Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY
OF NEW ZEALAND

KB national library
of the netherlands

WEB CURATOR TOOL

IIPC GA 2019 - WCT Workshop

Hands-on Setup Guide

Welcome to the Hands-on Setup Guide for the 2019 IIPC Web Curator Tool Workshop.

The reality of any hands-on workshop is that things will break. We've tried our best to provide a stable virtual machine that can let you walk through the basics of using the Web Curator Tool.

If you have any questions, please get in touch via our Slack group (*workshop-iipc-2019 channel*) or have a read of the WCT documentation.

Setup

Using our pre-configured VM

The WCT team have prepared a pre-configured virtual machine (VM), that has working WCT, OpenWayback and Heritrix 3 instances.

To reduce the setup time during the Hands-on session, it is recommended that you download and install Oracle VirtualBox prior to the workshop.

VirtualBox

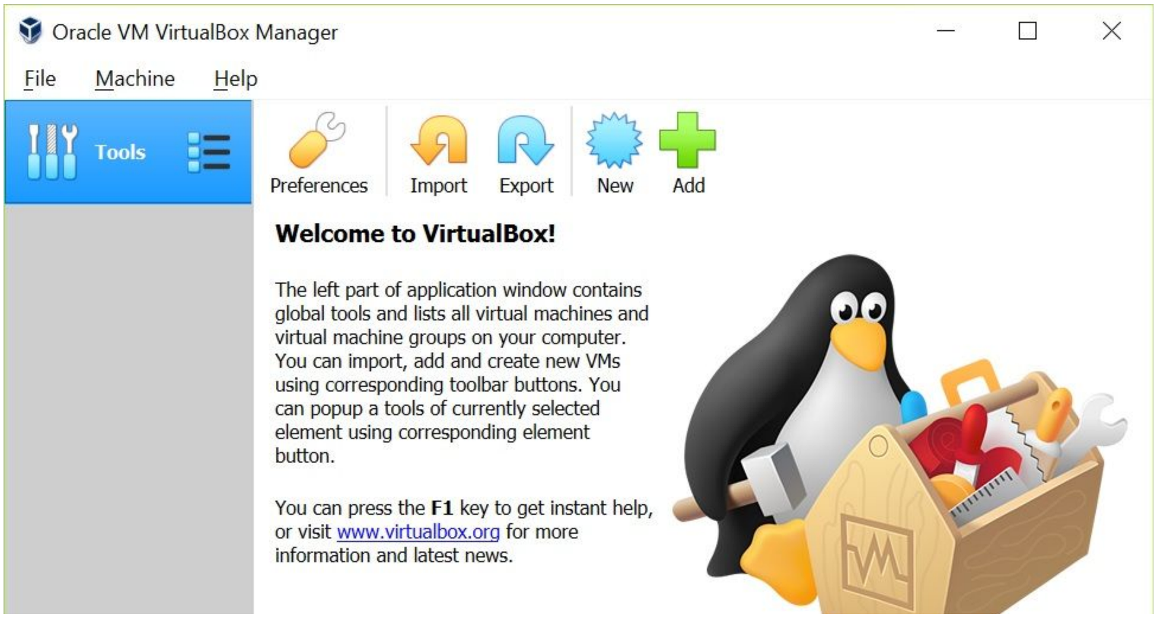
This section requires that you install the latest version of Oracle VirtualBox to run the provided VM (.ova file). VirtualBox is available for Windows, OSX and Linux. The version for your operating system can be downloaded here, and installation instructions found here. This VM has been tested on version **5.1** and later of VirtualBox.

Downloading the VM

A copy of the VM will be provided on a USB stick during the workshop. *Should you wish to download and familiarise yourself with the VM beforehand, a copy can be found here.*

Opening the VM

Once you have downloaded/copied the VM to an appropriate location on your computer, you can open it using VirtualBox.

