VRE for regional Interdisciplinary communities in Southeast Europe and the Eastern Mediterranean

Data Discovery Service for VI-SEEM

Vladimir Dimitrov
Institute of Information and Communication Technologies (IICT-BAS)
Agenda

- VI-SEEM Data Discovery Service (DDS)
- Software implementation
- Hardware implementation
- Features
- DDS front end
- Example screens
- Data Synchronization Tool
- Conclusion

Total of 18 slides.
The VI-SEEM Data Discovery Service provides flexible search functions for *(meta)data(sets)* which are used in the project.

Main access point:

https://search.vi-seem.eu

Backup access point:

https://dds.avitohol.acad.bg

Use case

To make such datasets searchable by means of associating meta data. The datasets are hosted at VI-SEEM Data repository and other places.
The Data Discovery service is based on a customized implementation of CKAN system (https://ckan.org)

CKAN is built with Python on the backend and JavaScript on the frontend, and uses The Pylons web framework and SQLAlchemy (database toolkit for Python, https://www.sqlalchemy.org) as its ORM (Object-Relational Mapping).

Its database engine is PostgreSQL and its search is powered by Apache search platform SOLR (http://lucene.apache.org/solr). CKAN has a modular architecture that allows extensions to be developed to provide additional features such as harvesting or data upload.

CKAN uses its internal model to store metadata about the different records, and presents it on a web interface that allows users to browse and search this metadata. It also offers a powerful API that allows third-party applications and services to be built around it.
Currently VI-SEEM DDS is installed on:

- One physical server with two CPU Intel Xeon 2.6 GHz, 4 physical cores each
- 16 GB RAM
- HDD SATA 250 GB
- 1 GBps UTP Ethernet external connectivity

Hosted and maintained by IICT-BAS.

The service does not seem to be experiencing a staggering hunger for HPC resources for now.
Features

- **User Interface**
  - Web based, for search queries, end-user and system administrators
  - Search and retrieval of items by browsing or searching the metadata
- **Workflow**
  - Enables differing searches for communities (groups) and organizations
  - Models "e-people" who have "roles" in the workflow of a particular Community (group) in the context of a given dataset collection
- **Integrated with VI-SEEM Data Repository**
  - Data Synchronization Tool
- **Access Control**
  - Most of the features are available to all users. However editing particular metadata records requires free registration and requesting a role for particular community or organization.
  - Integrated with the VI-SEEM Login Service (work in progress)
- **Metadata Schema**
  - Follows the guidelines for VI-SEEM data repository
  - Possibility to add new metadata schemas, different than the above
The service is comprised of Organizations and Groups

**Organizations** are the institution members of VI-SEEM project. However, it is easily to add departments, labs, schools etc.

**Groups** correspond to the main VI-SEEM **communities**: Climate Sciences, Digital Cultural Heritage and Life Sciences.

Also there are a Software projects group and common purposes **Generic group**.

Finally there are **groups and organizations** (named Zen*) with testing purposes only. Could be deleted at any time.
Example screen: Organizations

**What are Organizations?**

Organizations are used to create, manage and publish collections of datasets. Users can have different roles within an Organization, depending on their level of authorization to create, edit and publish.

18 organizations found

  - 0 Datasets
- **Cyl** - The Cyprus Institute Cyprus [https://www.cyi.ac.cy](https://www.cyi.ac.cy)
  - 1 Dataset
- **GRENA** - GEORGIAN RESEARCH AND EDUCATIONAL NETWORKING ASSOCIATION Georgia
  - 6 Datasets
- **GRNET** - Greek Research and Technology Network S.A. Greece [https://www.grnet.gr/en](https://www.grnet.gr/en)
  - 1 Dataset
- **IICT** - Institute of Information and Communication Technologies
  - 1 Dataset
Example screen: Groups
Example screen: simple browsing datasets
Example search
Example dataset view
Example: User registration form

```
Username: username
Full Name: Joe Bloggs
Email: joe@example.com
Password: ********
Confirm: ********
```
Example: Add/create a new dataset
VI-SEEM datasets synchronization tool

- Synchronizes VI-SEEM repo with VI-SEEM search service.
- Written in Python3.
- Uses only modules from The Python Standard Library.
- Runs in either create or update mode.
  - create is the default behavior and filters already synchronized items, uploading only new items from repo to search.vi-seem.eu(create);
  - update mode – if specified the metadata of all items will be checked and updated.
- Every 24 hours, 2 cron jobs for each mode automatically synchronize data.
- Records every unsuccessful upload in a log file.
- Developed and supported by IICT-BAS.
Synchronization with VI-SEEM Data repository
brief workflow
The tool is conversing the metadata tags of each item in the following way:

<table>
<thead>
<tr>
<th>Data Repository (DSpace)</th>
<th>Data Discovery Service (CKAN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>title</td>
<td>Name</td>
</tr>
<tr>
<td>lastModified</td>
<td>modified</td>
</tr>
<tr>
<td>dc.identifier.uri</td>
<td>url</td>
</tr>
<tr>
<td>dc.type</td>
<td>type</td>
</tr>
<tr>
<td>dc.contributor.author (multiple)</td>
<td>author</td>
</tr>
<tr>
<td>dc.creator (multiple)</td>
<td>notes</td>
</tr>
<tr>
<td>dc.description</td>
<td>description</td>
</tr>
<tr>
<td>dc.description.abstract</td>
<td>sponsorship</td>
</tr>
<tr>
<td>dc.description.sponsorship</td>
<td>organization</td>
</tr>
<tr>
<td>dc.publisher</td>
<td>rights</td>
</tr>
<tr>
<td>dc.rights</td>
<td>rights_uri</td>
</tr>
<tr>
<td>dc.rights.uri</td>
<td>sponsorship</td>
</tr>
<tr>
<td>dc.description.sponsorship</td>
<td>collection</td>
</tr>
<tr>
<td>collection</td>
<td>subject</td>
</tr>
</tbody>
</table>
The Data Discovery Service is a CKAN based search tool for indexing datasets and metadata which are used during the VI-SEEM project.

It uses custom developed Datasets Synchronization tool for indexing on a regular basis of VI-SEEM Data Repository and possibly different external sources.

Thank you for your attention. Questions?