UNICORE Data Management: Recent Advancements

K. Benedyczak T. Rękawek J. Rybicki B. Schuller

July 8, 2011 | Jędrzej Rybicki

Data wave is coming!

The coming years will be marked by an increasing amount of data produced and processed ("wave of data"[1]):

- global, diverse, valuable and complex data
- science is both producer and consumer of this data

[1] "Riding the wave: How Europe can gain from the rising tide of scientific data"

Data wave is coming!

The coming years will be marked by an increasing amount of data produced and processed ("wave of data"[1]):

- global, diverse, valuable and complex data
- science is both producer and consumer of this data
- . . . humanities also collect the data

[1] "Riding the wave: How Europe can gain from the rising tide of scientific data"

Data wave is coming!

The coming years will be marked by an increasing amount of data produced and processed ("wave of data"[1]):

- global, diverse, valuable and complex data
- science is both producer and consumer of this data
- . . . humanities also collect the data

Examples:

- The Virtual Human Brain: 50 billion neurons, a neuron can possess up to 15,000 synapses

[1] "Riding the wave: How Europe can gain from the rising tide of scientific data"

Data wave is coming!

The coming years will be marked by an increasing amount of data produced and processed ("wave of data"[1]):

- global, diverse, valuable and complex data
- science is both producer and consumer of this data
- . . . humanities also collect the data

Examples:

- The Virtual Human Brain: 50 billion neurons, a neuron can possess up to 15,000 synapses
- Medical data amounts to 30 % of the data produced
- 2.5 PB of mammograms are stored in the U. S. alone

[1] "Riding the wave: How Europe can gain from the rising tide of scientific data"

UNICORE

UNICORE (Uniform Interface to Computing Resources)

UNICORE

UNICORE (Uniform Interface to Computing Resources)

Question

How UNICORE can handle large amounts of data and support data-oriented scientific processes?

UNICORE

UNICORE (Uniform Interface to Computing Resources)

Question

How UNICORE can handle large amounts of data and support data-oriented scientific processes?

What are the UNICORE capabilities to:

- store,
- transfer,
- and manage large amounts of data.

Data storage

Problem: The data must be stored somewhere

Somewhere is the crucial word here. The user usually doesn't care as long as a seamless access to the data is granted.

Data storage

Problem: The data must be stored somewhere

Somewhere is the crucial word here. The user usually doesn't care as long as a seamless access to the data is granted.

UNICORE solution: Distributed Storage (**dSMS**)

- hides the complexity from the user: well-known SMS abstraction
- single "access point" for the users
- . . . which can be replicated for redundancy and load balancing
- flexibility (in adding new resources)

Data transfer

Problem: How to move the data from one place to the other?

As usually:

- the user doesn't care: she wants to just move the data quickly from one place to the other
- the admin doesn't care: she doesn't want to change anything (on firewall)

Data transfer

Problem: How to move the data from one place to the other?

As usually:

- the user doesn't care: she wants to just move the data quickly from one place to the other
- the admin doesn't care: she doesn't want to change anything (on firewall)

UNICORE solution: **UFTP**

- dynamic firewall port opening using a pseudo FTP connection
- parallel input/output streams

Data transfer

Problem: How to move the data from one place to the other?

As usually:

- the user doesn't care: she wants to just move the data quickly from one place to the other
- the admin doesn't care: she doesn't want to change anything (on firewall)

UNICORE solution: **UFTP**

- dynamic firewall port opening using a pseudo FTP connection
- parallel input/output streams
- new feature in UNICORE: scheduled transfers

Problem: How to organize and manage the data?

Where are the results of my simulation from 16/05/2005? I need them quickly!