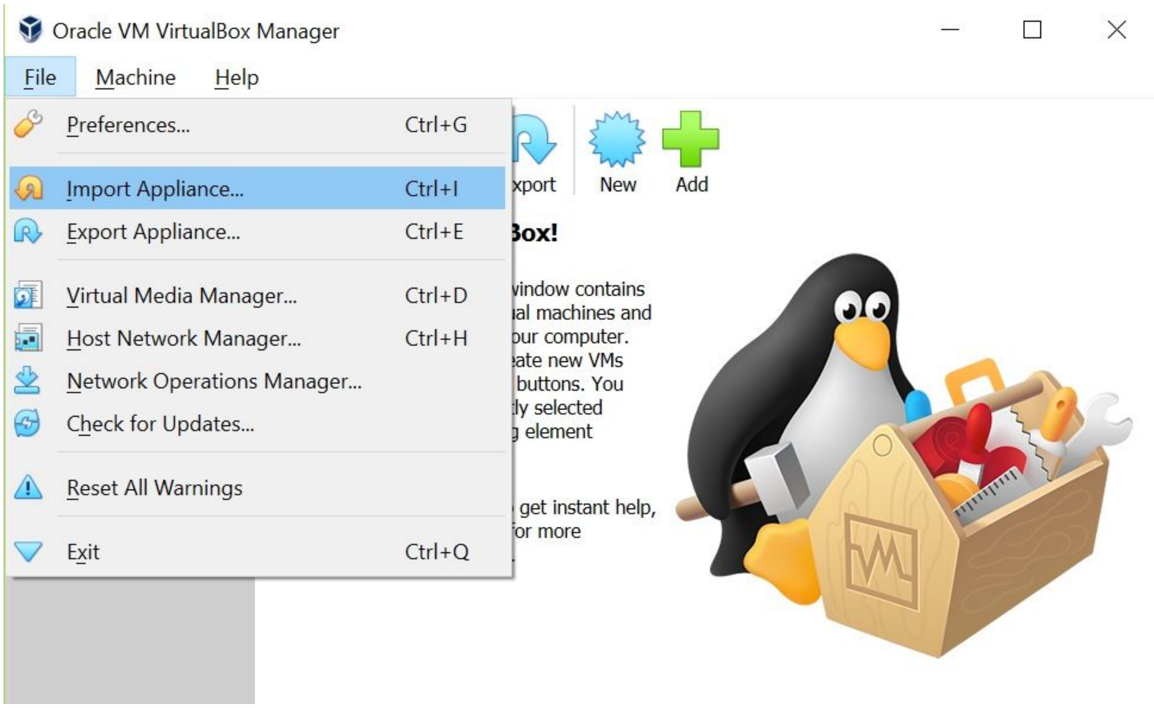


Welcome to VirtualBox!

