Dziedzinowo zorientowane usługi i zasoby infrastruktury PL-Grid dla wspomagania Polskiej Nauki w Europejskiej Przestrzeni Badawczej

Lightweight Metadata and Data Management with DataNet

Daniel Harężlak[1], Marek Kasztelnik[1], Maciej Pawlik[1], Bartosz Wilk[1], Marian Bubak[1,2]

[1]ACC Cyfronet AGH
[2]AGH University of Science and Technology, Institute of Computer Science AGH

Presentation Plan

- Motivation behind DataNet
- Metadata Management Requirements
- Architecture Description
- PL-Grid Deployment
- Conclusions

DataNet – Rationale and Objectives

- Rationale
  - Current data **discoverability** and **reproducibility** for scientific results is poor
  - Data management is a common requirement in computational sciences
  - Workflow and scripting engines provide only little support (GridSpace, Taverna) and make a tight coupling with the enactment engine
  - Each application is different and requires a dedicated metadata/data model

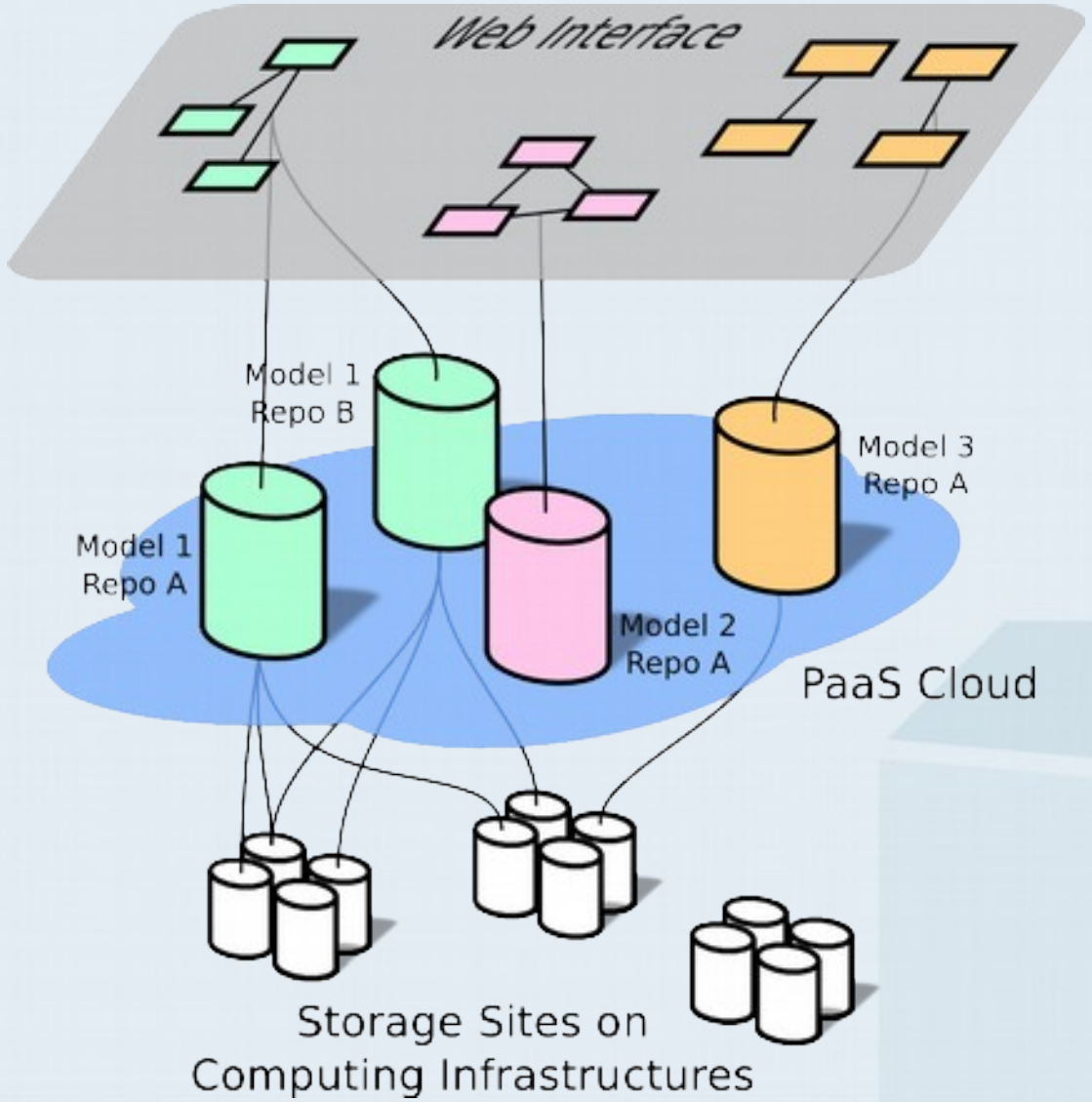
- Objectives
  - Provide means for **ad-hoc metadata model creation** and deployment of corresponding storage facilities
  - Create a research space for **metadata model exchange and discovery** with associated data repositories with access restrictions in place
  - Support **different types of storage sites** and **data transfer protocols**
  - Support the exploratory paradigm by making the models evolve together with data

