

Data Journalism Developer Studio 2012LX

Download Data Journalism Developer Studio 2012LX

M. Edward (Ed) Borasky

Borasky Research Journal

February 19, 2012

I Keep Six Honest Serving Men

I Keep Six Honest Serving Men

I keep six honest serving-men

(They taught me all I knew);

Their names are What and Why and When

And How and Where and Who.

Why A Data Journalism Developer Studio?

Hint: journalists today work in a world dominated by two trends

- ▶ Real-time many-to-many communications platforms
- ▶ Large sets of complex data with stories waiting to be told
- ▶ **Communications of the ACM, October 2011: "Computational Journalism"**

Real-time Many-to-Many Communications Platforms

- ▶ Hundreds of millions of Twitter accounts almost everywhere in the world
- ▶ Millions of *active* Twitter users
- ▶ Complex patterns of interactions around people, places and events
- ▶ Intricate and changing connection patterns between people
- ▶ *People break and discuss the news in real time on Twitter.*

Large Sets of Complex Data with Stories Waiting to be Told

- ▶ Government data - national, regional, and local
- ▶ Business and financial data
- ▶ Political fundraising, census/redistricting and vote tabulation data
- ▶ Environmental, weather and climate data
- ▶ Social network data
- ▶ And yes - traffic and sports data too

What Is The Data Journalism Developer Studio?

- ▶ 100 percent open source technologies!
- ▶ A complete Linux appliance, plus
- ▶ Tools for collecting, managing, analyzing and presenting data, plus
- ▶ Optional tools for
 - ▶ productivity
 - ▶ creating, editing and producing digital media
 - ▶ advanced numerical data visualization / exploration / presentation
 - ▶ advanced data collection / extraction / parsing / scraping
 - ▶ advanced finance / econometric / time series analysis
 - ▶ advanced server frameworks

Who Is It For?

Hint: the next generation of journalists

- ▶ Data journalism hackers and their friends
 - ▶ Journalism students – high school, community college and beyond
 - ▶ Freelance journalists, researchers, reporters, editors, publishers

Why 100 Percent Open Source?

- ▶ Open source software is robust
- ▶ Open source software is low cost

Robustness

Proven technologies in wide use.

- ▶ Crafted by highly-motivated self-regulated communities of experts
- ▶ Security flaws, functionality defects and performance issues are rapidly found and fixed
- ▶ Peer review process yields software that is usually more efficient than commercial counterparts

Low Cost

- ▶ The software in the Data Journalism Developer Studio is freely downloadable without legal restrictions
- ▶ Functionality that would cost thousands of dollars in commercial licenses is available for the cost of a download

Host System Requirements

- ▶ 64-bit Intel / AMD compatible
- ▶ Windows 7 or later
- ▶ Two or more cores
- ▶ 4 GB of RAM

Step 1: Download and Install Oracle VM VirtualBox®

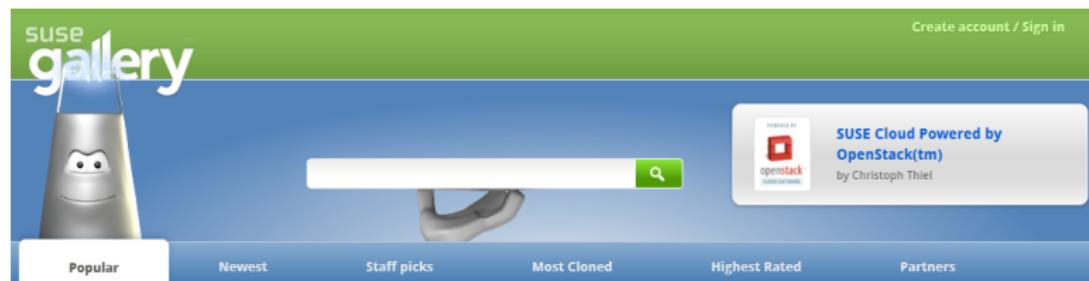
<https://www.virtualbox.org/>

Step 2: Download and Install 7-Zip

<http://www.7-zip.org/>

Step 3: Sign up for SUSE Gallery

<http://susestudio.com/browse>



Step 4: Download the Appliance

<http://j.mp/DJDS2012LX>



Data Journalism Developer Studio 2012LX



Published by M. Edward (Ed) Borasky Based on openSUSE 12.1 32-bit x86
Homepage at <http://znmeb.github.com/Data-Journalism-Developer-Studio/>

Data Journalism Developer Studio 2012LX is a 100% open source Linux™-based appliance designed for data journalism developers. It can be run as a desktop or a server and is designed for development of data journalism desktop and server applications.

The Data Journalism Developer Studio is *modular*. The core appliance consists of the operating system, desktop, browser, and data acquisition / cleaning / analysis / visualization tools. The core appliance is available as a virtual machine in Open Virtualization Format (OVF).

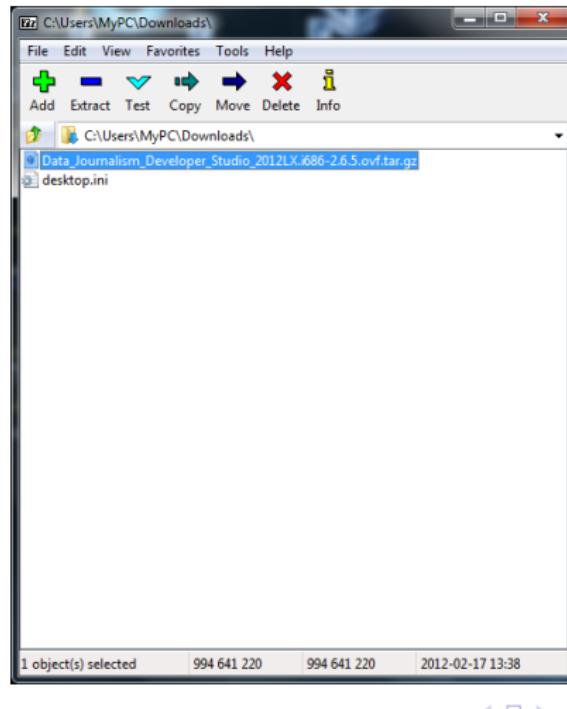
Additional packages can be installed as required using installation scripts provided. Each additional functional package can be independently installed to match the needs of the users.

