How Private are Home Directories Really?

Carlos Maltzahn
Computer Science Department, University of California, Santa Cruz
carlosm@cs.ucsc.edu

It is widely assumed that home directories contain mostly private files. But is this assumption correct? The data in Figure 1 provides some evidence indicating otherwise[1]: the fraction of shared or "sharable" files in home directories is quite high. This would have significant implications for the design of data and metadata management systems (such as Graffiti [1]), digital archiving and data de-duplication strategies, as well as data safety and security approaches.

The question arises on how to determine whether a file is shared: just comparing home directories among a group of volunteers will either over-estimate sharing by not properly excluding system or application files, or under-estimate sharing by not including files that is shared with people outside the sample group.

In this survey we try a different approach: we ask people to *subjectively* categorize their files into four levels of sharing. We intentionally do not distinguish between files that are actually shared and files that the user believes should be shared since we want to capture the potential of sharing. We also ask users to "skip" files (i.e. exclude them from this categorization) that would not benefit from any additional data management. Examples of skipped files are system or application support files, or files that are already managed by another sharing mechanism such as source code management systems like cvs. The advantage of this approach is that the amount of sharing does not depend on the sample nor on available technologies. Instead it is based on what users know about their data.

The four levels of sharing that we track in this survey are:

The file never leaves this computer: The user wants to manage this file but does not expect to share it with any other computer (including other computers he or she owns or has access to).

The file is private: The file is suitable for sharing among the user's computers but not with other users.

The file is restricted to a group: The file can not be shared with the world but with a limited group of people such as family, friends, or an organization.

The file is public: Typically these are files that are downloaded from the web or published some other way.

[1]To reviewers: we will have a significantly larger sample at the time of presentation.

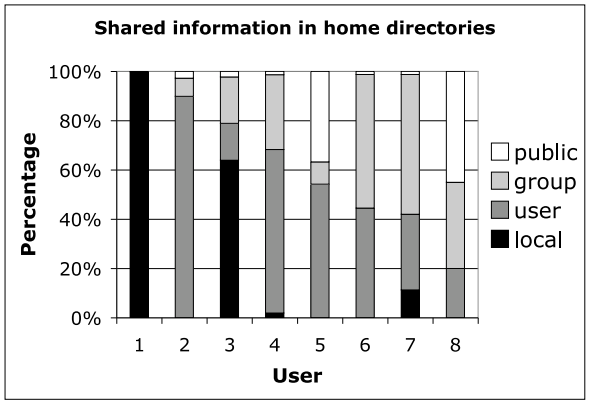

Figure 1: *Eight users classified their files in terms what they thought could be either shared publicly, only within a group, only among their own computers ("user"), or not at all ("local"). The users are sorted by sharing potential across users ("public" + "group") with an average of 38%.*

We designed an application that significantly reduces the tedium of categorizing files. The results are statistics about the total number of files in a home directory, the number of skipped files, and the number of files for each sharing category. As an additional benefit to the user the application leaves a categorized list of path names which can be used for other management purposes (but is not included in the survey). The application as well as the survey results so far are available at http://www.cs.ucsc.edu/˜carlosm/Survey. We encourage users to participate in this ongoing survey, and we will continue to update the web site as more results become available.

References

[1] C. Maltzahn, N. Bobb, M. W. Storer, D. Eads, S. A. Brandt, and E. L. Miller. Graffiti: A framework for testing collaborative distributed metadata. *Proceedings in Informatics*, 21:97–111, 2007.