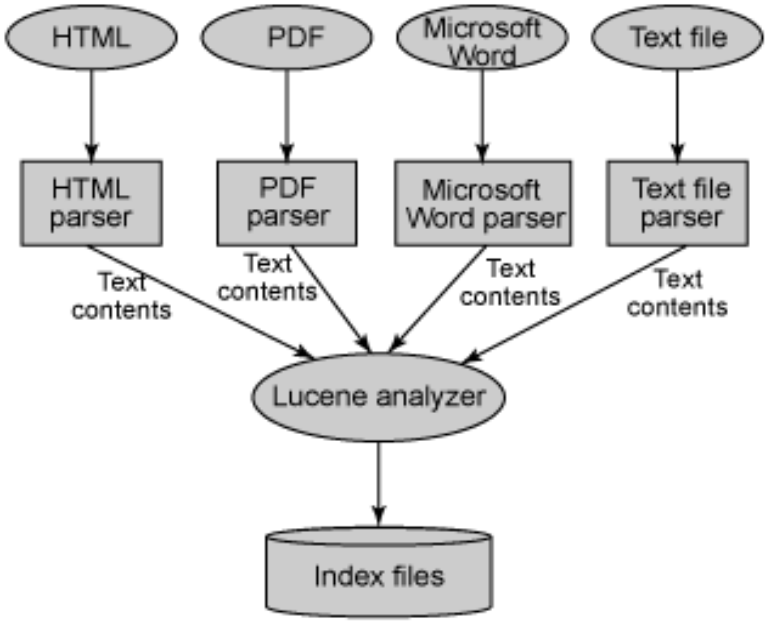
Lucene Manager

- Responsible for indexing and retrieving the data
- Provides search interfaces

Apache Lucene 2.9.0 is used for indexing data [4].

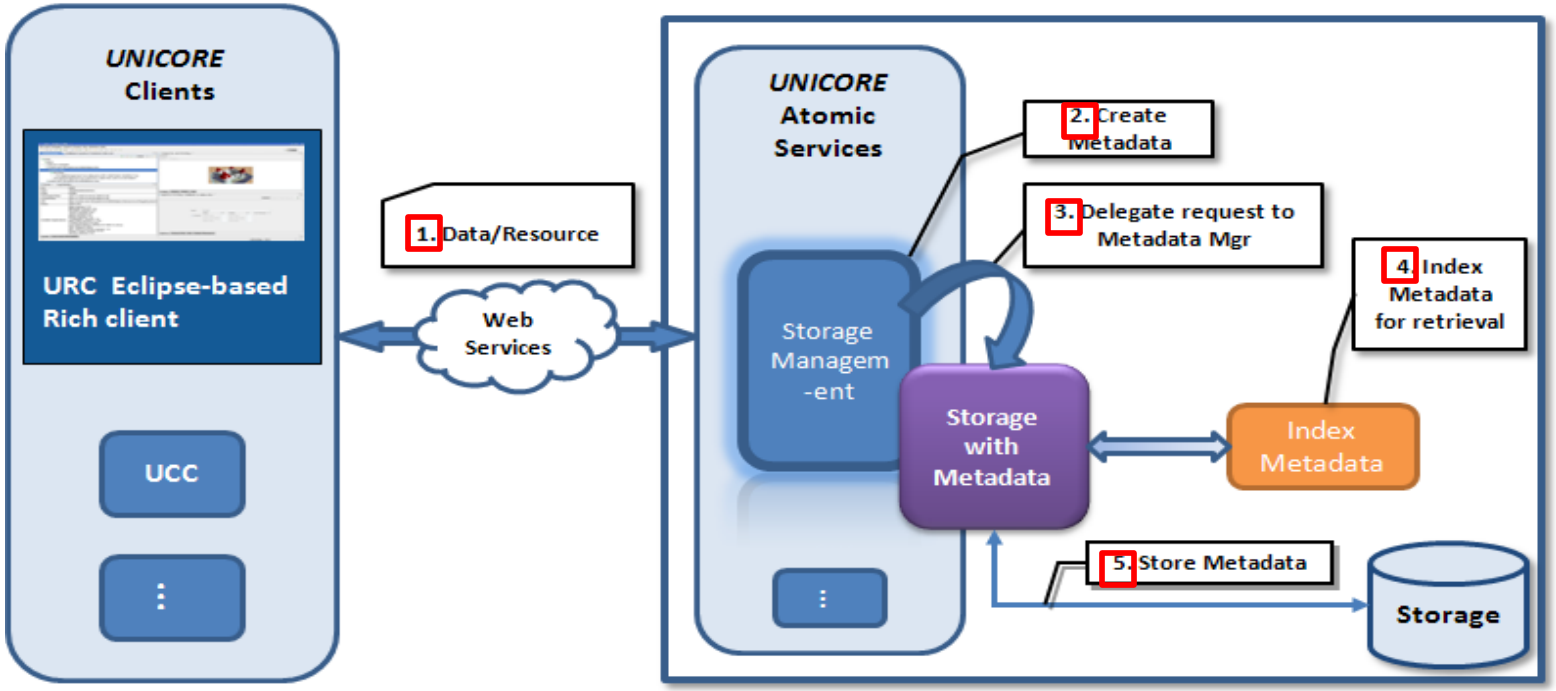
Apache Lucene (2.9.0)

- Open Source, cross-platform full text search engine library
- Indexing, searching and retrieving
- Supports many types of queries
- Scale to millions of documents
- Easy to integrate and use
- Inverted Indexing
  - <docId, term, termFreq, position>



