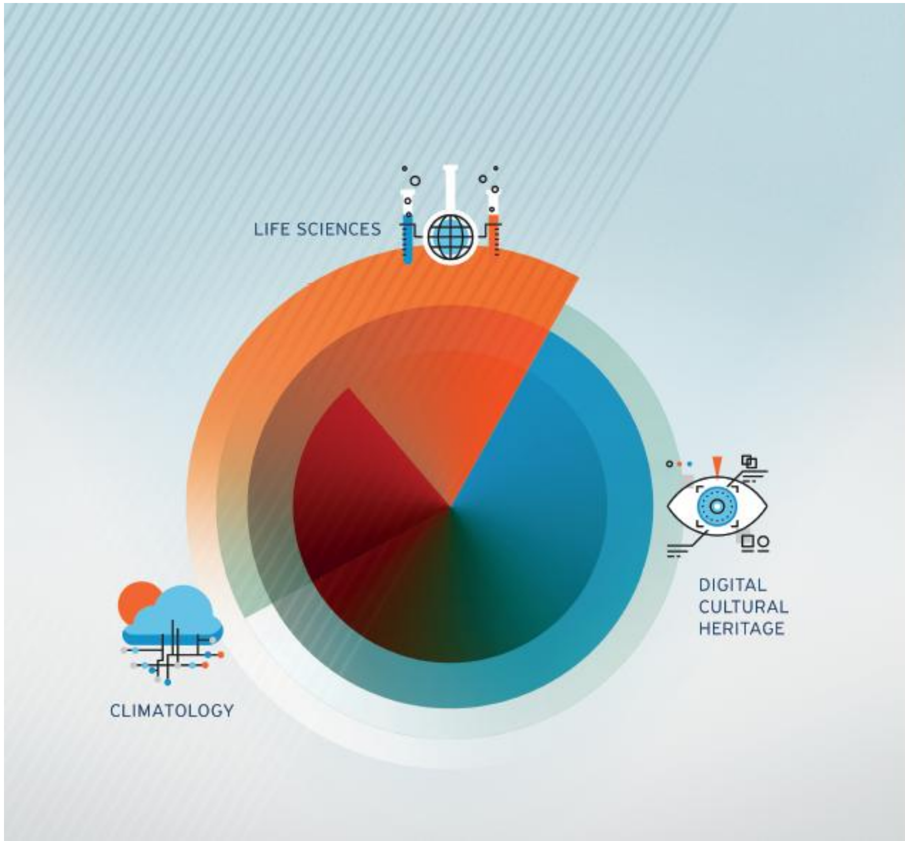
VRE data discovery service : Solutions for Data Discovery Service in a Virtual Research Environment

Vladimir Dimitrov,
Stilyan Stoyanov, IICT-BAS

International conference
“e-Infrastructures for excellent science in Southeast Europe and Eastern Mediterranean”
(**e-Infra4SEEM’18**), 15-16 May 2018, Sofia, Bulgaria



The VI-SEEM project initiative is co-funded by the European Commission under the H2020 Research Infrastructures contract no. 675121

Outline

- ❑ Motivation
- ❑ Data Discovery Service in VRE
- ❑ Software implementation
- ❑ Data Synchronization tool
- ❑ Performance evaluation
- ❑ Conclusion

(10 slides)

Motivation

- Scientific computing requires many and large volumes of complex structured data and metadata that are scattered across data centers.
- Traditional search engines, such as Google, are not effective in most of these cases.
- Some of the scientific data are confidential and are not publicly indexed.
- ✓ This presentation introduces the Data Discovery Service (DDS) solutions designed to serve the Virtual Research Environment (VRE) during the VI-SEEM project.

Data Discovery Service in VRE

- 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`

  (hosted and supported by IICT-BAS)

- Use case
  - To make such datasets searchable by means of associating meta data. The datasets are hosted at VI-SEEM Data repository and other sites oriented to hosting research data.