Additional resources:

- [Data Journalism Developer Studio 2012 blog](#)
- [Data Journalism Developer Studio 2012LX Overview](#)
- [Data Journalism Developer Studio 2012LX On Github](#)
- [Borasky Research Journal](#)
- [Support Data Journalism Developer Studio 2012LX On Fundry](#)

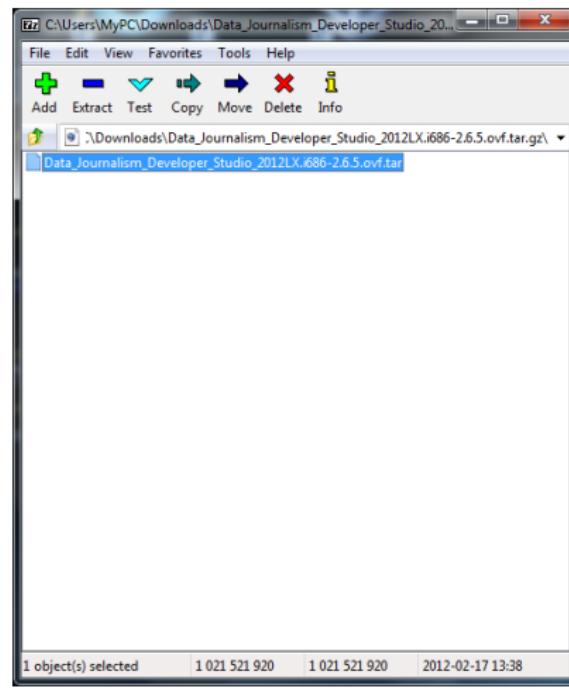
Step 5a: Open in 7-Zip and Remove Gzip Compression

Select appliance and press "Enter."



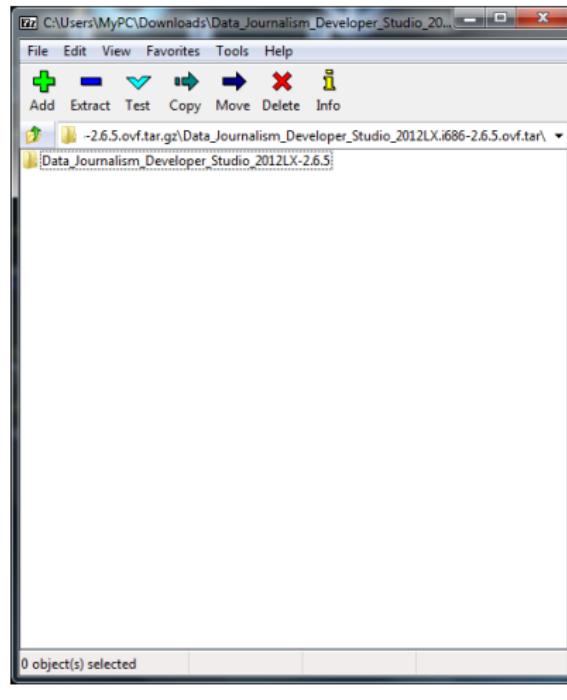
Step 5b: Extract Tar Archive

Select appliance and press “Enter” again.



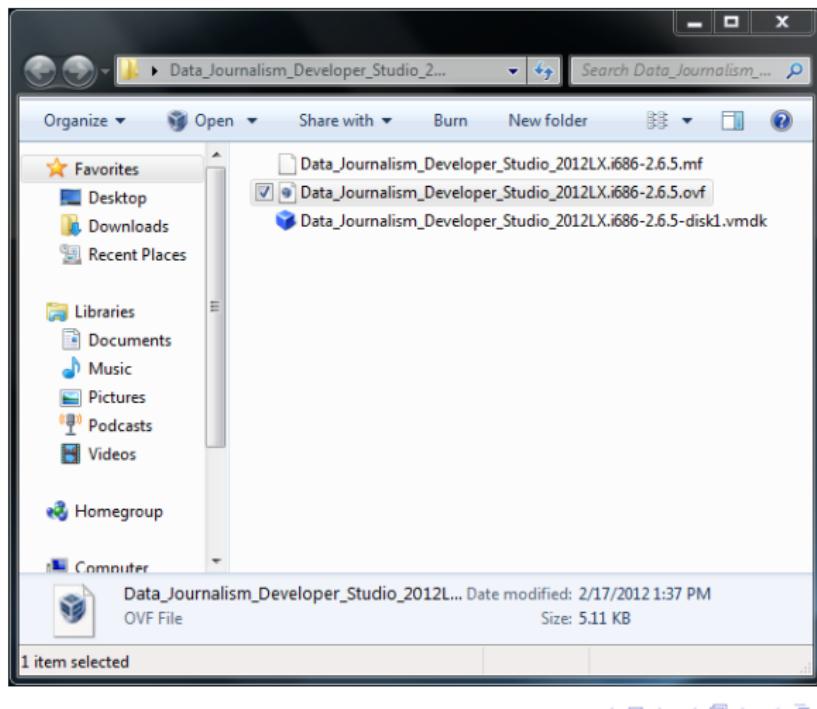
Step 5c: Appliance Unpacked

Drag appliance folder to desktop and close 7-Zip.