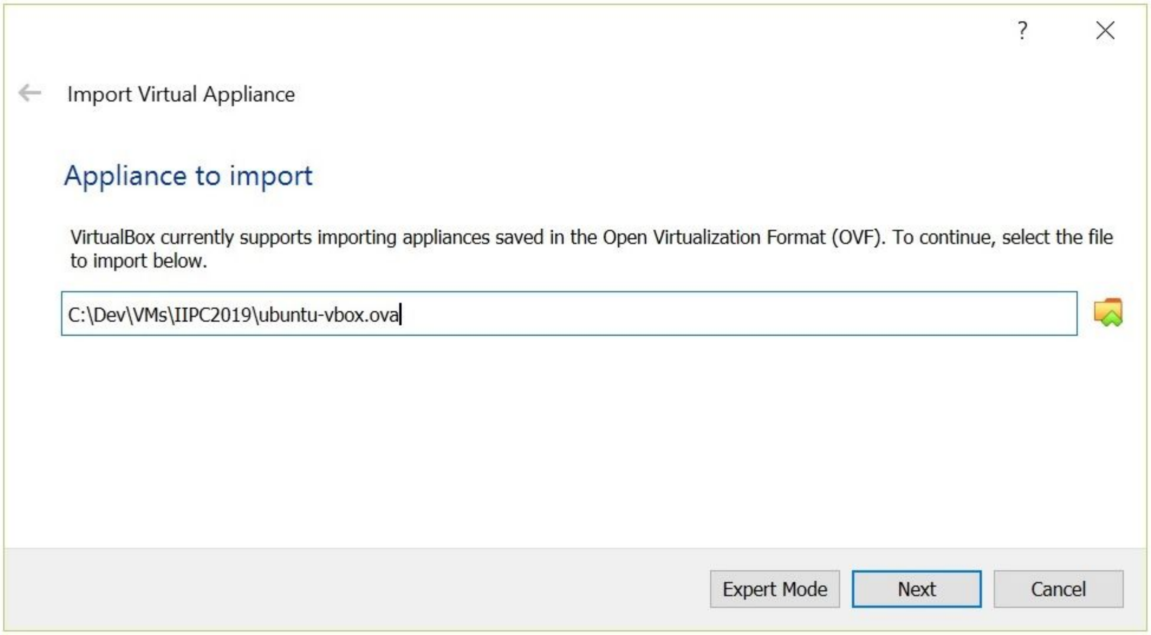
The left part of application window contains global tools and lists all virtual machines and virtual machine groups on your computer. You can import, add and create new VMs using corresponding toolbar buttons. You can popup a tools of currently selected element using corresponding element button.

You can press the **F1** key to get instant help, or visit www.virtualbox.org for more information and latest news.