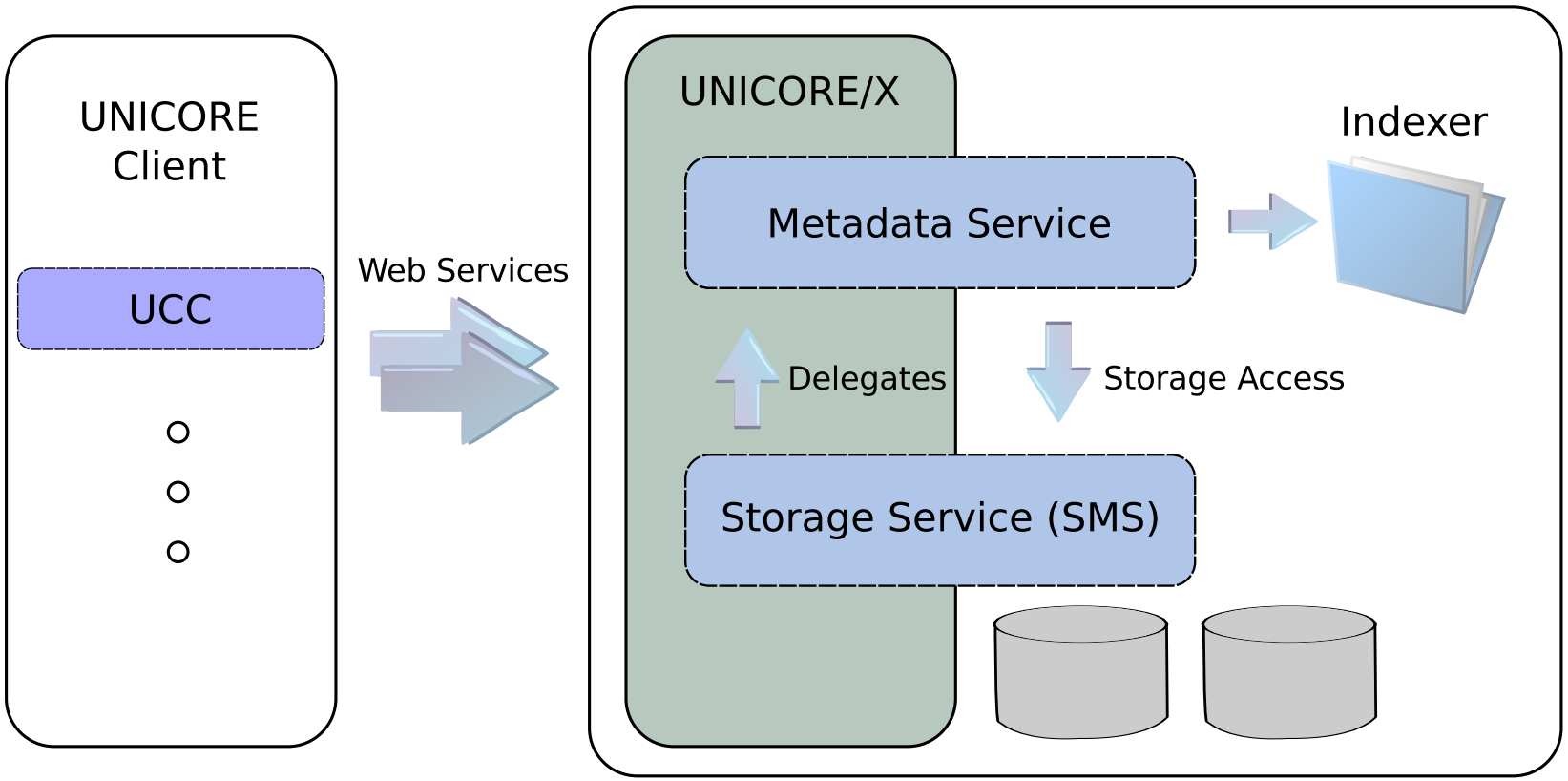
Problem: How to organize and manage the data?

Where are the results of my simulation from 16/05/2005? I need them quickly!

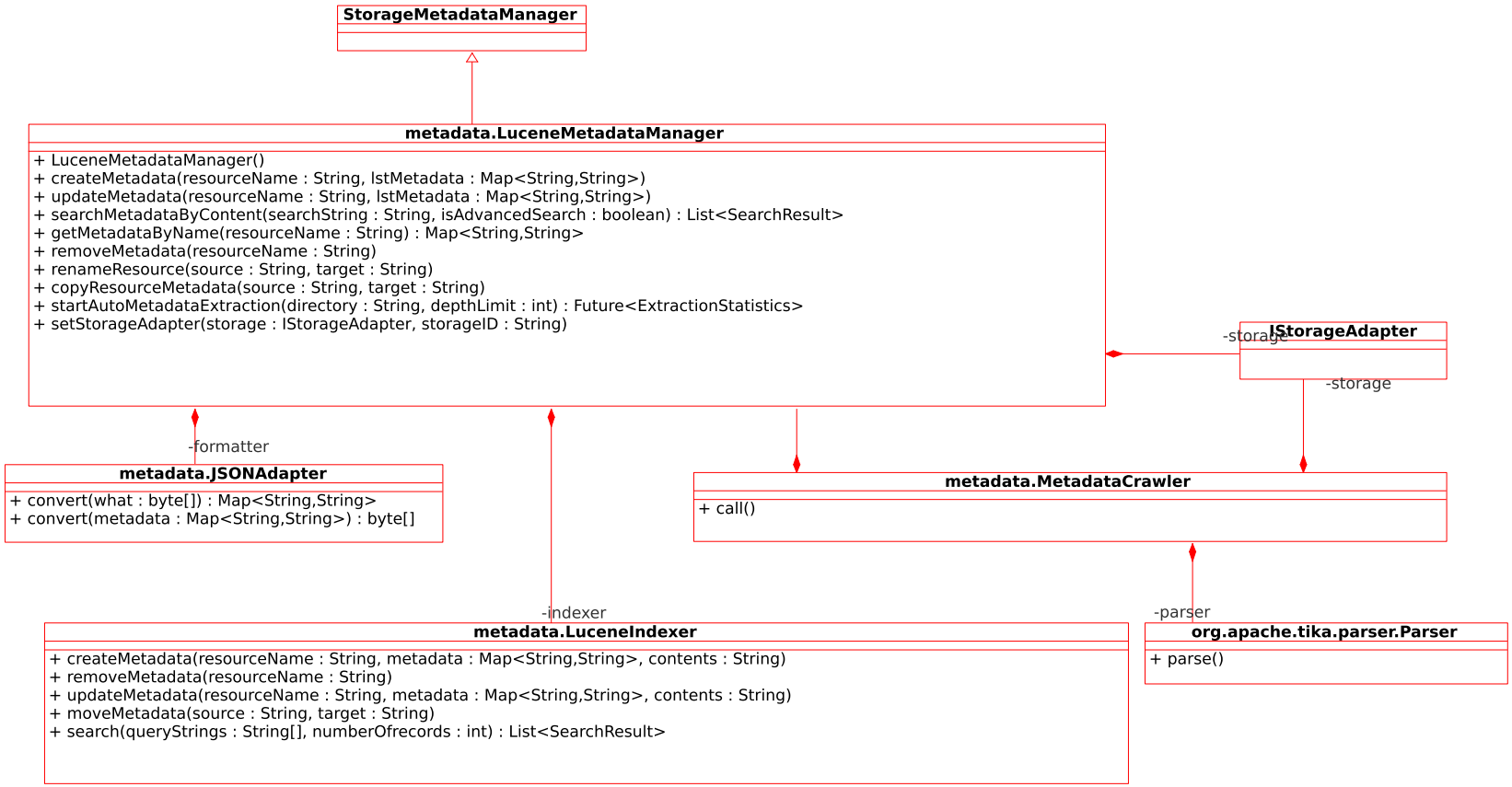
UNICORE solution: Flexible Framework for Metadata Management (**MMF**)