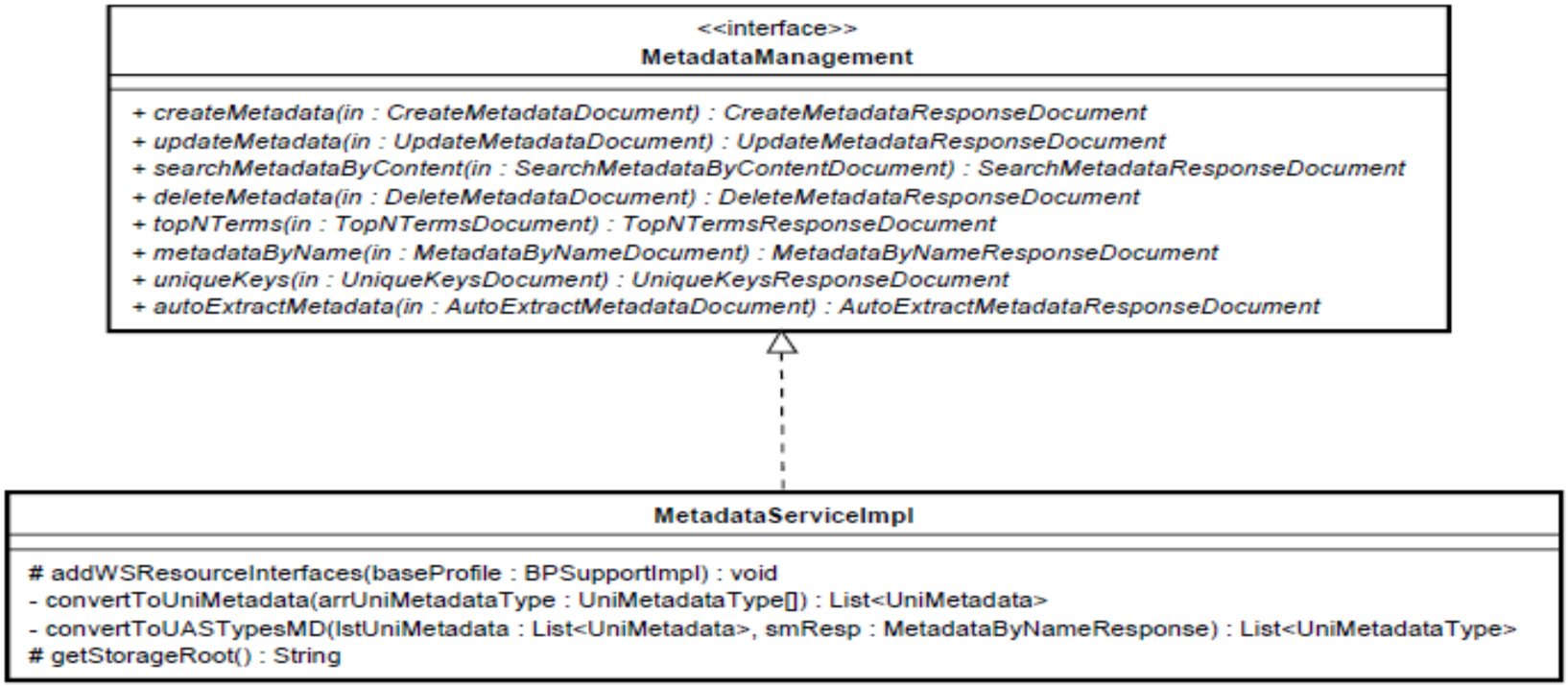
Metadata Management Architecture

Metadata with reference to UNICORE storage service



UCC: UNICORE commandline client

Metadata Management Interface



Implementation

Metadata Service Implementation

- As a *Web Service (WSRF)*
- Apache Lucene for indexing
- Only *English* documents are indexed
- JSON (JavaScript Object Notation) as data structure
- Metadata in UNICORE‘s Storage Management Service (SMS)
- Code added in UNICORE Sourceforge code repository as "**contribution**" code

Metadata Service Implementation

Sample JSON representation of metadata.

```
{
    "resourceName": "/./.TimeSeriesData",
    "type": “txt“,
    “createdDate”:Date/Time,
    “owner”: “<username>”,
    “experimentPlace”: “FZJ”,
    “key...”: “value...”,
    “description”:”The multiple dipole construction method using MUSIC”,
    “tags”: “Time series, parallel approach, estimation, noise,  auditory
      experiment”
}
```

Performance & Evaluation

Evaluation – Search Comparison

Search options as comparison with UNICORE' SMS.

Search by	Storage Management Service (SMS)	*Extended* Storage Management Service
Basic File Metadata (filename, creation date etc)	✓	✓
Boolean Query	Partially (AND, OR)	✓
Wild Characters search on text	Partially (filename only)	✓
Range Query	Partially (date only)	✓
Fuzzy Query	X	✓
Full Text Search	X	✓

Evaluation – Testing Environment

The following configuration is used for evaluation and performance.

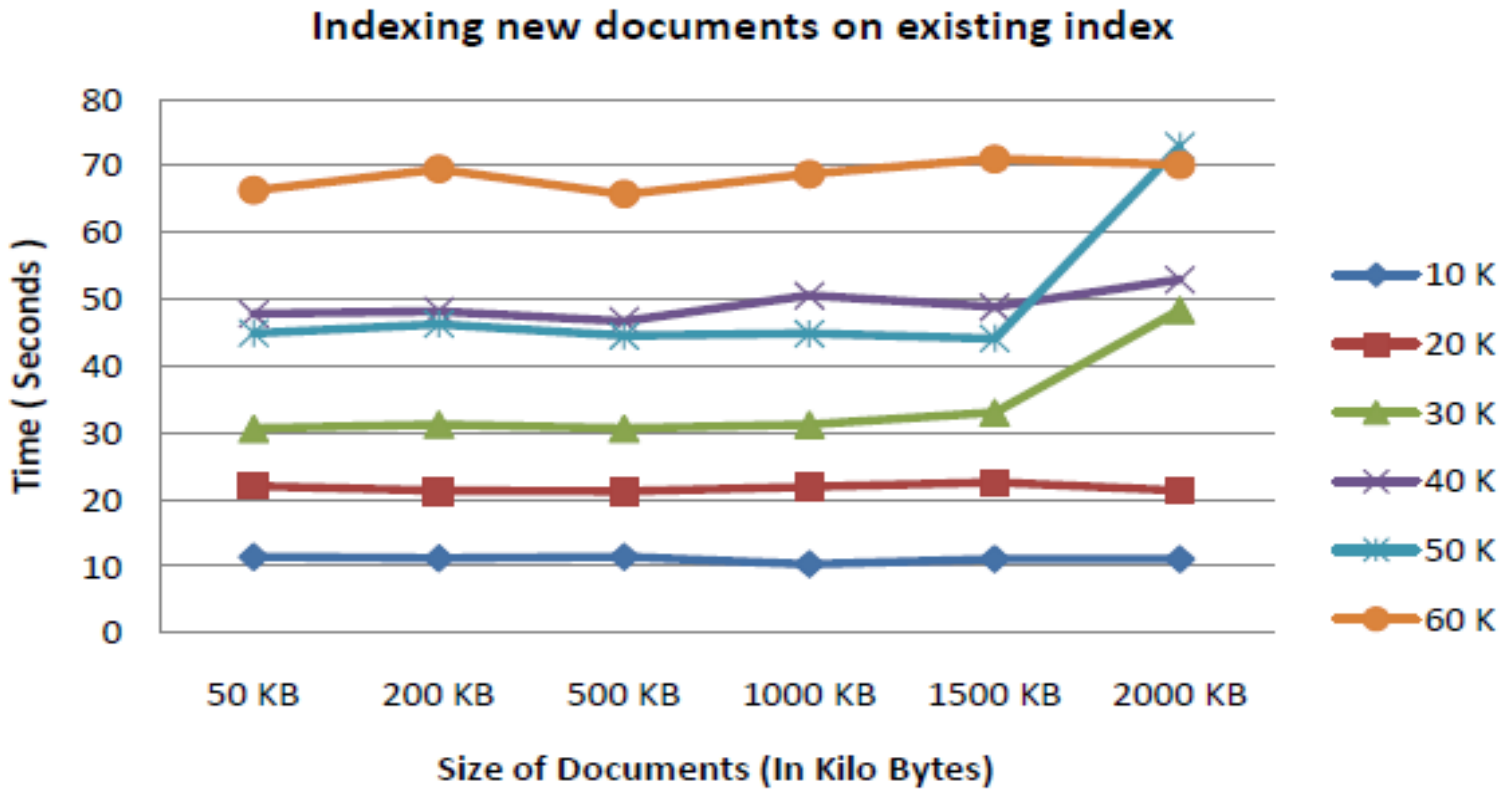
Hardware environment	
CPU:	Intel(R) Core(TM)2 Quad CPU Q9400 @ 2.66GHz
RAM:	4GB
OS:	OpenSUSE 11.2
Software environment	
Java Version:	1.6
Lucene Version:	2.9.0
Analyzer:	StandardAnalyzer
General settings	
Index Storage Place	Local system hard drive:
Documents Language:	English
Dedicated machine for index:	No

