iCMS 2015
INTERNATIONAL CONFERENCE ON COMPUTING, MATHEMATICS AND STATISTICS

Bridging Research Endeavour in Computer and Mathematical Sciences

For more information, please visit http://www.icms2015.org

Organized by : FACULTY of COMPUTER & MATHEMATICAL SCIENCES UiTM KEDAH

Jointly organized by : RESEARCH & INDUSTRIAL LINKAGES

4th - 5th November 2015

Langkawi Island, MALAYSIA

UNIVERSITI TEKNOLOGI MARA CAWANGAN KEDAH

REPRODUCIBLE RESEARCH WITH R

Ciprian Alexandru

R-omania Team | www.r-project.ro

Keynote Speaker

Antoniade-Ciprian Alexandru

- Economics Faculty Ecological University of Bucharest, Romania - Dean, Associate Professor
- National Institute of Statistics, Romania - Expert trainer in data analysis and databases
- R-omania R-team member

alexcipro@yahoo/gmail.com

LinkedIn

R-project.ro

GitHub

R-omania Team

- Our Team acts as user community for development of R project among the Romanian individual persons, institutions and commercial and noncommercial organizations.
- The Romanian Team promote the R project for statistical computing to provide a free and open source software environment for data analysis and graphics in Romania.
- Support the further development of R and related open source software projects in Romania.
- Initiate, promote and coordinate research projects, support communication within the R user community, and organize or sponsor courses.
- Organize R-related scientific conferences and workshops, participate at relevant R conferences sponsored by others, and promote the use and development of R and R-related software in Romania.
- Publish manuals, journal articles and other R-related documents in printed and electronic form.
- Promote the using of R environment in universities and offer support for curricula development in the field of statistical software.

source: https://www.r-project.ro/

R-omania Team – Conferences & Workshops

- **2016 - Conference - New Challenges for Statistical Software - The Use of R in Official Statistics**
- 2016 – Workshop - Applied R to Social Sciences
- 2015 - Conference - New Challenges for Statistical Software - The Use of R in Official Statistics
- 2014 - Workshop - New Challenges for Statistical Software - The Use of R in Official Statistics
- 2013 - Workshop - State-of-the-art statistical software commonly used in applied economics

source: https://www.r-project.ro/

R-omania Team - Courses

- 3 to 5 days Courses:
  - "Statistics with Applications in R"
  - "Data Analysis in Statistics with R"
  - "Introduction to Statistics - Applications in R"
  - "Introduction in Small Area Estimation Techniques with Applications in R"
- One-day courses:
  - "R Statistical Software – Presenting Advantages of its use for Data Analysis"
  - "Introducing Statistics, the Need for Official Statistics"
  - "Statistical Analysis – from Theory to Practice"
  - "Concepts, Models and Techniques for Data Analysis"

source: https://www.r-project.ro/

R-omania Team - Courses

- Over 400 trained people from:
  - National Institute of Statistics – over 200
  - National Bank of Romania
  - Vodafone
  - Orange
  - UniCredit Bank
  - Ministry of Finance
  - Romanian Academy
  - Universities



R-omania Team - Research

- "R cu aplicatii in statistica", [en. "R with applications in statistics"], Published in 2015
- More than 25 research papers
- "R pentru Incepatori", the Romanian version of "R for Beginners" by Emmanuel Paradis, translated by Ana-Maria Dobre, Aug, 2013

source: https://www.r-project.ro/

Topics

- Research where they are going?
- Replication or Reproducible?
- Why Reproducibility?
- What is necessary for Reproducibility?
- Research Pipeline
- Research about Reproducible Research
- Missing Reproducibility
- Literate (Statistical) Programming
- Research Pipeline in R
- Development Tools
- Presentation Tools
- Basic principles
- Markdown

Research where they are going?

- A lot of
  - Computing
  - Data analysis
  - Data manipulation or processing
  - Big Data
  - Communicating the results
  - Reconstructing
  - **Reproducible**

Replication or Reproducible?

