OpenDocument, OpenXML and the others

how file formats can be used to favor (or hamper) innovation, active citizenship and really free markets

by Marco Fioretti
http://mfioretti.net
http://digifreedom.net

Seminar Agenda

- *Introduction*
- *Basic concepts and definitions*
- *Format wars in Science, Culture, Industry and Welfare*
- *Formats in your private life*
- *Paper or bits? Pros and cons of digital archives*
- *Characteristics of open file formats*
- *OpenDocument (ODF)*
- *Office Open XML (OOXML)*
- *Public support for open file formats*
- *Conclusions*

Author introduction

Marco Fioretti

Freelance writer, activist and teacher about open digital standards, Free Software, digital technologies and the their relations and impact on education, ethics, civil rights and environmental issues

Author of the Family Guide to Digital Freedom (http://digifreedom.net)

Member of:

OpenDocument Fellowship

Digistan.org

Eleutheros.it

RULE Project

Preamble (2)

Purpose of the seminar

- *In general:*
  - *Explain what file formats really are*
  - *Explain why and how they impact on life and pockets of all citizens*
- *Here at LEM:*
  - *Explain why file formats are relevant in a Laboratory studying "management and corporate strategies, public choice and public policy, innovation and industrial history"*

Introduction to File Formats

What are file formats?

Why are they important?

*How do they support (or limit) the way we **all** learn, cooperate and work?*

The answer starts right here in Pisa... four centuries ago!

Marco Fioretti (marco@digifreedom.net)
http://mfioretti.net
http://digifreedom.net

How are we able to learn from the past?

DIALOGO
DI
GALILEO GALILEI LINCEO
MATEMATICO SOPRAORDINARIO
DELLO STVDIO DI PISA.
E Filosofo, e Matematico primario del
SERENISSIMO
GR.DVCA DI TOSCANA.
Doue ne i congressi di quattro giornate si discorre sopra i due
MASSIMI SISTEMI DEL MONDO
TOLEMAICO, E COPERNICANO;
Proponendo indeterminatamente le ragioni Filosofiche, e Naturali tanto per l'vna, quanto per l'altra parte.
CON PRI VILEGI.
IN FIORENZA, Per Gio: Batista Landini MDCXXXII.
CON LICENZA DE' SVPERIORI.

ca 1600:

the great Pisan scientist Galileo Galilei proves Aristotle was wrong, dropping two balls of different weight from the leaning tower of Pisa. Is this a legend? It doesn't matter.

2009:

*What matters is that, four centuries later, we could still learn physics from the **ORIGINAL** version of his work. Even if it had been never been republished, the first edition is still completely readable*

Marco Fioretti (marco@digifreedom.net)
http://mfioretti.net
http://digifreedom.net

What, exactly, does Galileo's book teach us?

- The Dialogo is not a special case! Every great work of the past, from the Bible to Newton's Principia and from Euclide's Elements to Shakespeare's plays, is equally accessible
- The cultural and economical benefits of this accessibility are obviously enormous
- What is the first, most basic thing that makes this technically possible, and why is it so important?

  (Hint: it is not copyright, or lack thereof)
- Does that thing still happens today?
- If not, what is the price we are paying, and what can we do about it?
- What impact can it have on the economy and job creation?

Basic concepts and definitions

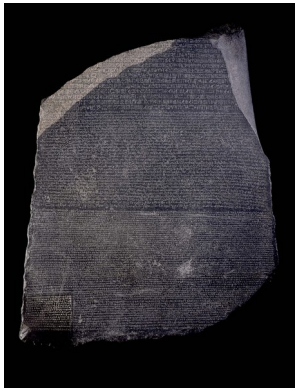
- Q: How do we create, access and preserve information?
- A: Thanks to three very different things:

Physical Support:	the material object containing the information
Data Format:	the rules by which the information is encoded on the support
User Interface:	the tools used to write and read the data according to the format

- ***almost always, Support, Format and Interface can (and should) be independent from each other***

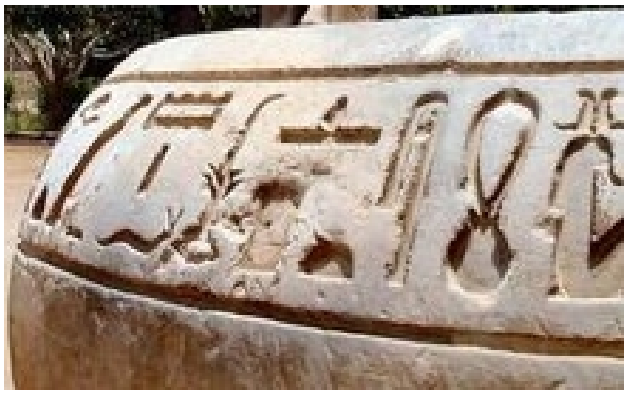
Support, Format, Interface: non electronic example

Support



The Rosetta Stone,
II Century BC

Format



Hyeroglyphs (which could also be written on paper, papyrus, wood...) and the meanings associated to each gliph

Interface

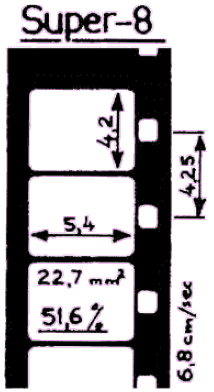
\+

Any manual writing instrument (don't forget quill and charcoal!) and your eye!