Import the VM using File->Import Appliance



Select the VM (.ova file) to import

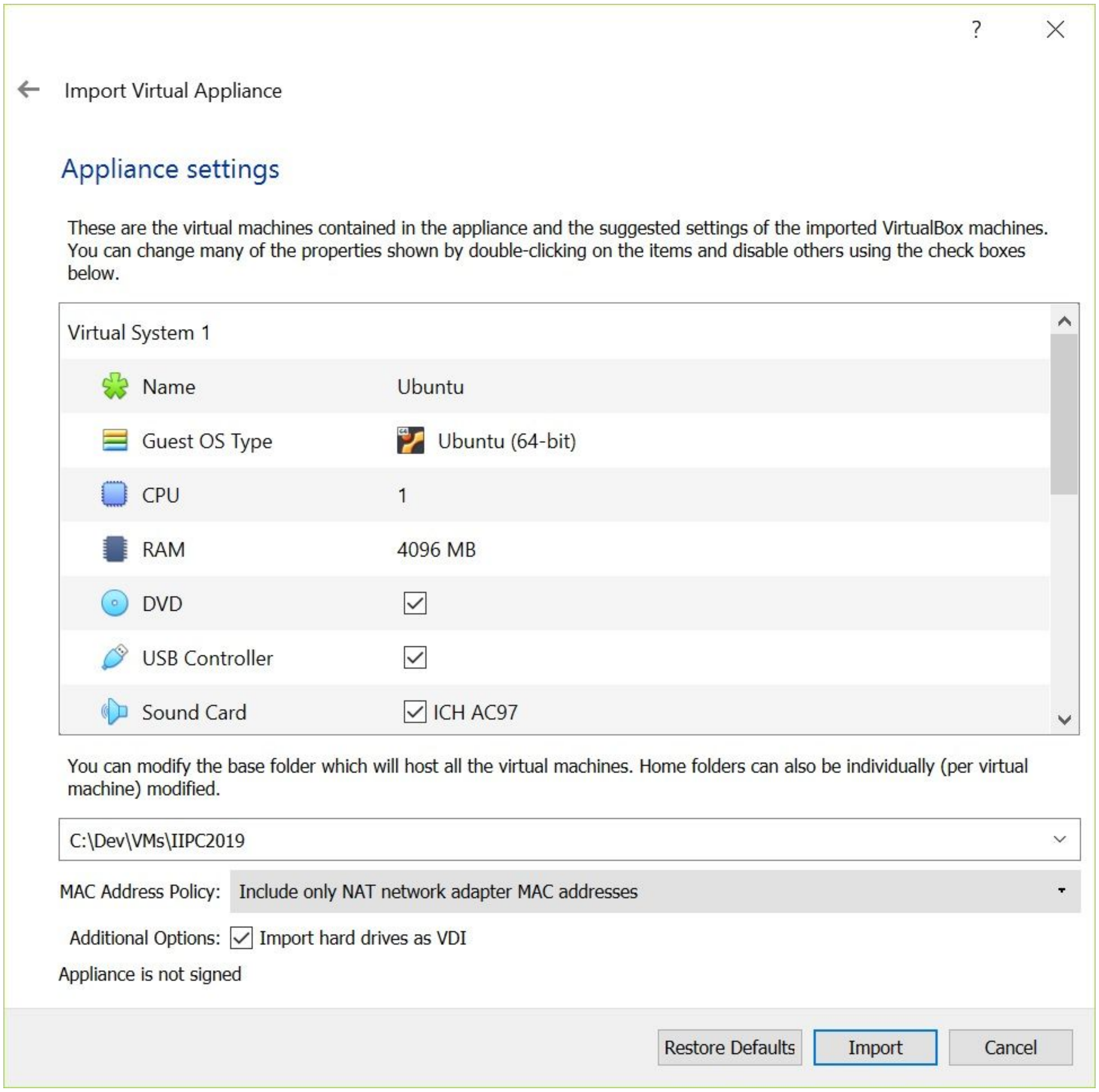


Appliance to import

VirtualBox currently supports importing appliances saved in the Open Virtualization Format (OVF). To continue, select the file to import below.

C:\Dev\VMs\IIPC2019\ubuntu-vbox.ova

Specify a name, and the directory to store the imported VM



Import Virtual Appliance

Appliance settings

These are the virtual machines contained in the appliance and the suggested settings of the imported VirtualBox machines. You can change many of the properties shown by double-clicking on the items and disable others using the check boxes below.

Virtual System 1

Name	Ubuntu
Guest OS Type	Ubuntu (64-bit)
CPU	1
RAM	4096 MB
DVD	☑
USB Controller	☑
Sound Card	☑ ICH AC97

You can modify the base folder which will host all the virtual machines. Home folders can also be individually (per virtual machine) modified.