- Replication – an independent study came with same conclusion as the original study

- Reproducible
  - original data and code are studied by an independent analysts obtaining the same results of the original study
  - *data analysis* can be repeated with same results

Reproducible ≠ Correctness

- a reproducible study could be wrong
- a wrong assumptions can be identified by an independent replication process

Why not Replication research?

- Lack of time
- Lack of financial resources
- Unique

Why Reproducibility?

- Enhancing scientific evidence
- It is the only guarantee about a study
- Ensures transparency
- Creates confidence
- Validation of the data analysis

What is necessary for Reproducibility?

- Publication of data sources and programs (R code)
  - analytic data, analytic code, documentation, standard tools for distribution
- Conducting studies using independent data sources

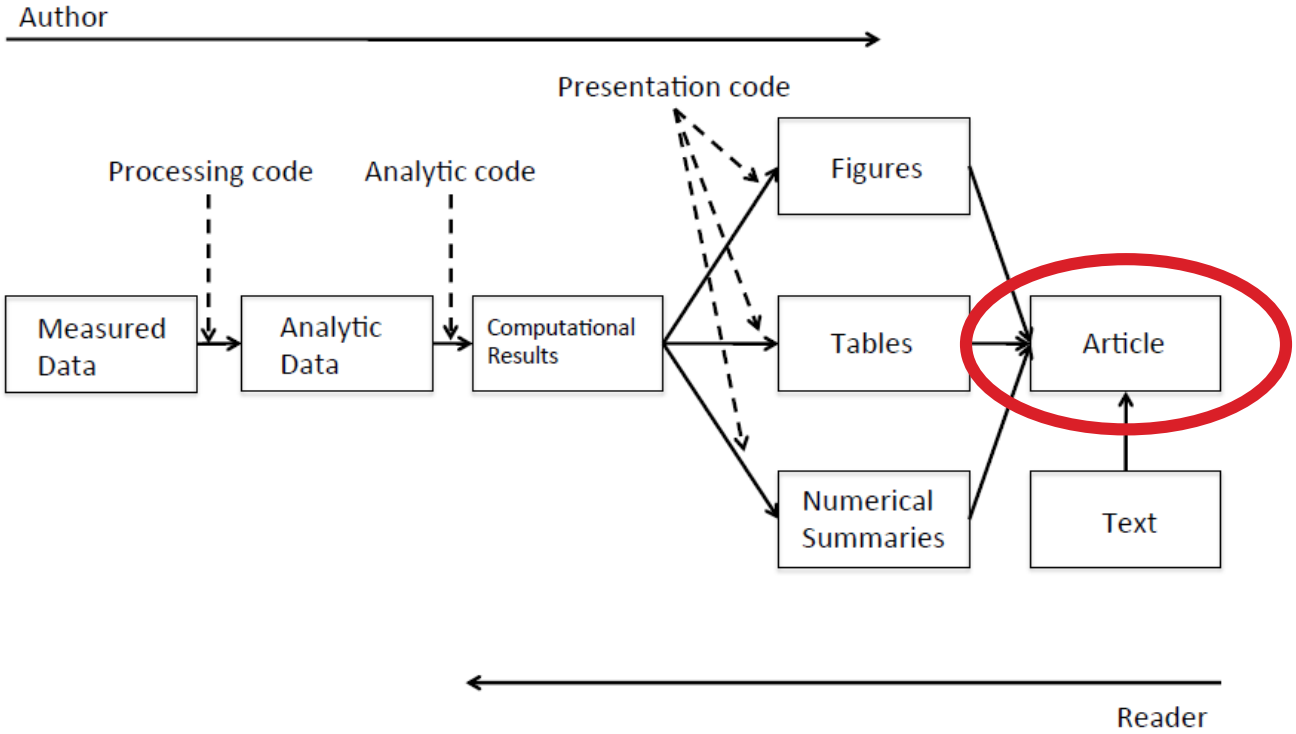
How easy can do our work reproducible?

