Year 4 Final Project Reports 2015-2016

A National Science Foundation University Cooperative Research Center

Transforming Data Adaptation Science and Service: An Innovative Visual Ontology Application

Table of Contents

Personnel

Principal Investigators:

Jane Greenberg, Ph.D., Drexel University

Xia Lin, Ph.D. Drexel University

Graduate Students:

Kai Li, doctoral student, Drexel University

Xuemei Gong, doctoral student, Drexel University

Executive Summary/Abstract

Motivation:

Data sharing efforts nationally and globally have increased data and software reuse across many different scientific domains. Although reuse demonstrates a positive return-on-investment (ROI) for the original scientific endeavor, tracking reuse only by citation is not sufficient for good science. Knowledge of data and software provenance and previous use conditions is key for making scientific conclusions. To date, there has been some standardization of reuse relationship terminology in selected metadata standards, although the scholarly communications community lacks applications that adequate track reuse and validate the types of reuse. Our CVDI work, "Transforming Data Adaptation Science and Service: An Innovative Visual Ontology Application" aimed to address this challenge by developing a prototype visual ontology and algorithm to track software reuse relationship types.

Objectives:

The objectives of our CVDI project were to develop:

1.) A prototype visual ontology application for capturing software reuse and adaptation in a target test domain.

2.) A platform for modeling adaptation science and service.

3.) An approach for CVDI partners to determine and strategically plan for greater impact of data, application, and algorithm outputs.

Methods:

We used a multi-method approach consisting of co-citation analysis, content analysis, and proof of concept. We focused on reuse of LAMMPS software, a popular software used in molecular simulation in materials science.

Results:

We found four reuse relationship: 1. reuse, unspecified, 2. modified reuse, 3. benchmark, 4. cite (for citation). We also developed a prototype online visualization application using R programming language. The prototype application includes an ontology for capturing and visualizing reuse relationships.

Conclusions:

This CVDI project, a one-year project, with a budget of roughly 30K, was exploratory on one level, however, we were able to achieve our goals and develop the prototype application and ontology. The approaches, tools and deliverables are transferable to other domains. The work demonstrates an innovative way to leverage data and algorithmic resources in an organization, and presents a more accurate view of software reuse. Most importantly, in-line with CVDI goals, the products of our work present an initial toolset that can be adapted to present a deeper understanding of ontological connections among knowledge assets.

Transforming Data Adaptation Science and Service: An Innovative Visual Ontology Application

Background:

Data sharing has become a national and international goal, motivated by a number of factors, including the digital data deluge, federal data sharing polices (Tenopir, et al, 2009), and new perceptions of science, such as the Jim Gray's notion of the 4th Paradigm (Hey, et al, 2009) and the move toward a data intensive science. Software and the underlying algorithms are very much a part of this new paradigm, given the need for software to gather, manipulate, store, and use and reuse data.

As part of this growth, the scholarly communications community supporting cyberinfrastructure has pursued solutions to track reuse, primarily developing citation standards and by establishing open repositories. Selected examples in these areas include:

Citation tracking:

- Thompson Reuter's Data Citation Index: http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/data-citation-index.html.
- Elsevier's Science Direct: https://www.elsevier.com/solutions/sciencedirect.
- Mendeley Data: https://data.mendeley.com/.

Selected open data repositories:

- Dryad: http://datadryad.org/
- FigShare: https://figshare.com/
- Knowledge for Biocomplexity Data: https://knb.ecoinformatics.org/
- Mendeley Data (also listed above, given the citation emphases)

These developments reflect and intersect with the open data and open science environment (Boulton, 2016; Uhlir & Schröder, 2007); they represent the digital data trend, including increased attention to archiving, preservation, access, use and reuse. To this end there is no shortage of research and scholarly work reporting data sharing results (Piwowar & Chapman, 2010) and the importance of this practice (Borgman, 2012). Attention to software and algorithm reuse is part of the new paradigm, and includes increased attention to software citation practices and workshops, such as the NSF supported software sustainability institutes (Timmes, et al, 2016; Katz, et al, 2015). All of these developments are important to scientific endeavors that build upon previous work and can impact conclusions. Additionally, these developments can have a bearing on next steps in industry relating to scholarly communications or any industry supporting data and software

reuse. A challenge is found with current infrastructure limitations that actually interfere with tracking reuse, particularly with scientific algorithms that are at the very core of certain disciplines.

