Towards a free matter economy (Part 5)

Discovering the future, recovering the past

Terry Hancock

> I think the health of our civilization, the depth of our awareness about the underpinnings of our culture and our concern for the future can all be tested by how well we support our libraries.—Carl Sagan, *Cosmos*

Libraries have been around for a lot longer than software, and librarians long ago learned many of the data management lessons that have only now begun to surface in the world of software and databases. By contrast, software is a young, rapidly changing field, and this has affected its outlook. Five years may seem like an eternity in software development, but in the archival business, it's just the blink of an eye.

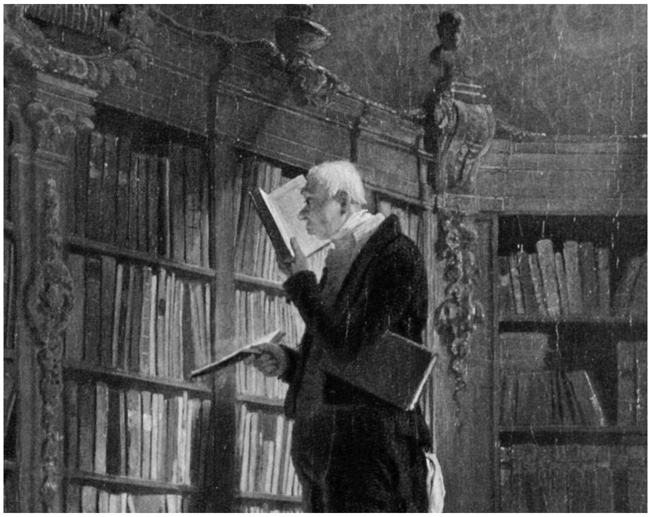
What libraries have not dealt with historically, however, is the dismaying array of data storage mechanisms and file formats that software data represents, the troublesome transience of the tools needed to access that data, and the overwhelming quantity of the data that is produced.

Finding your way

For uncontrolled data sources, such as the world wide web at large, we have no real choice except to use machine indexing systems such as Google, but if we want a free-design database to be a maximally reusable resource, we need to make use of metadata.

Metadata, "information about information" is the key to library science. There are many decisions about data sources that cannot be easily made by machine, and metadata tagging allows the librarian to classify works so that they can be retrieved later. Typically these include the familiar "subject", "title", and "keyword" indexes that were once the mainstay of library card catalogs, and now of electronic OPACs used by modern libraries [1].

Libraries had been in use for a long time before the advent of computers, and librarians have therefore learned important lessons about how information should be cataloged. As our databases become more full-text, more diverse in format, and broader in content, we increasingly encounter the same challenges that have formed the basis of library and information science for centuries (*Der Bücherwurm*, by Carl Spitzweg from Wikimedia Commons (PD))



Schemas

For a long time, libraries have used the highly rigorous but somewhat difficult MARC database system [2], which is not exactly a relational or object database, but is somewhat in between, and requires special handling to be managed properly. Since then, however, computer science development has led to more manageable database design styles, and the increasing quantity of digital and multimedia library resources has begun to make the conventional book-publishing orientation of MARC obsolete. Librarians are, therefore, developing more streamlined, agile metadata database systems, such as the FRBR [3], which is a pure relational database schema designed to encapsulate the minimal (but complete) requirements for library use, and the even more streamlined Dublin Core (DC) metadata system [4].

MARC, FRBR, and DC provide "schemas" or lists of the types of metadata elements that can be recorded for a work.

Vocabularies

A perhaps less obvious need is the "controlled vocabulary". There are many ways to express a single subject—is the study of extra-terrestrial life to be called "bioastronomy", "exobiology", or "xenobiology"? People have used all three terms, and, through their usage, established different emphases. Should each be given its own category, or should they be treated as synonyms and stored together? Which term will we use for that category?