Evaluation – Indexing Time

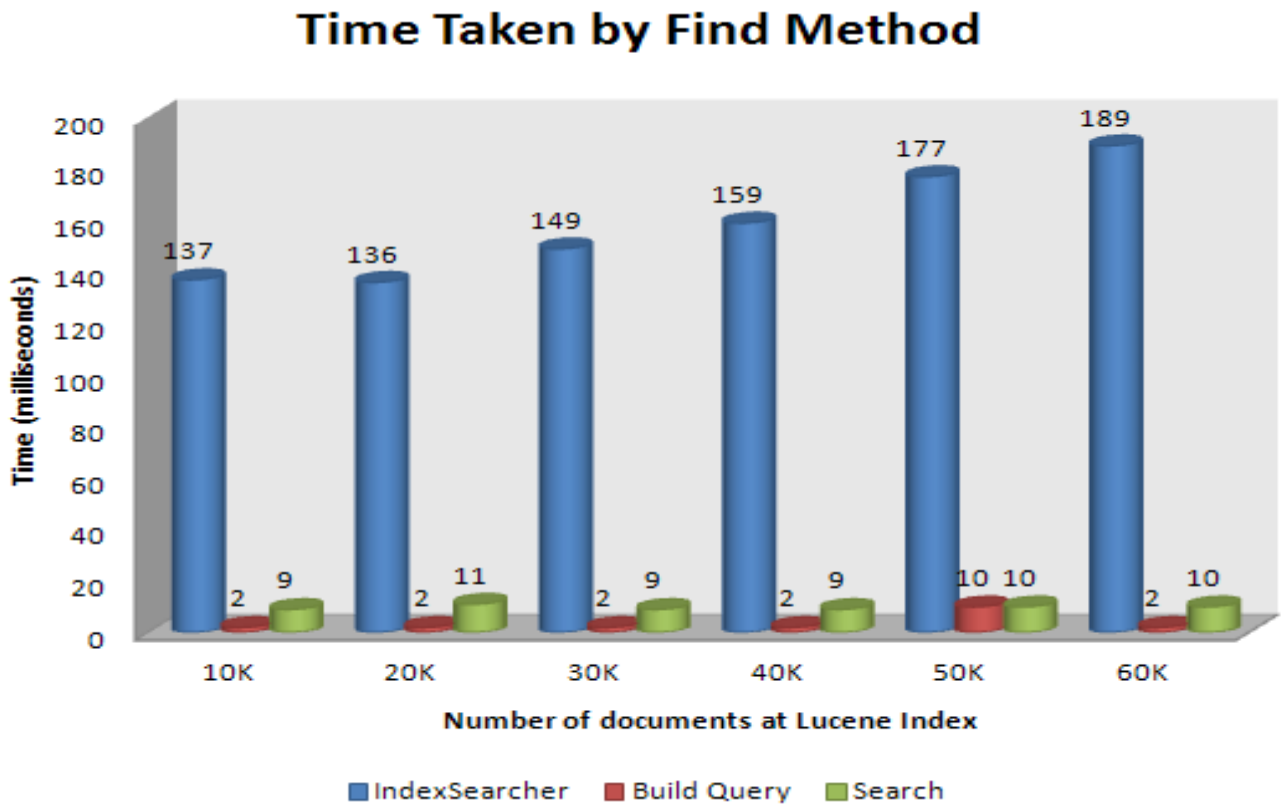
- # of documents varied, size of documents same (~48 KB).