Researchers track popular algorithm reuse (Kusashima, 2016); however, knowledge is limited on exactly how these algorithms are being reused. This knowledge is important to scientific conclusions. Research in this area can also enrich citation work and enable more robust metadata. These observations motivated the work pursued in our CVDI proposal and the goal to create a visual ontology prototype for capturing data reuse relationship automatically.

Research Approach and Test Environment

To pursue our objectives and identify reuse relationship in software reuse, we conducted a co-citation analysis, a two-phased content analysis, and developed a proof of concept. First, the co-citation analysis helped us to develop an algorithm to analyze the text around the initial citation for reuse, and create an initial classification for further analysis. Second, the content analysis phases allowed us to examine the texts surrounding the citation for LAMMPS software and the contexts of why and how the software was cited. These two steps were instrumental in confirming an ontological framework, building the visual application prototype, and demonstrating a proof of concept.

Our test environment was defined by the body of research papers that have cited LAMMPS software. LAMMPS, short for large-scale atomic/molecular massively simulator, is a molecular dynamics program created via an agreement involving Sandia National Laboratories, Lawrence Livermore National Laboratory and three other companies. LAMMPS was initiated in the mid-1990s, with the coding effort led Steve Plimpton from Sandia Laboratories. ("LAMMPS History," n.d.). We selected LAMMPS because of its historical and current significance, and the extensive body of research papers that use the software. The published research citing LAMMPS captures a broad perspective on software reuse. To further note, there is an official instruction of citation ("Citing LAMMPS in your papers," n.d.), in order to cite LAMMPS, one needs to cite the paper titled "Fast Parallel Algorithms for Short-Range Molecular Dynamics" written by Plimpton in 1995 (Plimpton, 1995).

Results: Defining the Ontology

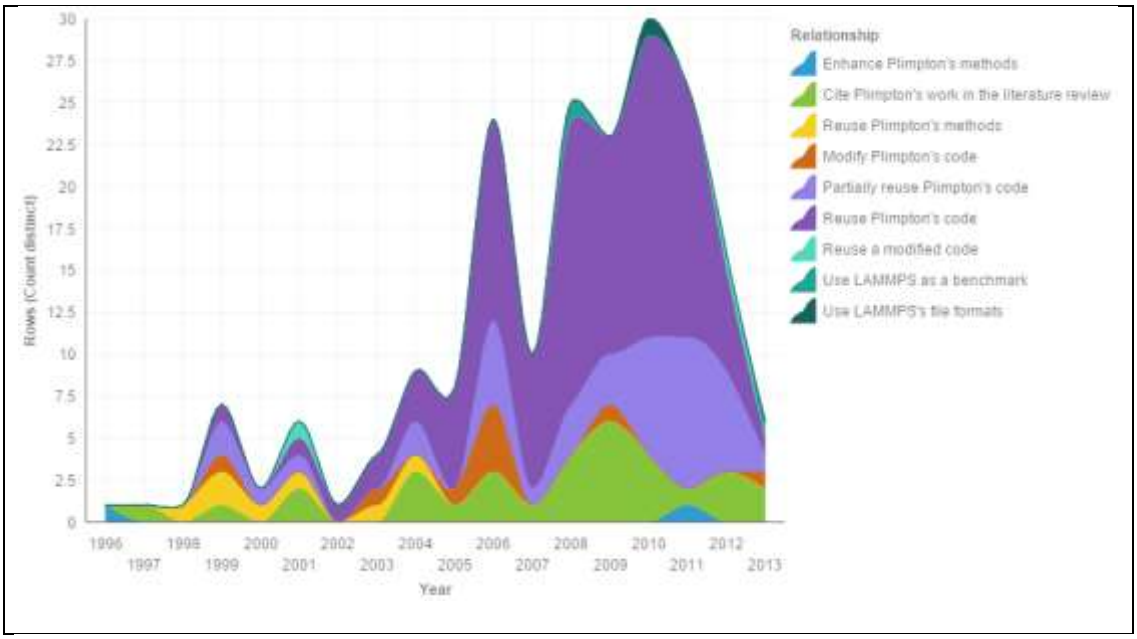
The first step, our initial co-citation analysis, followed by the baseline content analysis allowed us to glean a first ontology. The initial work was based on an analysis of first 200 papers resulted in the following eight categories of reuse types:

- Cite Plimpton's works in the literature review
- Reuse Plimpton's code
- Reuse Plimpton's methods
- Partially reuse Plimpton's code
- Modify Plimpton's code
- Enhance Plimpton's methods
- Reuse a modified code
- Use LAMMPS as a benchmark
- Use LAMMPS's file formats

The work included machine-learning technique was implemented in JAVA language, and the code is published below in Functionality of Innovation.

