Ten Quick Tips for Making Things Findable

Sarah Lin[1,*], Greg Wilson[1]
1 RStudio, Inc.

* sarah.lin@rstudio.com

Abstract

Information ecosystems consist of users, context, and content, all three of which must be addressed to make information findable and usable. Library science principles are a framework for doing this: they help researchers improve findability by defining the full scope of their users' needs and leveraging the structural elements of the software that creates, stores, and accesses data, research findings, or academic communication stored locally or shared on the internet.

Author summary

Sarah Lin is the Information Architect and Digital Librarian at RStudio, PBC. Greg Wilson works in the education team at RStudio, PBC.

Introduction

Researchers have always had to manage information, but the exponential growth of electronic data has both required and fostered the creation of new ways to do this [1, 2]. The problem is not just finding particular information when needed: information may be stored in many formats, exist in multiple versions, and need to be shared with varied audiences for both research and teaching.

Library science offers ways to work through the maze of information generated in professional life, and librarians' skills can be applied by any researcher who feels overwhelmed. The ten quick tips in this paper build on the fact that all information ecosystems have users, context, and content [1]. To solve the information retrieval problem, researchers must therefore think broadly about who needs that information and the context within which it is created as well as its actual content.

1. Design for a wide range of users.

The first step in making information findable is to determine who will be doing the finding. This includes everyone who might contribute to your work, expand upon it, or re-share information through their own networks [3]. While you might think your audience is small and knows your field well, complete novices will inevitably find your work if it is publicly available, thereby making your actual user base both larger and more diverse. Thinking about who your users are and how they interact with you and your work is the foundation of all ten tips.

The information you wish to convey and the way it is currently organized may make perfect sense to you, but its meaning for your users is determined by what *they*

interpret from the information they encounter and the way it's arranged. This means that the organizational structures you employ are a communication channel in their own right. To illustrate this, Borges created a classification of animals whose categories included "those belonging to the Emperor," "embalmed ones," "suckling pigs," "those included in this classification," "those drawn with a very fine camel hair brush," and "those that look like flies from far away" [4]. While this was deliberately ridiculous, it illustrates the fact that every way of organizing knowledge embodies choices by the organizer, which may or may not align with those of the audience.

More prosaically, consider the website of a faculty member coming up for tenure: She created the website as post-doc to publicize her papers and to make it easier to fill out grant applications by listing professional activities in one place, but:

- Colleagues will come to the site looking for un-paywalled copies of her papers, and to find out what she's currently working on or where she is next going to present her work.

- Tenure committee members might peruse her accomplishments to determine her work's impact.

- A librarian (or a program written by a librarian) might scrape that site for journal articles to include in the university's institutional repository.

- A student might come to the site looking for course information.

Reaching out to just a handful of users to ask for findability feedback will help you see any gaps in organizational alignment.

2. Figure out what "done" looks like.

Given how easy it is to create digital information and the plethora of software and formats you may employ, you almost certainly have lots of information in lots of different formats and locations. The second step in making things findable is therefore to catalog what you have, and then determine what should go where.

Figuring out what "done" will look like can be personally motivating, but once again you must determine what your users will consider a good outcome, which may not align with what you would do if you were the information's only consumer. However, remember that your future self is also one of your users: everyone is forgetful, so anything you do for others will likely pay off for yourself eventually.

Remember too that the question of how to find something can mean different things. Users may need to find your work on the web, find a specific item within a website, and/or find a particular piece of information within a specific file or webpage. Depending on your content, you may have challenges in all three areas; the remaining tips on context and content will help you address them.

Returning to the faculty website from the previous tip as just one example of varied needs:

- Tenure committee members will want human-readable descriptions of related sets of papers along with links to presentations or posters that discuss them.

- The librarian will want bibliographic information for your publications in machine-readable form (e.g., BibTeX, MARC, or MODS), either for individual items or in bulk for addition to an institutional repository.

- Colleagues will want all of the above, plus pointers to the software and data used in order to reproduce particular results.

- Students will want prominent links to the university's learning management system (LMS), which is where the course information they're searching for is actually located.

Each of these users might want content organized chronologically or topically. It's easy to provide both if the website is generated programmatically using a tool such as Blogdown [5] or Wordpress [6], but increasing the number of navigation options may make it harder for users to determine how the things they find are related to each other.