Documents (K)	Time to index (Seconds)
0.1	0.37
1	1.7
10	3.3
50	10
100	17
500	132

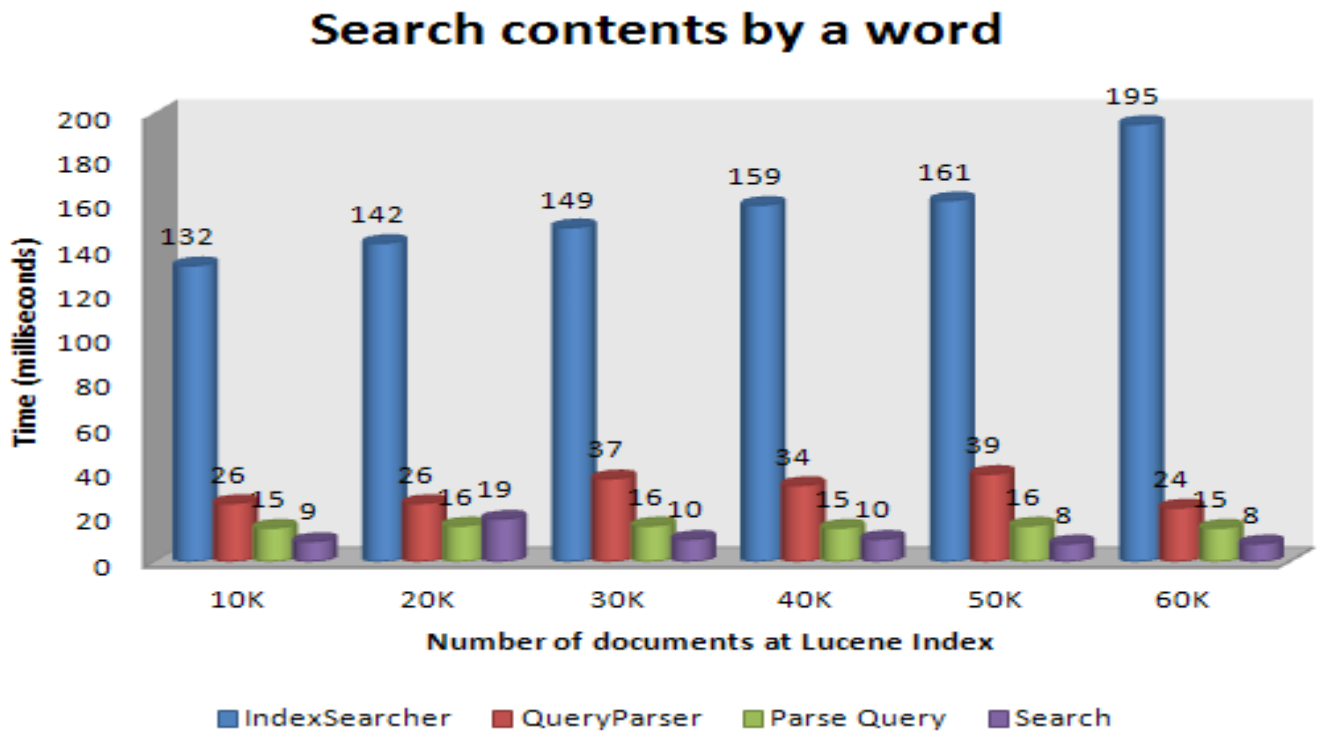
Evaluation – Re-open Index



Evaluation – Search by key/value pair

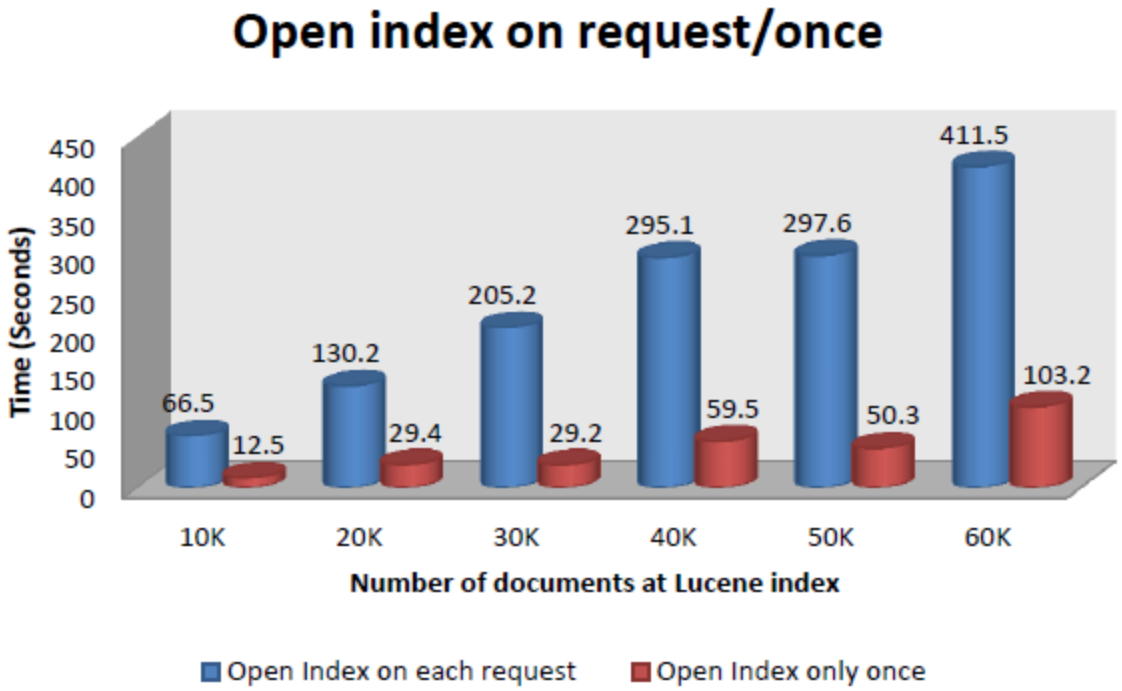


Evaluation – Search by content




```
SELECT documents FROM index where index contains 'word'
```

Evaluation – Performance enhancement



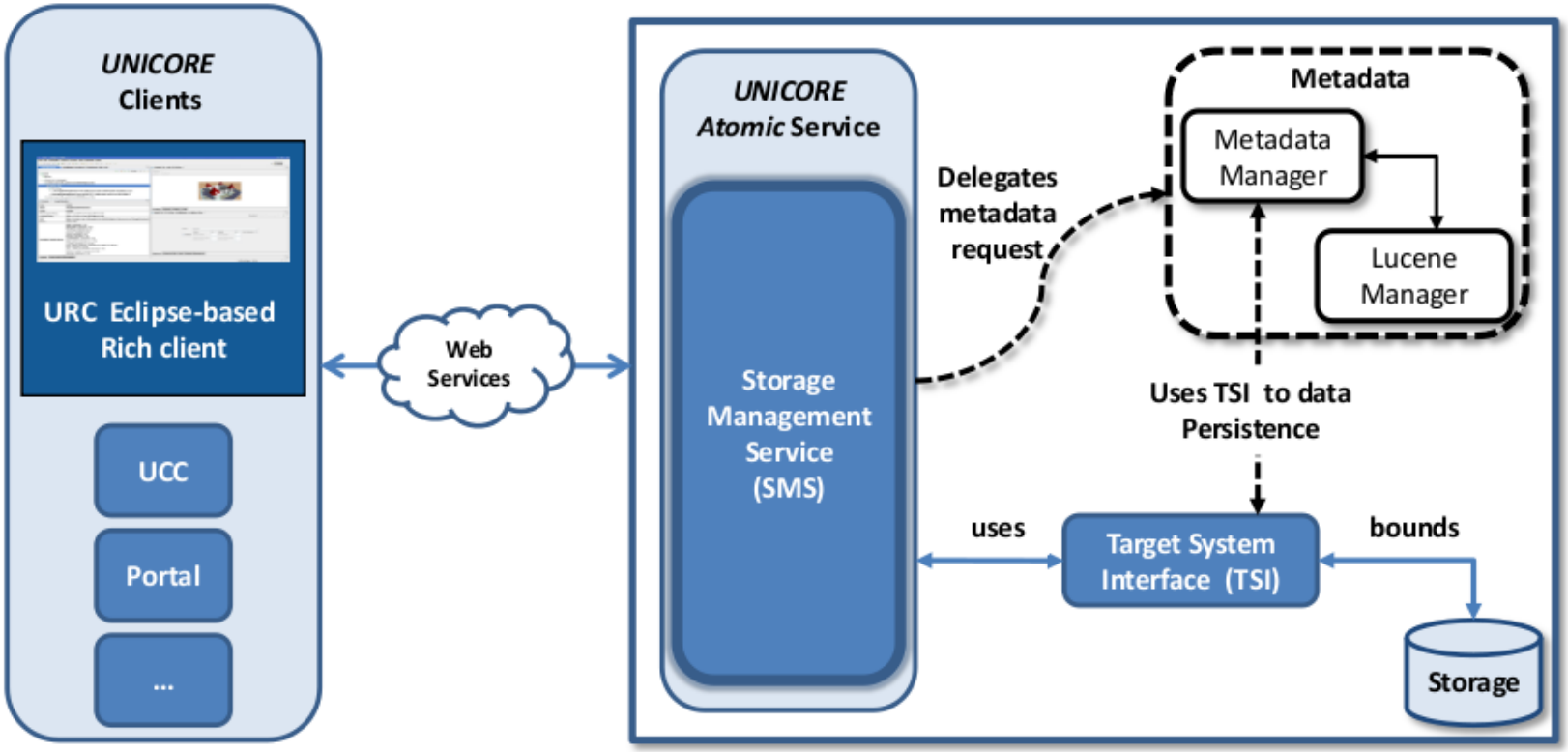
- 50 KB
- 200 KB
- 500 KB
- 1000 KB
- 1500 KB
- 2000 KB

Approx. 4-5 times indexing time is reduced.

