FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Deutscher Titel

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Software Engineering and Internet Computing

eingereicht von

Bernhard Gößwein
Matrikelnummer 01026884

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Andreas Rauber
Mitwirkung: Dr. Tomasz Miksa

Wien, 1. Jänner 2001

Bernhard Gößwein	Andreas Rauber

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Designing a Framework gaining Repeatability for the OpenEO platform

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Software Engineering and Internet Computing

by

Bernhard Gößwein
Registration Number 01026884

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Andreas Rauber
Assistance: Dr. Tomasz Miksa

Vienna, 1st January, 2001

Bernhard Gößwein

Andreas Rauber

Erklärung zur Verfassung der Arbeit

Bernhard Gößwein
Vorderer Ödhof 1, 3062 Kirchstetten

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 1. Jänner 2001

Bernhard Gößwein

Danksagung

Ihr Text hier.

Acknowledgements

Enter your text here.

Kurzfassung

Ihr Text hier.

Abstract

Enter your text here.

Contents

Kurzfassung **xi**

Abstract **xiii**

Contents **xv**

1 Introduction **1**

1.1 Problem Description 1

1.2 Aim of this Work 2

2 Related Work **3**

3 Methodology **5**

4 Proof of Concept **7**

5 Conclusion **9**

List of Figures **11**

List of Tables **13**

List of Algorithms **15**

Index **17**

Glossary **19**

Acronyms **21**

Bibliography **23**

CHAPTER 1

Introduction

1.1 Problem Description

Over the last decades remote sensing agencies have increased the variations of data processing and therefore the amount of resulting data. To preserve the data for further usage in the future it is necessary to have citable data and processes on the data to ensure repeatability in a long-term.[RMD12] Already most of the the data used in earth observation sciences are retrieved or provided via Service Oriented Architecture (SOA) interfaces. Provider like Google Earth Engine and EODC provide an Web API for retrieving and processing data. Due to a different range of functionality and a difference between the endpoints of the providers it is hard to create a workflow for more than one provider. The OpenEO project has the goal to be an abstraction layer above different EO data providers. The underlying structure consists of three parts:

- Client Module : Is written in the program language of the user and transfers the users commands to the backends.
- Core Module: A standard on how the communication should take place between client and backend.
- Backend Module: The provider of the data and the services, which gets the instructions from the clients and returns the results.

Further information on the software architecture of the project is defined in the project proposal ([PWS$^+$17]).Until now there is no consideration of repeatability verification of workflows for users in the OpenEO architecture.Generalised layers have the opportunity to be implemented in a way that makes processes and data scientifically verifiable and reproducible, because it handles data and processes on the data in a standardised way on different providers. Even though the range of functionality and the API endpoints are well defined in the OpenEO coreAPI the contributing content providers (OpenEO backends) will have different underlying software types and versions. The underlying technology of an OpenEO backend will also change over time and can lead to different results on the same workflow executions. Consider the following: A scientist runs an experiment using OpenEO as his research tool and gets results. The same scientist runs the same experiment with the same input some months later and gets slightly different results. The question occurs, why are the results different? Has the used data changed, has the user accidentally submitted different code or has some underlying software inside the backend provider changed. Adding a possibility for the users of OpenEO to gain this information is an important feature for the scientific community. The aim of this thesis is to provide a possibility for users of OpenEO to verify and validate a job re-execution on different underlying technologies of an OpenEO backend provider.[PWS+17]

Read through and improve above text.

1.2 Aim of this Work

The expected outcome of this thesis is to discover and develop a possible framework for providing repeatability in the OpenEO project. This enables users to re-execute workflows and validate the results, so that differences on the process or data are accessible for the users. To achieve this goal a model for repeatability within the project has to be discovered and implemented to evaluate the ability of the model. The model shall then conclude recommendations for the OpenEO project on how to improve re-execution validation for the user and how it can be achieved. Therefore the following research questions can be formulated:

- **How can an OpenEO job re-executed be applied like the initial execution?**
  - How can the used data be identified after the initial execution?
  - How can the used software of the initial execution be reproduced?
  - What data has to be captured when?
  - How can the result of a re-execution in future software versions be verified?
- **How can the equality of the OpenEO job re-execution results be validated?**
  - What are the validation requirements?
  - How can the data be compared?
  - How can the re-execution be validated after changes of the OpenEO backend environment?
  - How can differences in the environment between the executions be discovered?

Read through above text and improve. 2

Add description of use cases.

CHAPTER 2

Related Work

Currently, there exists no concrete solution to add the ability of repeatability to the OpenEO project. However there are concepts of adding repeatability in computer science.