Support, Format, Interface: analog electronic example

Support

Format



Interface

Support and format are mixed here: Photograms can only be impressed on a specific type of tape, in a way not usable with other cameras and projectors

Camera and Projector that are useless with any other tape

NOTE: this is the very popular Super 8mm home movie format, released on the market in 1965 by Eastman Kodak, not widely used since the 1980's

Support, Format, Interface: digital, finally!

Support

Hard drives, floppies, CD-ROMs, DVDs, Compact Flash drives... usable with ***different*** hardware

They all contain **the same bits** that can represent wildly different types of information: text, images, audio...

Format

CHARACTER ENCODING:

the meaning associated to each bit sequence:

EX: “01000001” means “A”

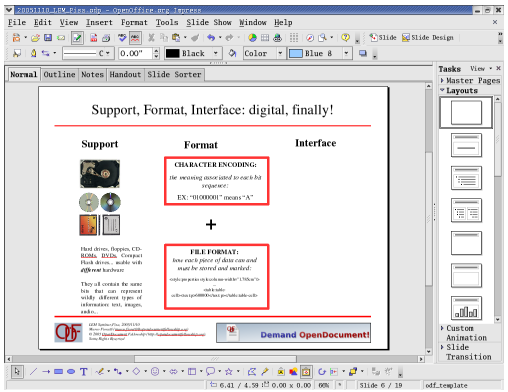
+

FILE FORMAT:

how each piece of data can and must be stored and marked:

<style:properties style:column-width="1.785cm"/>
...
<table:table-cell><text:p>600000</text:p></table:table-cell>

Interface



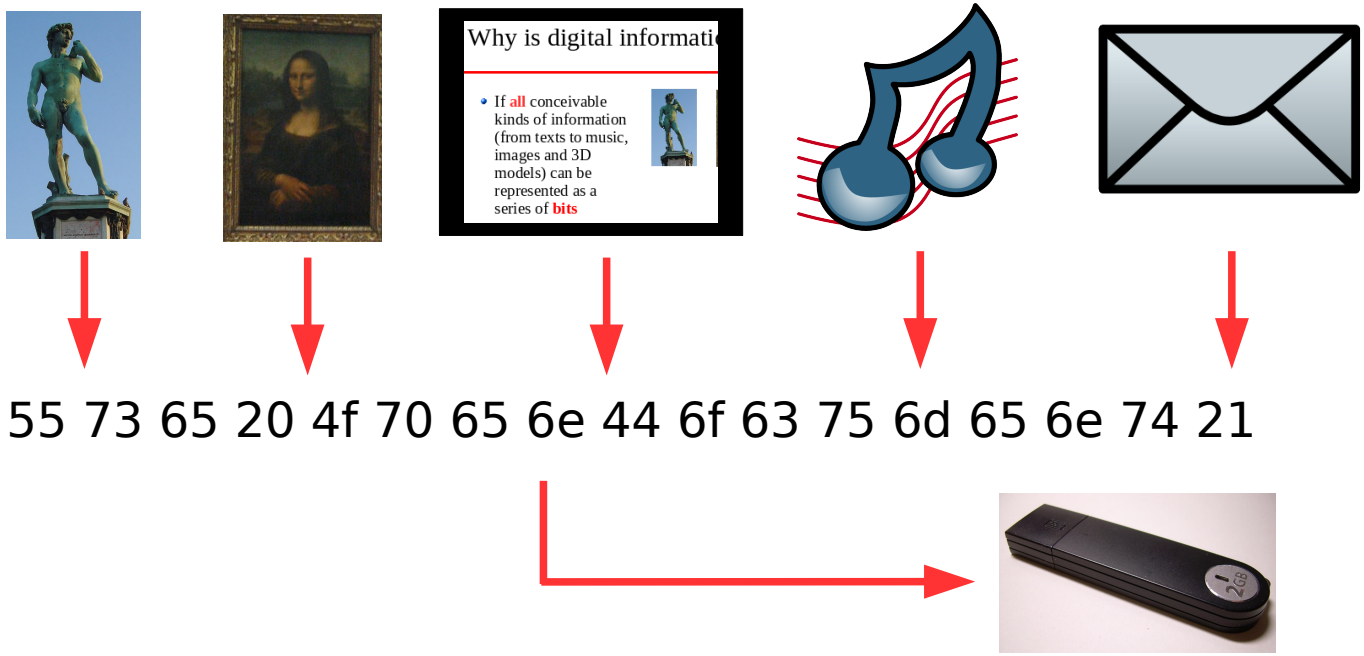
Any software program aware of the file format, **regardless** of :