It's a common mistake to underestimate the importance and difficulty of selecting appropriate taxonomic vocabularies. This is because we are all biased towards our own fields of endeavor and tend to have only a vague idea of the structure of other disciplines. Consider, for example, the domain-specific controlled vocabulary represented by the "Trove" system [5], which you use if you look for software projects on the Sourceforge site. It's an excellent system for finding software, provided that the paradigms of computing don't change too much. However, should entirely new software types evolve, or should the system be used for things outside of the realm of software, the Trove categories become much less useful.

Fortunately, library associations have done a lot of work on broad, inter-domain classification. For example, in the English language, there is the AACR [6], used in Canada and the United States. It seems like a wise idea to use these standards whenever possible.

Agility and human nature

One of the problems in applying professional library methods to software works and the results of community based production, is that they are fairly labor intensive, and rely on a class of professional experts to do the classifying. Considering the quantity of data on a site like Sourceforge, it's not hard to see that hiring librarians to manage the problem would be a daunting prospect.

Fortunately there are other ways to assign metadata to files.

Creator tagging

Perhaps the most obvious solution is to have authors assign metadata to their own works. It's an obvious solution; it's the typical starting point for most systems; and, for things like title, attribution, and licensing it is really the only way to do it, because the creator is the one who chooses those properties.

It's not without problems, however. An author is not always the best person to trust about their own data. Vendors tend to puff up their projects; authors can have greatly inflated (or deflated) egos; and people are often just lazy or inept with the submission mechanisms [7]. It seems clear that we can't always rely on the people who create works to be good at classifying them.

Expert tagging

The library solution is to have items cataloged by professionals who train in doing just that. Even when creators determine their own subject headings, it is usually the case that the schema (what data to record) and the vocabulary (what options are available for each field) are decided by experts.

Social or consensus tagging

A recently successful model, employed by community sites such as Wikipedia, and used by companies such as Amazon, is to rely on reader feedback to improve the metadata as it is being used. Many properties of such sites—the ability to edit them, the feedback forms ("Is this page useful?"), and other features—make this feasible. Despite some claims to the contrary, Wikipedia is a remarkably successful case of well-organized self-organization. Even if it were to fall short in quality and accuracy compared to a professionally cataloged encyclopedia, such as Encyclopedia Britannica, it would still represent a considerable achievement—and yet, studies have suggested that it does not fall far short of such works. [8]

Artificial intelligence

For purely objective data such as file format, checksum, or size, automatic cataloging is the obvious solution. Advances in search technology and artificial intelligence have allowed us to go much further than this, though. AI-based text analysis and data-mining has been a popular theme for some time, and Google's search engine benefits from some of this technology.

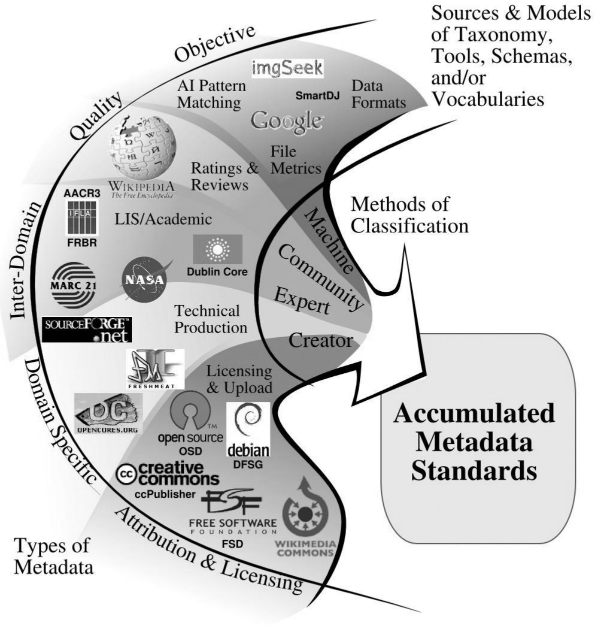
However, in the last year or so, we have started to see more ambitious and original AI techniques being used on non-text data. For example, the free software package "SmartDJ" [9] is a playlist manager that find songs that "sound alike", by analyzing the recordings themselves, and "imgSeek" [10] is a tool that recognizes similar images, so that a rough sketch of an expected image can be used to search for the image. No doubt these applications are still primitive, but they show a promising possibility for future indexing systems.