Usage Scenarios

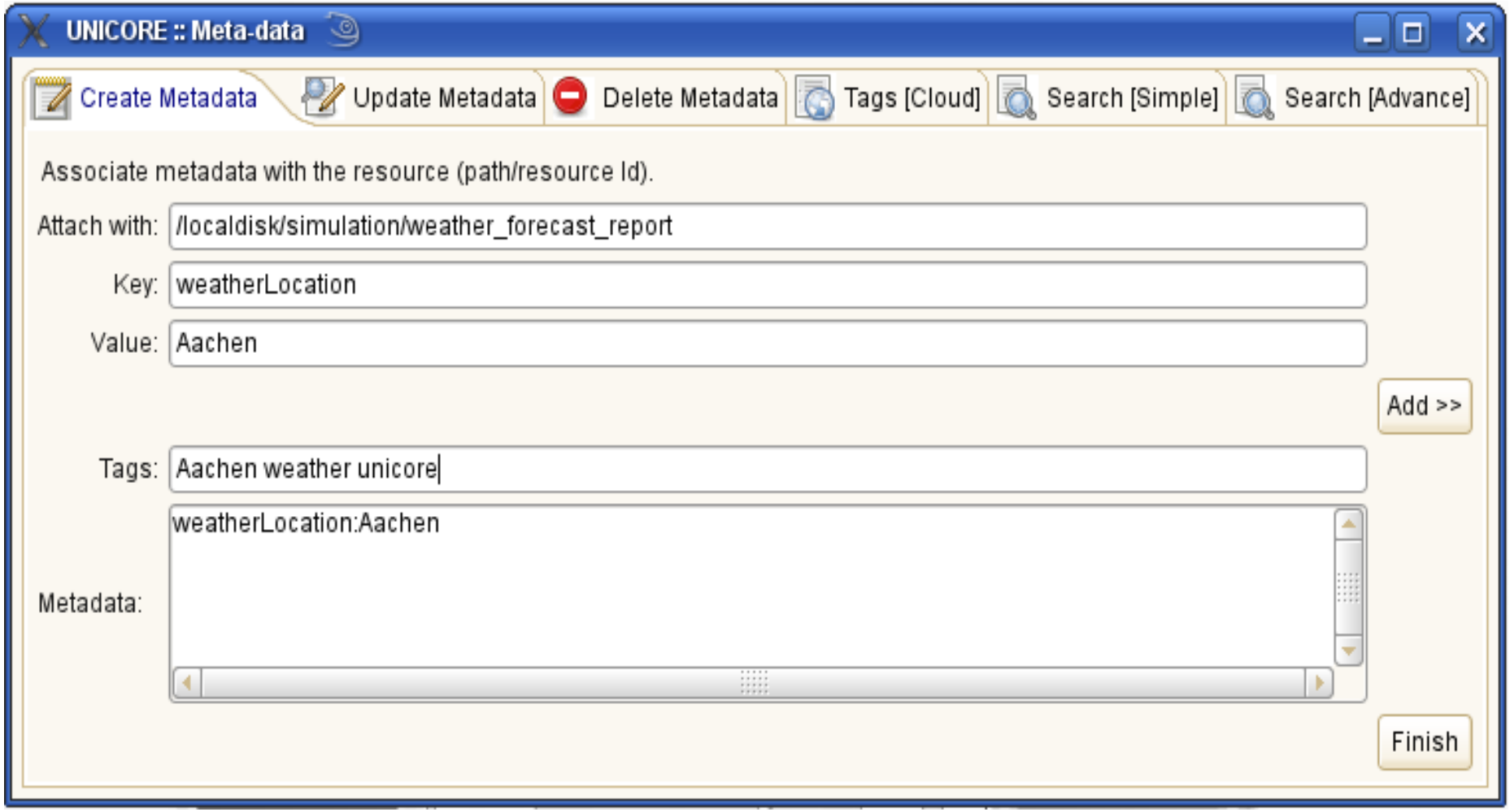
Usage Scenario 1 – Within UNICORE

Metadata Management integrated in UNICORE‘s Storage Service.