- the hardware it runs on: x86 or Apple computer, cell phone, DVD player, remote server...
- Its license of use

Marco Fioretti (marco@digifreedom.net)
http://mfioretti.net
http://digifreedom.net

Why is digital information good?

- If **all** conceivable kinds of information (from texts to music, images and 3D models) can be represented as a series of **bits**



- We only need:
  - **ONE** class of generic storage devices: *bit containers* which can change shape and technology without particular problems and are very cheap
  - **ONE** (ok, very large...) class of telecom networks, ie *bit transporters*
- *And all these data can be preserved or distributed with much less money, time and effort than before!*

Why is digital information bad?

- If **all** conceivable kinds of information (from texts to music, images and 3D models) can be represented as **bits sequences stored in bits containers**, we have (at least) two big problems:

- Bit containers are much more fragile than non-digital media: parchment lasts millennia, hard drives a few years
  - This problem has a relatively easy solution (make many copies of information, refresh them frequently) and is outside the scope of this seminar

- The second problem is that, **even when the container works perfectly,** the sequences of bits are absolutely useless if they are locked and we lose the key and cannot buy one:

Why file formats are a powerful weapon

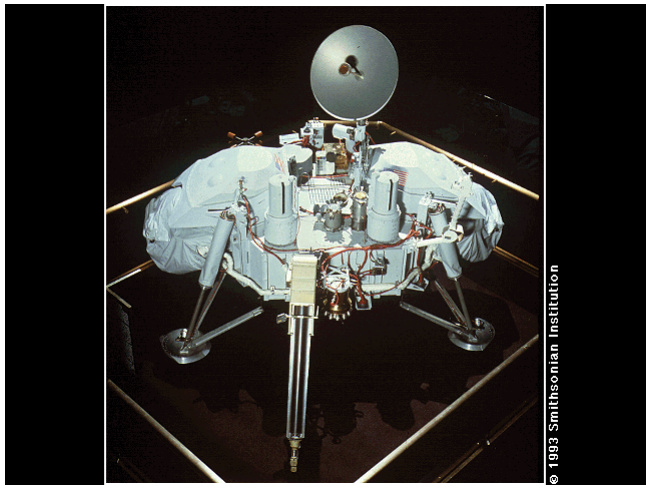
- Almost all software programs are worthless without information to process, store and display. For example, you could use a word processor to write letters or video editing suites to edit footage of your girlfriend at the beach.
- **Never forget** that here “information” means any kind of creative work: blog entries, private movies, essays, government reports, court rulings, road projects, laws, contracts...

- Sometimes, locking the information into some secret format is the easiest way to keep selling copies of a program, without really improving it...
- Sometimes it just happens because of ignorance and carelessness

  ...but in both cases, it can create lots of far-reaching problems

- Let's see some examples...

Format wars: Mars, 1976



July 20th, 1976:

the Viking Lander is the First Spacecraft to Operate on the Surface of Mars

2003

"All the programmers had died or left NASA"

"It was hopeless to try to go back to the original tapes"

"With the data in an unknown format, [it was necessary] to track down printouts and hire students to retype everything"

(www.cbsnews.com/stories/2003/01/21/tech/main537308.shtml)

Format wars: the BBC images, 1986

- In 1986 the British Broadcasting Corporation started a “computer-based collection of photographs, writings and other snapshots of life”
- In 2003 this digital gallery already needed customized software and unreliable hardware
- JPEG and other digital images are degraded every time that they are modified or converted to other formats
- Periodical format conversion has been compared to “preserving a Picasso by repainting it every few years”
- Solutions like ”imitating old platforms to run old software” are only temporary, not scalable patches

SOURCE: www.cbsnews.com/stories/2003/01/21/tech/main537308.shtml

Format wars: Italy, 1992

- **1992:** all the very complex equations for an Electromagnetic Compatibility thesis are written with Expressionist from Maple Soft

- **1999:** Expressionist is sold: *"Maplesoft has sold the MathView and Expressionist product line to Web Primitives, LLC... Technical Support for these products will be provided by WebPrimitives (http://www.livemath.com) effective June 25, 1999"*