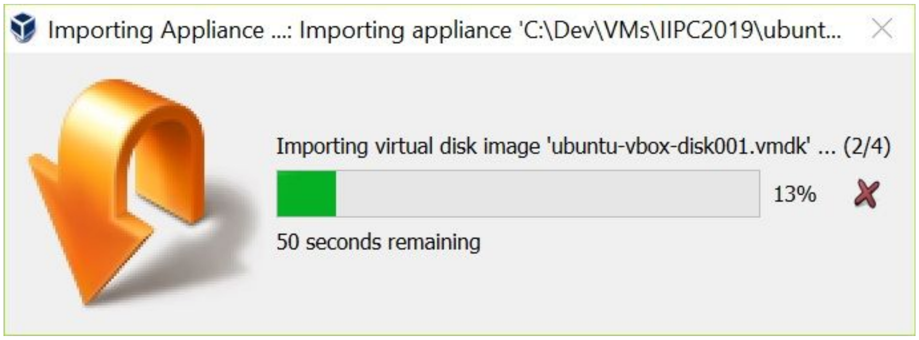
C:\Dev\VMs\IIPC2019

MAC Address Policy: Include only NAT network adapter MAC addresses

Additional Options: ☑ Import hard drives as VDI

Appliance is not signed

Import the .ova file, this can take some time



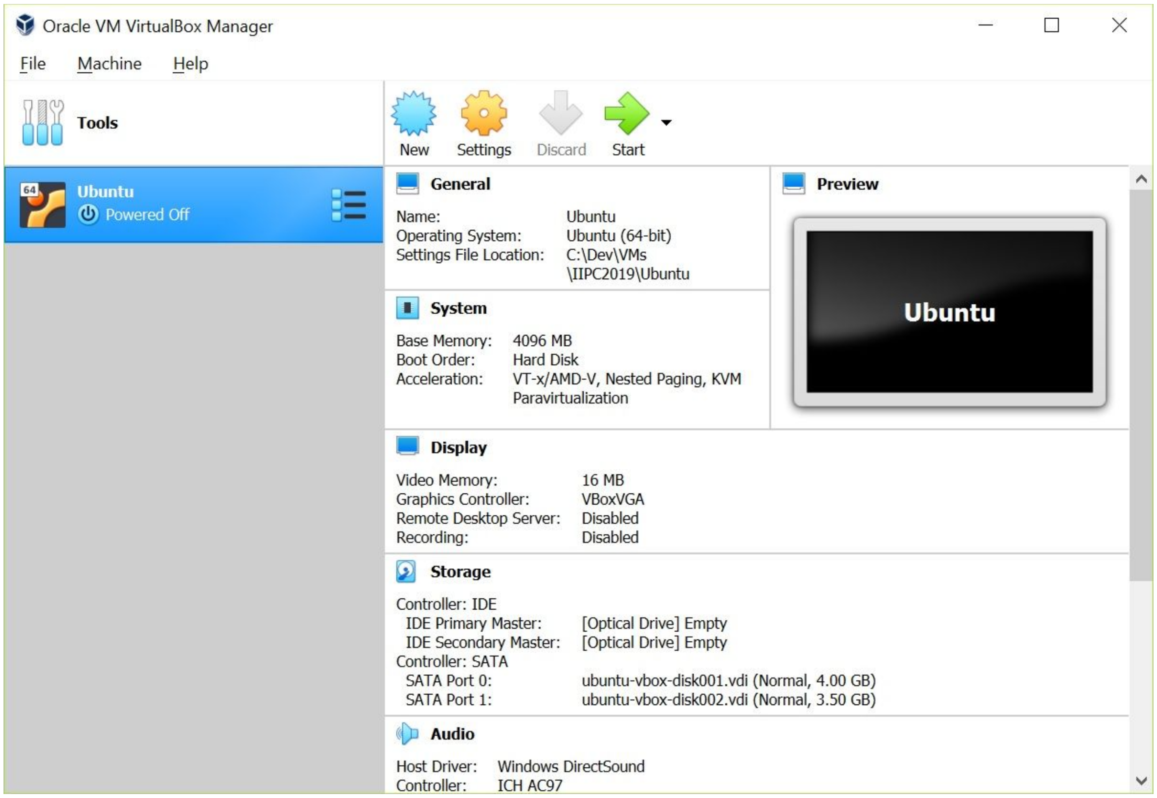
Importing Appliance ...: Importing appliance 'C:\Dev\VMs\IIPC2019\ubunt...

Importing virtual disk image 'ubuntu-vbox-disk001.vmdk' ... (2/4)

13%

50 seconds remaining

Successfully imported VM



Oracle VM VirtualBox Manager

General

Name: Ubuntu
Operating System: Ubuntu (64-bit)
Settings File Location: C:\Dev\VMs\IIPC2019\Ubuntu

System

Base Memory: 4096 MB
Boot Order: Hard Disk
Acceleration: VT-x/AMD-V, Nested Paging, KVM Paravirtualization

Display

Video Memory: 16 MB
Graphics Controller: VBoxVGA
Remote Desktop Server: Disabled
Recording: Disabled