- integrated in the UNICORE Atomic Services
- flexible and extensible
- schemaless
- searchable
- supports automatic extraction

Architecture



Model


StorageMetadataManager
metadata.LuceneMetadataManager
+ LuceneMetadataManager()
+ createMetadata(resourceName : String, lstMetadata : Map<String,String>)
+ updateMetadata(resourceName : String, lstMetadata : Map<String,String>)
+ searchMetadataByContent(searchString : String, isAdvancedSearch : boolean) : List<SearchResult>
+ getMetadataByName(resourceName : String) : Map<String,String>
+ removeMetadata(resourceName : String)
+ renameResource(source : String, target : String)
+ copyResourceMetadata(source : String, target : String)
+ startAutoMetadataExtraction(directory : String, depthLimit : int) : Future<ExtractionStatistics>
+ setStorageAdapter(storage : IStorageAdapter, storageID : String)
IStorageAdapter
-storage
-storage
-formatter
metadata.JSONAdapter
+ convert(what : byte[]) : Map<String,String>
+ convert(metadata : Map<String,String>) : byte[]
metadata.MetadataCrawler
+ call()
-indexer
-parser
metadata.LuceneIndexer
+ createMetadata(resourceName : String, metadata : Map<String,String>, contents : String)
+ removeMetadata(resourceName : String)
+ updateMetadata(resourceName : String, metadata : Map<String,String>, contents : String)
+ moveMetadata(source : String, target : String)
+ search(queryStrings : String[], numberOfrecords : int) : List<SearchResult>
org.apache.tika.parser.Parser
+ parse()

JÜLICH
FORSCHUNGSZENTRUM

Demo

Functionality I

UNICORE MetadataService

- whole storage interactions via IStorage interface (integration)
- metadata are stored both in:
  1. Storage (as files with `.metadata` extension)
  2. Lucene Index
- JSON representation without schema

Apache Lucene

- high-performance, full-featured text search engine
- advanced queries: wildcard, range, compound, proximity

Functionality II

Apache Tika

- toolkit for detecting and extracting of metadata and structured text content from various documents
- supports: `html/xml`, `doc/odt`, `pdf`, `rtf`, `zip`, `midi`, `mp3`, `tiff/jpg`, `flc`, `java`, `dwg`, `ttf`
- extensible (very simple interface)

Data Management in UNICORE

- Store in **dSMS**
- Transfer with **UFTP**
- Describe and search with **MMF**

Data Management in UNICORE

- Store in **dSMS**
  - ⇒ proper handling of sensitive data
  - ⇒ keeping data close to computing resources
- Transfer with **UFTP**
  - ⇒ automatic transfer protocol negotiation
  - ⇒ deployments
- Describe and search with **MMF**
  - ⇒ convenient ways to provide own parsers
  - ⇒ URC integration


UNICORE

Thanks

j.rybicki@fz-juelich.de