Requirements for a Next Generation Personal File Manager

[Position Paper]

Leonard Shapiro, Lois Delcambre, Warren Harrison, Tyler Hayes, Bart Massey
Portland State University
P.O. Box 751
Portland, OR 97207-0751
{len,lmd,warren,bart}@cs.pdx.edu, tylerhayes80@gmail.com

ABSTRACT

Scientists, engineers, knowledge workers and others need help managing their personal data files and the programs that manipulate their data. The current generation of software for supporting their needs, which we call Personal File Managers (PFMs), is not adequate. We propose five requirements that a next generation PFM should satisfy. We have created a mockup of a PFM which satisfies these requirements.

Categories and Subject Descriptors

H.4m [**Information Systems Applications**]: General

Keywords

File management, information management, personal information, data organization, file organization, version control, document management

1. INTRODUCTION

Our target audience uses and creates multiple data files and perhaps programs that manage the data files. Our users typically work with these files in a personalized desktop environment, creating new files and altering existing files; they need help organizing their files. Our users include scientists, engineers, knowledge workers and others who work on tasks or studies run primarily by a single person. Most of them do not use tools to organize their personal files because, as we shall see, they have little motivation to use existing tools. Further, they find that existing tools are not easy to use. Users that work in technical environments such as software development, or in large organizations with technical infrastructure, are not our primary target: they typically already have tools to manage files.

The purpose of a Personal File Manager (PFM) is to help our users better organize their data and program files. We feel that the current generation of PFMs is inadequate for this purpose. We explain why, and we propose several requirements that we feel a next generation PFM must meet to satisfy user needs.

2. MOTIVATION

Bill Howe, a Senior Scientist at the University of Washington, is charged with helping scientists organize their data. At the 2011 meeting of SSDBM[1] Howe commented that he often asks scientists the question "What percentage of your time do you spend managing your data vs. doing science?" He reported that the most common answer to his question is "Ninety percent" [15].

We have surveyed scientists and engineers in the Portland, Oregon area about data reuse [6]. By data reuse we mean users' ability to use their own or others' data (or programs or scripts) in subsequent studies or applications. The consensus was that reuse is very desirable but rarely done, because users' data and programs are poorly documented and poorly organized. Users do not take the time to document and organize because some of their work is largely exploratory work; it is not clear, until they are nearly finished, which steps and which programs will contribute to the final versions of the datasets or other results. Thus, until the time they are ready to complete their work, organization and documentation seem to be superfluous overhead. This lack of record-keeping means that they are not able to trust their work or other users' work; they are often forced to waste time redoing work, such as data cleaning or preparation.

With the support of grants from Microsoft Research and the Gordon and Betty Moore Foundation, Carly Strasser is leading a project [16] to build an add-in to Microsoft Excel for supporting digital curation. In one of her surveys of scientists [17], she asked what the add-in should do. 47% answered "Help me organize my data".

Google is a powerful organizer of information because it uses the data inside web pages to find relevant information. However usersr' data and program files may not contain useful, naturally-occurring keywords or patterns. A plausible remedy is for the user to provide external metadata that will enable a search engine to organize usersr' files. The challenge is to convince the user to take the time to provide useful metadata, documentation and keywords.

A primary motivation for documenting what our target users

[1]The 23rd Scientific and Statistical Database Management Conference, Portland, OR July 20-22, 2011

do is so that other users can reuse (possibly after tweaking) their datasets and/or programs, including the possibility of a user reusing their own datasets or programs at a later time. A primary barrier to reuse is lack of access to the details of how the datasets were cleaned or otherwise prepared and modified by programs. Thus the kind of documentation that we are talking about will allow new (or old) users to find out exactly what was done. This will lead to trust in the datasets and results produced by their programs, and is a crucial requirement for reuse in other studies or investigations.

To recap: our target audience recognizes the need for better data organization, and a key to better data organization is better documentation. If those goals could be achieved, the potential benefit is great. For example, if the 90% figure Bill Howe quotes could be reduced even to only 70%, that would triple the amount of time scientists could spend doing science.

