Software Citation, Reuse and Metadata Considerations: An Exploratory Study Examining LAMMPS

Kai Li
College of Computing and Informatics, Drexel University
3141 Chestnut Street, Philadelphia, PA 19104
kl696@drexel.edu

Xia Lin
College of Computing and Informatics, Drexel University
3141 Chestnut Street, Philadelphia, PA 19104
xlin@drexel.edu

Jane Greenberg
College of Computing and Informatics, Drexel University
3141 Chestnut Street, Philadelphia, PA 19104
jg3243@drexel.edu

ABSTRACT
Scientific software is as important to scientific studies as raw data. Yet, attention to this genre of research data is limited in studies on data reuse, citation, and metadata standards. This paper presents results from an exploratory study that examined how scientific software’s reuse information is presented in the current citation practice and natural language descriptions in research papers. We selected LAMMPS, popular simulation software used in material science, for this study. Both descriptive metadata elements and the types of reuse are examined from a sample of 400 research papers. The results indicate that both descriptive metadata elements and reuse types about LAMMPS are presented in incomplete and inconsistent ways, and this interferes with the values of scientific software, as a type of research data. Our findings necessitate future studies on the metadata standards to facilitate the identification of information related with scientific software reuse.

Keywords
Research data, scientific software, research data reuse, research data citation, metadata standard

INTRODUCTION
During the past decade, advancing means for reusing research data has become a strong focus in scholarly communication studies. Two major factors contributed to this research trend. First, the “data deluge” has been posing great influences on the scientific practice across nearly every knowledge domain (Hey, Tansley, & Tolle, 2009). Second, the increasing needs for reproducible research in scientific studies and academic discourse (Borgman, 2012; Rung & Brazma, 2013).

There has been a good deal of progress in this area, with the development of metadata and citation standards for research data, for example, the adoption of Digital Object Identifier for research data (Paskin, 2005), DataCite’s metadata schema for research data (DataCite International Data Citation Metadata Working Group, 2015; Starr & Gastl, 2011), and Force 11 (Martone, 2014). These developments are important for facilitating data discovery and enabling reuse; although they do not fully document why and how the research data is reused.

The absence of this specific information is even more apparent with scientific software, which has traditionally not been a source of citation, although views on this are rapidly changing. That is, software is seen as being just as important as the raw data for the outcomes of scientific studies, and researchers claim that it should be included in the research data complex (Ahalt et al., 2015; Herterich & Dallmeier-Tiessen, 2016; Wynholds et al., 2012). As part of this change, researchers also report that the position of software and other parts of the research process, such as code and technical documentation, in the overall landscape of research data, are still being debated (Kratz & Strasser, 2014; Parsons & Fox, 2013). Furthermore, given the unique characteristics of software, the standards for research data ‘may’ or ‘may not’ be applicable for software in the first place, not to mention to study how it is reused.

Given these circumstances, and the goal to support data reuse and reproducible research, it seems more important than ever to study software citation and reuse. The research reported on in this paper considers this need, and
researchers at Drexel University’s Metadata Research Center in connection with the Center for Visualization and Decision Informatics (CVDI), working with materials scientist have been studying this topic. A chief focus is the simulation software LAMMPS, which is heavily cited and reused in scientific studies. LAMMPS was selected because it is one of the most used software is molecular simulation studies. The research reported in this paper was undertaken to establish baseline in this area and to consider metadata requirements for scientific software reuse.

The following paper presents our work. The next section reviews literature on data reuse, data citation, and metadata elements about research data and scientific software. Next, the research questions are presented, followed by a review of our methods, the sample of data, and procedures. The key findings are reported, followed by a discussion, and the paper wraps up with a conclusion. The conclusion highlights how this study can contribute to our knowledge about the applicability of metadata for scientific software, and how this work is informing future research plans.