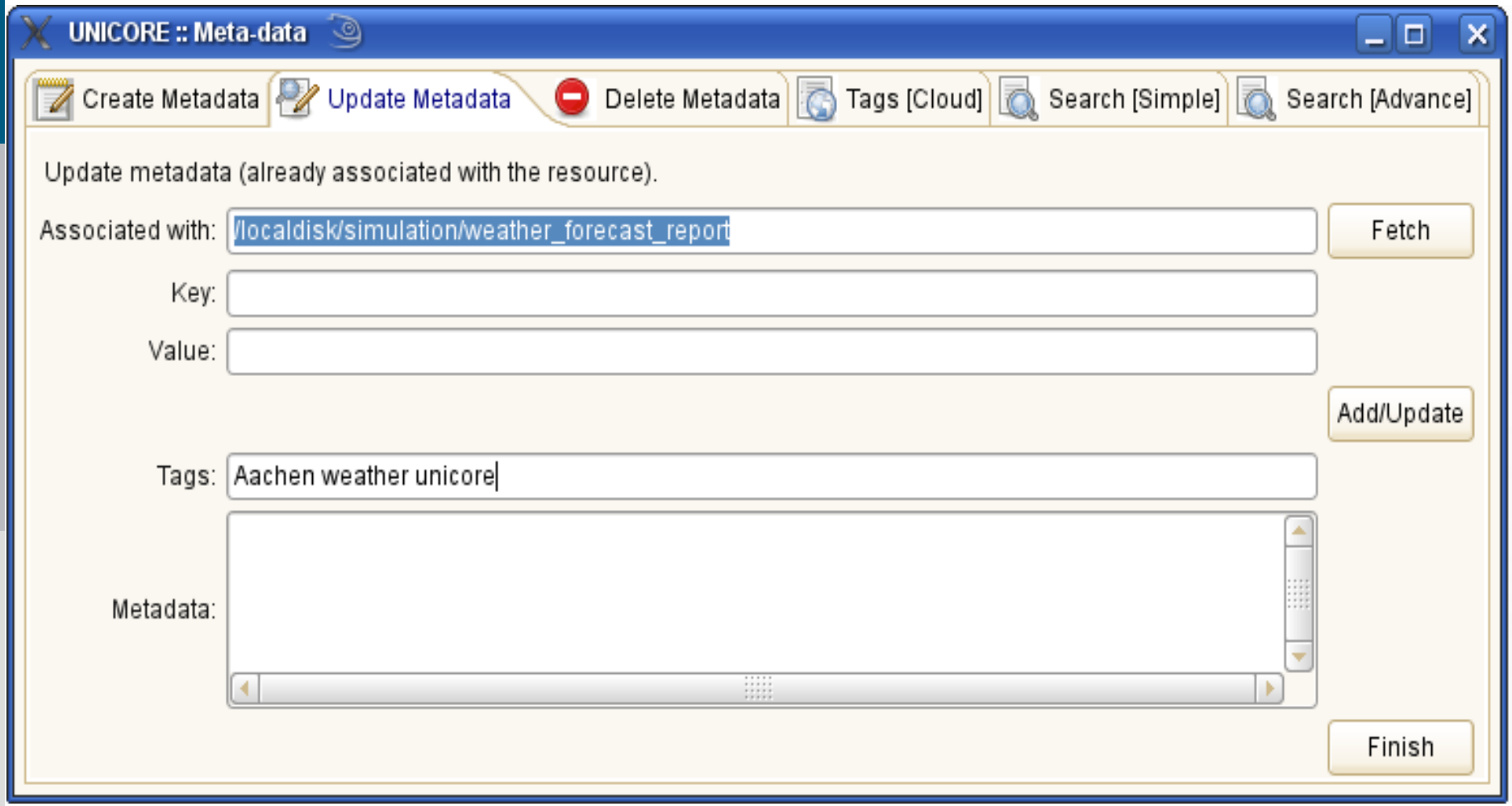


UCC: UNICORE commandline client

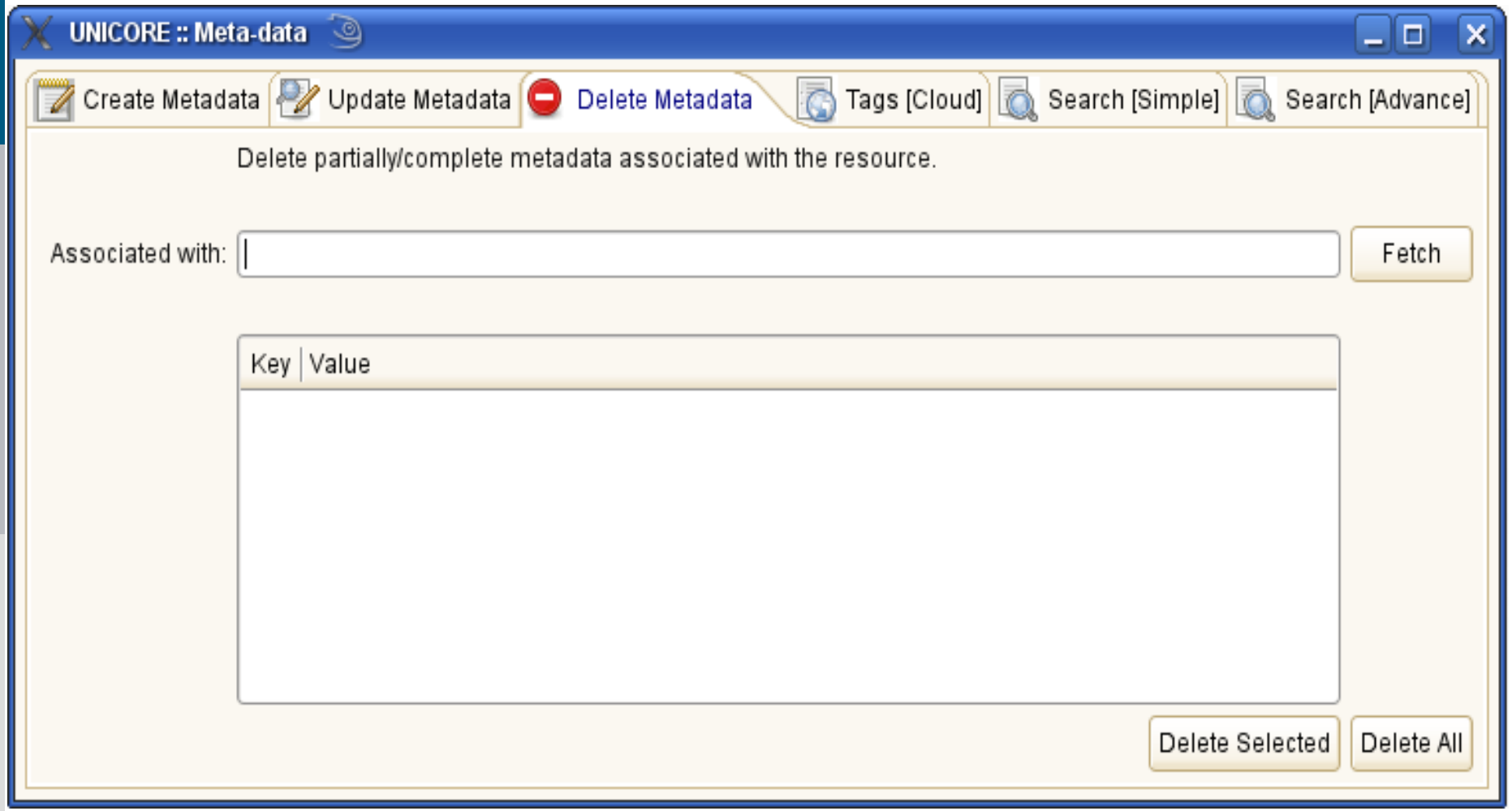
Usage Scenario 1 – Create Metadata



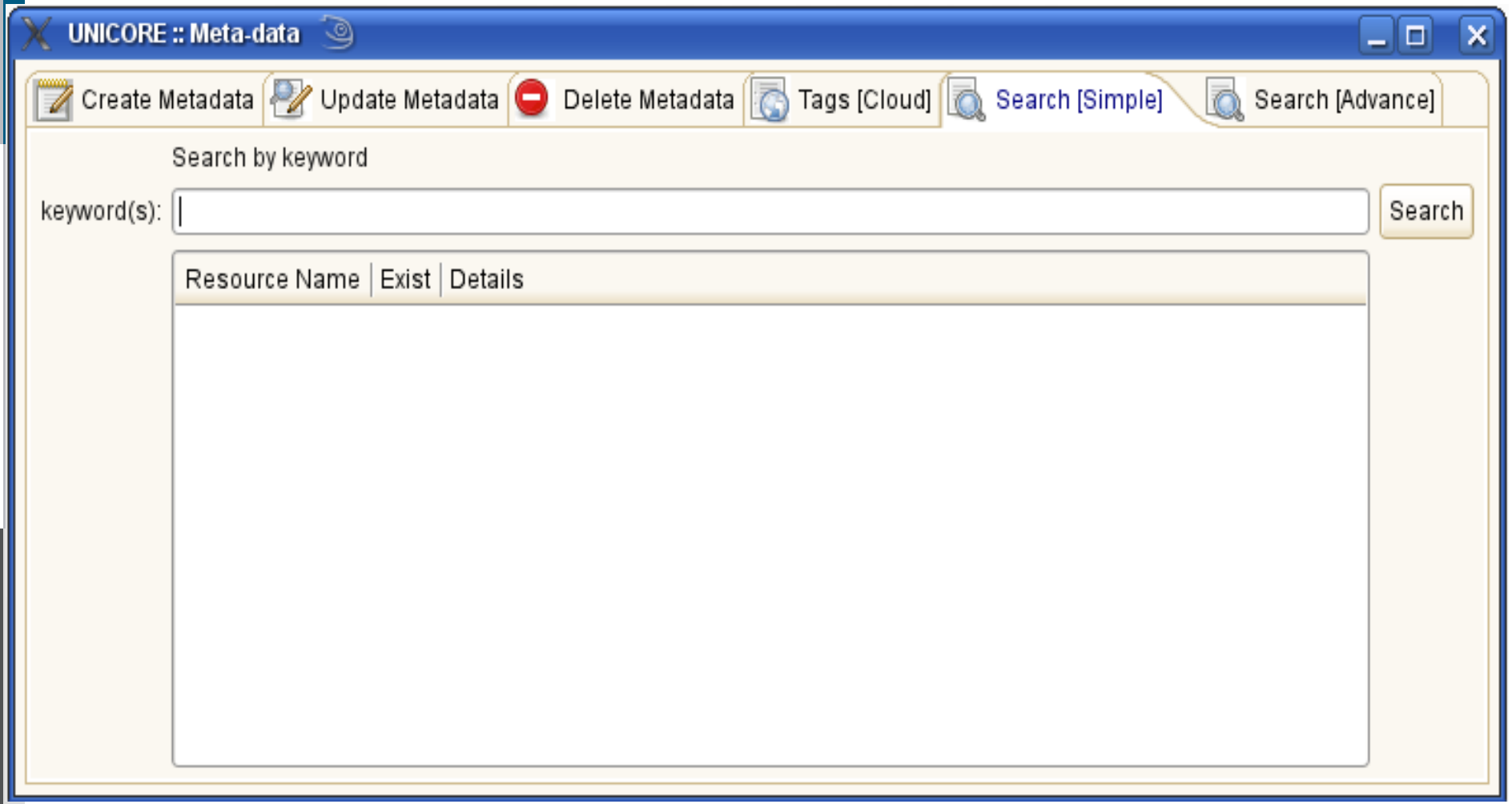
Usage Scenario 1 – Update Metadata



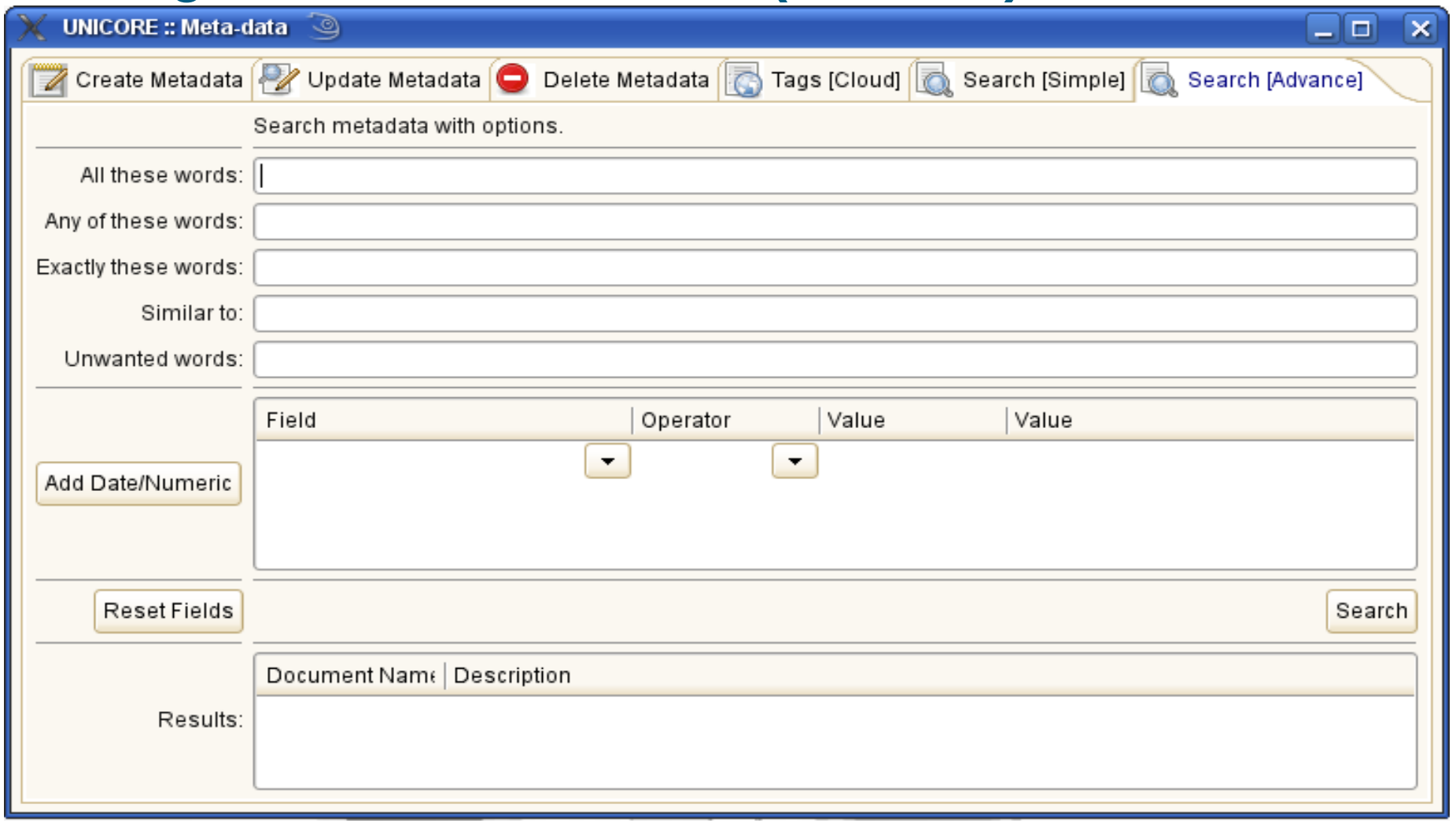
Usage Scenario 1 – Delete Metadata



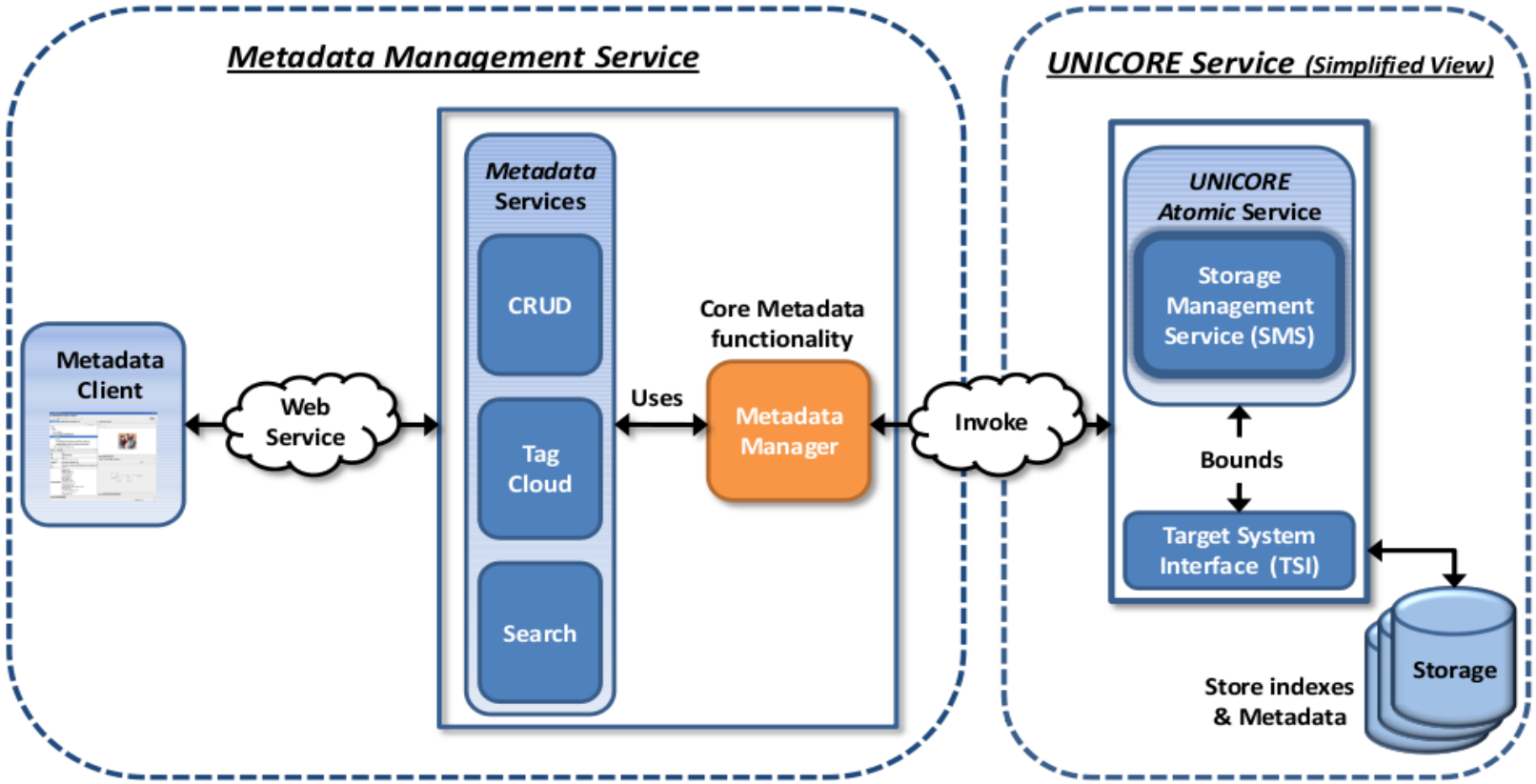
Usage Scenario 1– Search (simple)



Usage Scenario 1– Search (advance)