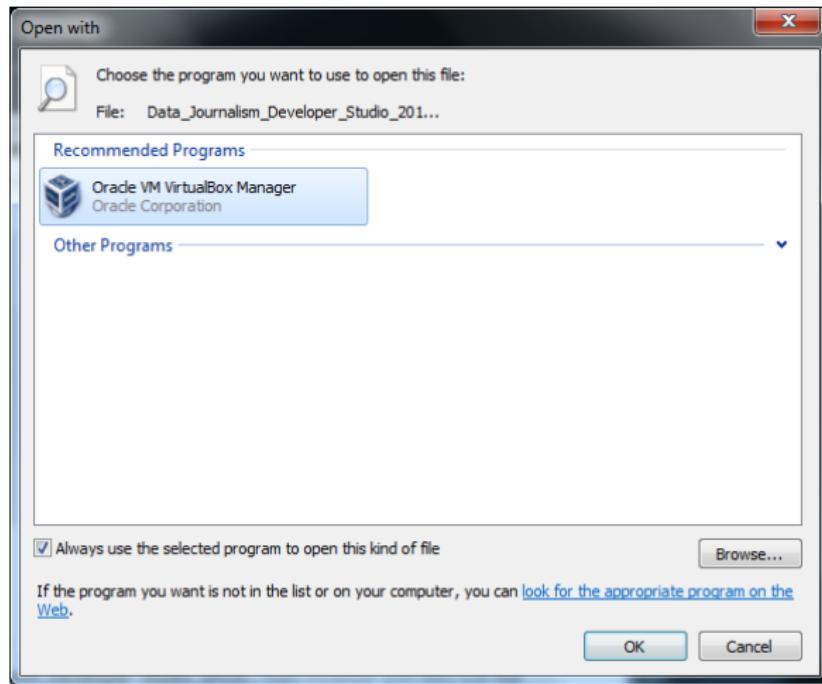


Step 6a: Open Appliance Folder

Right-click on “ovf” file and select “Open with”

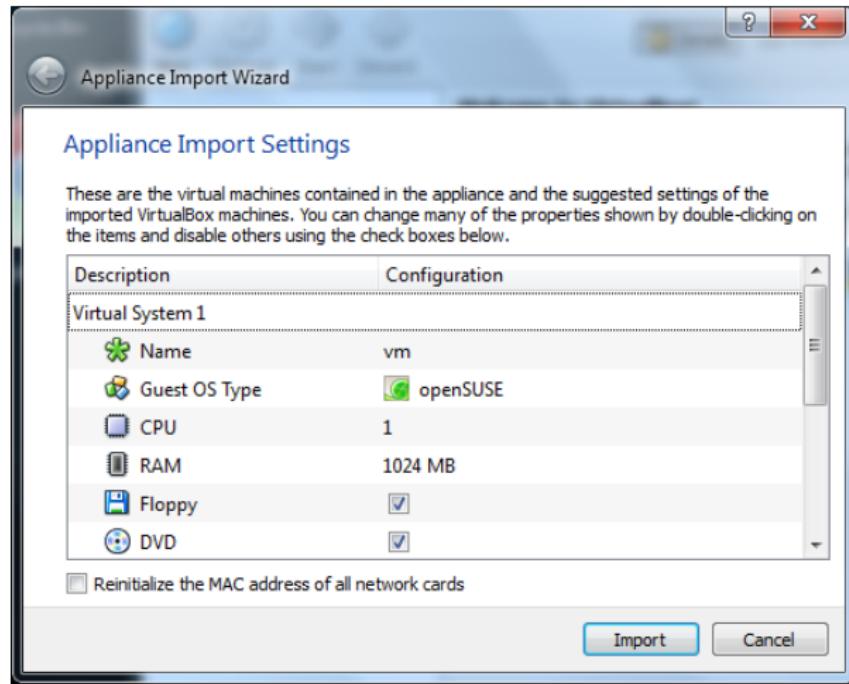


Step 6b: Open With Oracle VM VirtualBox Manager



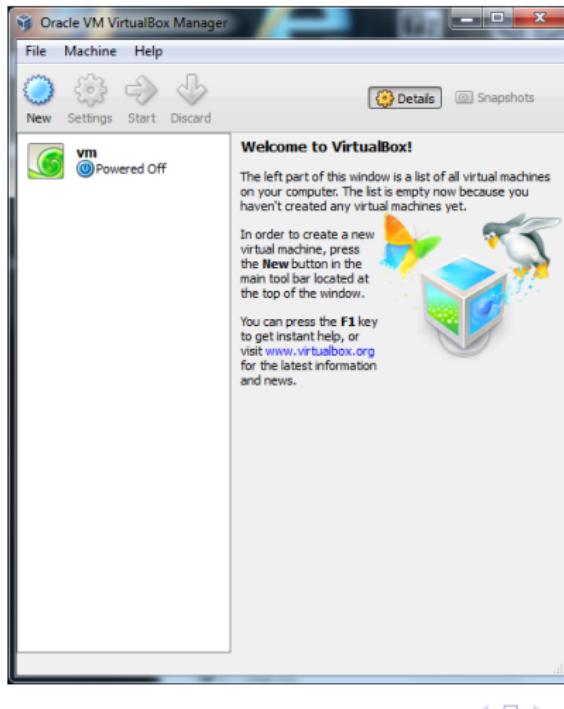
Step 6c: Appliance Import Wizard

Press "Import."



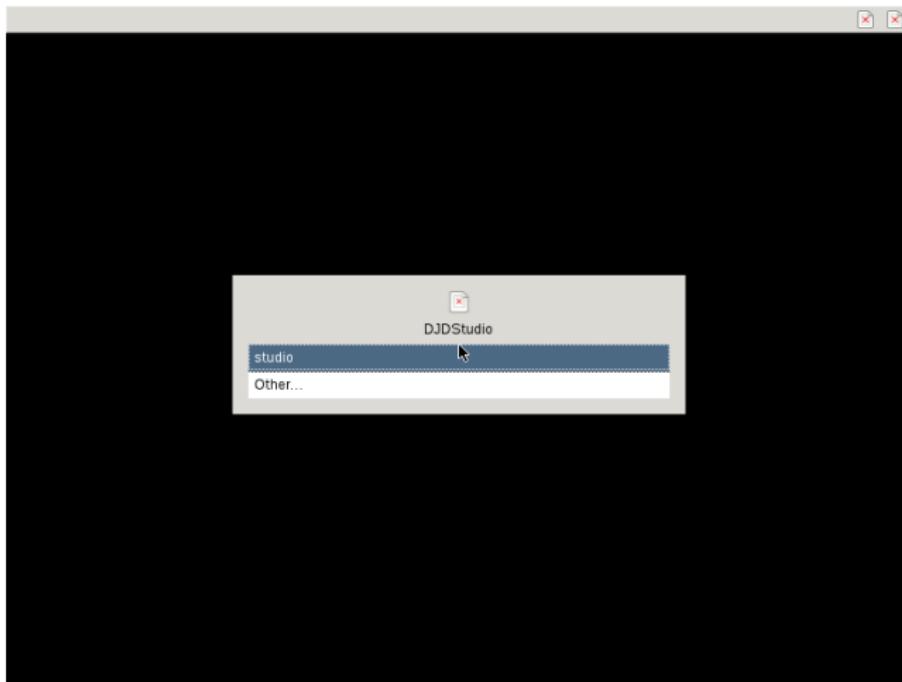
Step 6c: Import Successful

Select the “vm” appliance and press the green “Start” arrow.