- preparation of data in a friendly format (standard)
- find a web server for uploading code & data with “almost” limitless availability
- using some tools to help readers (well known or well documented, eventualy)
- making available a guidelines for future use

Conclusion:

a threshold between more “present time” consuming and saving for the future

Research Pipeline



Roger D. Peng, *Reproducible Research: Concepts and Ideas*

Research about Reproducible Research

Data Replication & Reproducibility

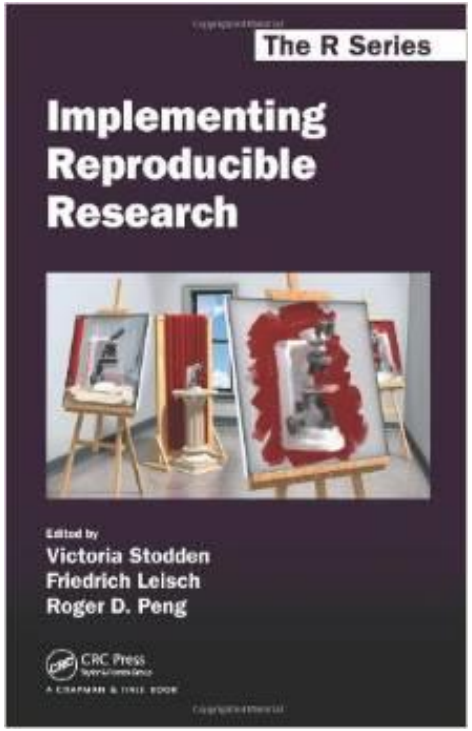
PERSPECTIVE

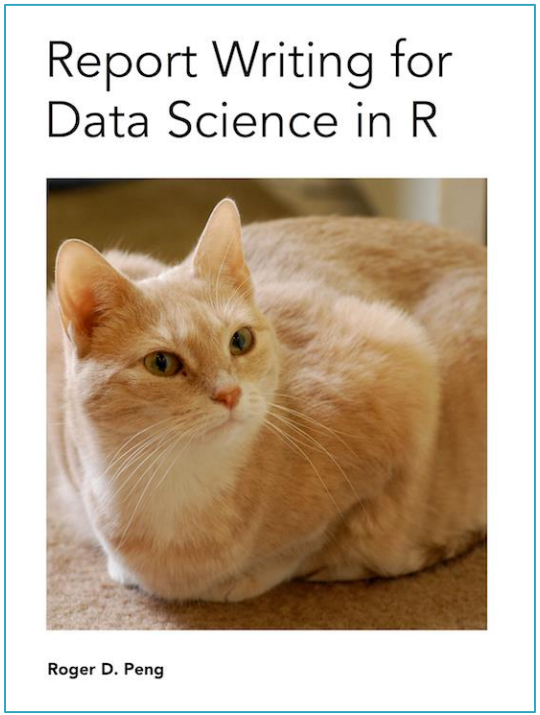
Reproducible Research in Computational Science

Roger D. Peng

Computational science has led to exciting new developments, but the nature of the work has exposed limitations in our ability to evaluate published findings. Reproducibility has the potential to serve as a minimum standard for judging scientific claims when full independent replication of a study is not possible.

