Data Management Practices for Programming [1]

Submitted by sophisticus on Fri, 2015-04-17 13:34 Thursday, July 16, 2015 - 08:30 to 10:00
Event: Summer Meeting 2015 [2]
Session Type: Breakout [3]
Expertise Level: Beginner [4]
Collaboration Area: Data Management Training [5]
Data Preservation [6]
Preservation and Stewardship [7]
Student Cluster [8]

Abstract/Agenda: The purpose of this session is to explore the manner in which the principles and practices of data management as discussed in ESIP “Data Management For Scientists Short Course” (Short Course) could be made applicable to software applications/scripts.

The session is meant to be differentiated from software carpentry or best practices for creating scripts. In other words, the session is not intended to teach attendees how to develop software applications or how to script efficiently. Rather, based on the key concepts presented in the Short Course, the session will seek to translate the principles and practices of data management to appropriate actions for individuals who are involved in developing software applications/scripts, such as computer scientists and scientists, for their projects. The aim is for the resulting software applications/scripts to be more “curation-ready”.

Sample of principles and practices to be explored include:

1. What does it mean to "curate" a software applications/scripts?
2. How can curating software program/scripts enhance my reputation?
3. What are the basic curation elements to consider in order to facilitate a deployable open-source software applications/scripts?
4. How to choose a program/file format to enhance interoperability?
5. What are the metadata content that should be created for the software applications/scripts?

Notes:

1. **How can curating software program/scripts enhance my reputation?**
    1. When the data and its related processes/methods/software cited in a paper are not traceable to an identifiable repository, the “credential” of the data could be questioned.
        1. This affects not only the trustworthiness of the data, but also the overall paper.
        2. The sources of the resources referenced in articles, including software, are crucial provenance that would influence the trustworthiness of the paper.
    2. As a result, the audience agrees that curating software program/scripts could enhance reputation by providing traceability and transparency of the provenance.

3. Other formats of digital objects that should also be considered for curation in order to enhance science reputations:

    1. Workflows

    2. Simulations/Models

    3. Images/Photos

2. **What are the basic curation elements to consider in order to facilitate a deployable open-source software applications/scripts?**

    1. Currently, although GitHub might be a common way used to share software scripts/codes openly, it is still important to note the following challenges:

        1. Description and assessment of the software utility is difficult.

        2. Reuse is also often challenging because the types and the levels of details of documentation associated and available with the software scripts/codes are often inconsistent.

3. **How to choose a program/file format and naming convention to enhance interoperability?**

    1. Changing file name is accepted when merging files with larger collection; however it is crucial to coordinate the changes and to ensure the changes are well documented.

    2. Sharing the extension of the program/file format can help in understanding the interoperability as well.

    3. Choosing program type versus file format might be based on different criteria. For instance, when choosing file format, it could depend on the available tools and support available to work with the file format. However, when choosing program type, it might depend more on the programmer’s knowledge and the tasks that the program needs to accomplish.

        1. Although it is desirable for the collaboration community to use the same programming language, It is not easy to impose or enforce the use of same programming language. It might also not be pragmatic to expect the entire collaboration community will use the same programming language.

        2. Some noted that as long as the input and the output are interoperable/standardized, the program types used to create the input and to generate/process the output might not need to be exactly interoperable.

        3. The nature of collaboration also influences the selection of the programming language.

4. **What are the metadata content that should be created for the software applications/scripts?**

    1. The metadata content should not be language dependent.

        1. The metadata content should provide at least the information regarding what the software is for and how it achieves its goals (i.e. what it is, how to use it, and how does it do what it does).

    2. XML schema implementation of the software’s metadata might be an option, such as the software description section of the Ecological Metadata Language (EML)..

        1. However, it is important to connect the software metadata with the software itself.

    3. Using RDF representation/ontology to describe software with URI might be another option.

    4. It is also important to consider the generation of DOI for the software and the inclusion of DOI information with the software’s metadata.

    5. It is crucial to capture the context of the software in the metadata as well.

5. **How to provide access to your software applications/scripts for a broader user community?**

    1. It is important to describe the full “ecosystem” or the dependencies when sharing software with a broader user community.

    2. GitHub might be useful currently, but it is not a long term repository, and might fall out of favor over time.

   1. For example, Google Code is being shut down, so it is important to consider a sustainable solution when sharing and archiving code.

3. It is also to consider the “versions” of the software to share with the community. Different options might include:

   1. The exact version that was used to generate specific data products.

      1. When the data products are referenced in papers, this version of software should also be cited.

   2. A “working” software that is constantly being worked on and enhanced.

      1. Different versions and the associated changes should be documented.

   3. Software that need to be migrated to different platform or languages over time.

      1. The migration process and any interoperability dependencies, including the descriptions of the the platforms, should be documented.

Attachments/Presentations: DataManagementForProgramming_Hou - ESIP 07-16-2015.pdf [9]
DataManagementForProgramming_Peterson - ESIP 07-16-2015.pdf [10]

Session Leads:

Name: Sophie Hou [11]
Organization(s): UCAR/NCAR [12]

Name: Fox Peterson [13]
Organization(s): H.J. Andrews Experimental Forest LTER [14]
Email: fox@tinybike.net [15]

Notes takers:

Name: Fox Peterson [13]
Organization(s): H.J. Andrews Experimental Forest LTER [14]
Email: fox@tinybike.net [15]

Name: Sophie Hou [11]
Organization(s): UCAR/NCAR [12]

Participants:
In Person: Fox Peterson, Sophie Hou, Audrey Mickle, Maksym Petrenko, Aaron Herbert, Aubrey Beach, Paul Foster, Doug Fils, Alan Righter, Sven Bohm, Donna Scott, Ruth Duerr, Greg Yetman, Nancy Wilkins-Diehr, Pam Mlynczak, Victor Zlotnicki, Wade Sheldon, Justin Goldstein, Mike McCann, Shelley Olds, Shannon Rauch, Jocelyn Elya, Ross Bagwell, Karl Benedict, Tom Cram

Remote: Patrick West

Teaser: Data Management Practices for Programming: A session proposed by the 2015 Fellows for the Summer Meeting

Accepted:
Keywords: data management training [16]
Software [17]
Community best practices [18]

Source URL: http://commons.esipfed.org/node/7968

Links:
[1] http://commons.esipfed.org/node/7968
[2] http://commons.esipfed.org/2015SummerMeeting
[3] http://commons.esipfed.org/session-type/breakout
[4] http://commons.esipfed.org/taxonomy/term/260
[5] http://commons.esipfed.org/collaboration-area/data-management-training
[6] http://commons.esipfed.org/collaboration-area/data-preservation
[7] http://commons.esipfed.org/collaboration-area/preservation-and-stewardship
[8] http://commons.esipfed.org/collaboration-area/student-cluster
[9] http://commons.esipfed.org/sites/default/files/DataManagementForProgramming_Hou%20-%20ESIP%2007-16-2015.pdf
[10] http://commons.esipfed.org/sites/default/files/DataManagementForProgramming_Peterson%20-%20ESIP%2007-16-2015.pdf
[11] http://commons.esipfed.org/node/7872
[12] http://commons.esipfed.org/taxonomy/term/2486
[13] http://commons.esipfed.org/node/7964
[14] http://commons.esipfed.org/taxonomy/term/1878
[15] mailto:fox@tinybike.net
[16] http://commons.esipfed.org/taxonomy/term/912
[17] http://commons.esipfed.org/taxonomy/term/917
[18] http://commons.esipfed.org/taxonomy/term/1875