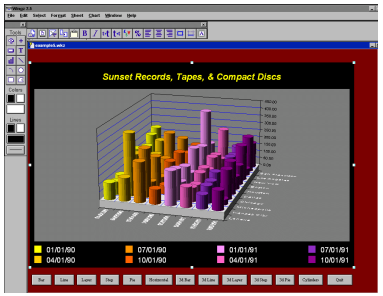
  - **2005:**the equations are still accessible... maybe...:

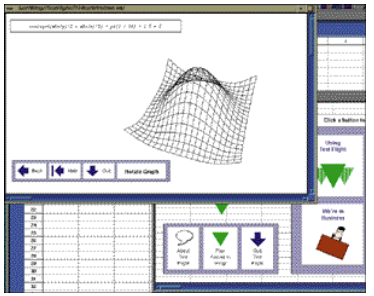
  - From LiveMath support personnel:
    *"MathEQ is presently available for Windows, Mac OSX and OS9 only. I've been **told** it will run under Windows emulation."*

  - **Exercise:** run a poll in your University to discover how many of your current teachers and other personnel is in the same situation today

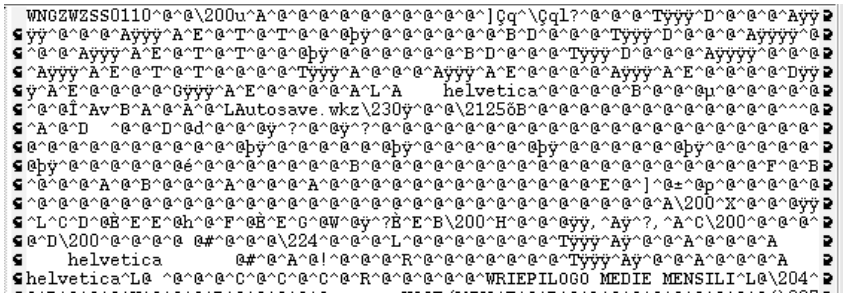
Format wars: Italy, 1994

- **1994:** a spreadsheet is created with the Wingz program (for Solaris and MacIntosh)





- Today that spreadsheet looks like this:



- **2005:** (answer obtained from Usenet Newsgroups)

You should still be able to [download and compile by yourself] a copy of Wingz for Linux or FreeBSD.

BUT: *You may need some pretty old libraries, as I think it predates GLIBC 2, and so almost certainly needs some way old versions of LIBC and Xlib. (Hey, the last binary release dates back to 1996! Lots of things have changed!)...You'd probably have better luck running it on Debian [since it] has somewhat more useful "legacy" support.*

Format wars 2007: time bomb at the UK National Archives

- The UK National Archives, which holds 900 years of written material, has more than 580 terabytes of data in **older file formats that are no longer commercially available**.
- Chief Executive of The National Archives , Natalie Ceeney, said society faced the possibility of "losing years of critical knowledge"
- “some digital documents held by the National Archives had already been lost forever because the programs which could read them no longer existed.”
- “**the issue of older file formats was a bigger problem** than reading outdated forms of media, such as floppy discs of various sizes and punch cards”.
- "We are starting to find an awful lot of cases of what has been lost. What we have got to make sure is that it doesn't get any worse."

- “**The root cause of the problem is the range of proprietorial file formats that proliferated during the early digital revolution**”.

Source: http://news.bbc.co.uk/2/hi/technology/6265976.stm

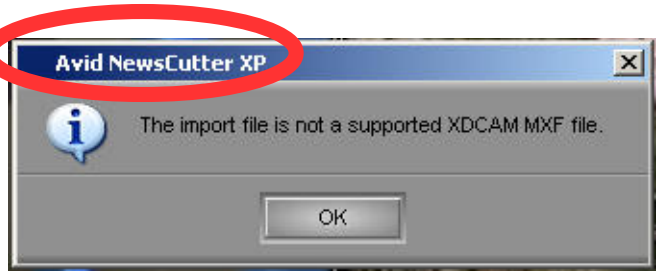
Test: how many other public archives are in the same situation?