Putting them together

Together, these methods can be usefully understood as a continuous spectrum, based on types of metadata and who should be most trusted to establish them (see figure). It's easy to imagine a system based on creating a metadata stub for a package, starting with creator-provided title, attribution, and licensing, and then allowing that stub to be further constructed by the actions of the different interested parties managing the overall database system.

A promising tool for this kind of task is "Resource Description Framework" (RDF) [11]. This is the basis of the so-called "Semantic Web" [12], and is useful because it provides a completely extensible and decentralized system for assigning metadata. Such projects as "IkeWiki" [13] are already providing ways to make such RDF tagging easy to apply to dynamic websites.

There are a wide variety of sources of metadata schemas and controlled vocabularies for classifying information that a free design based economy will need to archive both as supplier and consumer. Some popular examples are referenced here. The data can be usefully thought of as occupying a spectrum from "most personal" information, which can only be supplied by the creator, to "completely objective" information which can be derived by machine. Other important methodologies have been created for the space in between, and the scope of what can be machine-processed has increased

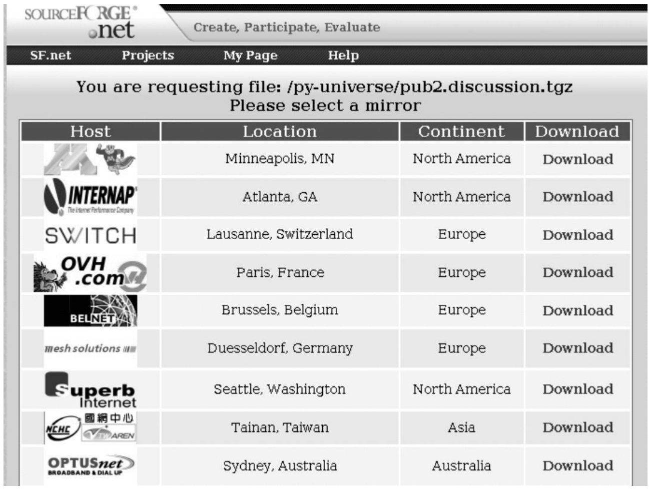


Signal to noise

Nearly all of the data that passes through our websites, mail servers, and even development projects is dross. Most data is only temporarily valuable, or even unwanted "spam". If we were to institute a policy of saving all of that data, we wouldn't only require exponentially increasing data storage mechanisms, but we'd also be hindering the recovery of data through an extremely low "signal to noise" ratio. What is needed is an effective information sieve that only captures the permanently valuable information, and allows the rest to spill through.

Community based rating systems, such as Slashdot's "moderator points" system are likely to be useful in solving this problem, although it really calls for more than a simple score. For example, an announcement of an upcoming meeting may be very important at the time, but have little

Sourceforge originally maintained all of its own download servers, but as the site grew, it had to find a way to scale. Their solution is shown here, with a series of commercial mirror sites providing bandwidth, and of course, getting a small banner ad in return. People who use this service are likely to appreciate the companies' direct contribution of bandwidth to their needs, so this is a very positive kind of advertisement



SOURCEFORGE .net Create, Participate, Evaluate

SF.net Projects My Page Help

You are requesting file: /py-universe/pub2.discussion.tgz
Please select a mirror

Host	Location	Continent	Download
	Minneapolis, MN	North America	Download
INTERNAP	Atlanta, GA	North America	Download
SWITCH	Lausanne, Switzerland	Europe	Download
OVH .com	Paris, France	Europe	Download
BELNET	Brussels, Belgium	Europe	Download
mesh solutions	Duesseldorf, Germany	Europe	Download
Superb Internet	Seattle, Washington	North America	Download
NCHC 國網中心 TWAREN	Tainan, Taiwan	Asia	Download
OPTUSnet	Sydney, Australia	Australia	Download

permanent value. A more specific rating system is probably called for, which tells why a post is important as well as how important it is. An RDF-based representation might be appropriate for this purpose.