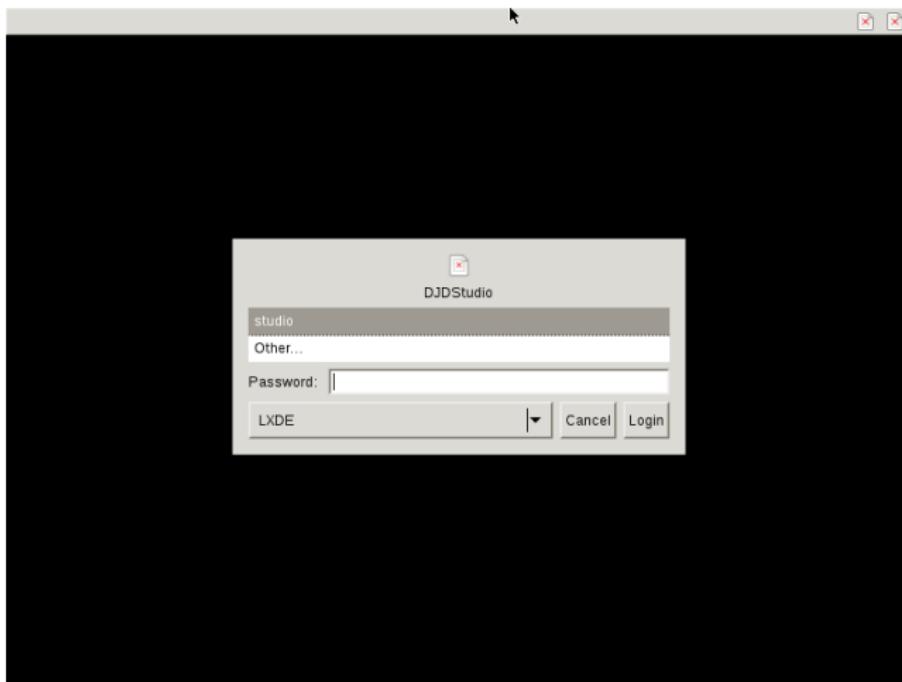
Step 7a: First Login as “studio”

Press “Enter.”



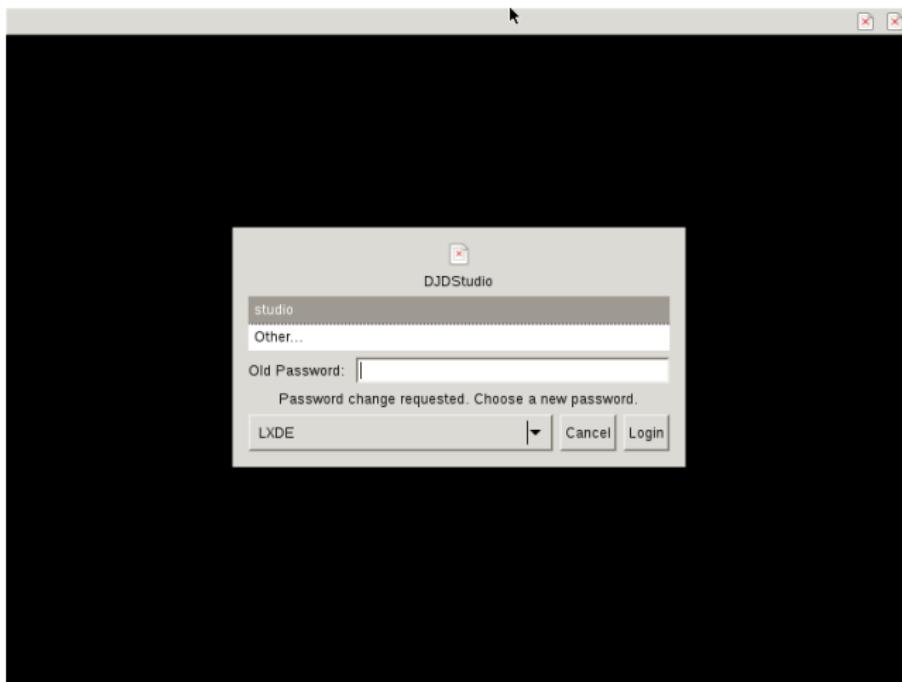
Step 7b: Enter Original “studio” Password

From the page where you downloaded the appliance: <http://j.mp/DJDS2012LX>.



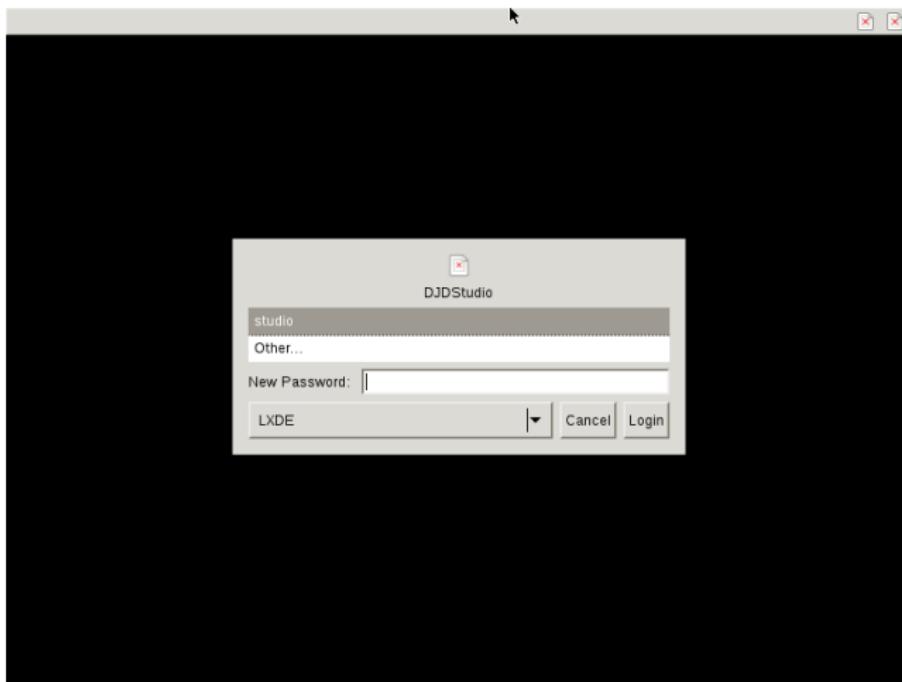
Step 7c: Change “studio” Password

“Old” password is the one you got from the gallery page.