LITERATURE REVIEW

Reuse of Scientific Software as Research Data
A growing body of studies has been done on the reuse of scientific data because of its importance (Borgman, 2012). For example, qualitative studies have been pursued across multiple fields, including archaeology (Faniel et al., 2013), astronomy (Wynholds et al., 2012), earthquake engineering (Faniel & Jacobsen, 2010), and ecology (Zimmerman, 2007; Zimmerman, 2008), just to name a few, to explore researchers’ intentions and behaviors of reusing research data, and the underlying factors. Moreover, increasingly more quantitative studies have also been conducted on topics like what factors encourage researchers to reuse research data (Curty, 2015) and influence the satisfaction of reusing social science data (Faniel, Kriesberg & Yakel, 2015).

A growing idea that is attracting more attention is that research data should include secondary research data objects, like scientific software, code, and technical documentation (Herterich & Dallmeier-Tiessen, 2016; Wynholds et al., 2012), rather than just raw data (e.g., Carlson & Anderson, 2007; Zimmerman, 2007). Even though, there are still many challenges in seeing scientific software as a type of research data. (Parsons & Fox, 2013)

Much of the literature on the reuse of scientific software focuses on user needs, licensing, review of code, and user awareness (Goble, 2013; Joppa et al., 2013; Morin et al., 2012). This concept is also covered by studies on software sustainability, most of which examine reuse from the developer’s perspective, i.e., how to take reusability into consideration when the software is being developed (e.g., Crouch et al., 2013; Stewart, Almes & Wheeler, 2010; Wilson et al., 2012).

To help contextualize the research presented in this immediate paper, we turn to Huang et al. (2013) who defined scientific software as “all kinds of software that are used to support scientific research.” This definition supports the research on LAMMPS presented in this paper.

Data and Software Citation
Data citation is an important topic in research studies addressing data reuse. Some of the most important topics include sharing research data can result in increased citing rate (Piwowar & Chapman, 2010; Piwowar, Day, & Fridsma, 2007), a lack of proper citation policies and infrastructure is preventing research data from being shared and reused (Mooney & Newton, 2012).

To address this problem, a series of policies and technical standards have been proposed. Starr et al. (2015) summarized four components of a research data infrastructure from the perspectives of implementation and system, namely, document data model, publishing workflows, common repository application program interfaces (APIs), and identifiers, metadata and machine accessibility.

However, researchers have identified a few problems in the current software citation practice. For example, Mooney (2011) identified that only 29 percent of the papers had a complete data citation in the reference list. Moreover, it is obvious that, without proper citation, there is no way that information about software reuse can be identified by whatever means. Ince, Hatton & Graham-Cumming (2012) discussed ambiguities in software citation created by the use of natural language description and a lack of context of deployment. As a result, they suggested that journals should establish policies concerning the degree of source code accessibility that is connected with the research paper.

Metadata Standards for Software Reuse
Metadata is central to the data citation framework and identifying and tracking data reuse. Metadata for research data has been the topic of extensive scholarly and practice efforts, though few of them took research data reuse and scientific software into consideration.

In practice, the Dublin Core Metadata Element Set has been commonly used for research data, where the common practice is to develop an application profile for a given context, the most notably of which is the Dryad Application Profile (Ball, 2009; Diamantopoulos et al., 2011). However, the perceived flat structure of Dublin Core model (Lagoze, 2000) presents challenges for some, in expressing the complex relationships among research datasets and between research datasets and other types of research products, some of which were discussed in Kratz & Strasser's literature review (2014), for example, deep and dynamic citation.

These challenges are part of the motivation for the development of DataCite Metadata Schema. Besides the
descriptive metadata elements, the standard is also able to describe the relationship between two datasets, for example, citing, supplements, and continuation, et al. (DataCite International Data Citation Metadata Working Group, 2015; Starr & Gastl, 2011) As exciting as this step is, its applicability to scientific software and research data reuse seems to not have been a key focus.

Although group efforts addressing software citation are more recent, it is important to note that it has, in fact, been a topic of interest at times, and there has been definite interest in the application of metadata standards. One key examples is the work of Garcia et al. (2006), where researchers developed a software measurement ontology, which covered software citation from the perspectives of software development. More recently, Hong (2014) developed a multi-level metadata framework to describe the reusability of scientific software for developers. The most relevant effort to our study is the project called “Code as a research object” organized by Mozilla, Figshare and GitHub, where researchers can upload their codes to GitHub and get persistent identifiers. This project has the goals to promote the reuse of academic code as a type of research data, and to integrate code and scientific software into the scholarly workflow. (Thaney, 2013) These and related developments mark interests in this topic from both academia and industry. The work that is presented in this paper is also along the same path, focusing on how to identify the information about research papers reusing the scientific software.