3. Use textual structure.

Findability at the document, post, or article level can be improved by taking advantage of the textual structures that information management programs provide [2]. For example, a key part of searching the web is scanning the text returned by search engines to see if it contains target information. Textual structure helps that process [7]: formatted headers (rather than just enlarged text), bulleted or numbered lists, and **highlighting** terms that are important all make both the information and its structure easier to understand. Similarly, headings and table of contents can be hyperlinked, which supports both scanning and navigation.

Textual content is created and aggregated in so many forms, using so many different programs, that it is difficult to specify strategies beyond headings, lists, and highlighting. However, specialists working in the same field tend to adopt the same tools, so it is worth exploring how your peers annotate information as well as creating, manipulating, or storing it. For example:

- GitHub allows users to add tags to issues and commit messages which can then be searched for across projects.

- Electronic lab notebooks can use XML schemas like Darwin Core, EML, or FITS [8].

- Using specific Google Docs heading levels creates a table of contents in real time, visible when the file is open.

- CSV files do not have a standard way to store metadata, but authors commonly created a README or MANIFEST file that describes the structure and content of the files in a collection. (See [9] for examples.)

On a practical level, templates for file creation, data collection, and lab notebooks makes it easier to be consistent and to spot inconsistencies.

4. Add metadata.

Just like people who end up with piles of photographs with nothing written on the back, we all have digital mounds of files and content with no metadata describing when it was created or what it contains. Even the most basic metadata provides extra clues for information retrieval; however, what you can add depends on the software you use to create, store, and access your information, and on the file formats that information is stored in.

Almost all modern operating systems allow you to add information to the Properties of a file or directory. Databases, word processors, and website construction programs also have built-in metadata capabilities, though they may be hard to find and harder to understand how to leverage. To make matters worse, the fact that metadata is often software-specific makes it easy for inconsistencies to creep in. For example:

- The tags used on a WordPress website may not be in step with the properties in the images on that site.

- Keywords added to a journal article when submitting to the publisher's site are not automatically added to the metadata in the PDF being submitted.

- When a citation is copied from an article database to a bibliography manager, the software may not copy over the structural information implied by the article's location in the database.

The most difficult thing about metadata, however, is getting into the habit of creating it in the first place. If you get to choose what software to use, it helps to pick one that makes simple things simple. For example, most website generators allow you to type tags into an article's header without having to define them first. This can lead to a proliferation of synonymous (or misspelled) tags, but some occasional cleanup is better than tackling a mountain of untagged information.

You should also examine how metadata can be transferred from an old system to a new one if you have the luxury of switching software (or have had a change forced on you). Some form of XML is usually the best option when doing this: it is likely to be with us for many years to come, and the same pedantry that makes it tedious for human beings to type and read ensures that programs can read it without having to guess what its creators actually intended.

5. Use search *and* browsing

Research on information seeking shows that people search *and* browse when they're trying to find information. As they browse a website or document or file, they build a mental map of the content they could possibly find, then search based on that map. "In the process, they modify their information requests as they learn more about what they need and what information is available from the system" [1]. You have probably seen or done something similar with a print book, trying to determine if it's one you want by looking at the table of contents and the back cover. These two functions work together because search allows users to find information they know they need, whereas browsing allows users to find information they don't know that they need [10].

You should therefore make information accessible both ways and make it easy to move from searching to browsing and back again. Tags and other metadata help with searching, while structural clues tell users about the content contained in the information they are looking at. That communication, "enables the answers to users' questions to rise to the surface and answer questions like, Where am I? What's here? Where can I go from here?" [1]. Similarly, "...the words you use in the navigation systems and headings of [your content] help you find what you're looking *for*, but they also help you understand what you're looking *at*" [11].

For example, when users don't know exactly what they need, the terms in a menu help them understand the vocabulary used in this domain and the boundaries of what is included (i.e., terms are listed) or excluded (i.e., no menu terms exist). At the same time, the headings in the documents they find act as topical markers: they help users summarize the information contained in the document, but also refine what they would search for based on the terms used in those headings. Navigation bars on websites function in a similar way: if the user knows exactly what they are looking for, they can scan the menu and select the option that matches their need.

6. Mimic real world directions.

The language we use in digital environments mirrors that used for physical directions: we "visit" or "go to" a website without actually changing our physical location. Using the navigational metaphor consistently helps users build the mental map mentioned in the previous tip. File paths and breadcrumb trails on websites give users a sense of where the information resides and suggest new paths they can take [7]. For example, the URLs of a website might all include the name of a section of the site, such as `/papers/` or `/blog/`.

