Research and Academic Software Projects at the Institute for Quantitative Social Science

Mercè Crosas, Ph.D.
Chief Data Science and Technology Officer
IQSS, Harvard University
twitter: @mercecrosas web: mercecrosas.com
The Big Picture

Identify a problem or need in research or academia
The Big Picture

Identify a problem or need in research or academia

Build a technology solution, easy-to-use, gives control to researcher
The Big Picture

Identify a problem or need in research or academia

Build a technology solution, easy-to-use, gives control to researcher

Generalizable, Open-source
The Big Picture

Identify a problem or need in research or academia

Build a technology solution, easy-to-use, gives control to researcher

Build a community that makes the technology better

Generalizable, Open-source
Example: Dataverse
Example: Dataverse

- How do we increase data sharing to improve research transparency and replication with incentives to researchers?
Example: Dataverse

- How do we increase data sharing to improve research transparency and replication with incentives to researchers?
- Provide a repository solution, where researchers have control of branding and access of their data, and get credit through data citation.
Example: OpenScholar
Example: OpenScholar

How do we enable scholars to build their academic web sites in a cost effective way?
Example: OpenScholar

- How do we enable scholars to build their academic web sites in a cost effective way?

  - Provide a web site builder with pre-set features for academics, where a single hosting serves thousands of sites.
Example: Zelig
Example: Zelig

How do we simplify using thousands of R statistical methods built by different authors?
Example: Zelig

- How do we simplify using thousands of R statistical methods built by different authors?

- Provide a statistical package that uses the same three commands for all methods, with consistent documentation.
Example: Consilience
Example: Consilience

- How do we make sense of thousands (or millions!) of texts?
Example: Consilience

- How do we make sense of thousands (or millions!) of texts?
- Provide an application that helps researchers explore many possible ways of categorizing documents.
The Process

- Research, standards & best practices
- Input from users, community, stakeholders
- Development, testing & releases

Dataverse Case Study
The Process

- **Research, standards & best practices**
- **Input from users, community, stakeholders**
- **Development, testing & releases**

metadata standards, harvesting protocols, data transfer, data citation, provenance, connecting to journals, integrating with cloud computing, ....
metadata standards, harvesting protocols, data transfer, data citation, provenance, connecting to journals, integrating with cloud computing, ….

usability testing, community calls, annual community meeting, pull requests

Input from users, community, stakeholders

Research, standards & best practices

Development, testing & releases

The Process
The Process Details

Dataverse Case Study
An agile process, integrating Waffle + GitHub + Jenkins, including these steps:
An agile process, integrating Waffle + GitHub + Jenkins, including these steps:

Backlog > Ready > Dev > Code Review > QA > Usability Test > Polishing > Done
An agile process, integrating Waffle + GitHub + Jenkins, including these steps:
Backlog > Ready > Dev > Code Review > QA > Usability Test > Polishing > Done
Not only Best Practices in Process, but also in Coding
Not only Best Practices in Process, but also in Coding

Best Practices for Scientific Computing

Published: January 7, 2014 • http://dx.doi.org/10.1371/journal.pbio.1001745
Not only Best Practices in Process, but also in Coding

1. Write programs for people, not computers.
Not only Best Practices in Process, but also in Coding

1. Write programs for people, not computers.
2. Let the computer do the work.
Not only Best Practices in Process, but also in Coding

1. Write programs for people, not computers.
2. Let the computer do the work.
3. Make incremental changes.
Not only Best Practices in Process, but also in Coding

1. Write programs for people, not computers.
2. Let the computer do the work.
3. Make incremental changes.
4. Don't repeat yourself (or others).
Not only Best Practices in Process, but also in Coding

1. Write programs for people, not computers.
2. Let the computer do the work.
3. Make incremental changes.
4. Don't repeat yourself (or others).
5. Plan for mistakes.
Not only Best Practices in Process, but also in Coding

Best Practices for Scientific Computing


Published: January 7, 2014 • http://dx.doi.org/10.1371/journal.pbio.1001745

1. Write programs for people, not computers.
2. Let the computer do the work.
3. Make incremental changes.
4. Don't repeat yourself (or others).
5. Plan for mistakes.
6. Optimize software only after it works correctly.
Not only Best Practices in Process, but also in Coding

1. Write programs for people, not computers.
2. Let the computer do the work.
3. Make incremental changes.
4. Don't repeat yourself (or others).
5. Plan for mistakes.
6. Optimize software only after it works correctly.
7. Document design and purpose, not mechanics.
Not only Best Practices in Process, but also in Coding

1. Write programs for people, not computers.
2. Let the computer do the work.
3. Make incremental changes.
4. Don't repeat yourself (or others).
5. Plan for mistakes.
6. Optimize software only after it works correctly.
7. Document design and purpose, not mechanics.
8. Collaborate.
Impact at Harvard

OpenScholar®

6,833 OpenScholar sites created

13,904 Registered users

75,378 Publications posted

24 Academic departments
Impact at Harvard

243 Dataverses from Harvard affiliates

1,226 Datasets by Harvard affiliates as authors

1,427 Registered Harvard users
Broader Impact
Dataverse world-wide impact
Dataverses by Category

- Research Projects (57.9%)
- Researchers (20.7%)
- Journals (12.7%)
- Organizations Institutions (5.3%)
Datasets by Subject

Social Sciences (48.7%)

Other (6.1%)
- Earth and Environmental Sciences (4.1%)
- Business and Management (4.0%)
- Agricultural Sciences (3.9%)

Arts and Humanities (5.4%)
- Law (2.4%)
- Astronomy and Astrophysics (2.7%)
- Computer and Information Sciences (1.4%)
- Chemistry (1.4%)
- Mathematical Sciences (1.0%)
- Physics (0.7%)

Medicine, Health and Life Sciences (13.9%)
53 Stats Models, easy-to-use
Thank you!

Presented by @mercecrosas mercecrosas.com

dataverse.org
dataverse.harvard.edu

zeligproject.org

openscholar.org
projects.iq.harvard.edu
scholar.harvard.edu

Coming soon!

iq.harvard.edu