**RESEARCH QUESTIONS**

The goal of this paper is to examine how relationships about scientific software reuse are presented in the current practice of scientific software documentation and citation, and to consider how this may impact future metadata work. Based on the discussions above, if a metadata scheme is to be developed in the future, it should not only describe the dataset's intellectual and physical attributes, but also its relationships with other research products, most important of which is the reuse relationship. We selected to study LAMMPS because it is known to be a popular molecular simulation software. Additionally, there is little empirical reporting on the reuse. For these reasons, we conducted an exploratory study to answer the following two questions:

- How is LAMMPS reused in research papers?
- What descriptive metadata elements about the reuse of LAMMPS can be identified in the natural language descriptions in the papers?

Zimmerman’s (2008) definition of data reuse helped us to frame our work. For our research, reusing scientific software is defined as the use of software to study a new problem. However, distinguished from scientific data, LAMMPS, like many software packages, are designed to be reused. As a result, many cases of the reuse of software can be called use of software in more common contexts.

**RESEARCH METHODS**

In order to investigate the above research questions, we conducted a two-phased content analysis, examining a sample of research papers that cited LAMMPS software. The first phase allowed us to examine the reuse types of LAMMPS as presented by the research papers, and the second phase targeted what metadata elements are embedded in the natural language descriptions in the papers. A content analysis method was selected because it allowed us to examine the texts surrounding the citation for LAMMPS software and the contexts of why and how the software was cited. LAMMPS, short for large-scale atomic/molecular massively simulator, is a molecular dynamics program created by agreement between Sandia National Laboratories, Lawrence Livermore National Laboratory and three other companies. This project began in the mid 1990s, whose coding works were led by Steve Plimpton at Sandia Laboratories. (“LAMMPS History,” n.d.) We selected this software because it is so popular that a rich body of research papers has cited it, which gives us the chance to have a broad perspective on software reuse. According to its 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). Below are the three main steps in our study.

**Data Preparation**

To prepare for our data, we first identified all the papers that cited the Plimpton’s 1995 paper on Google Scholar. We searched the title of Plimpton’s paper to have a list of the papers, and selected the top 400 papers as our sample based on the number of citation they received. Full text PDF files of all these papers were downloaded manually.

**Classification**

In order to determine the different ways in which scientific software are reused, a classification scheme was developed using a qualitative and inductive content analysis method. In other words, we did not have any pre-determined scheme before extracting the categories. The first 200 papers were used to extract the categories by the first coder. An eight-category scheme was developed, which was later integrated into a four-category scheme after the discussions between the project members, including a faculty member in the Material Science program at Drexel University. Our final scheme is summarized in Table 1 below.

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reuse, unspecified</td>
<td>The paper reuses LAMMPS as whole in the main study or does not specify which other types of reuse it is.</td>
</tr>
<tr>
<td>Modified reuse</td>
<td>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.</td>
</tr>
<tr>
<td>Benchmark</td>
<td>The paper only uses LAMMPS (original)</td>
</tr>
</tbody>
</table>
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. |

**Table 1. Final scheme of the reuse types of LAMMPS**

This classification scheme was triangulated using the latter 200 papers by the same coder. Moreover, the categorization of all the papers and the scheme itself were further tested by the second coder. Even though some papers were classified differently, no revision to the model was found to be necessary. Discussions were held to solve the differences in the classification.

**Metadata Element Extraction**

All the sentences that are about LAMMPS or Plimpton’s paper were examined to extract attributes of how LAMMPS was reused in these papers using content analysis method. Our intention was not to collect a full list of metadata elements; rather, we focused on three metadata elements that are most common and useful for describing data reuse to our knowledge in these papers, namely, version, parallel codes, and models of calculation.