Research about Reproducible Research

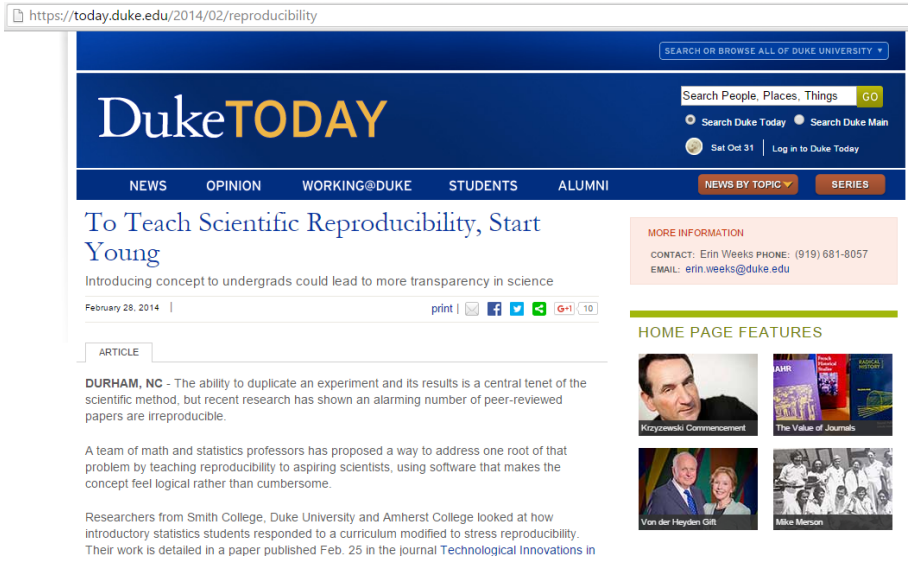




Missing Reproducibility

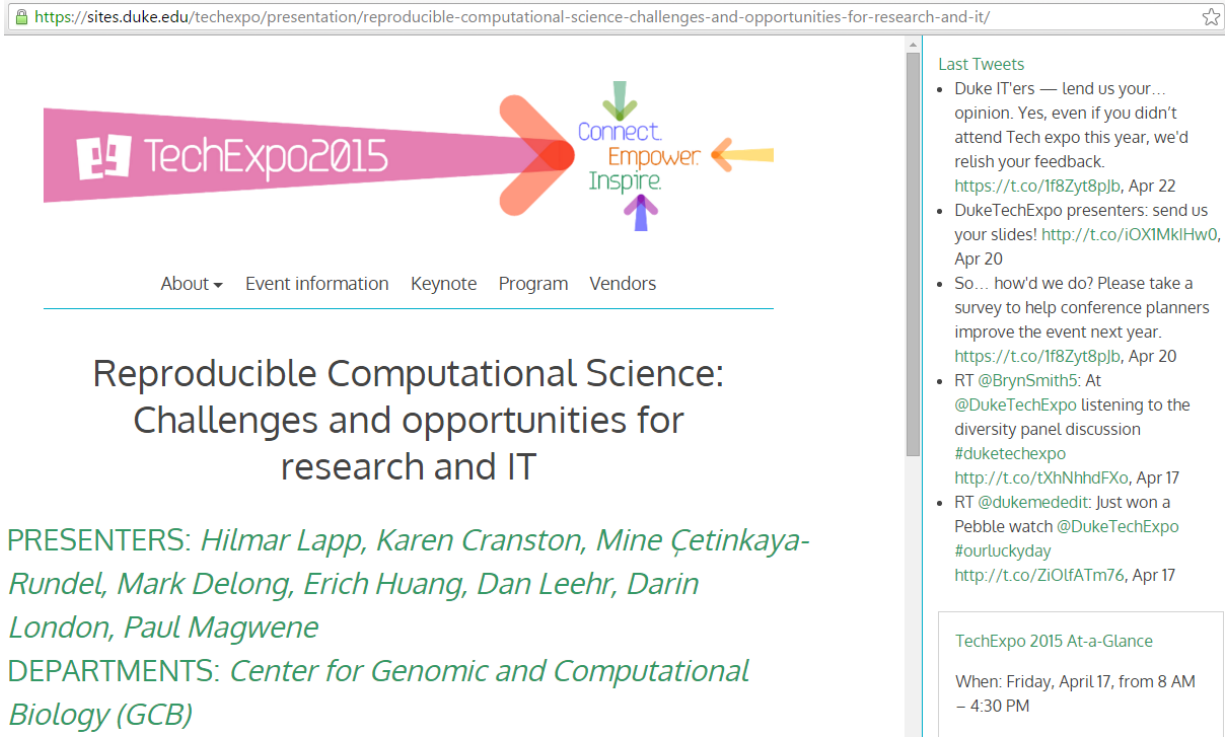
- Duke University

Lesson Learned



https://today.duke.edu/2014/02/reproducibility

Lesson Learned



https://sites.duke.edu/techexpo/presentation/reproducible-computational-science-challenges-and-opportunities-for-research-and-it/