Data Discovery Service in VRE
Software implementation

Vi-SEEM

- **The Data Discovery Service is based on a customized implementation of CKAN system (**Comprehensive Knowledge Archive Network**, https://ckan.org)**

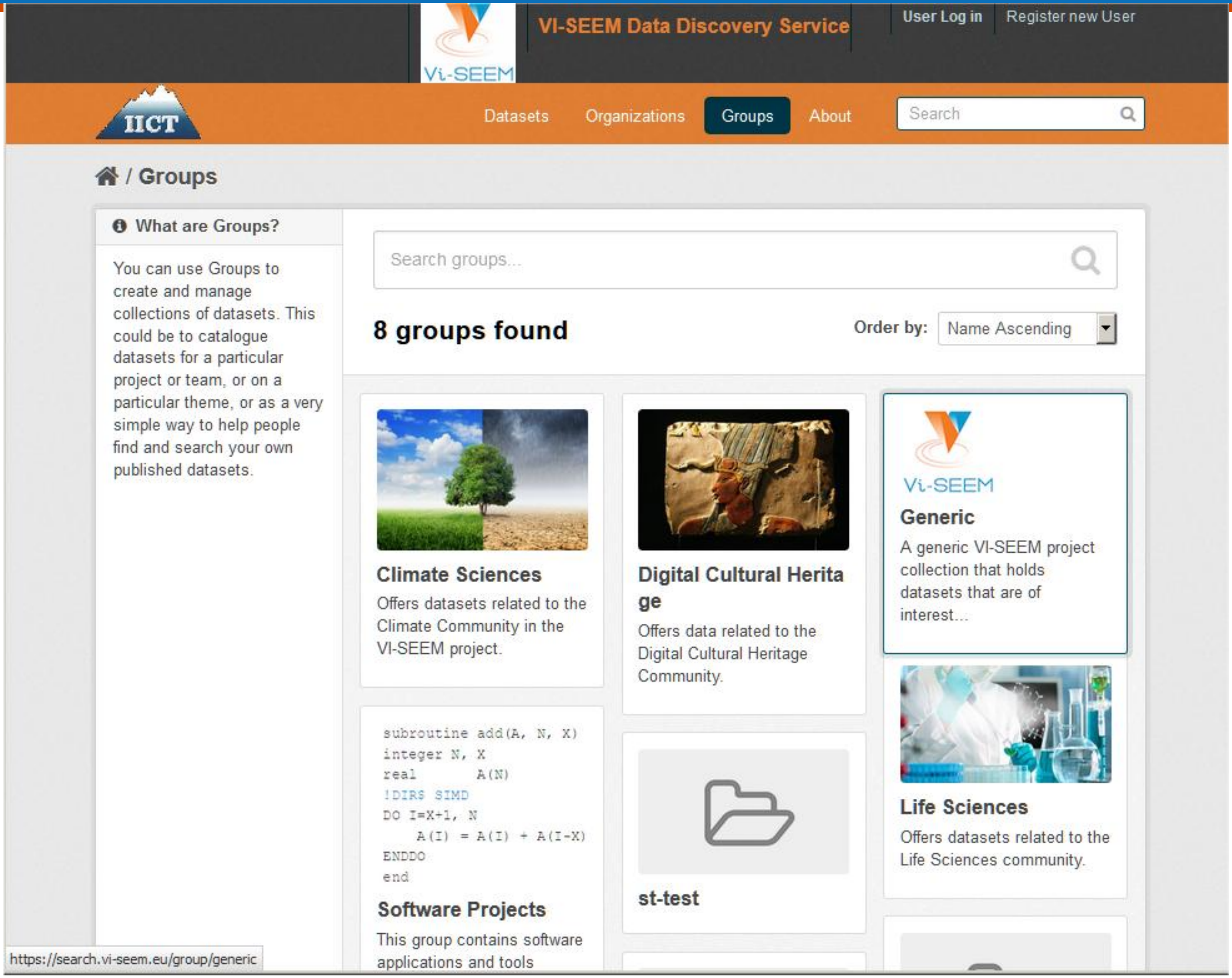
- **CKAN system:**
  - **Python** on the backend
  - **JavaScript** + **HTTPS** on the frontend
  - and depends on: **Pylons** web framework, **SQLAlchemy**, **PostgreSQL**
  - Search engine: Apache search platform **SOLR**
  - Allows third party or custom **modular extensions**.

- **CKAN uses its internal model to store metadata about the different records, and presents it on a convenient web interface that allows users to browse and search this metadata.**

- **CKAN offers a powerful and well documented API that allows third-party applications and services to be built around it.**

Data Discovery Service in VRE

Example frontend screen



What are Groups?

You can use Groups to create and manage collections of datasets. This could be to catalogue datasets for a particular project or team, or on a particular theme, or as a very simple way to help people find and search your own published datasets.

8 groups found

Climate Sciences

Offers datasets related to the Climate Community in the VI-SEEM project.