Usage Scenario 2 – Catalog Service



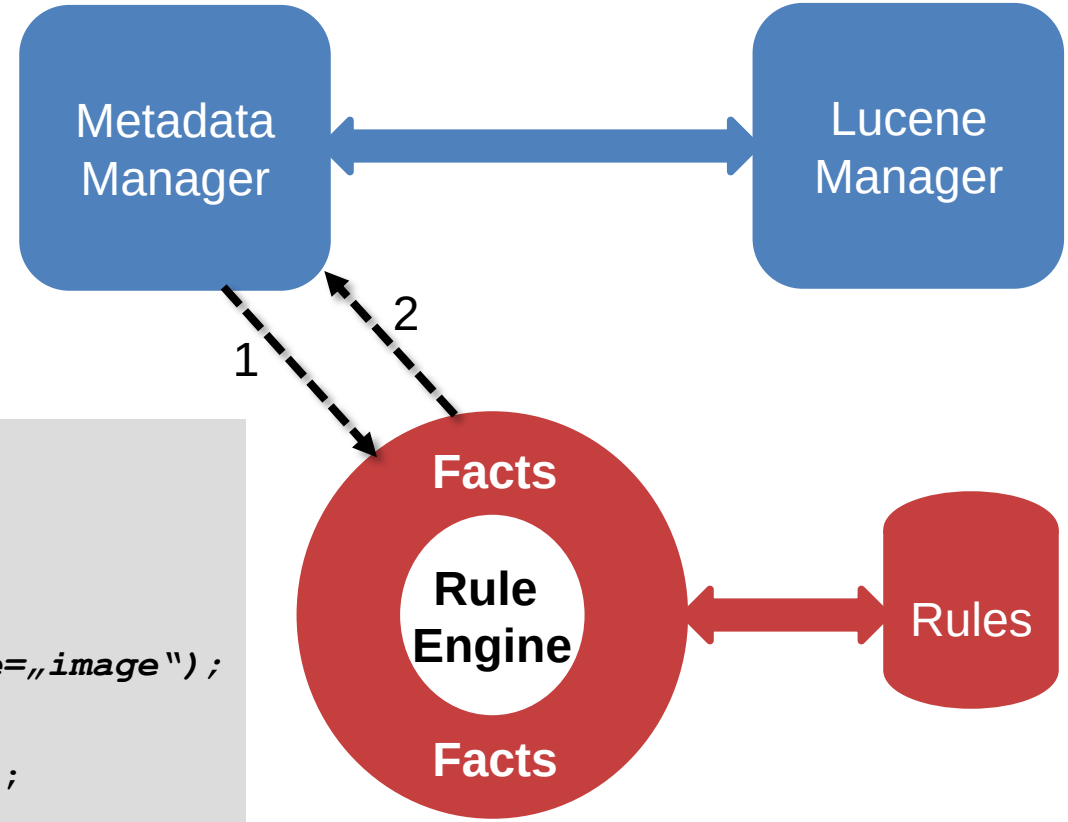
Future Work 1

- More types of languages support for indexing.
- Asynchronous process for indexing data.
- Integration of Metadata Graphical Client (MGC) with UNICORE Rich Client (**URC**).
- Enabling metadata support for other UNICORE services.

Future Work 2

Rule based selection of data to be indexed.

- 1. Forward request (what kind of data to be indexed?)
- 2. Based on *rules*, rule engine respond to Matadata Manager.




```
// Rules

Rule „file_type_image“
     when
          file : new File(type=„image“);
     then
          return „Not Allowed“;
end
```

References

[1] http://uvos.chemomentum.org/index.html
[2] http://amga.web.cern.ch/amga/
[3] http://www.fz-juelich.de/jsc/grid/#GridRDProjects
[4] http://lucene.apache.org/

Questions ?