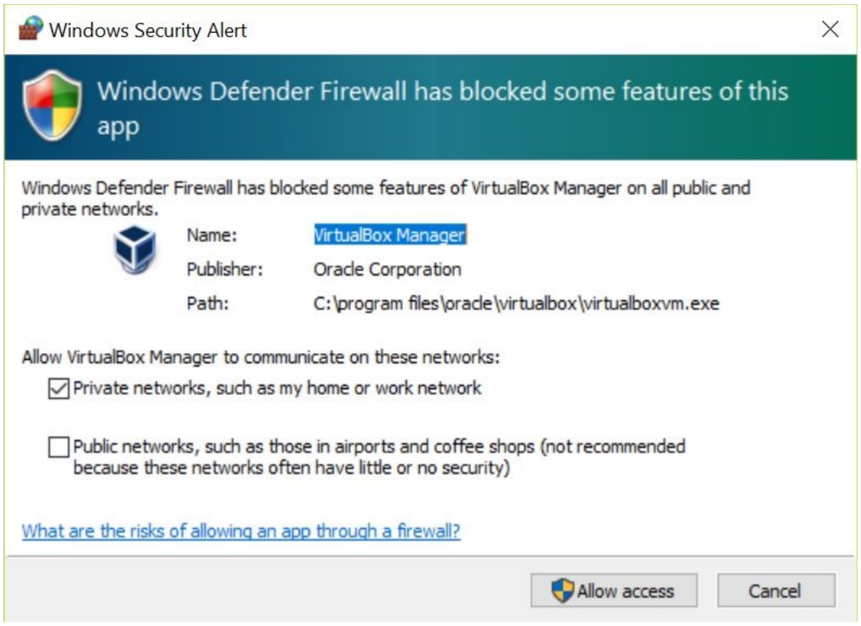
Storage

Controller: IDE
IDE Primary Master: [Optical Drive] Empty
IDE Secondary Master: [Optical Drive] Empty
Controller: SATA
SATA Port 0: ubuntu-vbox-disk001.vdi (Normal, 4.00 GB)
SATA Port 1: ubuntu-vbox-disk002.vdi (Normal, 3.50 GB)

Audio

Host Driver: Windows DirectSound
Controller: ICH AC97

If prompted, allow Firewall access for Virtual Box



Windows Defender Firewall has blocked some features of this app

Windows Defender Firewall has blocked some features of VirtualBox Manager on all public and private networks.

Name: VirtualBox Manager
Publisher: Oracle Corporation
Path: C:\program files\oracle\virtualbox\virtualboxvm.exe

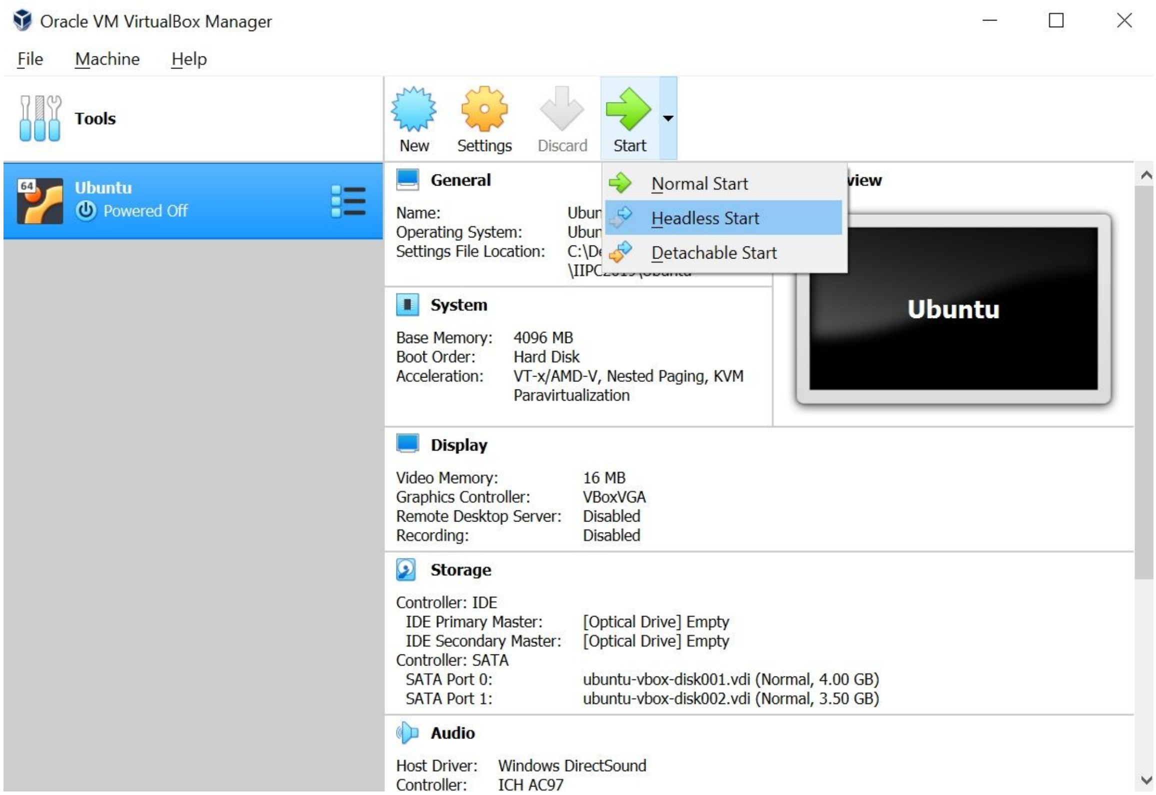
Allow VirtualBox Manager to communicate on these networks:

