Interview: Merce Crosas – Dataverse Network

1. Tell me about the product or service that you provide in the role that you play at your library?

Our product is the Dataverse Network - a data repository for sharing, citing, archiving research data (see http://thedata.org). Our Dataverse Network software is also integrated with another one of our products: the Zelig software (an R statistical package that provides a wide range of statistical models). Both software applications are open source and developed by the Data Science team at the Institute for Quantitative Social Science (IQSS) at Harvard University. I'm the director of the Data Science team. The development of the Dataverse software started in our group in 2006, as a need from social science researchers to share data sets with their collaborators and be able to reference them permanently from their published work. The project benefited from the expertise of the former Harvard University Library Director, Sid Verba, work on research replication from the IQSS director, Gary King (King 1995, Replication, Replication), and work on data citations from Altman and King, 2007. The Dataverse Network has now expanded to all research fields and is used in multiple institutions across the world. More on the Dataverse can be found at Crosas, 2011, "The Dataverse Network: An Open-source Application for Sharing, Discovering and Preserving Data." and Crosas, 2013, "A Data Sharing Story".

2. How can other librarians use this product or service?

There are two main ways librarians around then world can use the Dataverse Network:

1) As a service – The Harvard Dataverse Network (http://thedata.harvard.edu) is open to all researchers in all scientific fields. Librarians can assist researchers publishing their data sets in the Harvard Dataverse Network, or assit them searching and accessing data sets from the repository. An individual researcher or a research project team can create a Dataverse at the Harvard Dataverse Network, customize it. and deposit and share their own data sets. In a Dataverse, data sets are organized in studies, where each study, in addition to the data, contains cataloging information and, if needed. documentation and code files that accompany the data. Each study gets automatically a data citation which can be used to referenced persistently that data set. The Dataverse Network provides a good solution as a Data Management Plan now required by some funfing agencies.

2) As a software installation: Institutions that wish to have their own Dataverse Network as their institutional data repository can download the software and install it in their servers. Note that at least a system administrator is required to install, host and maintain a Dataverse Network, providing enough storage, setting up daily backups of the data, and upgrading the software as needed.

3. How has your library reached out to your institutional community and how have you earned support for this particular service?

The Dataverse Network development and maintenance are supported partially by Harvard University and partially by grants (NSF, Sloan, and IMLS in the past). We've reached out to the community by providing training, workshops and outreach materials in thedata.org web site. We have a support team that provides individual support when needed and time permitting. We also work with a newly formed group at Harvard, the Research Data Collaborative led by Gosia Stergios, which includes

members of the Harvard Library, Harvard IT, provost office, and office of sponsored research. This group helps with data services and outreach across our institution.

4. What skills or experience do you think librarians need to acquire to meet the needs of escience and data management and can you provide examples of the skills and services that you or your other staff have in this particular area?

In this new escience area, I think that librarians can be an essential resource to provide support for new tools that are available to researchers to work in this data intensive research environment, and encourage open, well-documented data-driven science. A number of new research tools focus on capturing the research workflow, from the original data sets to the methods to the derived data, and ensure that the research is reproducible, transparent and accountable. The librarians can become familiar with data science tools for data storage, analysis and visualizations, and be part of the outreach, support and services around these tools. Also, they can provide services around curating and documenting the data and code used through the research workflow, so that all the research objects and processes become a reusable resource to others. They can also provide support and services around provenance and preservation of these research objects.

As an example, here at Harvard we are organizing with members of the Harvard Library a Week of Data to mix librarians, researchers and students to learn about data tools, data curation, management and analysis. Another example, as part of an on-going project to integrate the Dataverse with Journals (through the Open Journal System), we are working with an information scientist to gather requirements from journals, understand how traditional publishing workflows need to be integrated with publishing data associated with the article, and provide outreach and support to journals on how to use the Dataverse.