3. MOTHERHOOD AND APPLE PIE

We, and our target audience, know that better organization and better documentation are of great benefit. But, like exercise or eating healthy food, everyone agrees they are good, but far too few do them. What is needed to achieve these goals is simplicity and motivation. We must make it easy for users to provide documentation for their files, and we must provide motivation (in the form of useful, immediate benefits) for them to do so.

4. EXISTING SOLUTIONS

Existing PFMs, tools to help users organize and document their data and program files, fall into two categories: Version Control Systems and Document Management Systems.

A Version Control System (VCS) focuses on managing changes to files. VCSs were originally used by software developers to manage program files, and are now also used in medium to large scientific organizations. The most popular VCS is the open source VCS Subversion [1]. It is a second generation VCS, in that its users can be situated over a network but its control is centralized. The most popular third generation VCS, whose control is distributed, is the open source system Git [4]. Popular proprietary systems include Microsoft's Visual SourceSafe [11] and IBM's ClearCase [9].

A Document Management System (DMS) encompasses collaboration tools, metadata, indexing and search, versioning, security, workflow and auditing capabilities.

Other solutions are related to PFMs. These include Content Management Systems (CMSs) such as Microsoft's Sharepoint [12]. CMSs are very complex systems which include some DMS capabilities, but those capabilities have not yet matched the document management capabilities of current DMS systems [3] . Workflow systems [5, 2] emphasize collaboration, not personalized information - they manage the flow of business or scientific processes through organizations or laboratories. Most importantly, workflow systems allow the user to describe (ahead of time) which workflow steps are to be executed. Our target users, on the other hand, are often engaged in a trial and error process, attempting to discover the appropriate workflow as they work.

5. REQUIREMENTS

Based on our review of existing solutions and our interviews with scientists and engineers, we propose several requirements that a next generation PFM should satisfy. We have also produced a mockup [`http://www.cs.pdx.edu/~len/WHIM.pptx`] of a system called WHIM (Work History Information Manager), which, if developed, would meet these requirements.

The most important requirement for a next generation PFM is that it simplify the process of users providing documentation for their files. Existing solutions do this in a simplistic way. They provide a tabula rasa interface: when a new file is created or a file is revised, a blank text box is created for the user to fill in. WHIM uses context-aware prompting techniques: instead of a blank text box, the user is presented with selection buttons, whose choices are based on the user's application domain or on the user's own inputs. Thus our first requirement is that **(1) the PFM should provide domain-specific or user-personalized metadata suggestions**. For example, in the mockup of WHIM, selection buttons such as "added files, cleaned data, integrated datasets" are presented to the user. Personalization, in another form, has been suggested previously for Personal Information Management systems (PIMs) [10]. Domain-specific metadata, in another form, has also been suggested previously for PIMs [13].

Making it simple to provide metadata includes making it simple to provide a rich variety of metadata parameters. These parameters might include data about the provenance of a file (for example, the original creator) or about its treatment (for example the number of standard deviations used in excluding outliers). Having more metadata parameters means that subsequent searches for information will be more successful. DMSs are more effective in this regard, providing multiple fields per file and change, whereas VCSs tend to provide only one text box per change. However, the more fields that are provided, the harder it is to motivate the user to fill them all in. The PFM can ameliorate this situation by filling in some parameters for the user. Thus another requirement is that **(2) the PFM should suggest metadata parameters based on previous user experience**.

After simplicity, the next most important requirement is that the PFM provide incentives for the user to provide documentation, in the form of immediate rewards for using the PFM tool. Our remaining requirements deal with these incentives.

A prime incentive to user documentation is **(3) a rich report writing capability**. This helps the user document previous work for work reports, and recalll the state of work after a hiatus. Neither VCSs or DMSs provide report writers as rich as those demonstrated in the WHIM mockup.