☑ Private networks, such as my home or work network

□ Public networks, such as those in airports and coffee shops (not recommended because these networks often have little or no security)

What are the risks of allowing an app through a firewall?

Start the VM in headless mode



General

Name: Ubun
Operating System: Ubun
Settings File Location: C:\De \IIPC

System

Base Memory: 4096 MB
Boot Order: Hard Disk
Acceleration: VT-x/AMD-V, Nested Paging, KVM Paravirtualization

Display

Video Memory: 16 MB
Graphics Controller: VBoxVGA
Remote Desktop Server: Disabled
Recording: Disabled

Storage

Controller: IDE
IDE Primary Master: [Optical Drive] Empty
IDE Secondary Master: [Optical Drive] Empty
Controller: SATA
SATA Port 0: ubuntu-vbox-disk001.vdi (Normal, 4.00 GB)
SATA Port 1: ubuntu-vbox-disk002.vdi (Normal, 3.50 GB)

Audio

Host Driver: Windows DirectSound
Controller: ICH AC97

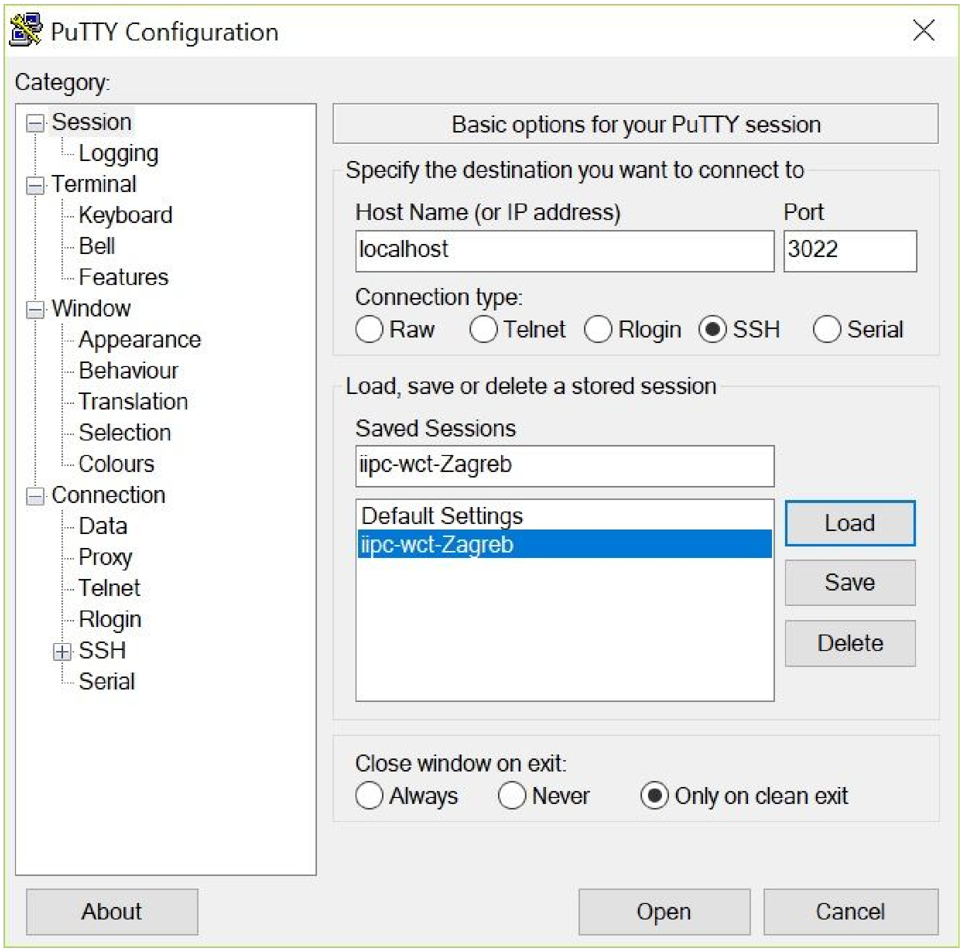
Accessing WCT

WCT can be accessed using a browser running on your regular operating system (i.e. outside of the VM). The location is http://localhost:3080/wct/.
Username: wct
Password: Zagreb19

Logging into the VM

You can log into the VM through your command line, using ssh: **ssh -p 3022 wct@localhost**.
Username: wct
Password: wct
This is also the password for sudo access.

Or using Putty



PuTTY Configuration

Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address): localhost
Port: 3022

Connection type:
○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

Load, save or delete a stored session

Saved Sessions: iipc-wct-Zagreb

Default Settings
iipc-wct-Zagreb

Close window on exit:
○ Always ○ Never ● Only on clean exit

Where things are

WCT

The WCT directories are all under **/opt/wct**.

- The harvest agent stores its files in **/opt/wct/harvest-agent**.
- The store uses **/opt/wct/store**.
- All components log to files in **/opt/wct/logs**.