While much has been written about web usability in general [3, 12], library science focuses on information-seeking behavior. For example, we know that users scan but don't read: they click on the first close thing they see and give up very, very quickly [10]. Your markers and directions should therefore be as consistent as highway signs with regards to appearance, style, and type of information. Wherever possible (and it's *always* possible), use mechanisms that users will have become familiar with elsewhere, such as the vertically nested folders of file browsers or the left-to-right arrangement of breadcrumb trails.

7. Use meaningful names.

The names of files and URLs of webpages are the one piece of metadata you cannot avoid creating, so always choose ones that are human-readable and that convey information about what they name, both when navigated to *and* when returned in search results. Returning again to the faculty member's website, it would be easy to name a paper `plos2020.pdf`, but since other people may also have published papers in PLoS in 2020, a more structured name such as `lin-findability-plos-2020.pdf` will both convey more information at a glance and retain that information after the paper has been downloaded and put in a folder with dozens of others.

There are many ways to develop a naming schema, largely related to the nature of the information you create. At the most basic level, "you should use consistent names for the same reason that you use good file organization: so you can easily find and use data later. Additionally, good naming helps you avoid duplicating information [8]. Researchers with multiple research projects or significant complexity in their data sources should establish and document a unified system of abbreviations for those projects or sources; these can be summarized in a data dictionary or README file. Consistency is key: standardizing on lower case, a preferred date format (YYYMMDD or YYYY-MM-DD will both sort chronologically), and filename suffixes (`.jpg` instead of `.jpeg`) will help everyone find what they need [13, 14].

Renaming existing files to be consistent with your standards after the fact can seem like a waste of precious time, but since the research cycle doesn't end with publication [8], there is a very high likelihood that someone will need to reuse your data and will have to try to figure out what files corresponded to what part of your research.

If you have things to name that are not files, such as projects, web pages, or document headings, remember that the more generic a term is, the harder it is to search for: naming a raw data file "raw" or a downloaded file "download" makes finding the information they contain nearly impossible. A quick test is to search for the name before adopting it: if dozens of unrelated pages come up, you may want to pick a different name. You should also think about nicknames or shortened versions of your names and make sure they are present in text or tags so that the content can be discovered by a search engine and a user.

8. Use tags.

After meaningful names, tags are the easiest and most effective metadata you can create. Almost all digital tools allow users to add arbitrary tags to items: file properties on Windows and labels on GitHub issues are just two examples. Additionally, almost all search tools leverage tags to narrow a query's scope. This means that you can now file a single thing in multiple "locations", which was not possible in the pre-digital era. Multiple tags also assist users from varied backgrounds because the terms can be customized to each type of user your information has.

When choosing tags, be consistent in your depth of topical term assignment (how specific your terms are) and your selection of terms for subject and format (the number of terms you use to describe each subject and format). For example, if you tag some items in an ecological data set with a species name, don't tag others simply as "reptile" unless the species is unknown, in which case you should:

- tag all items "reptile", "bird", "mammal", and so on for high-level searches, and

- tag all items with a species, which might be "unknown" or "NA" (not available).

What should you tag? The answer is "everything" from informal personal notes to data sets submitted with publications or included in repositories, because it is all material you will want to be able to find later. The benefit of tagging comes from doing it in all of those situations, not just when a journal submission requires it.

If you are certain something is for purely personal use, you can create your own taxonomy of subject keywords, which is called a *folksonomy*. Folksonomies are what you see with tags on Flickr: early content creators assign terms as they see fit, and later contributors can use those or add their own. If you take this route, it's worth reviewing new tags regularly to look for synonyms, misspellings, differences in capitalization, singular/plural discrepancies, and other inconsistencies.

What terms you use as tags for personal consumption may not matter much, but work that is shared with colleagues should use particular terms or tags that conform to certain standards [15]. These terms typically come from taxonomies, thesauri, and ontologies: taxonomies and thesauri generally have built-in subject hierarchies that can help you create navigational structure, while ontologies map relationships between ideas. Crucially, all three are *controlled vocabularies*: they are a defined list of terms created and maintained by experts rather than being crowdsourced like a folksonomy.

There may or may not be relationships built between terms in a controlled vocabulary, such as equivalencies ("CA" for "California"), broader/narrower terms (United States/California), and/or replacement (weed *use* marijuana). Established subject terms will match article databases, data repositories, and library catalogs that you and your users might already be familiar with, which will again aid search and navigation. Well-known examples in the United States include the National Cancer Institute (NCI) Thesaurus [16] and the Medical Subject Headings (MeSH) [17].