DataNet – Additional Requirements

- PLGrid infrastructure – supporting different e-Science domains
  - Various applications coming from different scientific communities operate on different metadata and data models
  - Common computational and storage resources are used to store raw data
- Deployment of model data as repositories in PL-Grid cloud
  - Robust enablement of a dedicated interface to compose metadata and data models
  - Access control capabilities to restrict access to data
  - Exploitation of available storage infrastructure for large file-based data sets
- Universal availability of the repository
  - Platform independent metadata and data recording
  - Facilitated by existing standards to support wide range of programming languages

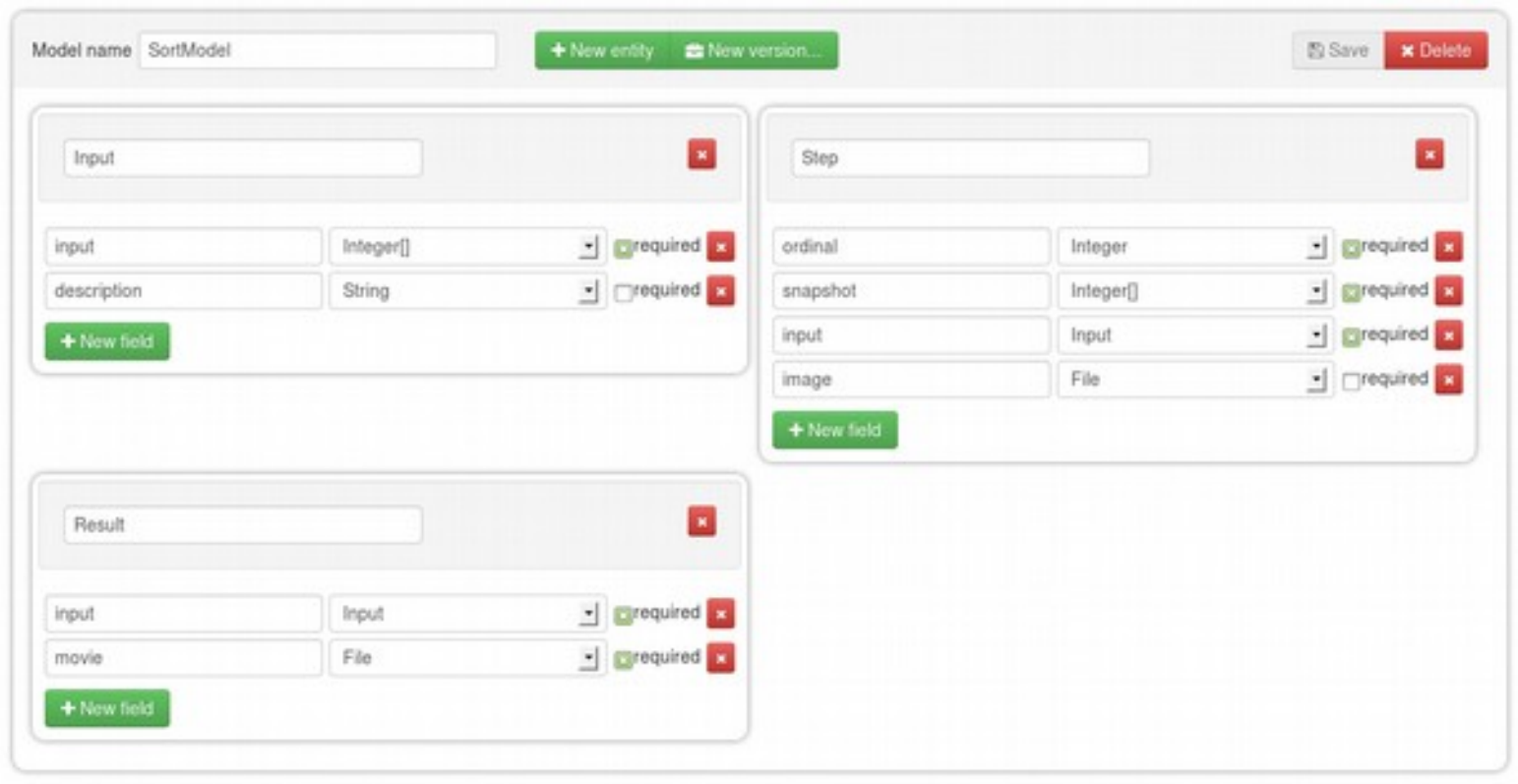
DataNet – Architecture

- **Web Interface** is used by users to create, extend and discover metadata models
- Model repositories are deployed in the **PaaS Cloud** layer for scalable and reliable access from computing nodes through REST interfaces
- Data items from **Storage Sites** are linked from the model repositories