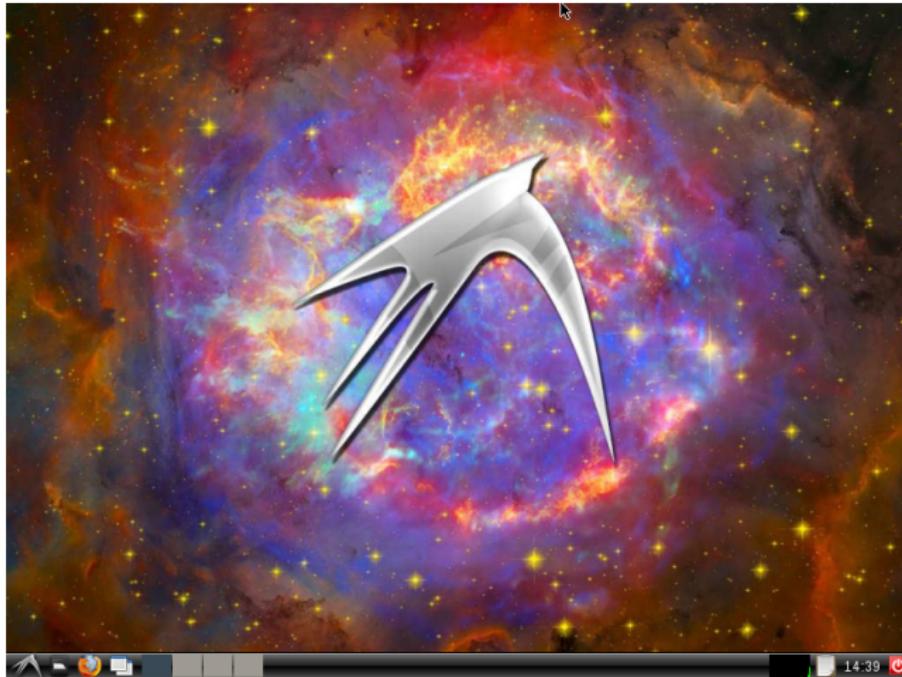
Step 7d: Choose a New Password

Easy for you to remember, impossible for others to guess, enter same password twice.



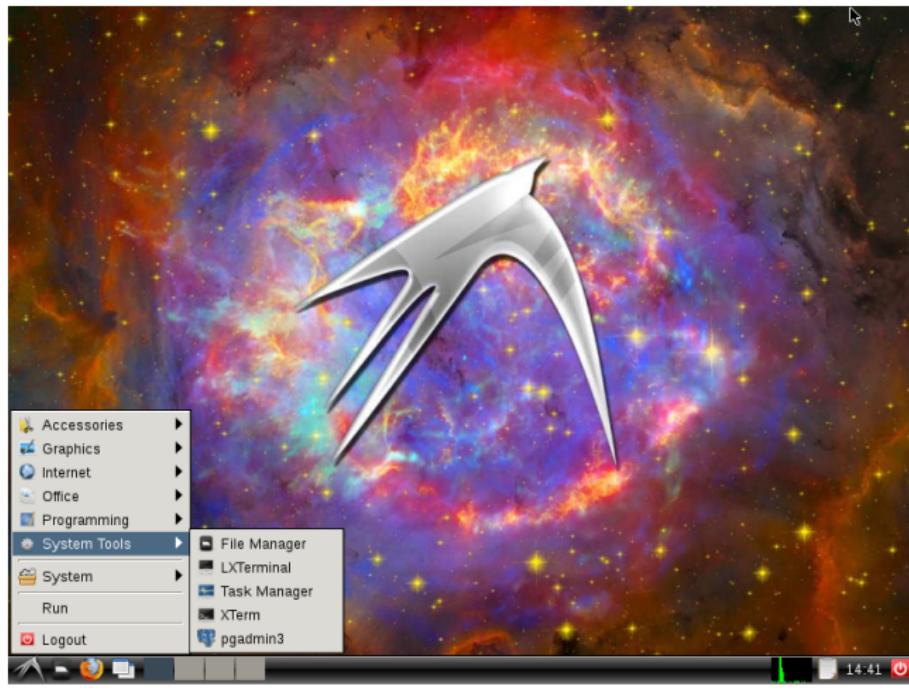
Data Journalism Developer Studio 2012LX Desktop

Press the swift-shaped “Start” button in the lower left corner.