Reproducibility in computational research

- Assign a unique ID for each version of released data & code
- Use open licensing for data & code
- Workflow tracking should happen during the research process
- In your publication, include a statement describing computing environment(s) and software version(s)
- Make data, code and methods available and accessible
- Version control
- Publish data, code and methods in non-proprietary formats (if possible)
- Cite any 3rd party data and code
- Follow data and code sharing guidelines for funded research

Jane Frazier, Data Librarian, Australian National Data Service, *Research Data & Output Management*

Stodden and Miguez. Best practices for computational science: software infrastructure and environments for reproducible and extensible research. Journal of Open Research Software, 2014. [DOI:10.5334/jors.ay]

Yale Law School Roundtable on Data and Code Sharing. Reproducible research: addressing the need for data and code sharing in computational science. Columbia University Academic Commons, 2010. [DOI:10.1109/MCSE.2010.113]

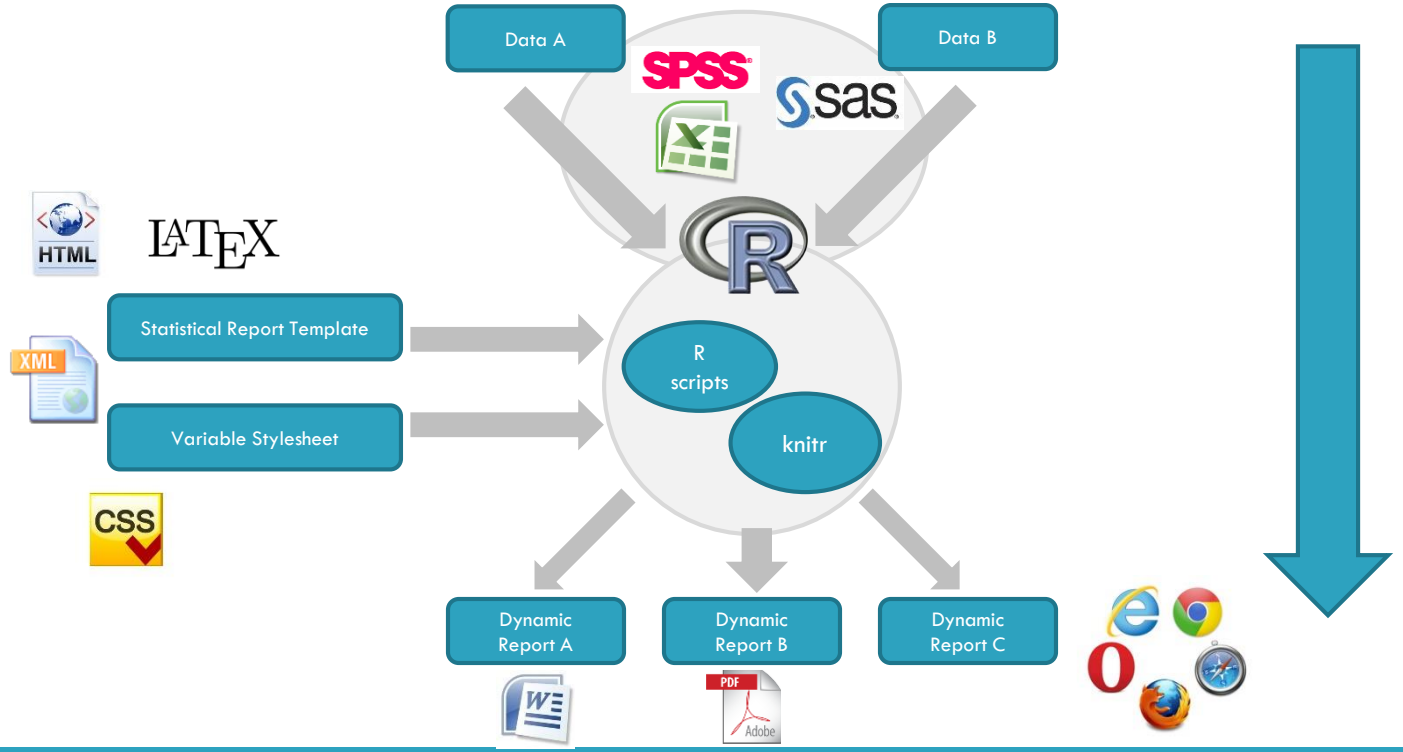
Literate (Statistical) Programming

- a research paper/an article = a stream of text and code chunks
- code chunks = loads and manipulate data, computes results, presents tables, figures etc.
- literate programs produce:
  - human-readable documents
  - machine-readable documents

Where do we start?

- Using some tools
- Basic principles

Research Pipeline in R



Development Tools

- Documentation language (human readable)
  - Sweave ($\LaTeX$) [http://leisch.userweb.mwn.de/Sweave/]
  - knitr ($\LaTeX$, Markdown, HTML) [http://yihui.name/knitr/]
- Programming language (machine readable)
  - R
  - Python

Presentation Tools

- GitHub - powerful collaboration, code review, and code management for open source and private projects [https://github.com/]

- RPubs - web publishing from R [https://rpubs.com/]

Sweave

- Included in R basic installation
- Documentation language $\LaTeX$
- Lacks features like caching, multiple plots, per chunk, mixing programming languages
- Not frequently updated
- Developed by Fritz Leisch, core member of R

http://leisch.userweb.mwn.de/Sweave/

knitr

- Contributed package (need installation)
- Documentation language:
  - LaTeX
  - Markdown
  - HTML
- Mix other programming languages
- Very popular package
- Developed by Yihui Xie while he was a graduate student at Iowa State