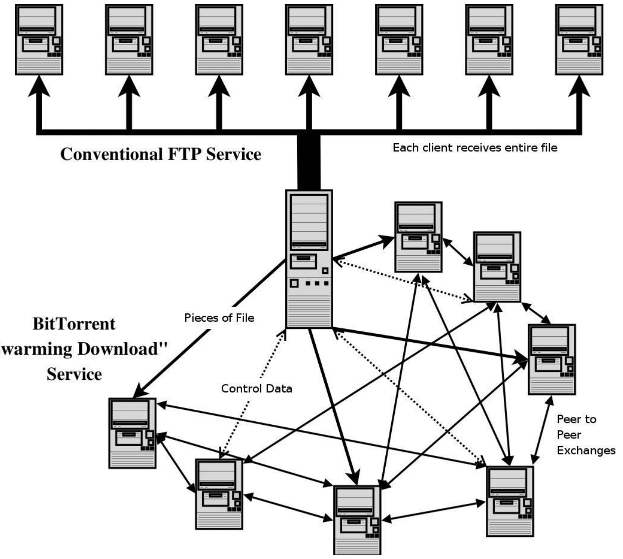
Technical hurdles

Internet bandwidth, server uptime, and data storage capacity are the primary physical costs of maintaining a large online archive of free design data, and if the data is to remain free, it is counter-productive to recover this cost through charging patrons directly for access to it. Fortunately, there are two existing solutions for these problems.

Conventional mirroring

The older technique, as demonstrated by Sourceforge is to simply have "big pipes" and provide for the necessary bandwidth. When the bandwidth gets too high, the site can recruit mirrors. Sourceforge attracts sponsors in this way, who use the opportunity to advertise to users (see figure), who, seeing the immediate benefit of the sponsorship are likely to be very positively influenced by the ads.

A "swarming download" system, like BitTorrent solves this bandwidth scaling problem in a different, more community-based way. Rather than relying on a few suppliers with "big pipes", the swarming downloads system creates a peer-to-peer (P2P) network, moderated by a controlling or "seed" site, which only has to supply comparatively few feeds of data to client computers, which then aid in the dissemination process by serving their piece of the data to the other requesting clients. In this way, swarming downloads are most efficient for data that is most widely requested, so they scale very well (an obvious objection is that some clients may be modified to "leech" or not serve data, but BitTorrent provides a simple "market" method of resolving this problem—good citizens are rewarded with better bandwidth)



Swarming downloads and peer-to-peer

A more recent development is the use of peer-to-peer networks and so-called "swarming downloads" which use the internet in a much more distributed way. In this model, patrons effectively pay for their own download bandwidth via their local ISPs, rather than burdening the server. The cost is so distributed, however, that it is not noticeable.

This system is exemplified by Bit Torrent [14], which allows a file to be "shared" to a peer-to-peer network from a "seed" site which takes the place of the conventional FTP server. This system scales particularly well, because it shows the highest gain for the packages with the highest demand.

Torrent feeds require special clients however, so they are not yet ubiquitous. Therefore, it is very likely that an archive will have to provide both methods, though it seems desirable to find ways to encourage the peer-to-peer method.

Free design sources

There are already a number of sources for free design data, so encouraging the growth of a free design community will involve making existing data more available as much as making more designs. Furthermore, since new designs build on old ones, building a good archive of design data is important to new innovation as well as to end users.

Government projects

Perhaps the most obvious source of public good works are public agencies. Under United States' law, any work which is done completely by government employees is automatically in the public domain. This includes much work of the data developed by such organizations as NASA [15], the Forest Service [16], the USGS [17], DOE [18], USDA [19], and even the DOD [20].