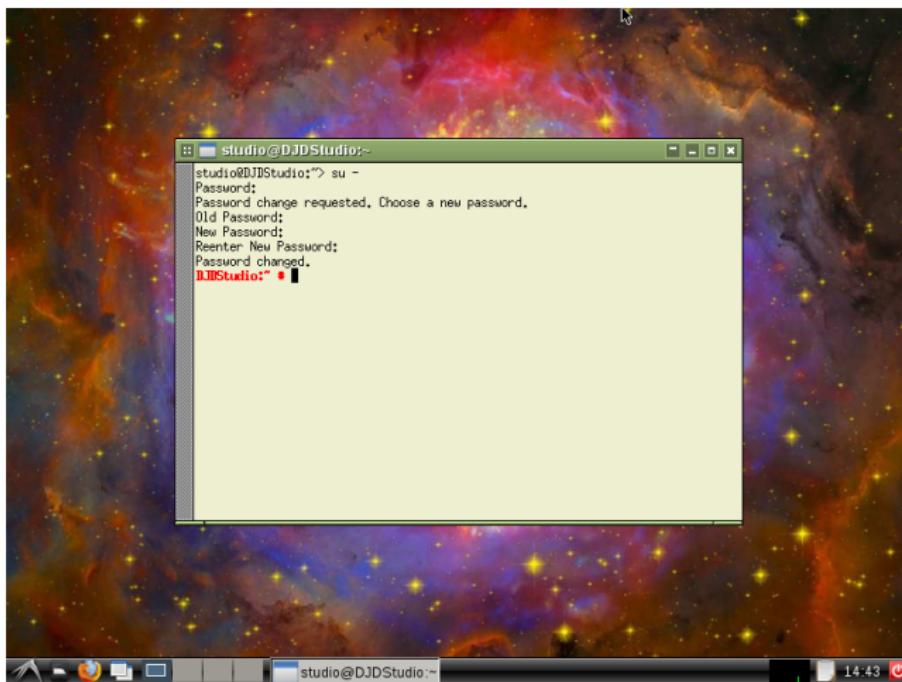
Step 8a: Open a Terminal Window

"Start -> System Tools -> XTerm"

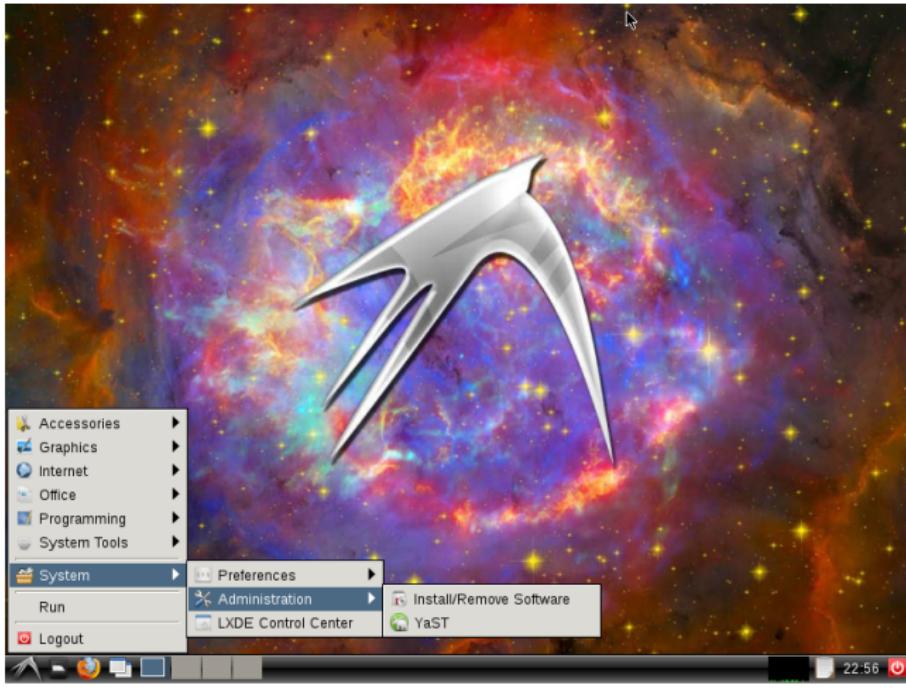


Step 8b: Change “root” Password

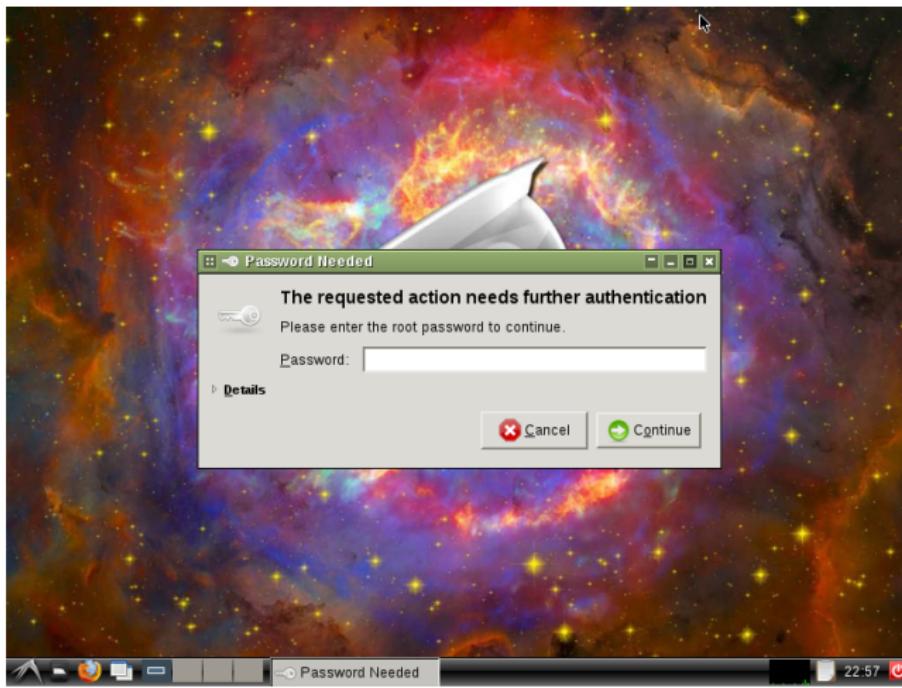
“Password” and “Old Password” are “root” password from <http://j.mp/DJDS2012LX>.