http://yihui.name/knitr/

Basic principles

"Golden Rule of Reproducibility: Script Everything",
Roger D. Peng

Basic principles - Data

- Create an analysis folder unique for each research
- Raw data
  - Saved data - stored in your analysis folder
  - Web data – include url, description, accessed date
- Processed data
  - Named clearly (files and variable)
  - Make it tidy

Basic principles - Figures

- Exploratory figures
  - In general include all data set
  - Explore your data until they told you everything
  - They are not pretty, but very useful

- Final figures
  - Use the most representative figures
  - Axes/Labels/Colors are set to make the figures clear and easy understandable
  - Multiple panels is preferred

Basic principles – R code

- Raw / unused scripts
  - Small comments is recommended (our memory is not a HDD)
  - Use multiple versions
  - Keep all intermediary analysis, until you are sure that they are useless
- Final scripts
  - Clear commented (small in row, large for sections)
  - Include processing details
  - Keep only relevant analysis, that are included in final paper
- R markdown files
  - Used for generate reproducible reports
  - Integrates Text and R code

Basic principles – Text

- README files
  - Optional for R markdown
  - Include all instructions for analysis

- Text of analysis / report
  - Title, introduction, methods, results and conclusions
  - Include only relevant analysis regarding the conclusions
  - References are important

Coding Standards for R

- Use text files and editor
- Indent the code, 4 spaces minimum, or 8 (better)
- Width of the code – limits to 80 columns
- Functions – keep it small

Markdown


```
Italics
*This text will appear italicized!*

Bold
**This text will appear bold!**

Headings

## This is a secondary heading
### This is a tertiary heading
```

Markdown


```
Unordered Lists

- first item in list
- second item in list
- third item in list

Ordered Lists
1. first item in list
2. second item in list
3. third item in list
```

Markdown

Links

```
[Johns Hopkins Bloomberg School of Public
Health](http://www.jhsph.edu/)

[Download R](http://www.r-project.org/)

[RStudio](http://www.rstudio.com/)
```

Markdown


```
Advanced Linking

I spend so much time reading [R bloggers][1] and [Simply
Statistics][2]!

[1]: http://www.r-bloggers.com/ "R bloggers"

[2]: http://simplystatistics.org/ "Simply Statistics"
```

