Free file formats and the future of intellectual freedom

Information as property may be served by closed file formats, but the freedom of information requires free formats

Terry Hancock

So far, proprietary formats have been maintained through a number of short-term tricks, but the advantages of free formats become clearer in the long run. Business and the computer industry have tended to be very shortsighted. However there are some important classes of technically proficient users with a much longer outlook, whose needs can only be met by free file formats. If we in the free software community want to see free formats take hold, we need to address the needs of *these* users. We need to do this in order to leverage their interest into long-term acceptance of free standards by the world at large. We also need to ensure free standards *exist* — because in many key areas they just don't. Fortunately, the record provides good evidence that free software developers will step forward to meet these needs, once they become aware of them.

> The need for ten to twenty years of data stability is routine in the sciences

My introduction to computers came as much through my interest in astronomy and spaceflight, as through any interest in computers for their own sake. One of the biggest differences is the perception of time — in the computer and IT world, three years may seem like an eternity, but there is still active research being done based on 30-year-old space technology (everything needed to get to the Moon), or even 300-year-old astronomical data (Galileo's drawings of the Great Red Spot on Jupiter). More mundanely, the need for ten to twenty years of data stability is routine in the sciences, and there is considerable likelihood that this need for stability will increase in future, as these fields push the theoretical limits of detection and measurement accuracy.

From that perspective, any proprietary software's entire existence - let alone the duration of *any* commercial file format, or indeed the *concept* of file formats itself — is fleeting ephemera. To such users, the difficulties of closed file-formats are more likely to be blamed on the electronic medium itself, which is itself still regarded as a new development, even after 30–50 years of use.

As a researcher, I've also spent a lot of time using libraries, which as any serious researcher knows, are still far better than Google, simply because there's an awful lot of data that isn't available electronically, let alone for free and posted on the web. The search engines I first learned to use were the electronic library catalogs that became ubiquitous in the 1980s. Those systems ran on the MARC database standard that endures in modern library systems, although more modern standards like FRBR and Dublin Core are being developed.

To this day, there remains a public perception that libraries and the internet are somehow opposing forces in the world, with librarians clinging to worn-out paper technology in the face of the inevitable onslaught of better electronic methods. Maybe in 1980 that perception was true of many librarians, but in 2005 it's total bunk. Many, many librarians are excited about and are fully embracing the idea of *electronic libraries* — systems which combine the best of the web technologies with the tried-and-true methods that librarians have been using in their cataloging systems for decades, and in the processes of *document imaging* whereby they can convert existing print media to be remotely accessible. But they are encountering resistance, not just from the natural difficulties of the technology, but also from the artificial obsta-

cles created by copyright laws which have been made more restrictive than ever in the form of the Digital Millenium Copyright Act (DMCA) and made essentially immortal by the several copyright extension acts that have been passed since 1978. Finally, the blow to intellectual freedom and personal privacy imposed by clauses in the USA PATRIOT act have librarians absolutely steamed! The fact that the basic mechanisms of file formatting, that make such full-text databases possible, are unstable and under attack by the same commercial and political forces, is *not* being missed by this group of people.

So when I began to research the task of applying the free-licensing model, which has worked so well for software, to the design processes needed for colonizing space, as we are doing at Anansi Spaceworks on the Narya Project, I immediately realized that stable, free data formats would be a necessity. Experiences with software like Microsoft's Word, Autodesk's AutoCAD, and RSI's IDL had shown me that vendor lock-in was a sure-fire way to kill any free development prospects.

Free design projects also involve a lot of different types of data to exchange: rich-text documents, yes, but also slide presentations, illustrations, software packages, 2D Computer Aided Design (CAD) drawings, 3D CAD models, Computer Aided Manufacturing (CAM) and Computer Numerical Controlled (CNC) machine control scripts, audio and video recordings, and a miscellany of less common data types.

What I found, is that the results are somewhat mixed. Some content types have good and obvious free-format choices, some have only proprietary formats or very poor free formats that can't compete, and still others are engaged in pitched battles between free and non-free standards. Each of them tells a piece of the story, and shows what we may expect from the years to come.

The writing is on the wall

The awareness of the free format issue is pretty high, and probably nowhere higher, than with word processing documents. The only serious proprietary contender here is the Microsoft Word DOC format. All of the other formats, including the Word Perfect WPD format are pretty much on the way out, and even Microsoft itself has capitulated to the degree of focusing on its more open RTF format, and promoting XML. Although, as has been argued elsewhere, XML is by no means a sure-fire way to a meaningfully free file format.