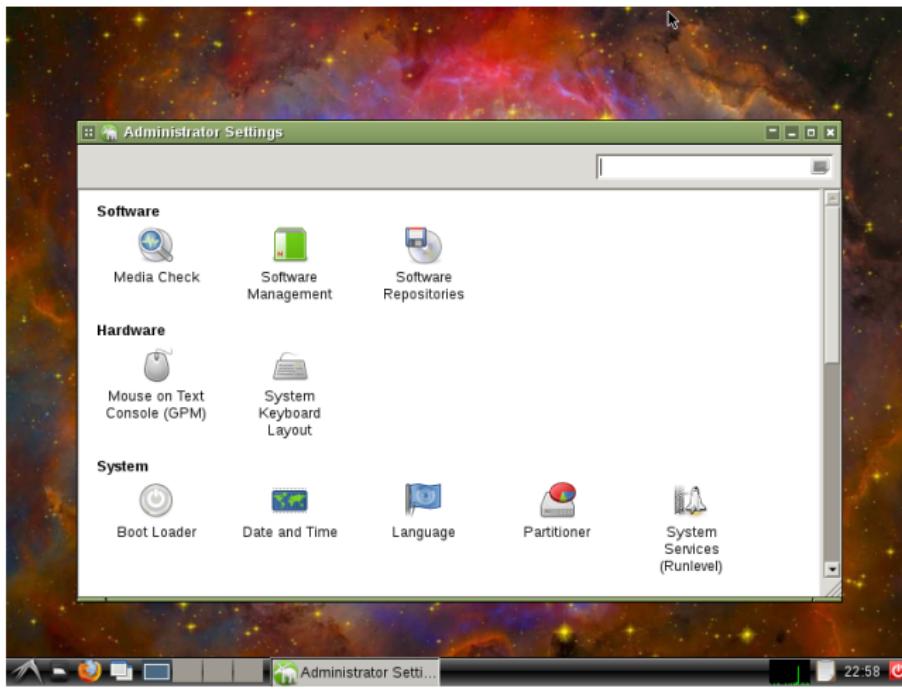
Step 9a: Set and Sync Clock



Step 9b: Enter New “root” Password



Step 9c: Single-Click “Date and Time”

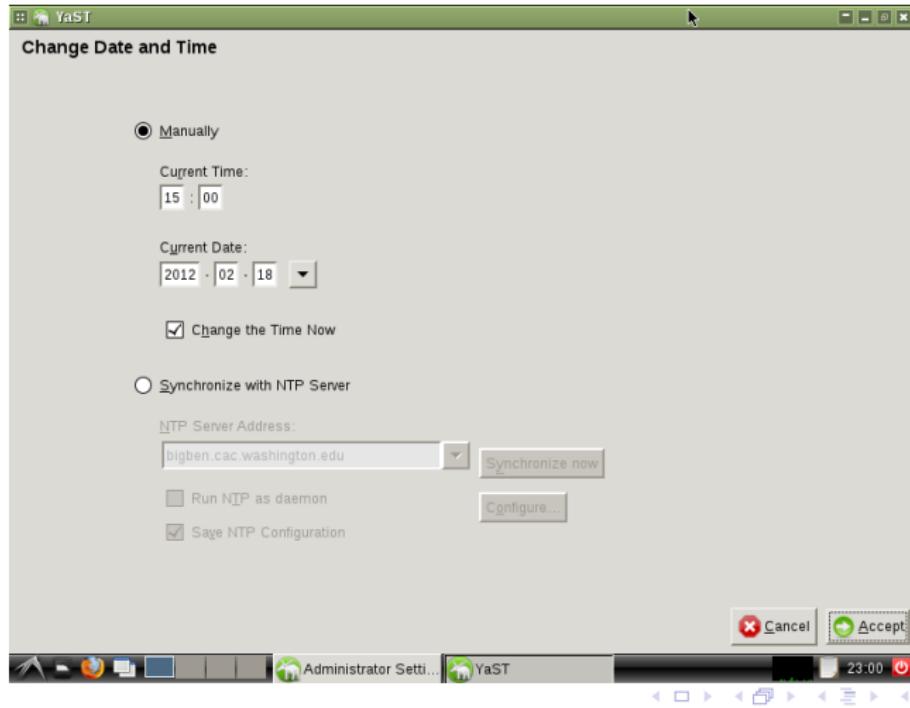


Step 9d: Choose Time Zone and Press “Change”

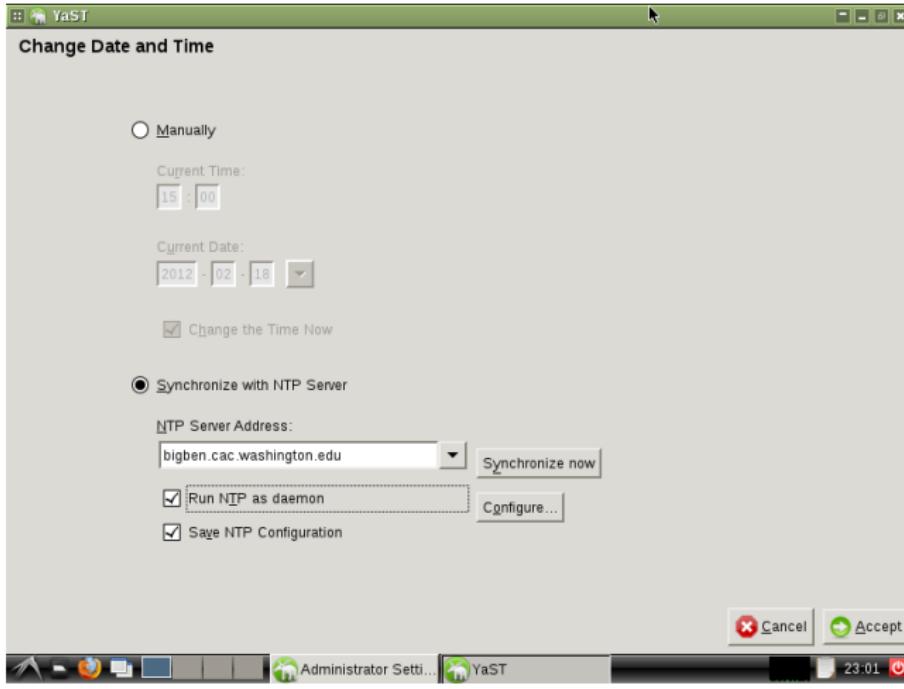
If time looks wrong, toggle “Hardware Clock Set To UTC” first!



Step 9e: Select “Synchronize With NTP Server” and “Run NTP As Daemon”



Step 9f: Press “Accept”



Step 9g: Press "OK"



Step 10: Install Guest Additions

I Keep Six Honest Serving Men

I Keep Six Honest Serving Men

I keep six honest serving-men

(They taught me all I knew);

Their names are What and Why and When

And How and Where and Who.

The Rest of the Story

Did you know there was more?

I send them over land and sea,

I send them east and west;

But after they have worked for me,

I give them all a rest.