Another helpful incentive to user documentation is **(4) an Origins feature**, the ability to point to a piece of data or a line in a program and have the PFM report where that item was last changed. Our survey indicates that users would find this tremendously helpful, since they often look at an item and wonder where it came from. VCSs provide this capability (called "blame" in Git and Subversion), but only

for lines of code, not for data items such as cells in Excel spreadsheets. No DMS provides this. The origins feature is a restricted form of data provenance [14] in that it finds only the most recently changed occurrence of an item.

The spreadsheet is by far the most popular tool for manipulating data. Its popularity derives from its ability to visualize data in a fashion easy for users to comprehend and control directly. In fact Visicalc, the first computerized spreadsheet, was a major factor in the adoption of the personal computer by businesses, the first "killer app" [8] . Today there is a broad variety of programs to visualize data [7], yet no current PFMs use these visualization tools to provide an improved user experience. Thus a vital requirement for a next generation PFMs is that it **(5) provide a plugin to incorporate data visualization tools**.

6. CONCLUSIONS

Scientists, engineers, knowledge workers, and others need help managing their personal data files and the programs that manipulate them. The current generation of PFMs does not meet their needs. We have proposed five requirements that the next generation of PFMs should satisfy in order to meet those needs. We have constructed a mockup of a PFM, WHIM, which satisfies these requirements.

7. ACKNOWLEDGMENTS

This work was supported in part by NSF Grant Number 0954268. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Shapiro is a Principal Researcher at the Else Institute [`http://www.elseinstitute.org`] and appreciates their support.

8. REFERENCES

[1] Apache Software Foundation. Apache subversion. `http://subversion.apache.org/`.

[2] A. Barker and J. Van Hemert. Scientific workflow: a survey and research directions. In *Proceedings of the 7th international conference on Parallel processing and applied mathematics*, pages 746–753. Springer-Verlag, 2007.

[3] A. Cohen. Juggling act. `http://tinyurl.com/7qhelde`, 2010.

[4] git the fast version control system. `http://git-scm.com/`.

[5] A. Haller, E. Oren, and S. Petkov. Survey of workflow management systems. Technical report, DERI, 2005.

[6] T. Hayes, L. Delcambre, and L. Shapiro. Can transportation researchers reuse project datasets and programs? Technical report, Portland State University, 2011.

[7] J. Heer, M. Bostock, and V. Ogievetsky. A tour through the visualization zoo. *Communications of the ACM*, 53(6):59–67, 2010.

[8] T. Hornby. Before the macintosh. *LowEndMac*, 2006.

[9] IBM Corporation. Rational clearcase. `http://www-01.ibm.com/software/awdtools/clearcase/`.

[10] A. Katifori, C. Vassilakis, I. Daradimos, G. Lepouras, Y. Ioannidis, A. Dix, A. Poggi, and T. Catarci. Personal ontology creation and visualization for a personal interaction management system. In *Workshop on Personal Information Management, in CHI*, volume 2008, 2008.

[11] Microsoft Corporation. Introducing visual source safe. `http://msdn.microsoft.com/en-US/library/3h0544kx%28v=VS.80%29.aspx`.

[12] Microsoft Corporation. Microsoft sharepoint 2010. `http://sharepoint.microsoft.com`.

[13] A. Rath, N. Weber, M. Kröll, M. Granitzer, O. Dietzel, and S. Lindstaedt. Context-aware knowledge services. *Personal Information Management: PIM*, pages 5–6, 2008.

[14] Y. Simmhan, B. Plale, and D. Gannon. A survey of data provenance in e-science. *ACM Sigmod Record*, 34(3):31–36, 2005.

[15] S. Spengler. Data scientists, data management and data policy. In *Scientific and Statistical Database Management*, pages 490–490. Springer, 2011.

[16] C. Strasser. Digital curation for excel project. `http://dcxl.cdlib.org/`.

[17] C. Strasser. Digital curation for excel project page 2. `http://dcxl.cdlib.org/?paged=2`.