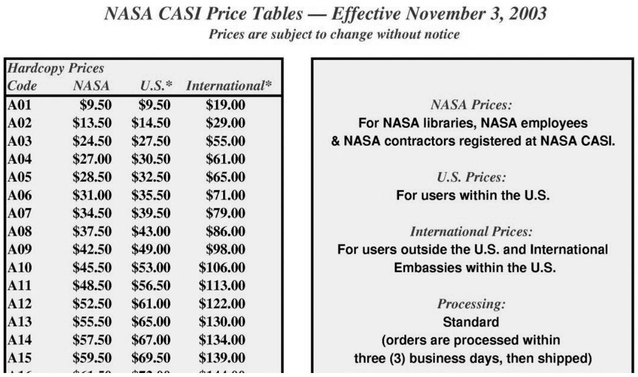
Other countries often have similar rules. In the European Union (EU), public funding may require publication of results under a free license, for example.

NASA, obviously of particular importance to our project, already provides some online access to search their documents [21], although many of them are not yet digitally imaged. So, it may be necessary to pay document processing fees to access the full text of the documents. One of the desirable possibilities for our community based project, would be to begin effectively mining these resources and making them more accessible to the free design community.

Community based projects

There is an increasing body of hardware development that is managed in community collaborative websites, such as the Open Cores project [22], which works on integrated circuit "IP Cores". These are relocatable elements of IC chips which can be used in large "Application Specific Integrated Circuits" (ASIC) and are therefore obvious candidates for creating commodity reusable designs. Among the projects successes are free versions of all the major simple gate chips (e.g. 7400 series) and complex projects such as RISC CPUs and micro-controllers. These are all critical areas of development, if we want to see "completely free" designs, since computer control systems are an important part of so many advanced hardware projects.

A recent search for a NASA contractor's report turned up an abstract, a price code (A02) for the document, and a link to this PDF pricelist (the document is not proprietary, this is a document-processing fee). If I knew I needed this document, or were I a contractor with plenty of money to spend on library research, this document processing fee would not be prohibitive. But a casually interested space enthusiast will never be inspired by it to contribute to a community-based development project. Imaging such documents and getting them into a searchable online form is an important project for anyone who wants to spur future innovation based on them



NASA CASI Price Tables — Effective November 3, 2003

Prices are subject to change without notice

Hardcopy Prices			
Code	*NASA*	*U.S.**	*International**
A01	$9.50	$9.50	$19.00
A02	$13.50	$14.50	$29.00
A03	$24.50	$27.50	$55.00
A04	$27.00	$30.50	$61.00
A05	$28.50	$32.50	$65.00
A06	$31.00	$35.50	$71.00
A07	$34.50	$39.50	$79.00
A08	$37.50	$43.00	$86.00
A09	$42.50	$49.00	$98.00
A10	$45.50	$53.00	$106.00
A11	$48.50	$56.50	$113.00
A12	$52.50	$61.00	$122.00
A13	$55.50	$65.00	$130.00
A14	$57.50	$67.00	$134.00
A15	$59.50	$69.50	$139.00

NASA Prices:
For NASA libraries, NASA employees
& NASA contractors registered at NASA CASI.

U.S. Prices:
For users within the U.S.

International Prices:
For users outside the U.S. and International
Embassies within the U.S.

Processing:
Standard
(orders are processed within
three (3) business days, then shipped)

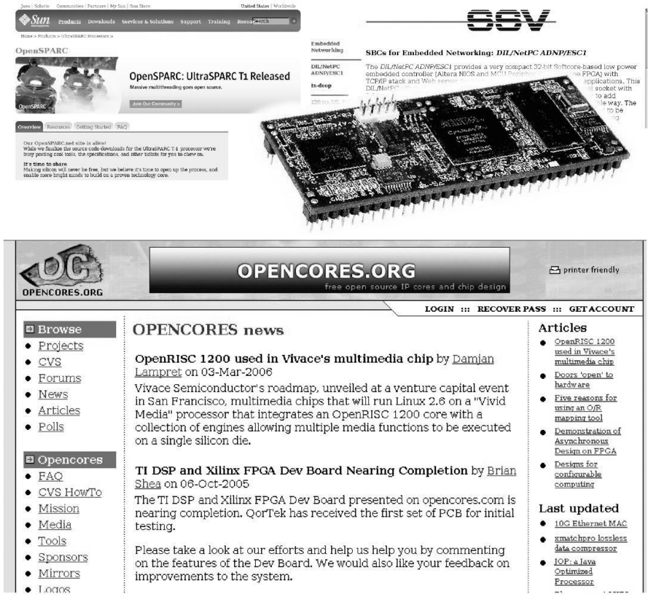
These technologies are early adopters of a free collaboration environment because the technologies lend themselves to software collaboration tools, because the people doing the work have seen successful free software projects, and because the complex designs benefit significantly from the kind of collaborative testing and development that goes into software.