Which is not to say that DOC is dead. That would clearly be wishful thinking, as I know from conversations with content providers like the National Space Society, which has consistently used MS Word DOC format in a misguided attempt to provide educational materials in a "common" format. It can be quite difficult to persuade authors and distributors of such information; even that the format is a thing worthy of serious thought, let alone try to explain why requiring all of their potential audience to have the latest version of MS Word to read their work, is a very bad idea.

Among people more in the know, such as librarians and serious researchers and publishers of content, the awareness is growing

Nevertheless, among people more in the know, such as librarians and serious researchers and publishers of content, the awareness is growing. You don't have to go through too many frustrating experiences trying to read files that aren't fully forward or backward compatible to get the idea that something must not be right. That's all while being faced with huge stockpiles of data that must be read from tape or CD and converted file-by-file and rewritten to other media. With this audience, the only real trick is to get them to realize that the problem is the closed format, rather than, say, an intrinsic failing of electronic media. In other words — help them to realize that the problem is artificial and solvable.

The most extreme reaction to this is that of the Project Gutenberg archive, which has opted (at least for most of its existence) to use only ASCII-encoded plain text to store their documents. Of course, this makes the documents much less usable, since only through human intervention is it possible to add the expected text formatting, but it has served them very well.

Acceptance of PDF is very deep: you can get all your tax forms this way, and most government sponsored research reports are released in PDF, or occasionally, in HTML, which can also be regarded as a useful rich-text file format, even if we do generally only associate it with the web itself. And although there are some misgivings about the PDF standard, seeing that it's driven entirely by its originator Adobe, in order to promote a proprietary product, the standard is generally considered open since it continues to be documented by a published specification.

In the technical science and engineering communities, of course, the older TEX and LATEX standards (which are definitely free) continue to be prominent. Combined with XML

based markup systems such as Docbook and MathML, and converters to Postscript and PDF formats, we have a fairly complete system for academic authors using these content-aware text-formatting systems.

In the more conventional word processing world, of course, there is another open standard, which has emerged, based on the OpenOffice.org project, and standardized by the OASIS standards body. This standard is not yet so widely accepted, but there is some likelihood that it will be.

With the caveat that we will probably need import methods for DOC and RTF formats for some time to come, it's pretty clear that the battle for richly formatted text data will be won by free formats. There's a lot of awareness towards moving the whole of business and technical communications forward to free and stable file formats. And this is not only in the free software community, but in a much broader community of technically-inclined producers, who are valuable as opinion leaders and early-adopters.

The same old song and dance

Exacerbated partly by the brouhaha over internet file-sharing and the MPAA and RIAA's ham-fisted response to this perceived threat, the formats for interchanging and storing audiovisual data have received a lot of attention. This is something of a red-herring, but it has at least drawn developer attention to the problem.

There are at least two serious cases of patent-encumberment to be found here — in the cases of the GIF image format and the MP3 sound-recording format. In both cases, software patents were used to insist that programs implementing these standards were automatically subject to royalties or other limitations. Although lax enforcement prevented either from being a really serious threat, these incidents exposed the potential threat of such patent-based attacks.

The free software community rallied very well in both cases - inventing the Portable Network Graphics (PNG) and the Ogg Vorbis (OGG) standards respectively. Both standards are technically equal or superior to their proprietary competition, and stand a good chance of taking over in the transmission of such data on the web. It's very likely that the lax enforcement of patent protection for both GIF and MP3 was motivated in part because of the ease of migration to PNG and OGG formats.

Dark and disturbing things are happening with the storage of music on fixed media, however, because of all the players I can find in my local Walmart, none supported the free Ogg Vorbis format, providing only MP3 and a particularly nasty new proprietary format from Microsoft called Windows Media Audio (WMA), distinguished primarily by its strong support for "digital rights management". This of course, exposes a new vendor lock-in tactic, which is software/hardware cronyism: the producers and sellers of digital equipment are highly vulnerable to the hard-ball business tactics that companies like Microsoft are so well-known for.

The only really widespread standard for digital video is the DVD standard, which is built on MPEG2 encoding, which is at least a well-documented format. However, Ogg Theora has been started as a true free format for compressed video, to stand alongside Vorbis.