9. Understand the difference between format and subject.

However you create tags, you need to address the distinction between format and subject. Format describes what your content *is*, while subject describes what it is *about* [18]. About-ness is the most common content analysis, but is-ness issues will probably affect people's ability to use your information, so you may want to add metadata to make it explicit.

A simple example of this is a blog post on a website. The post is *about* a subject, but it *is* a blog post rather your biographical details, your bibliography, or thumbnails of the images you have used. Going back to your users, what subjects are important to them? And do those topics carry over or change between differences in format? For a librarian, this is basically a question of combined terms: are your format terms uniquely matched to topics (e.g., blog posts are always about news) or do you have multiple topics in each format (e.g., blog posts and tutorials on the same subject)?

Similarly, you can rely on filename suffixes to distinguish computational notebooks from PDF files, tabular data sets, or slide decks, but should use tagging, a filename convention, or a description in a README to tell people whether the contents are raw information, tidied-up data, or an aggregation of several underlying datasets. This enables users to search by topic, format, or both.

Since dissemination sometimes changes a file's format (e.g., printing slides to a PDF), naming and metadata conventions tend to be more robust as well as more informative than relying on file types. Once again, structural clues can help: a folder specifically for conference presentations may contain one sub-folder for each presentation, which in turn contains the PowerPoint and PDF versions of the presentation with exactly the same names but different filetype suffixes. Likewise, journal articles you store will need a naming or structural convention to distinguish articles you have written from those you have downloaded for your own use.

10. Do not abbrvt.

Acronyms and abbreviations make communication between those who know them more efficient at the price of making them less accessible to newcomers. Spelling out acronyms and abbreviations that you take for granted (or hyperlinking to their definitions) therefore makes information easier to find and newcomers feel more welcome. When doing this, remember that acronyms are often repurposed by different professions or disciplines: what seems obvious to you is probably not obvious to people from other communities. Since every discipline has some common abbreviations, write them all out in full the first time they appear or create or point to a term dictionary.

Conclusion

Changing work habits is hard, so remember that while perfection isn't possible, progress is. Start by deciding whether to begin your next project with a new set of information organizing principles or to go back and alter existing artifacts [8]. Whichever you choose, the "ways you enforce your way of doing things changes how users think about the place[s] you made and perhaps ultimately, how they think about you" [3].

References

1. Rosenfeld L, Morville P, Arango J. Information Architecture. O'Reilly Media Inc.; 2015.

2. Hedden H. The Accidental Taxonomist. Information Today; 2016.

3. Covert A. How to Make Sense of Any Mess. CreateSpace Independent Publishing Platform; 2014.

4. Borges J. Selected Non-Fictions. Penguin; 2000.

5. Xie Y, Thomas A, Hill AP. blogdown: Creating Websites with R Markdown. Chapman and Hall/CRC; 2017.

6. Williams B, Damstra D, Stern H. Professional WordPress: Design and Development. 3rd ed. Wrox; 2015.

7. Krug S. Don't Make Me Think, Revisited. New Riders; 2014.

8. Briney K. Data Management for Researchers. Pelagic Publishing; 2015.

9. The Pudding. Data sets created for stores on The Pudding; 2020. https://github.com/the-pudding/data.

10. Bates MJ. Toward an integrated model of information seeking and searching. New Review of Information Behaviour Research. 2002;3:1–15.

11. Arango J. Living in Information. Two Waves Books; 2018.

12. Group NN. UX research and consulting firm; 2020. `https://www.nngroup.com/`.

13. Wilson G, Aruliah DA, Brown CT, Hong NPC, Davis M, Guy RT, et al. Best Practices for Scientific Computing. PLoS Biology. 2014;12(1). doi:10.1371/journal.pbio.1001745.

14. Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK. Good Enough Practices in Scientific Computing. PLoS Computational Biology. 2017;13(6):1–20. doi:10.1371/journal.pcbi.1005510.

15. FAIRsharing. FAIRsharing.org; 2020. `https://fairsharing.org/`.

16. Institute NC. NCI Thesaurus; 2020. `https://ncit.nci.nih.gov/ncitbrowser/`.

17. American Society of Indexers. Online Thesauri and Authority Files; 2020. `https://www.asindexing.org/about-indexing/thesauri/online-thesauri-and-authority-files/`.

18. Joudrey DN, Taylor AG, Miller DP. Introduction to Cataloging and Classification. Libraries Unlimited; 2015.