Industry commoditization

Perhaps the most surprising trend has been initiated by electronics and computer manufacturing companies, who have decided to free the design of older model hardware. This has been done in order to provide a "future proofing" value proposition to customers, or simply to develop goodwill, especially among customers who already see the benefit of using free software such as embedded Linux systems.

There are also a few high-profile cases, such as Sun Microsystem's recent decision to offer the Verilog source code and other design information for the UltraSPARC T1 CPU under the GPL [24]. This move is presumably meant to bolster Sun's hardware platform as a "commodity" design, just like the enormously successful Intel architecture machines.

Open hardware, like this FPGA “softcore” single board computer from SSV Embedded Systems (middle) [23], is becoming increasingly popular. There are both industry releases, such as Sun’s recent decision to go open hardware with its new SPARC T1 processor [24], and community-based production projects, like Open Cores, which has a wide variety of “IP Cores” or hardware designs which may be combined to master integrated circuit chips



Java | Solaris | Communities | Partners | My Sun | Sun Store
United States | Worldwide
Sun
Products
Downloads
Services & Solutions
Support
Training
Research
Search
Home > Products > UltraSPARC Processors >
OpenSPARC
OpenSPARC
Join Our Community »
Overview
Resources
Getting Started
FAQ
SSV
Embedded Networking
DIL/NetPC ADNP/ESC1
In-deep
OPENCORES.ORG
free open source IP cores and chip design
printer friendly
LOGIN ::: RECOVER PASS ::: GET ACCOUNT
Browse
Projects
CVS
Forums
News
Articles
Polls
Opencores
FAQ
CVS HowTo
Mission
Media
Tools
Sponsors
Mirrors
Logos
Articles
OpenRISC 1200 used in Vivace's multimedia chip
Doors 'open' to hardware
Five reasons for using an O/R mapping tool
Demonstration of Asynchronous Design on FPGA
Designs for configurable computing
Last updated
10G Ethernet MAC
xmatchpro lossless data compressor
IOP: a Java Optimized Processor

OpenSPARC: UltraSPARC T1 Released

Massive multithreading goes open source.

Our OpenSPARC.net site is alive!
While we finalize the source code downloads for the UltraSPARC T-1 processor we're busy posting cool tools, the specifications, and other tidbits for you to chew on.

It's time to share
Making silicon will never be free, but we believe it's time to open up the process, and enable more bright minds to build on a proven technology core.

SBCs for Embedded Networking: *DIL/NetPC ADNP/ESC1*

The *DIL/NetPC ADNP/ESC1* provides a very compact 32-bit Softcore-based low power embedded controller (Altera NIOS and MCU Periphe