The dominant format for presentation of graphics on the web remains conventional HTML used with animated GIF images, although a proprietary format, Macromedia Flash, has really threatened to take over. Nevertheless, there is gradual progress on the Scalable Vector Graphics (SVG) standard, which, combined with Javascript, and given real browser support, promises to do many of the same tasks — an XML format backed by the Worldwide Web Consortium (W3C).

Planning for the future

After all of this good news, however, entering the world of computer-aided design and manufacturing is something of a disappointment. CAD software has always been very expensive and targeted to relatively few users, tightly bound to the manufacturing industries it supports, and the licensing cost has gone without much complaint, seeing as it sits alongside many more expensive hardware capital costs.

At the very highest levels of industry and government, there has been a definite recognition for the need for CAD interchange formats, and there have been standards such as IGES and the much newer STEP (ISO 10303-1) for doing that. IGES was somewhat limited in the types of data it could exchange, and although STEP is in principle much more capable, it is so complicated that no *full* implementation, proprietary or free, yet exists. You might say that STEP is actually a set of dozens of CAD drawing standards in one unified framework. Moreover, STEP is actually only a schema, leaving the actual data representation open. An XML-based storage format is reportedly under development, but is apparently not complete.

Nevertheless, these standardization efforts are promising, and there is at least one attempt at a free software implementation capable of reading and writing STEP data, called Open Cascade.

This situation is pretty daunting, both for the small-time CAD user and for the free software CAD developer. So it's not surprising that for many, the only seriously used format remains AutoCAD's DWG, which is proprietary and undocumented. Autodesk once promoted the DXF format, currently supported in free software by QCad, for drawing exchange, but it is considered too limited a format for serious CAD work. Also worthy of note is the XML format for 2D CAD used natively by PythonCAD.

It might be worth considering whether free software developed for computer graphics modeling, such as Blender, might be adaptable to 3D CAD applications, although it's clear that this might be far from trivial. Nevertheless, Blender provides a native format, which may be regarded as free, seeing that a free implementation exists.

Interested developers should probably check out one or more of these projects to see what can be done to create a useful 2D or 3D CAD standard, suitable for free software users and would-be free hardware designers to use.

A taxing situation

Perhaps the saddest example of proprietary lock-in gone mad, government cluelessness, and blind support for vendors to the detriment of the people is the IRS "e-file" system.

The goal is a very laudable one — simplify and reduce the IRS's costs in processing tax returns by making the entire process electronic. But of course, somebody gets hurt by this: the large number of commercial tax preparation businesses, whose business is driven entirely by the difficulty of preparing taxes to comply with the US's constantly changing income tax laws. As a result, after the government's usual process of working with "stakeholders", we have a completely proprietary e-file system, which locks users into using commercially provided, proprietary tax preparation software, if they want to get the advantages of e-file.

My latest tax return booklet tells me I'm probably eligible to use "free e-file software". However, not only is the software non-free, but it relies on services which demand that you agree to having your personal income tax information used for marketing purposes. It's hard to imagine a more complete disregard for personal privacy! As a result, I still file my taxes on paper, and I'll do so until they get it right.

What's "right"? Probably adopting a standard like the Tax ML standard in design at OASIS, precisely for this purpose. It's hard to imagine any natural reason for the IRS wanting anything else: an XML standard for expressing your tax return would allow for easy automatic verification and cross-checking, and the whole process could be easily and completely automated. This exposes another proprietary lock-in weapon, though: the government is often completely clueless about the importance of free interchange standards for promoting a free market, instead swallowing industry propaganda, which promotes proprietary standards.

The government is often completely clueless about the importance of free interchange standards for promoting a free market, instead swallowing industry propaganda, which promotes proprietary standards

Unlocking the door

From all of these examples, we can see the methods, which have been used to create and promote proprietary formats and the vendor lock-in:

- The oldest trick is the one that initially made the DOC format so difficult to follow: simple obfuscation. Don't document it and change it constantly so that it's hard to reverse-engineer.
- The next twist is to claim a software patent on the format. You can do this because there are essentially no standards for denying software patents, leaving this a complete loose cannon in all sorts of situations — including file formats.
- Next, the use of encryption technologies, ostensibly introduced to protect the user's file data, but implemented in ways that also obscure the format. This wouldn't be much of a problem, except that the DMCA has made it a criminal act to decrypt such a file and reverse-engineer it, under some interpretations of the "circumvention device" language it includes.
- Business-to-business dealings can deal free formats out and closed ones in — just like the consumer audio equipment that can read the hardly-used proprietary WMA format and not the free OGG format.
- Sadly, the government practice of consulting "stakeholders" when making policy, tries to "promote commerce" by essentially trying to do whatever the industry tells them to do, without serious consideration of the *public* stake in the technology.