Digital Cultural Herita ge

Offers data related to the Digital Cultural Heritage Community.

Generic

A generic VI-SEEM project collection that holds datasets that are of interest...

```
subroutine add(A, N, X)
integer N, X
real       A(N)
!DIR$ SIMD
DO I=X+1, N
    A(I) = A(I) + A(I-X)
ENDDO
end
```

Software Projects

This group contains software applications and tools

st-test

Life Sciences

Offers datasets related to the Life Sciences community.

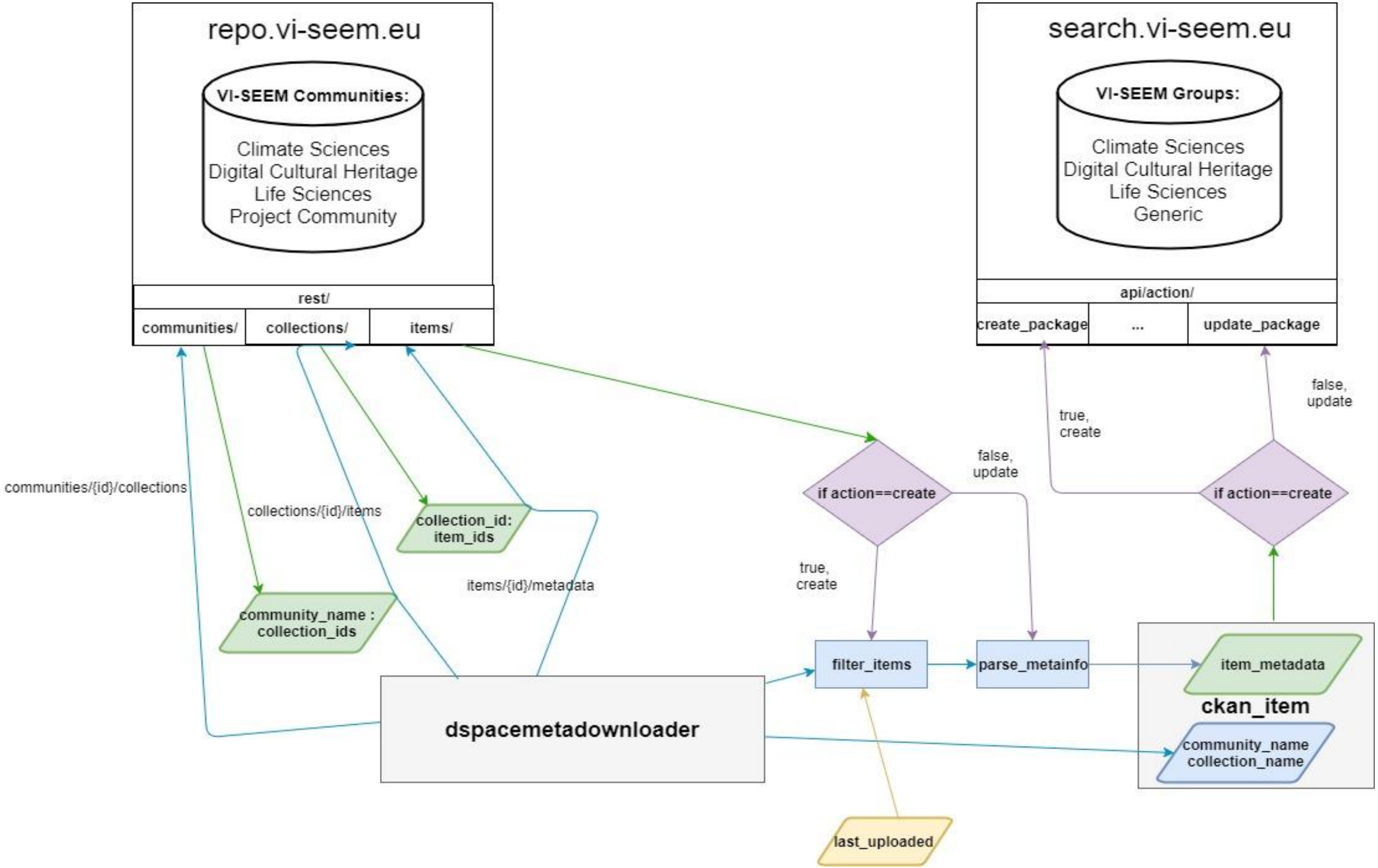
Data Synchronization tool

- Synchronizes and updates VI-SEEM DDS with VI-SEEM data repository.
- Written in Python 3 using modules from the Python Standard Library only.
- Runs in either create or update mode.
  - Create mode is the default behavior and filters already synchronized items, uploading only new items from the data repository
  - Update mode – if specified the metadata of all items will be checked and updated.
- Two cron jobs for each mode automatically synchronize data every day.
- Records a detailed log file.
- Developed by IICT-BAS.