Format wars 2008: the most expensive (digital) Black Hole

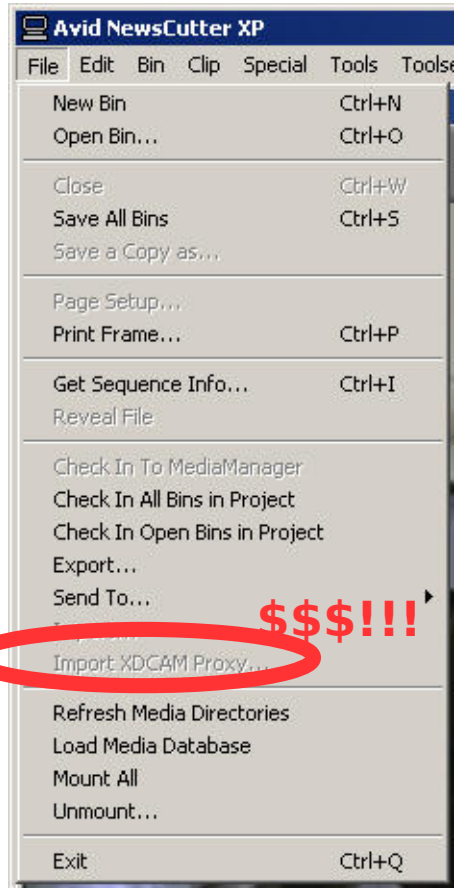
- *"Due to ever-shifting platforms and file formats, much of the data we produce today could eventually fall into a black hole of inaccessibility."*
- How much data? *"at last count, 369 exabytes worth of data, including electronic records, tax files, e-mail, music and photos, for starters. (An exabyte is 1 quintillion bytes; a quintillion is the number 1 followed by 18 zeroes.)"*
- Losing these data would be like burning money *"because we would lose the huge economic investment [governments], libraries and archives have made digitizing materials to make them accessible"*
- *"Software companies have seen the benefits of locking people into a platform and have been very resistant to change"*

Source: interview to Jerome P. McDonough, assistant professor in the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign, October 2008 (http://news.illinois.edu/news/08/1027data.html)

Format wars 2009: behind the curtains of Digital Tv



= Sony?,
No, Sorry!

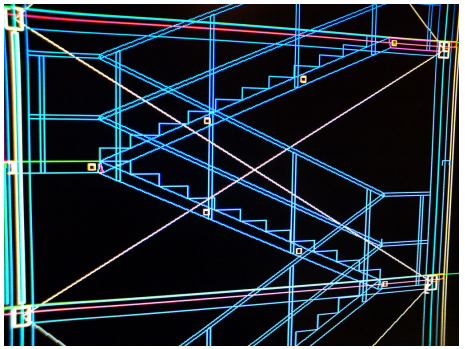


- (synthesis of a conversation with a TV broadcast professional)
- The format used where I work for all video editing is MXF (www.digitalpreservation.gov/formats/fdd/fdd000013.shtml)
- This is a container file format. Inside it there are a video essence, audio tracks and assorted metadata
  - The video essence is always the same (DvCam 4:2:0)
  - Metadata and indexing information is different for every vendor (even inside the same standard!)
    - Specs of the AVID version (OP-1) are available
    - Those of Sony's aren't. Even differents Sony productshave incompatible formats
    - Most video capture equipment is Sony
    - Most video editing software is Avid, which doesn't recognize Sony's MXF unless you purchase an expensive plugin
    - *Things seemed better a couple of years ago, thanks to the OMF initiative.* ***The passage to MXF, which locks our video, happened just when Open Source alternatives were finally becoming viable***

Marco Fioretti (marco@digifreedom.net)
http://mfioretti.net
http://digifreedom.net

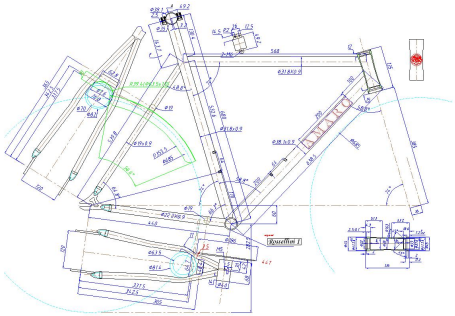
Format wars: Autocad vs Engineering

- Computer Aided Design (CAD) has been THE way to design or model complex any kind of 3D products for decades



- One of the most popular CAD products is Autocad from Autodesk:

AutoCAD is found in 85% of the businesses and schools that design, document and manufacture.. it is used in architecture, interior design, shop fit-outs, construction, engineering, landscape design, product design and manufacture, naval and aeronautical design, piping and cabling...



Source: www.jidaw.com/certarticles/autocadcareer.html and www.wikinvest.com/stock/Autodesk_(ADSK))

Format wars: Autocad vs Engineering (2)

- In the 1990's, more than **two billions private and public projects (mechanical parts, furniture, buildings, bridges...)** were already stored in the DWG file format of AutoCAD
- In 1998, several competitors launched cheaper products based on an equivalent format
- AutoDesk's advertising campaign focused on reminding that only AutoDesk's products were 100% capable of keeping **existing** projects completely accessible
- Besides the cost, the Autodesk software for reading, writing, and displaying DWG files:
  - came under a "selective licensing" program
  - supported the writing of only the most recent versions of AutoCAD DWG
  - had no public specification that would permit independent development

Source: www.opendesign.com/about/whtpaper/whynot.htm from the Open Design Alliance