2.0.1 eScience

The eScience has the potential to enable a boost in scientific discovery by providing approaches to make digital data and workflows citable. In [RMMP15] is a common way of reaching this goal formulated. It describes an approach to look at whole research processes, other than only data citation by introducing Process Management Plans. The capturing, verification and validation of the needed data for a computational process is also demonstrated within the paper.[RMMP15]

2.0.2 Data Citation

Since the earth observation community use a high amount of satellite data and also within the OpenEO project a lot of big data sets are being used, there needs to be a solution to cite the used data in a workflow. The Research Data Alliance (RDA) working group on data citation provides a 14 step recommendation of data citation. It contains solutions not only for static, but also for dynamic data, so data that changes over time. Using the guideline for data citations from the RDA makes the data scientifically citable. [RAvUP16] In earth science there is also a strategy of ESA and NASA to achieve a content standard for data preservation.[RMD12]

2.0.3 Provenance Data

The re-execution of an OpenEO workflow not only needs data citation, but also the information of how the workflow was executed. Therefore provenance data has to be captured.[RMH+11]There are already several provenance models defined in the scientific community. One of the existing models is the PROV model, which was published in 2013 by the World Wide Web Consortium Provenance Working Group and consists of recommendations and guidelines for provenance data.[MGC+15] Another model is the VFramework, designed for the purpose of redeployment including the verification of a re-execution of the same workflow. [MPM+13]

Read through above text and improve.

CHAPTER 3

Methodology

1. **Literature review**
   The background and other approaches on repeatability have to be considered for an implementation in earth observation data science. Especially for the knowledge of earth observation data, a base of information has to be gathered. Since the thesis is related to the OpenEO project and especially the Backend Module, information about their structure is important.

2. **Create concept for OpenEO**
   In the second part, a concept of repeatability for the OpenEO project gets created. The information gathered by the literature review leads to design decisions and approaches to achieve this. Data citation and workflow capturing are the key elements of the model. Another important component of it is how the re-execution can be validated and viewed from the users perspective.

3. **Implementation of a prototype**
   A software for the capturing of the data and environment has to be implemented for OpenEO job executions. The implementation also includes the validation of the OpenEO job re-execution.

4. **Analysing Results**
   In this step the implemented software is build into an OpenEO instance and gets tested and evaluated. It also includes the discussion of the results and thoughts about further steps or improvements.

Enter your text here.

CHAPTER 4

Proof of Concept

Enter your text here.

CHAPTER 5

Conclusion

Enter your text here.

List of Figures

List of Tables

List of Algorithms

Index

distribution, 5

Glossary

editor A text editor is a type of program used for editing plain text files.. 5

Acronyms

CTAN Comprehensive TeX Archive Network. 11

FAQ Frequently Asked Questions. 11

PDF Portable Document Format. 6, 10, 11, 15

SVN Subversion. 10

WYSIWYG What You See Is What You Get. 9

Bibliography

[MGC+15] Luc Moreau, Paul Groth, James Cheney, Timothy Lebo, and Simon Miles. The rationale of prov. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35:235 – 257, 2015.

[MPM+13] Tomasz Miksa, Stefan Pröll, Rudolf Mayer, Stephan Strodl, Ricardo Vieira, José Barateiro, and Andreas Rauber. Framework for verification of preserved and redeployed processes. In *Proceedings of the 10th International Conference on Digital Preservation, iPRES 2013, Lisbon, Portugal, September 2 - 6, 2013*, 2013.

[PWS+17] Edzer Pebesma, Wolfgang Wagner, Matthias Schramm, Alexandra Von Beringe, Christoph Paulik, Markus Neteler, Johannes Reiche, Jan Verbesselt, Jeroen Dries, Erwin Goor, and et al. Openeo - a common, open source interface between earth observation data infrastructures and front-end applications. Nov 2017.

[RAvUP16] Andreas Rauber, Ari Asmi, Dieter van Uytvanck, and Stefan Pröll. Identification of reproducible subsets for data citation, sharing and re-use. 2016.

[RMD12] H. Ramapriyan, J. Moses, and R. Duerr. Preservation of data for earth system science - towards a content standard. In *2012 IEEE International Geoscience and Remote Sensing Symposium*, pages 5304–5307, July 2012.

[RMH+11] David De Roure, José Manuel, Kristina Hettne, Khalid Belhajjame, Raúl Palma, Graham Klyne, Paolo Missier, José Enrique Ruiz, and Carole Goble. Towards the preservation of scientific workflows. In *In Procs. of the 8th International Conference on Preservation of Digital Objects (iPRES 2011). ACM*, 2011.

[RMMP15] Andreas Rauber, Tomasz Miksa, Rudolf Mayer, and Stefan Pröll. Repeatability and re-usability in scientific processes: Process context, data identification and verification. In *DAMDID/RCDL*, 2015.