OPENCORES news

OpenRISC 1200 used in Vivace's multimedia chip by Damjan Lampret on 03-Mar-2006

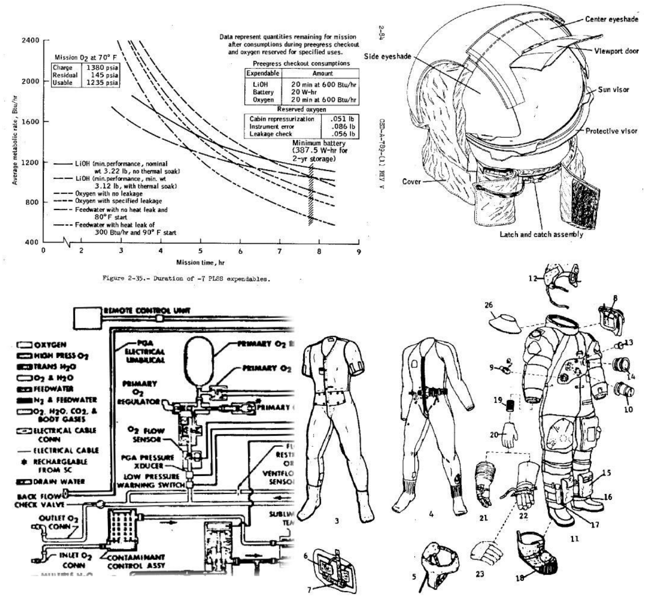
Vivace Semiconductor's roadmap, unveiled at a venture capital event in San Francisco, multimedia chips that will run Linux 2.6 on a "Vivid Media" processor that integrates an OpenRISC 1200 core with a collection of engines allowing multiple media functions to be executed on a single silicon die.

TI DSP and Xilinx FPGA Dev Board Nearing Completion by Brian Shea on 06-Oct-2005

The TI DSP and Xilinx FPGA Dev Board presented on opencores.com is nearing completion. QorTek has received the first set of PCB for initial testing.

Please take a look at our efforts and help us help you by commenting on the features of the Dev Board. We would also like your feedback on improvements to the system.

Much of the design data for developing a future in space is already available under free licensed or public domain terms, either because of government funding, or patent expiration. This set of illustrations from an Apollo procedures manual (which is available online [26]) gives a taste of what is available. Unfortunately, much of this data is still only available in hard copies in only a few places in the world. The process of digitally imaging and extracting, or adding metadata to make it searchable, is an enormous project. But, it is one which is increasingly possible, given the availability of community-based production, artificial intelligence tools, an interested community, and cooperation from the present custodians of the data



A rich public domain

It is startling to think that all of the progress in launch vehicles, spacecraft, and spacesuits, developed by NASA during the 1960s and 1970s that got us to the moon and built the reusable Space Shuttle, is now in the public domain, regardless of whether it was developed by government agencies or contractors. However, patents, unlike copyrights, have relatively short durations of 20 years or less in most countries (including the US), so anything developed before 1986 is effectively “fair game” at this point.

This data is old, and there are in many cases smarter, more modern ways to design new equipment, based mainly on improvements in computerized micro-controllers and materials science. Nevertheless, the older designs are often an excellent starting place. For example, the requirements and industrial materials available for the creation of spacesuits have not changed significantly since 1970, and most of that data is already available to the public domain, even if the sources on the subject matter remain somewhat inaccessible. The primary need, therefore, is to get that information into a more usable form through document imaging, text extraction [25], metadata tagging, and cataloging.

Bricks and mortar

Even today, libraries reach many more people than the internet, and have a more direct impact on many more. Technology users, who are not primarily in computer science fields, are not as well represented online, and they include many of the people we are interested in, both as developers and consumers of new free design. There’s also the possibility of bringing people to the free design community via internet-connected computers in libraries.

So, it’s important for a major free design project to embrace the existing library standards and institutions, particularly if the project wants to appeal to a broad audience. What we should do is provide a “library interface” that allows such

Even today, when card catalogs are becoming a thing of the past, books are not obsolete, and neither are libraries. The library network reaches many more people than does the internet alone. A truly effective free-design archive should strive to cooperate as much as possible with existing brick-and-mortar library systems, by leveraging such hooks as "Interlibrary Loan" [27] and making information usable from kiosk-type systems in libraries. (Joe Crawford, Wikimedia Commons (CC-By))

an archive to act like an ordinary brick and mortar library as well as an internet resource.

Imagine this scenario: a library patron in a far away library would be able to search the archive via their own library's OPAC or website. The archived materials would appear as books in a remote library collection. Using the Interlibrary Loan mechanism [27], the patron could then request the "book" from the librarian, who would then request it from the archive. The patron would pay the cost to have the book delivered, based on processing fees. Using appropriate print-on-demand technology, the book is printed digitally and sent to the library. The book might then belong to the patron, or become part of the local library's collection.