The initial results are presented in Figure 1. The prototype was, as noted in above and CVDI Quarterly reports, developed studying LAMMPS (also known as Plimpton's code or Plimpton's methods).

Figure 1, Reuse of LAMMPS software, 1996-2013



The initial reuse categories (listed above, and illustrated in figure 1) were reviewed by CVDI team researchers, including a faculty member in the Material Science program at Drexel University. The ontology was refined four classes (reuse types), which was reviewed by second coder. Table 1 displays the ontology relationship types, including their definitions:

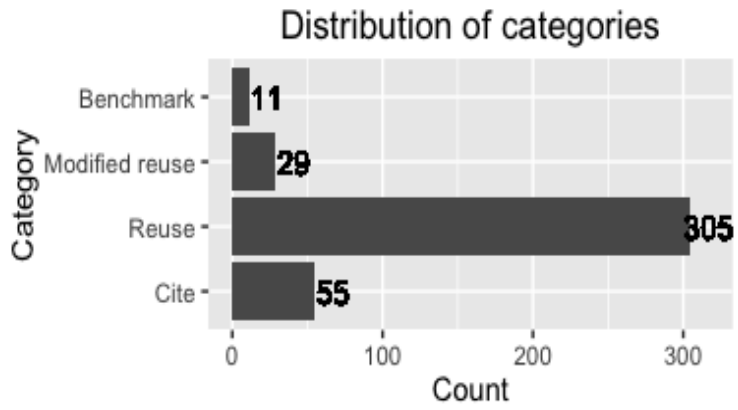
Table 1: Ontology Relationships-Defined

Relationship Type	Definition
Reuse, unspecified	The paper reuses LAMMPS as whole in the main study or does not specify which other types of reuse it is.
Modified reuse	The paper uses a modified version of LAMMPS in the main study. The specification of modification may or may not be specified in the paper.
Benchmark	The paper only uses LAMMPS (original or modified version) in the background study.
Cite	The paper does not use LAMMPS per se, but just cites either the software or Plimpton's paper, including those papers that just use the method represented in the original paper.

Working with this ontology, each paper in our sample, was designated to single reuse type, based on the most prominent connection with the software. The work was reviewed manually by the new coder reading and reviewing the PDF files.

The ontology was vetted in a second phase with an additional 200 papers added to the initial sample. Efforts were also made to automate this workflow, including text extracting text and classification, although more work is needed with the automated aspects. The coder worked extensively using the established ontology. Descriptive statistical analysis was used to study the results. Finally, a text analysis was also conducted to calculate the term frequencies in the extracted sentences and identify metadata elements that are related to LAMMPS. Results for the full sample of 400 papers distributed across the ontology are presented in Figure 2.

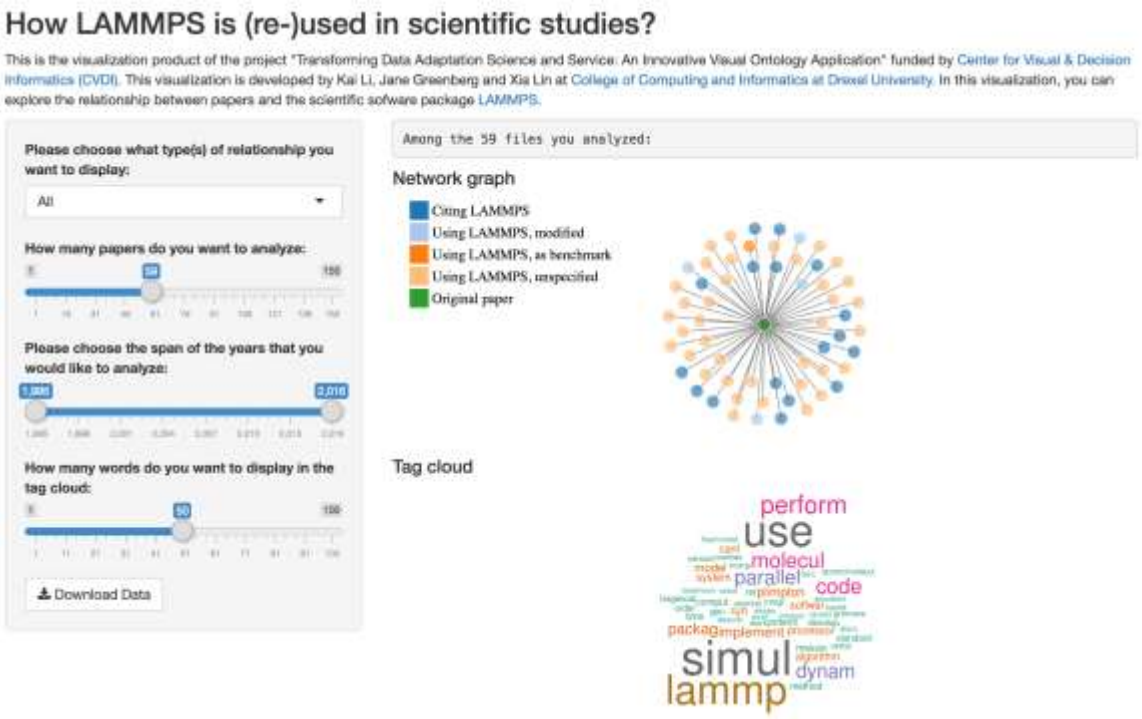
Figure 2: Ontology Relationships-Distribution