**RESULTS**

The sections below present the results of this study. First we present the four categories of the classification scheme, including more detailed explanations and examples of each category. In the second part, we identify metadata elements that were found in the natural language descriptions.

**LAMMPS’ Reuse Types**

Four categories of the reuse of LAMMPS were discovered in our study, namely, unspecified reuse, modified reuse, benchmark, and cite. The numbers of papers belonging to each category are summarized in Figure 1 below.

**Distribution of categories**

- Reuse: 305
- Modified reuse: 29
- Benchmark: 11
- Cite: 55

**Figure 1. Occurrences of the four categories in the reuse scheme**

**Reuse, unspecified**

This category includes all the papers that used the software LAMMPS in their main studies, but didn't specify if a modified version of LAMMPS was used. It is reasonable to doubt that at least some of the papers in this category actually belong to the other two reuse-related categories, but just that the information was not specifically addressed in the paper, especially because of the fact that a lot of the citing sentences belong to this category are very short and uninformative. Below are two examples of this category:

- "LAMMPS was used for all MD simulations. [37]" (McMahon, Cheung & Troisi, 2011)
- "LAMMPS [28,29] (Large-Scale Atomic/Molecular Massively Parallel Simulator), developed at Sandia National Laboratories, was used to model [0001] oriented ZnO NWs with diameters ranging from 5 to 20 nm." (Agrawal, Peng & Espinosa, 2009)

**Modified Reuse**

This category includes those papers that used a modified version of the software LAMMPS in the main study. Below are two examples of this category:

- "The annealing simulations were performed with LAMMPS (large-scale atomic/molecular massively parallel simulator) code from Plimpton at Sandia (modified to handle our force fields)." (Jang et al., 2004)
- "We would like to thank E. Charlaix and P.-F. Gobin for introducing us to this subject, and Dr. S.J. Plimpton for making publicly available a parallel MD code, [25] a modified version of which was used in the present simulations." (Barrat & Bocquet, 1999)

**Benchmark**

This category includes all the papers that used the software LAMMPS in the background study and/or to compare with the results in the main study. However, it doesn't include comparative studies, which used LAMMPS and some other software to do the same simulation in the main study. Two examples are as below:

- "To demonstrate what one should expect of a precise MD trajectory, the same simulation run is performed using LAMMPS again, but this time on four processor cores in parallel. ... We have compared our GPU implementation against LAMMPS running on a fast parallel cluster, see Fig. 8, and we have shown that the GPU performs at the same level as up to 36 processor cores." (Anderson, Lorenz & Travesset, 2008)
- "In order to compare our GPU version to a well-optimized sequential code, we have also compared our CUDA implementation to LAMMPS." (Liu et al., 2008)
Cite
This category involves those papers that did not use the software LAMMPS in the study, but just cited the original paper written by Plimpton in 1995, the methods discussed in Plimpton's paper, and/or the software LAMMPS. This is a non-reuse category, and most of the papers in this category have the citation of the paper in the literature review section, if the papers are not literature review themselves. Below are two examples of this category:

- "Research by Plimpton [38], Plimpton and Hendrickson [37], and Hwang et al. [19] shows that this method provides a better speedup than RD, and can be used with good speedups up to hundreds of processors." (Kale et al., 1999)
- "Software such as LAMMPS [212], IMD [213] and DL_POLY [214] are publicly available to perform large-scale MD simulations on parallel platforms." (Mishin, Asta & Li, 2010)

Metadata Elements about LAMMPS’ Reuse in the Natural Language Descriptions
The following analysis was pursued to answer the second question, to identify the metadata elements related with the reuse of LAMMPS that could be gleaned from the natural-language descriptions in the papers. As mentioned in the Research method section, we focused our analysis on the following three descriptive metadata elements, version, parallel codes of LAMMPS, and models of calculation, because they are the most commonly discussed elements in our sample.

Version
On its website, all the versions of LAMMPS were listed from ParaDyn (a parallel code which will be discussed later), LAMMPS 99, 2001 to the most current version (“LAMMPS History", n.d.). Version is an important attribute of software, given the potentially significant differences between versions, and their influences on the results of the experiments. However, among all the 305 papers, only five instances of the version information were identified.