- Update 2009: *"we have exactly these problems with DWG, and the Open Design Alliance does not help, because their DWG toolkit is not open source, they have restrictive licensing terms and better terms are not affordable for us" Source: Benjamin Ducke, Oxford Archaeology Consultant*

Format wars: Autocad vs Engineering (3)

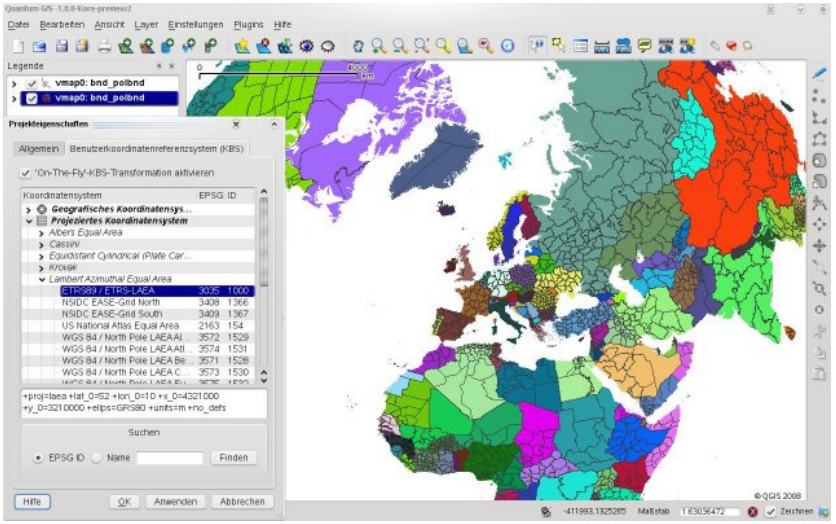
"AutoCAD DWG and DXF formats change with every new release. Both are extremely complex, containing the actual drawings plus the whole layer structure and all attributes; and only AutoDesk's own software can handle them perfectly well. ***This means we can forget about teamwork, sharing or archiving our CAD data (usually archaeological site plans) in the long term.***

This also makes our migration from GIS to CAD extremely hard. Many GIS nowadays offer some level of DXF support, but I have never seen one that could read more than 20-30% of our CAD files".

Source: B. Ducke

Format wars: the GIS world

- **"Who controls the map controls how people perceive the world"** ("Mapping Hacks",O'Reilly)
- Geographic Information System (GIS) are software systems for creation, correlation, display and interactive analysis of geography-related data.
- GIS link generic data to (or through) real places.



- GIS make it much easier and less expensive to:
  - find all the data which share a location, or
  - which objects or statistical phaenomena in a region are closer to which others
  - find how very different, apparently unrelated classes of events influence each other when they *happen close to each other*
- *"[In November 2006], global market size of GIS products arrives at **US$1 billion**. Business revenue brought by GIS-related software, hardware and service reaches **US$ 10 billion each year"**. (www.ccwresearch.com.cn/en/research_text.asp?articleId=13937&Columnid=1008)*

Format wars: the GIS world (2)

- Practically all open source GIS at least save to a plain ASCII or XML file, so it's easy to write converters
- Many organizations (including academic institutions which get big discounts on license prices) demand project files in the proprietary, binary ESRI formats (MXD) from their GIS partners.
- **The result:** *"ESRI has managed to lock that market segment, plus anything that depends on it, into their MXD format. ESRI software can read those files, but offers no freely available conversion facilities for MXD whatsoever, forcing anybody willing to use other software to manually recreate GIS projects, layer by layer"*

- ***ESRI MXD ArcView project files:***
  - the actual GIS data formats, such as ESRI Shapefiles, GeoTIFF etc. are well-documented, but...
  - a GIS project is more than just the plain data: styling map layers, arranging and producing map layouts etc.
  - *inside ESRI MXD project file, all these extra data are saved in a proprietary format*
- ***ESRI Geodatabase:***
  - *"ESRI uses the Microsoft Access MDB container format as a Geodatabase, inaccessible for just about any free GIS out there. ESRI itself is now dropping this format due to its limits (max size 2 gigs!), but for a completely new GeoDB format (of course, proprietary again): all those who converted their Shapefiles to MDB Geodatabases should start all over!"*
    (source: B. Ducke)
- ArcView 3.X saved project files as plain ASCII and the new ArcView 8.X line had built-in support to convert to ArcView 3.X.

"ESRI must have seen that this would provide a bridge for users to break out of the MXD lock and has removed the converter in recent versions..."