Proof of Concept

The last phase of our project included the development of a prototype online visualization system. The system was designed using R programming language. Figure 3 present a visual display for a search in our prototype system displaying the ontology reuse relationship for the LAMMPS software.

Figure 3—Prototype-Ontology Reuse Relationship for the LAMMPS Software



Functionality of Innovation(s)

Innovations:

There are 3 innovations that have results from this work:

1. Code that can be used to extract text and build an ontology capturing reuse.

```
Process for automatically extracting sentences (Java):
Extract text from pdf.
Identify the reference id of Plimpton’s article.
Extract sentences that contain the reference id, “Plimpton” or
“LAMMPS” in the text.
Process for automatically extracting sentences (Java):
Extract text from pdf.
Identify the reference id of Plimpton’s article.
Extract sentences that contain the reference id, “Plimpton” or
“LAMMPS” in the text.
Java Code README
*****Libraries Required*****
javacsv-2.0.jar
pdfbox-app-2.0.0-RC1.jar
spark-assembly-1.5.2-hadoop2.6.0.jar

*****Application Needed*****
winutils.exe (is placed in "/bin")

*****Classes*****

***Sentence Extractor***
Sentence Extractor extracts sentences from articles in the given
folder in the following steps:
For each article:
1. read text from pdf file
2. find the reference id of Plimpton's article
3. identify the sentences containing reference id, "plimpton" or
"lammps"
Output format: filename \t referenceID. sentences
(referenceID is 0 when no reference id is found.)

***Reuse Type Analyzer***
ReuseTypeAnalyzer has two methods:
1. getWordCount counts word frequency in each category, output the
data in csv file to be used in IBM Watson Analytics.
2. createTrainingDatasetForCategoryPredictor creates training dataset
to be used by CategoryPredictor and CategoryPredictorEvaluation

***Keyword Extractor***
```

```
KeywordExtractor extracts keywords for each category in the following
steps:
1. converts documents into vectors of terms weighted by tf*idf
2. calculate the centroid of each category
3. identify 10 keywords with the highest weights in the centroid for
each category

***Category Predictor***
CategoryPredictor categories articles using the logistic regression
method. It trains a model from the training data set and then predict
the category for each article in the test dataset.
Format of the training dataset: articleID \t categoryID \t sentences
Format of the test dataset: articleID \t sentences

***Category Predictor Evaluation***
CategoryPredictorEvaluation splits the input dataset into training and
test datasets, trains a model with the training data set, predict the
category of each article in the test dataset, and evaluate the
precision of the prediction.
Format of the dataset: articleID \t categoryID \t sentences

***Data Manager***
DataManager provides methods for reading and writing files in various
formats.

***Labeled Document***
LabeledDocument is used by CategoryPredictor and
CategoryPredictorEvaluation.

***Document***
Document is used by CategoryPredictor and CategoryPredictorEvaluation.

In CategoryPredictor, CategoryPredictorEvaluation and
KeywordExtractor, hadoop.home.dir is set to the current directory of
the project:
System.setProperty("hadoop.home.dir",
"current/directory/of/the/project");
```

2. A baseline ontology presented above in table 1, and here for convenience purposes

Table 1: Ontology Relationships-Defined (from above)

Relationship Type	Definition
Reuse, unspecified	The paper reuses LAMMPS as whole in the main study or does not specify which other types of reuse it is.
Modified reuse	The paper uses a modified version of LAMMPS in the main study. The specification of modification may or may not be specified in the paper.
Benchmark	The paper only uses LAMMPS (original or modified version) in the background study.
Cite	The paper does not use LAMMPS per se, but just cites either the software or Plimpton's paper, including those papers that just use the method represented in the original paper.