This scenario would require a number of individual technology problems to be solved, but none is a show-stopper. The FRBR provides an interface to the OPAC system (MARC records can be generated). There is apparently nothing stopping an electronic archive from joining an interlibrary loan organization. Systems developed for free software documentation have been developed to automate document preparation. Print-on-demand service has become a viable business model [28], with a number of vendors providing the service; and finally, the electronic commerce and shipping industries are entirely capable of handling the transaction fees and shipping.

It's not enough to innovate. We must also remember what we innovated and forget the irrelevant details so they don't pollute the ocean of information

Joining a library association is also a good idea politically. Librarians are perceived as a mild-mannered group, but they can be fierce in the protection of free speech and free expression rights. There is clearly a lot to be gained by the internet community at large, community-based information production projects, and library associations joining together as defenders of the free interchange of knowledge.

Libraries of the future

It's not enough to innovate. We must also remember what we innovated and forget the irrelevant details so they don't pollute the ocean of information our data must be found in. Accurate metadata, achieved by combining a variety of different methods, based on the most-reliable sources for each, is essential to ensuring the long-term accessibility of the data we need. There is already a substantial volume of free design data in existence, from community, industry, academic, and government sources, but it is often underutilized because of short-falls in document imaging and recognition, and (most importantly) metadata tagging. Twenty-first century developments in artificial intelligence and community-based technologies are making it possible to construct means of solving the technical problems, though. So, this puts us in a good position to start building the digital design libraries of the future.

Bibliography

[1] Online Public Access Catalog (`http://en.wikipedia.org/wiki/OPAC`)

[2] MAchine Readable Cataloging (`http://www.loc.gov/marc`)

[3] Functional Requirements for Bibliographic Referencing (http://www.ifla.org/VII/s13/frbr/frbr.htm)

[4] Dublin Core (http://dublincore.org) initiative

[5] Trove (http://sourceforge.net/softwaremap/trove_list.php) categorization system

[6] Anglo-American Cataloging Rules (http://www.collectionscanada.ca/jsc)

[7] Cory Doctorow Metacrap (http://www.well.com/˜doctorow/metacrap.htm)

[8] Jim Giles Internet encyclopaedias go head to head (http://www.nature.com/news/2005/051212/full/438900a.html), **Nature**, 2005

[9] SmartDJ (http://rudd-o.com/projects/smart-dj)

[10] imgSeek (http://www.imgseek.net)

[11] Resource Description Framework (http://www.w3.org/RDF)

[12] Semantic Web (http://www.semanticweb.org)

[13] IkeWiki (http://ikewiki.salzburgresearch.at)

[14] Bit Torrent (http://www.bittorrent.com)

[15] US National Aeronautics and Space Administration (http://www.nasa.gov)

[16] US Forest Service (http://www.fs.fed.us)

[17] US Geological Survey (http://www.usgs.gov)

[18] US Department of Energy (http://www.energy.gov)

[19] US Department of Agriculture (http://www.usda.gov)

[20] US Department of Defense (http://www.defenselink.mil)

[21] NASA Technical Report Service (http://ntrs.nasa.gov)

[22] Open Cores (http://www.opencores.org)

[23] SSV's DIL/NetPC ADNP/ESC1 (http://www.dilnetpc.com/dnp0046.htm) Single-Board Computer (SBC)

[24] OpenSPARC (http://www.sun.com/processors/opensparc)

[25] Optical Character Recognition (http://software.newsforge.com/article.pl?sid=05/12/15/1848236)

[26] Apollo Apollo Extravehicular mobility unit. Volume 1: System description (http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19730064705_1973064705.pdf)

[27] Interlibrary Loan (http://en.wikipedia.org/wiki/Interlibrary_loan)

[28] Print on Demand (http://en.wikipedia.org/wiki/Print_On_Demand), *e.g.*Lulu (http://www.lulu.com/)

Copyright information

About the author

Terry Hancock is co-owner and technical officer of Anansi Spaceworks (http://www.anansispaceworks.com), dedicated to the application of free software methods to the development of space.