(source: B. Ducke)

Format wars: an example from the GIS world

- *Drainage networks and associated drainage basins form complex functional entities not only for hydrological processes but also for environmental processes at large... JRC's Catchment Characterisation and Modelling (CCM) activity responded to this need through the development of a pan-European database of river networks and catchments...*

- *...CCM data are made freely available for non-commercial use at:*

  http://desert.jrc.ec.europa.eu/action/php/index.php?action=view&id=23

- Unfortunately, this very useful, high-quality data created with taxpayer money is locked down in an ArcGIS Personal Geodatabase. Why?

Formats wars in archeology and other research fields

Some reasons why [archeologists] should be enthusiastic about FOSS

...

Archivability: Stable and long-lived data formats; free and open standards instead of "industry" standards; no pressure to deprecate older software or data formats.

Problems with commercially licensed software [in our company]

...

Hard to archive: a new file format every year means constant pressure to upgrade and convert

Source:

Benjamin Ducke, Senior Geospatial Consultant at Oxford Archaeology (oadigital.net)
"Use of GIS in Archaeological Settlement Research - Facts, Problems and Challenges"
Workshop of the Romano-Germanic Commission of the German Archaeological Institute, Frankfurt, September 26th 2008
ftp://88.208.250.116/ducke-frankfurt-foss-gis-arch.pdf

(nuclear) Format wars: Great Britain, 2005

- The UK Atomic Energy Authority (UKAEA) has begun an £8bn project to dismantle 26 atomic reactors used for research, and bury the waste in concrete bunkers
- It is necessary to preserve information about the bunkers and how to manage them for all the time that those waste will be dangerous: millennia
- the UKAEA, assuming that the software and hardware used to store any supporting documentation will not last that long, was forced to look for a better alternative
- (note that they confused software, hardware and... formats all together...)
- So they choose the closest thing we've got to... Egyptian Papyrus. Permanent paper is equally acid free, so won't discolour, or rot over the years

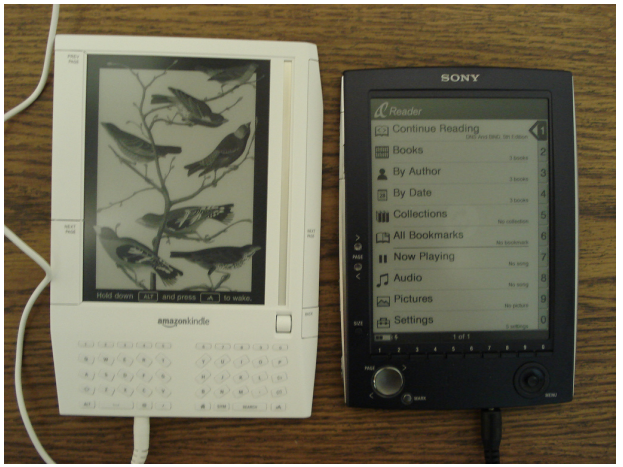
- Conclusion: 423 documents for a total of 11,718 sheets of paper - will be kept in special copper-impregnated bags, and packed in long-life archive boxes
- So much for the benefits of the Digital Age...

SOURCE: www.theregister.co.uk/2005/08/09/papyrus_nuclear_waste/

Format wars: can you read your own e-books?

- Traditional book formats: only one, that is paper, ink and alphabet
- Number of e-book formats, as of 2009/02/10, according to http://wiki.mobileread.com/wiki/E-book_formats: **about ~40!**
  - Many popular ones, like LRF (Sony) or Amazon (AZW) are proprietary
- There are open formats like OPF (Open Publication Format) but:

  "[they are] **not meant to be delivered to the reading device of the customer**. The OPF publication must first be compiled into a binary eBook." (www.mobipocket.com/dev/article.asp?BaseFolder=prcgen&File=mobiformat.htm)



The Amazon Kindle and Sony e-book readers

- Being forced to worry about compatibility between a book and your way to read it is like having to worry if the paper books on your bookshelf will still be readable if you change your brand of glasses

- Why does this happen? Because...

  "The Consumer Electronics Association estimates that 538,000 e-readers were shipped in 2008, reflecting $154 million in revenues and 235% growth over 2007... Citigroup analysts... project Kindle-related revenues to reach $1.2 billion in 2010" (http://money.cnn.com/2009/02/06/technology/ebooks.fortune/index.htm?postversion=2009020612)

December 2008: Goodbye Mr President

- December 2008: The National Archives has put into effect an emergency plan to handle 100 terabytes of electronic records from the Bush White House, that is 50 times as much as was left by Clinton in 2001
- The collection will include top-secret e-mail tracing plans for the Iraq.
- The transition poses “unique challenges” because of the huge volume of electronic records, **some of them in “formats not previously dealt with.”**
- The archives said it had “a high level of confidence” that it could bring the e-mail into its electronic record-keeping system and retrieve messages in response to requests from Congress and the courts.
- Thomas S. Blanton, director of the nonprofit National Security Archive, a plaintiff in several lawsuits seeking Bush administration records, said the National Archives’ track record did not justify such a claim. “Their confidence is inexplicable,” Mr. Blanton said.