<table>
<thead>
<tr>
<th>Citation</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voronov, Papavassiliou &amp; Lee, 2006</td>
<td>17 Jan 2005</td>
</tr>
<tr>
<td>Karayiannis &amp; Mavrantzas, 2005</td>
<td>2005</td>
</tr>
<tr>
<td>Sachs, Crozier &amp; Woolf, 2004</td>
<td>2003</td>
</tr>
<tr>
<td>Zhang &amp; Greenfield, 2007</td>
<td>2001</td>
</tr>
<tr>
<td>Daub et al., 2010</td>
<td>&quot;Current version&quot;</td>
</tr>
</tbody>
</table>

Table 2. Version information identified in the sample
"17 Jan 2005" and "2001" are the only two full version names based on LAMMPS' website. For the rest of the examples, it can only be assumed that 2003 and 2005 might be the years when the researchers downloaded the software, not to mention the very little information that is implied from the last example. However, in the lights of controlled vocabulary, all the examples that are not full version names are less than ideal practices, and are not so useful for both human and machine users of the texts.

Codes that are part of LAMMPS package
During the history LAMMPS' development, some parallel codes that were developed by Sandia have been gradually integrated into the package, including ParaDyn, Warp, and GranFlow. (“LAMMPS History", n.d.) These codes represent both parts of and specific versions of LAMMPS, and are also significant information about the reuse of LAMMPS.

WARP is the most commonly mentioned code in our sample, being used by 13 papers in our sample. ParaDyn was mentioned to be used in 5 papers. And GranFlow failed to appear in any papers. It is shown by the table below that all these papers were written by a small group of researchers and published between 2005 and 2008. Yet to be examined by further studies, this phenomenon might be because of that scholars tend to reuse the same piece of software for the sake of efficiency and consistency. All the papers are summarized in Table 3 by the group of authors.

<table>
<thead>
<tr>
<th>Name of the code</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>WARP</td>
<td>Ji &amp; Park, 2006; Park, 2006; Park, Gall &amp; Zimmerman, 2005, 2006; Park &amp; Zimmerman, 2005, 2006</td>
</tr>
<tr>
<td></td>
<td>Liang &amp; Zhou, 2006</td>
</tr>
<tr>
<td>ParaDyn</td>
<td>Cao &amp; Ma, 2008; Cao &amp; Wei, 2006, 2007a, 2007b; Cao, Wei &amp; Ma, 2008</td>
</tr>
</tbody>
</table>

Table 3. Papers citing parallel codes of LAMMPS organized by group of authors

Models of calculation
In many papers we collected, the models that were used to in the calculation or simulation were also discussed. Because of the user-contributed nature of LAMMPS, a large number of functions have been added to this software package since it was developed. As a result, the specific models as well as the corresponding parts of LAMMPS used in the studies are also an important metadata element to describe the reuse of LAMMPS.

Because of the difficulties to determine if the model that was mentioned in the paper is a specific model or just a general name, our efforts were not intended to record all the models that were mentioned in the papers. Rather, we kept track of the commonly mentioned, and clearly articulated ones. Table 4 is the summary of the 5 top models in our sample and their occurrences.

<table>
<thead>
<tr>
<th>Model</th>
<th>Occurrence</th>
</tr>
</thead>
</table>

5
Adaptive Intermolecular Reactive Empirical Bond Order (AIREBO) Potential 15
Embedded atom method (EAM) potential 15
Nose-Hoover thermostat 10
Reactive force field (ReaxFF) 7
Velocit-verlet algorithm 6

Table 4. Model information identified in the sample

AIREBO is an extended version of the original REBO potential developed by Brenner (1990). It was proposed by Stuart, Tutein & Harrison (2000) to study hydrocarbon system, which is unable to be simulated by the original model. Embedded atom method (EAM) potential was first developed by Daw & Baskes (1983, 1984) to study metallic systems including the parameters of fractures, surfaces, impurities, and alloy additions.