- Heritrix scripts for use within the WCT script console can be placed in **/opt/wct/h3scripts**.

Apache Tomcat

The Tomcat directory is **/opt/wct/apache-tomcat-9.0.19**. Tomcat is started automatically as a systemd service (called 'tomcat') when the system boots.

MYSQL

You can log into MariaDB using **mysql -u root -p**. The password is "password". A basic WCT database has been provided, containing everything needed to start configuring and running crawls.

Heritrix 3

Heritrix can be found in **/opt/wct/heritrix-3.4.0-20190418**. Like Tomcat, Heritrix is also started as a systemd service (called 'heritrix'). You can access the web interface at https://localhost:3443/engine.
Username: admin
Pasword: admin

OpenWayback

In this setup WCT uses OpenWayback to display harvest results. This OpenWayback has been deployed in the same Tomcat as WCT itself. It expects its WARC files in **/opt/wct/wayback**. It can be accessed at http://localhost:3080/owb/wayback/.

Digital repository

This WCT has been configured to use a filesystem-based digital repository, which amounts to just a directory to copy WARC files to. This directory is **/opt/wct/warc-repository**.

Crawling

The Heritrix 3 instance is configured with the following user agent string:
Mozilla/5.0 (compatible; heritrix/3.4 +https://docs.google.com/document/d/1SZq2EV5Q0PxCMiU0hcll158gita9LU5CqJ5xKBUtAr0)

This URL will be valid until the end of this workshop. If you wish to continue using this VM after the workshop, please update the user agent string within the WCT profiles to a URL relevant to your institution.

Installing your own WCT

Should you wish to install WCT yourself, in your own environment, there are several guides available:

- a quick start guide to installing WCT.
- a comprehensive System Administrator guide for installing WCT.

Resources

Documentation
Slack
Github