3. CODE base for R software and the prototype software supporting a visual ontology display, presented in Figure 3, and accessible at: https://nalsi.shinyapps.io/cvdi_original/.

Functionality:

All 3 of these innovations can be used in their current state. They can also be modified and adapted as necessary for other domains. Our focus was on software reuse, LAMMPS specifically, but the work framework and innovations can also be applied to data.

Conclusions and Recommendations

The work pursued here provides a platform for automatically tracking reuse relationship among data and software, drawing from text association with citation. Additionally, our results underscore the lack of consistency in software citation. However, we anticipate improvement in this area given noted software citation workshops and sustainability institutes that have been gaining attention of the last couple of year (Timmes, et al, 2016; Katz, et al, 2015).

These factors have bearing on our results and the ontology relationship that have been incorporated into our prototype, and they very likely do not reflect the full scope of relationships that could be useful to scientific endeavors supporting

reuse. That said, the work presented here provides an initial and valued starting point, given the significance of citation as a source for tracking the progression of scientific research. Finally, our methods provide an approach and a prototype that can be modified for different communities.

Impact and Uses/Benefits

The impact and benefits of our work include the following:

- A more accurate view of data and algorithm reuse.
- Platform to enable radical, new adaptation combinations, documenting reuse of data and algorithms.

Specific to industry, our work can help industry provide services that support better science and informed decision making. The actual impact on better science is hard to measure, although the growth in digital data and data intensive research provides opportunities to address society's grand challenges in ways that have been previously unimaginable. The cost of data gathering and software development is not trivial, and the reuse of these resources is being mandated and encouraged by federal agencies. Industry also recognizes the value of these approaches in efforts such as the recent launch of the NSF Big Data Regional Hubs. The work pursued and achieved in our CVDI project leads to a better return on investment (ROI) of resources allocated to data and software creation, use, archiving, by enabling reuse that is accurate and resourceful. The work may also procure deeper understanding sustainable knowledge of ontological connections among knowledge assets. Finally, we believe the work can lead to better effort to explore predictive capabilities in the future, although more research is needed in this area.

List of References

Borgman, C. L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), 1059–1078.

Boulton, G. (2016). Reproducibility: International accord on open data. Nature, 530(7590), 281-281.

Citing LAMMPS in your papers. (n.d.). Retrieved March 17, 2016, from http://lammps.sandia.gov/cite.html.

Hey, T., Tansley, S., & Tolle, K. M. (2009). The fourth paradigm: data-intensive scientific discovery (Vol. 1). Redmond, WA: Microsoft research.

Katz, D. S., Choi, S. C. T., Wilkins-Diehr, N., Hong, N. C., Venters, C. C., Howison, J., ... & de Val-Borro, M. (2015). Report on the second workshop on sustainable software for science: Practice and experiences (WSSSPE2). arXiv preprint arXiv:1507.01715.

Kusashima, N., Nogami, T., Takahashi, H., Yokomakura, K., & Imamura, K. (2016, May). Talk Algorithm with Frequency Reuse for LTE Based Licensed Assisted Access in Unlicensed Spectrum. In 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring) (pp. 1-5). IEEE.

LAMMPS History. (n.d.). Retrieved March 17, 2016, from http://lammps.sandia.gov/history.html.

Piwowar, H. A., & Chapman, W. W. (2010). Public sharing of research datasets: a pilot study of associations. Journal of Informetrics, 4(2), 148–156.

Plimpton, S. (1995). Fast Parallel Algorithms for Short-Range Molecular Dynamics. Journal of Computational Physics, 117(1), 1–19. http://doi.org/10.1006/jcph.1995.1039.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., ... & Frame, M. (2011). Data sharing by scientists: practices and perceptions. PloS one, 6(6), e21101.

Timmes, F., Turk, M., Ahalt, S., Wang, S., Idaszak, R., Brower, R., ... & Gustafson, K. (2016). 2016 Software Infrastructure for Sustained Innovation (SI2) PI Workshop.

Uhlir, P. F., & Schröder, P. (2007). Open data for global science. Data Science Journal, 6, OD36-OD53.