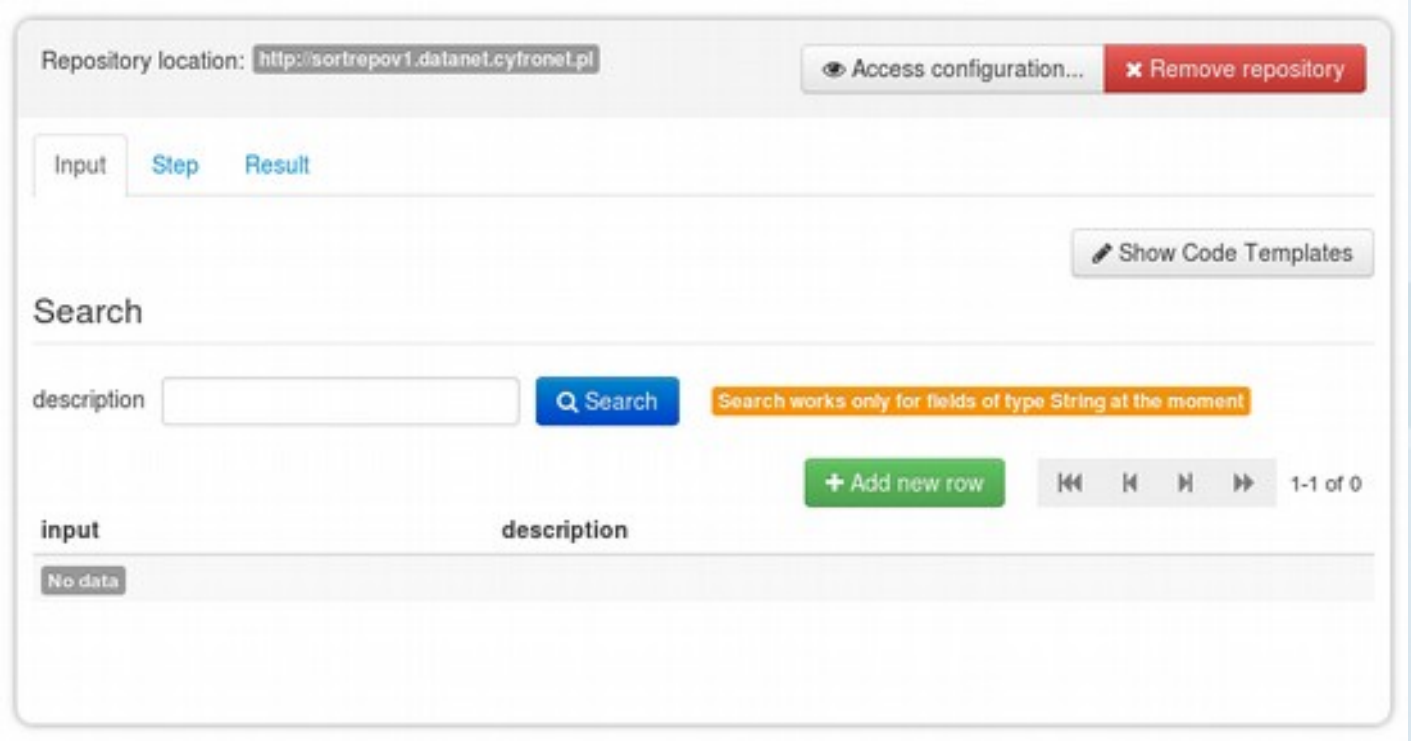
DataNet – Data Model

- Set of entities with fields of different types
  - Simple types
  - Array types
  - File type
  - Relations



DataNet – Repository

- Repositories are accessed through REST
  - Data view through a web application
  - Configurable Access control
    - Public
    - Private (within a group of users)



DataNet – Repository Access

- Data sent over with JSON or FORM
  - REST methods
    - POST – submit new data
    - PUT – modify data
    - DELETE – remove data
    - GET- retrieve data
    - Queries with URL

```
require 'rest-client'
require 'json'

datanet=RestClient::Resource.new('http://a:a@repo.datanet.cyfronet.pl')
datanet.get
def get_user(first_name, last_name)
  {first_name: first_name, last_name: last_name}.to_json
end
get_user "marek", "kasztelnik"
datanet['user'].post get_user("Marek", "Kasztelnik")
datanet["user"].get
10.times {datanet['user'].post get_user("Marek", "Kasztelnik")}
datanet["user"].get
datanet["user/519dbfed2fbb0c79f400000b"].delete
datanet["user"].get
```

```
import requests as req
import json

headers = {'content-type': 'application/json'}
resp = req.post('http://test5.datanet.cyfronet.pl/Hello',
        data = json.dumps({'name': 'hello1'}), auth = ('', ''), headers = headers)
```

Contact details

- Acknowledgements
  - This research has been partially supported by the European Regional Development Fund program no. POIG.02.03.00-00-096/10 as part of the PL-Grid PLUS project

- Contact us and help make DataNet better

- Visit http://dice.cyfronet.pl for more information

- See DataNet in action at https://datanet.cyfronet.pl (PLGrid account required)