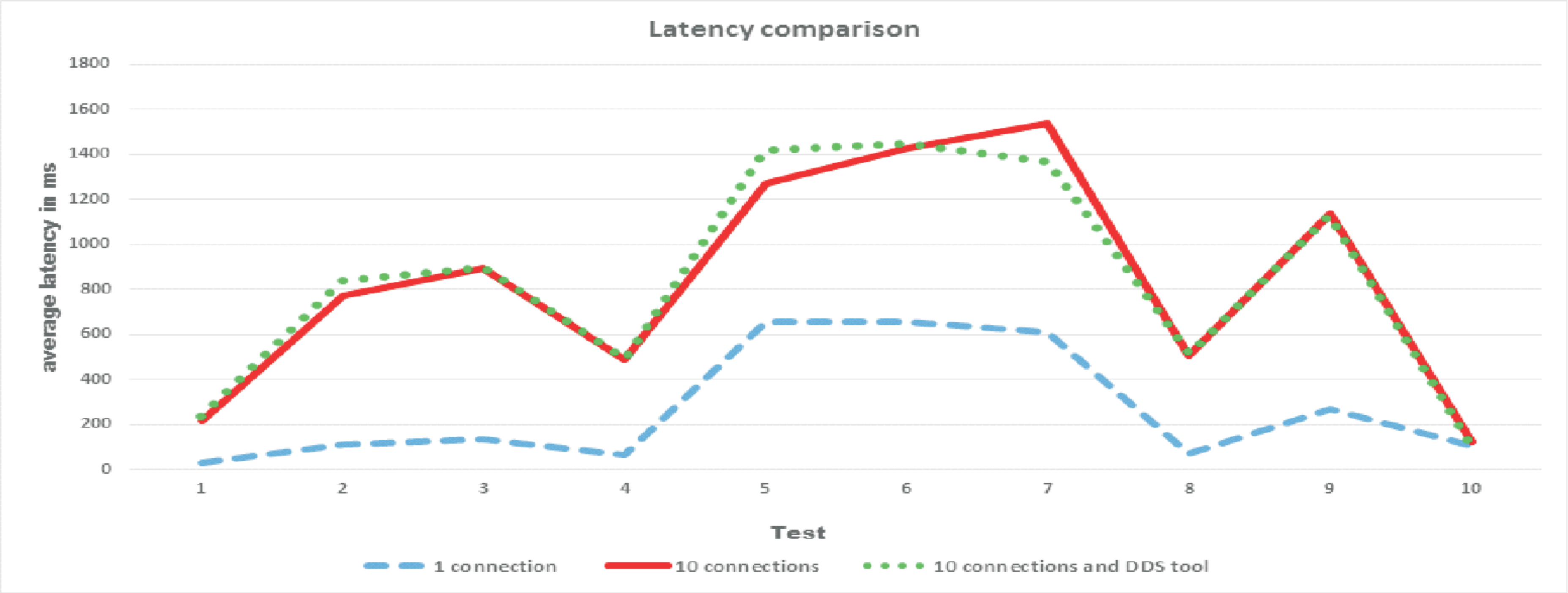
➢ The dataflow between VI-SEEM Repository Service (VRS) and DDS is shown on the next slide >>>

Synchronization between services

Dataflow



Performance evaluation



- **Blue dashed line**: Average latency on 1 active connection to 10 API calls to DDS server.
- **Red solid line** represents 10 simultaneous connections on the same tests.
- **Dotted green line** presents the impact of the used resources by data synchronization tool while the server has significant load at the same time.

The performance differs by a small margin at some of the tests and it does not have effect on user experience at all.

Conclusion

- ❑ **The Data Discovery Service is a CKAN based search system 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.**

- ✓ **Part of this work is accepted for publishing in**
  Scalable Computing: Practice and Experience (SCPE),
  Scientific International Journal for Parallel and Distributed Computing, 2018

❖ **Thank you for your attention. Questions?**