

# Data Journalism Developer Studio 2012LX Overview

*Download Data Journalism Developer Studio 2012LX*

M. Edward (Ed) Borasky

*Borasky Research Journal*

February 7, 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

## 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

## Base Appliance Media Creation

- ▶ The Sigil WYSIWYG eBook Editor
- ▶ The Maqetta 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 Option

- ▶ The LibreOffice Productivity Suite
- ▶ Evolution Email, Address Book and Calendaring
- ▶ 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

## Add-On Server Construction Frameworks

- ▶ Node.js Scalable Network I/O Platform / NowJS Real-Time Web Application Framework / CoffeeScript
- ▶ NoSQL Databases
  - ▶ CouchDB
  - ▶ MongoDB
  - ▶ Redis
  - ▶ Riak
- ▶ Django

# 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!*

## 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