The existence of these models in the papers suggests that LAMMPS have been used for various purposes in scientific studies. And their importance can be validated by the fact that there are more than 1,000 papers on Google Scholar with AIREBO along with LAMMPS as the searching terms. The information about these models adopted may help researchers and other stakeholders better understand how the scientific software as research data help to achieve the goals of each study project.

CONCLUSION

Paradise Lost: What is Missing from the Current Software Citation Practice?
The results presented above show that how LAMMPS is cited in the research papers cannot help people learn and keep track of much meaningful information about how it is reused in the studies. This general conclusion is true for both descriptive metadata elements and reuse relationships based on the sample of papers citing LAMMPS.

Descriptive Metadata Elements
In terms of descriptive metadata elements, the conclusion is that there is no position for these elements in the status quo. According to the official instruction of citing LAMMPS, Plimpton's 1995 paper and the link to the software’s website should be cited as the substitutes of LAMMPS. However, the only descriptive metadata element about the software itself in these two recommended text strings is the link to the website, which can be interpreted as a type of identifier based on DataCite's specification, version 3.1. (DataCite International Data Citation Metadata Working Group, 2015) None of the other required, recommended or optional metadata elements in DataCite schema can be found in the recommended citation practice of LAMMPS.

However, beyond the structural citation of LAMMPS, there are a few metadata elements describing the reuse of LAMMPS that can be identified in the natural language descriptions in the papers, most notably, version, parallel codes, and model of calculation. However, it is also shown by our study that the inclusion of these metadata elements in the papers is rare and highly inconsistent.

LAMMPS’ reuse relationship
As addressed in the literature (DataCite International Data Citation Metadata Working Group, 2015; Kratz & Strasser, 2014), the relationship between research datasets and other types of research products is critical to a robust identification system of scientific software in scholarly works. However, just like descriptive metadata elements, the reuse relationship of LAMMPS is hardly included in the current citation practice, because there is no position of this information in the structured citation practice, no matter what type of resource that is cited.

Moreover, it is surprising that a very large number of papers belong to the unspecified reuse category, and that a large number of these papers have a very short description of reusing LAMMPS. These findings have also prompted us to question whether or not some papers just failed to mention if they were using a modified version of the software, if not something else. However, this irregularity can only be examined in our future studies, or if we were able to interview authors of all the papers examines, but this is impossible give the time span.

Based on our findings and Ince, Hatton & Graham-Cumming's comments (2012) that natural language descriptions are insufficient and unintentionally misleading to describe the software citation, it is our suggestion to start exploring the possibilities to use a structured and comprehensive way to include metadata elements and relationship around scientific software in the citation. A large segment of important information is absent from the current citation practice, and it is a barrier to the realization of reproducible studies.

Limitations
As an exploratory study, our work reveals a few limitations that should be noted. First of all, by choosing the 400 most cited papers out of more than 10,000 papers that cited Plimpton's 1995 paper, our sample could be potentially less representative than desired. Second, LAMMPS might be a special case even among the scientific software used in the field of engineering, because of its extensibility and popularity. Last, our reuse type scheme represents an "etic" or "outsider" perspective, which needs to be further validated by evidences collected from researchers but especially software programmers. To address some of these limitations, a materials scientist/engineer, who is part of our larger metadata team, validate the relationships, but it’s possible that a larger sample would reveal different result. Despite these limitations, we believe our research provides important baseline data for further studies, including methods on how to study this topic.

Future Directions
This paper presents the methods and results of our exploratory study on how the information about the reuse of
ACKNOWLEDGMENTS
The authors would like to thank Dr. Garrett Tuckr from College of Engineering at Drexel University and Dr. Xuemei Gong at Quixey for their important contributions to this project. We would also like to acknowledge support of this NSF/Center for Visualization and Decisions Informatics (CVD) OCI#1160958.

REFERENCES
Diamantopoulos, N., Sgouropoulos, C., Kastrantas, K., & Manouselis, N. (2011). Developing a metadata application profile for sharing agricultural scientific and scholarly research resources. In Research Conference on...


