Statika: managing cloud resources, bioinformatics tools and data

Alexey Alekhin, Evdokim Kovach, Pablo Pareja-Tobes, Marina Manrique, Eduardo Pareja, Raquel Tobes and Eduardo Pareja-Tobes

Oh no sequences! Research Group. Era7 bioinformatics, Granada, Spain.

Introduction

Statika is a set of Scala libraries

- which allows building *well-structured* module systems
- where dependencies are *correct by construction*
- and you know it *before you run* anything

Basic notions

Bundle

A **bundle** is a thin wrapper for a tool, library, resource or any other component of your system:

- it may have dependencies on other bundles
- it may do something in runtime, e.g. install a program or download some data

Bundles can be *applied*, i.e. deployed to an EC2 instance

Distribution

A **distribution** is a bundle, which can deploy other bundles (its members):

- it represents some environment, where you're going to use your bundles
- being a member of a distribution means to work fine with this environment
- distribution takes care of installing member dependencies first, and then the member itself

Distributions abstract over the *cloud infrastructure*

Availability

AGPLv3 Free Software

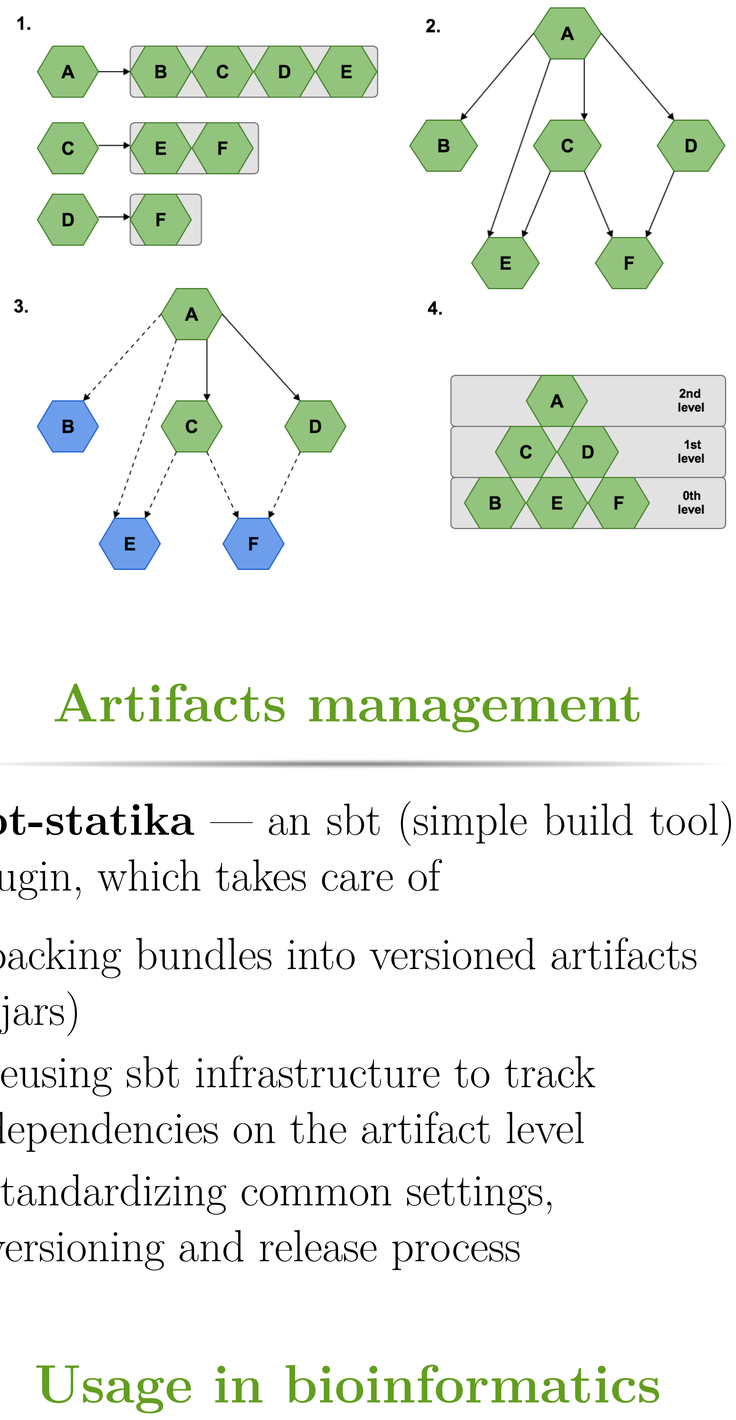
Free as in Freedom

Statika is free and open-source under the AGPLv3 license

See http://ohnosequences.com/statika

Abstract module system

- Bundles are represented as Scala types.
- Their dependencies on each other are validated by the compiler — i.e. **statically**.
- Statika linearizes the types graph to get them in the right order.



Artifacts management

sbt-statika — an sbt (simple build tool) plugin, which takes care of

- packing bundles into versioned artifacts (jars)
- reusing sbt infrastructure to track dependencies on the artifact level
- standardizing common settings, versioning and release process

Usage in bioinformatics

First of all, there is a Statika distribution for bioinformatics tools, such as Velvet, Cufflinks, Tophat and Bowtie(2). See github.com/statika/bioinfo-dist/.

nispero

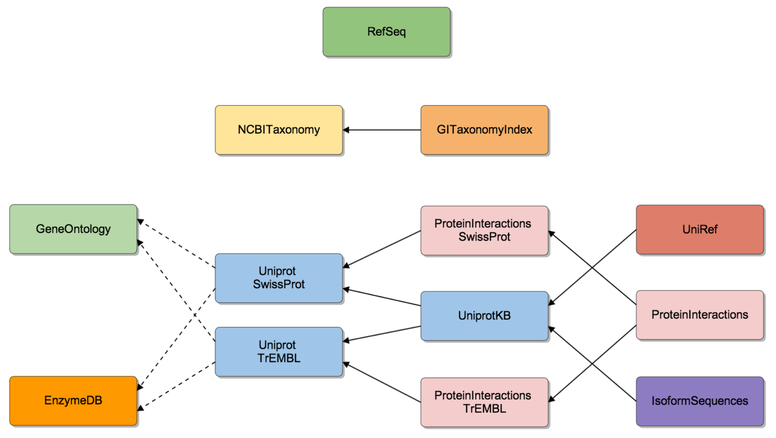
Cloud-computing System

Nispero is a toolset for declaring scalable cloud-based systems for bioinformatics computations. It has pretty complex inner commponents structure, which is managed with the help of Statika.

Usage in bioinformatics

Bio4j **Graph Database**

Bio4j is a bioinformatics graph database which integrates data from a lot of different sources:



Every module has some inner structure:

- raw data from a data source
- data importing process to the graph database
- nodes and relationships type definitions
- some abstract interface representing what you can do with this data

So with Statika Bio4j achieves

- simple data-import process
- automized dependencies management
- easy and robust deployment to AWS

Acknowledgments

Statika is developed by the R&D team of the Era7 Bioinformatics company

Era7 bioinformatics

ohnosequences!

This project is funded in part by the ITN FP7 project INTERCROSSING (Grant 289974)

INTERCROSSING MARIE CURIE ACTIONS