The Rest of the Story

I let them rest from nine till five,

For I am busy then,

As well as breakfast, lunch, and tea,

For they are hungry men;

The Rest of the Story

But different folk have different views:

I know a person small -

She keeps ten million serving-men,

Who get no rest at all!

The Rest of the Story

She sends 'em abroad on her own affairs.

From the second she opens her eyes -

One million Hows, two million Wheres,

And seven million Whys!

Base Appliance

- ▶ The openSUSE 12.1 Linux operating system
- ▶ The Firefox Browser
- ▶ The The R Project for Statistical Computing
- ▶ The RStudio Integrated Development Environment
- ▶ The R Commander GUI
- ▶ The Node.js Network Application Development Framework

Base Appliance Media Creation

- ▶ The [Sigil WYSIWYG eBook Editor](#)
- ▶ The [Maquette WYSIWYG HTML5 User Interface Authoring Tool](#)
- ▶ The [Calibre eBook Library Manager](#)

Base Appliance Internet Data Collection Libraries

- ▶ Perl, Python and Ruby Scripting Languages
- ▶ Google Refine and Google Tesseract OCR Optical Character Recognition Engine
- ▶ Perl Net::Twitter API Interface
- ▶ Perl WWW::Mechanize Web Mining Tools
- ▶ Perl AnyEvent::Twitter::Stream Streaming API Interface
- ▶ The PostgreSQL Advanced Open Source Database
- ▶ The SQLite3 Open Source Database

Base Appliance R Library Packages

- ▶ Textir Sentiment Analysis and Topic Modeling
- ▶ tm Text Mining Package
 - ▶ Email plugin
 - ▶ Sentiment analysis plugin
 - ▶ Web mining plugin
- ▶ GGPlot2 Publication-Quality Graphics
- ▶ googleVis Google Visualization API Interface

The Data Journalism Developer Studio Is Modular

- ▶ The base appliance provides developer-level tools for building desktop or server applications
- ▶ The base appliance provides a browser-based desktop interface to cloud-based collaboration suites
- ▶ Add-on installation scripts provide end-user desktop and advanced specialized analysis tools

Add-On Productivity Suite Options

- ▶ Semantic Desktop
 - ▶ Evolution Email, Address Book and Calendaring
 - ▶ Mozilla Thunderbird Email Client
 - ▶ Tracker Semantic Data Storage / Indexing / Search
 - ▶ R Language `rrdf` and `SPARQL` Library Packages
- ▶ The LibreOffice Productivity Suite
- ▶ VoIP and Messaging
 - ▶ Ekiga Voice and Video Conferencing
 - ▶ Empathy Instant Messaging

Add-On Data Journalism Tools

- ▶ Digital media creation and editing
- ▶ Data collection, management and analysis
- ▶ Numerical data / visualization / exploration
- ▶ Financial and economic analysis
- ▶ Geospatial / mapping
- ▶ Natural language processing, machine learning and text mining

Add-On Digital Media Creation / Editing / Production

- ▶ Blender 3D Content Creation Suite
- ▶ The GIMP GNU Image Manipulation Program
- ▶ Scribus Open Source Desktop Publishing
- ▶ Inkscape Vector Graphics Editor
- ▶ Audacity Free Audio Editor and Recorder
- ▶ PiTiVi Video Editor

Add-On Numerical Data Visualization / Exploration / Presentation

- ▶ GGobi Exploration / Visualization System
- ▶ Mondrian Interactive Statistical Visualization System
- ▶ Rattle: Gnome Cross Platform GUI for Data Mining using R
- ▶ SciViews-R GUI

Add-On Financial and Economic Analysis

- ▶ QuantLib free/open-source library for quantitative finance
- ▶ R Empirical Finance Task View
- ▶ R Computational Econometrics Task View
- ▶ R Time Series Analysis Task View

Add-On Geospatial / Mapping

- ▶ PostGIS Geospatial Information System / Database
- ▶ GRASS Geospatial Information System
- ▶ GDAL Geospatial Data Abstraction Library
- ▶ PROJ.4 Cartographic Projections Library
- ▶ R Spatial Data Analysis Task View
- ▶ BARD (Better Automated ReDistricting)

Add-On Natural Language Processing / Text Mining

- ▶ WordNet English Lexical Database
- ▶ Python Natural Language Toolkit
- ▶ MALLET MAchine Learning for LanguagE Toolkit
- ▶ R Natural Language Processing Task View
- ▶ R Machine Learning Task View

Social Network Analysis

- ▶ Python NetworkX Network Analysis and Visualization Package
- ▶ Statnet tools for analysis, simulation and visualization of network data

Add-On Server Construction Frameworks

- ▶ NoSQL Databases
 - ▶ CouchDB
 - ▶ MongoDB
 - ▶ Redis
 - ▶ Riak
- ▶ Django

Links

- ▶ [Support Data Journalism Developer Studio On Fundry](#)
- ▶ [Download Data Journalism Developer Studio 2012LX From SUSE Gallery](#)
- ▶ [Data Journalism Developer Studio 2012LX On Github](#)
- ▶ [Borasky Research Journal](#)

Recommended Reading

- ▶ 'Facts are Sacred: The power of data (Guardian Shorts)' by Simon Rogers
- ▶ 'Multimedia Journalism: A Practical Guide' by Andy Bull
- ▶ 'R Through Excel' by Richard M. Heiberger
- ▶ 'Crafting Digital Media' by Daniel James
- ▶ 'ggplot2: Elegant Graphics for Data Analysis (Use R)' by Hadley Wickham
- ▶ 'Interactive and Dynamic Graphics for Data Analysis' by Dianne Cook

Recommended Reading - Advanced

- ▶ Finance: '*Asset Price Dynamics, Volatility, and Prediction*' by Stephen J. Taylor
- ▶ Geospatial: '*Applied Spatial Data Analysis with R (Use R!)*' by Roger Bivand
- ▶ Natural Language Processing: '*Natural Language Processing with Python*' by Steven Bird
- ▶ Machine Learning: '*The Elements of Statistical Learning*' by Jerome Friedman
- ▶ Web Data Mining: '*Web Data Mining*' by Bing Liu