Markdown

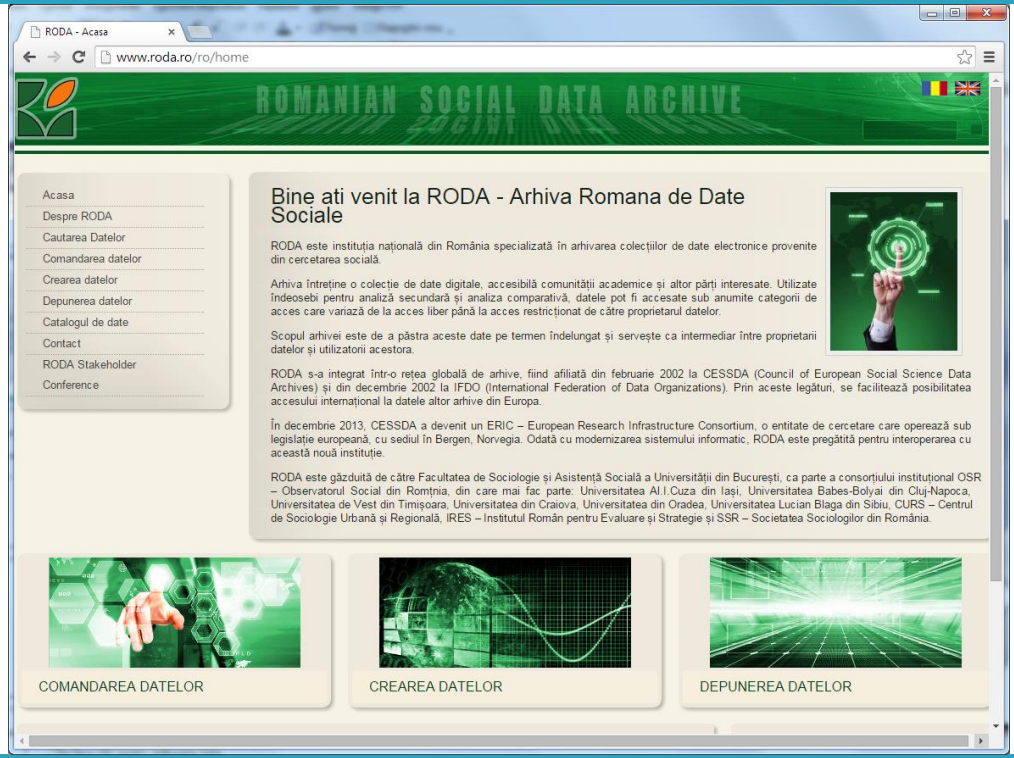

```
Newlines require a double space after the end of a line.

First line
Second line

First line Second line

First line
Second line
```

Markdown

- The Offical Markdown Documentation [http://daringfireball.net/projects/markdown/basics]
- Github's Markdown Guide [https://help.github.com/articles/github-flavored-markdown/]

Romanian Social Data Archive

- www.roda.ro



Reproducible Research - Economics

RePEc + RODA ARHIVA ROMÂNĂ DE DATE SOCIALE = Reproducible Research Papers

GitHub

- https://github.com/

References

- The Real Reason Reproducible Research is Important [http://simplystatistics.org/2014/06/06/the-real-reason-reproducible-research-is-important/
- Reproducible Data Analysis [http://www.biomedicale.univ-paris5.fr/SpikeOMatic/ReproducibleDataAnalysis/ReproducibleDataAnalysis.html]
- Get your R education going with GitHub [http://www.r-bloggers.com/get-your-r-education-going-with-github/]
- Everyone loves R markdown and Github; stories from the R Summit, day two [http://www.r-bloggers.com/everyone-loves-r-markdown-and-github-stories-from-the-r-summit-day-two/]
- Introducing the Reproducible R Toolkit and the checkpoint package [http://www.r-bloggers.com/introducing-the-reproducible-r-toolkit-and-the-checkpoint-package/]

Thank you!

Ciprian Alexandru
alexcipro@yahoo.com
https://github.com/alexcipro/iCMS2015_Keynote