The first three problems are problems with using proprietary formats — to be avoided by using free formats instead.

In order for this to work, though, the free software community has to ensure that free standards do exist, either through standards bodies, or through well-documented *de facto* commercial standards. We also need to settle on a widely accepted definition of what a free format is — I can suggest our own Texas OSI's definition and the guidelines put forth by the Open Data Format Initiative as starting points for that, but there is no widely accepted "Free Format Definition" to compare with the FSF's Free Software Definition, which is a major stumbling block in promoting free formats.

The last two are bigger problems — the means by which free formats can be shut out: either by corporate cronyism, or by government agency fiat. Greater support for free hardware development may address the absence of good hardware for handling free formatted data, but the government problems can only really be fought with legislation.

Fortunately, free format laws, should be, as has been pointed out by the Free Software Foundation and the Electronic Frontier Foundation, much easier to defend than laws to preferentially support free software. The nature of public documentation makes it clear that transparency, public design accountability, and openness to public audit should be goals for the government of any free nation, and this is the tack to take in promoting free formats for government documents. So far, the proprietary world has yet to come up with a serious answer for that — the value of free formats is simply too obvious to defend against.

We are not alone

In the end, the free software community is still too small to directly make the kind of widespread change that serious adoption of a free format requires. So, in addition to creating viable standards, we also have to promote the use of the formats to the groups that need them most, and are most willing to adopt them: technical, non-developer users. Scientists, mathematicians, and engineers have a long-standing relationship with free-software, so promoting free-formats to them will be preaching to the converted.

More recently, the troubles that librarians face are forcing them to examine the issues related to file-formats more carefully. They don't always realize that this is the right way to frame the problem, and that's what we need to communicate. It's very important to show that the problems are *not* intrinsic to all electronic formats, only to proprietary ones.

Most of all, there needs to be an openness in the members of the free software community, to speak to people from other disciplines about these key issues which are of benefit to us all. That hasn't always been one of the strengths of the community, but it's probably what's required.

The problems are *not* intrinsic to all electronic formats, only to proprietary ones

Once such power-users of file formats are converted to using free formats, they will have a very nonlinear effect on promoting change. After all, the biggest reason people adopt formats is: that the format is what they need to access their favorite information. It's only after a format has come into wide use that it begins to encourage other authors to switch to it. So promoting the idea to technically-capable early adopters just makes sense. But we have to be aware that these users need more than just a new word processing format — technical users have a wide range of data format needs.

Finally, we have to have patience. It won't happen overnight, and most disciplines move more slowly than the computer industry does. But it seems very likely that it will happen. And it must — after all, the inability to share information, driven by superstition and guild secrecy (the first incarnation of so-called intellectual property), is what put the "dark" in "Dark Ages". None of us want to go there, and that's the point we've got to sell as a community.

Bibliography

[1] MARC 21 Standard (http://www.loc.gov/marc/)

[2] FRBR Standard (http://www.ifla.org/VII/s13/wgfrbr/wgfrbr.htm)

[3] Dublin Core Initiative (http://dublincore.org/)

[4] DMCA information (http://anti-dmca.org/)

[5] Public Knowledge (http://www.publicknowledge.org/)

[6] Anansi Spaceworks (http://www.anansispaceworks.com)

[7] Narya Project (http://www.narya.net)

[8] Project Gutenberg (http://www.gutenberg.org/)

[9] OASIS Standards Body (http://www.oasis-open.org)

[10] PNG Standard (http://www.libpng.org/)

[11] Ogg Vorbis (http://www.vorbis.com/)

[12] Ogg Theora (http://www.theora.org/)

[13] Scalable Vector Graphics (http://www.w3.org/Graphics/SVG/)

[14] IGES CAD format (http://www.nist.gov/iges/)

[15] STEP CAD format (http://www.steptools.com/)

[16] Open Cascade library (http://www.opencascade.org/)

[17] PythonCAD project (http://www.pythoncad.org/)

[18] Blender program (http://www.blender.org/)

[19] Texas OSI free format definition (http://open.narya.net/)

[20] Open Data Format Initiative (http://odfi.org/)

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.