Source: www.nytimes.com/2008/12/27/washington/27archives.html

Format wars: our public archives

- *From: Virginia State Laws on Optical Images (www.archiveindex.com/laws/law-va.htm):*
  - About "Public Records For Permanent Retention":
    - *Electronic records are not acceptable for permanent storage at this time. The Library of Virginia can not be responsible for maintaining the necessary hardware and software necessary*
    - *The Library of Virginia* ***cannot accept records for permanent storage on digital media*** *at this time due to the lack of hardware and software standards*
    - *Electronic records identified as permanent... must be converted to archival quality microfilm or alkaline paper before being transferred to the Library*
- **But this is much more expensive and means no electronic indexing and no other benefit of the digital age (like Internet access to data, copies at almost null cost...)**
- ...and without any guarantee that microfilms will be readable, without preserving special hardware

Format wars: did you have heart surgery or not?

- Electronic Health Records (EHRs) are patient's medical history in digital format, which help doctors to make the best decision for your health in the fastest and cheapest way possible.
- 2008: The administration proposes to invest $10 billion a year for five years in order to "move the U.S. health care system to broad adoption of standards-based electronic health information systems, including EHR." Success in this endeavor could save up to $77 billion in annual health care costs” (from Barack Obama's Obama technology and innovation platform)
- But only if all invoved data formats are all 100% open!
- Main involved organizations:
  - Health Information Technology Standards Panel (www.ansi.org/hitsp/)
  - Health Level 7 (HL7, www.hl7.org/)
    - European Institute for Health Records (www.eurorec.org)
      - openEHR Foundation (www.openehr.org/home.html)

Source:www.consortiuminfo.org/bulletins/dec08.php

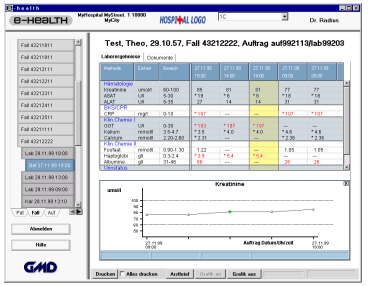


Image Sources:
www.infocamere.it/osservatorio.htm,
www.infocamere.it/doc/denardi.ppt

Formas in your life: I shot a wonderful movie, but where is it?

- The .TOD video file format was created by JVC, to be used by their high-definition camcorder range (I.E JVC Everio).
- Comments from several users, October 2008:
  - *#1 HELP! I'm desperately trying to find a way to convert .TOD files so I can edit them in adobe premiere pro 1.5, but nothing seems to work. Renaming files doesn't work, neither converting them to .AVI ...*
  - *#2 How can I convert my video to something more compatible with common video editing applications?*

Source: http://dotwhat.net/tod/8934/

- What about the Panasonic SDR-H60?
- January 2009:*"The camcorder itself worked well, [but]* ***records in some funky proprietary format that is not compatible with software editing programs other than the clunky one it comes with****... In the end, I returned the camcorder... and plan to purchase a different make/model.*

Source: http://reviews.cnet.com/digital-camcorders/panasonic-sdr-h60/4864-6500_7-32815213-2.html

Formats in your life: Smartphones and PDAs

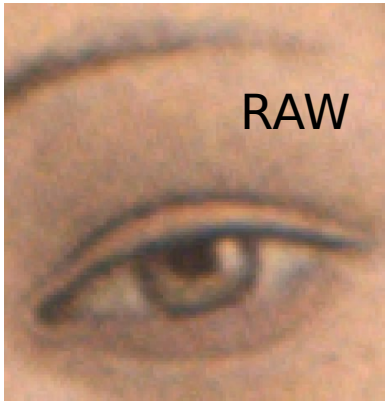
- *The .IPD (Inter@ctive Pager Backup) file contains Blackberry backup data, be it your emails, calendar or even the configuration data.*
- *This format was designed for backups and restores only, and was not intended to be accessed [directly]... at the time of writing, the only software application that will let you do this is ProcessText Group's "ABC Amber BlackBerry Converter" which is free to try for 30 days and then $19.95 to purchase a single user licence.*

- **Effects:**
  - User #1: "Blackberry file copied in .ipd. This was listing of all phone numbers. I retired from company. I have on CD and can not open with home computer"
  - User #2: "how do i open .ipd files? trying to view address book, etc"
  - User #3: "I backed up my blackberry with BB Deskyop Managr and it is in a .ipd file and I can not open it. Is there a way to open and view this file? Please help."
  - Many other similar questions got no answers

Source: http://dotwhat.net/ipd/853/

Formats in your life: look at my pictures! No, wait...

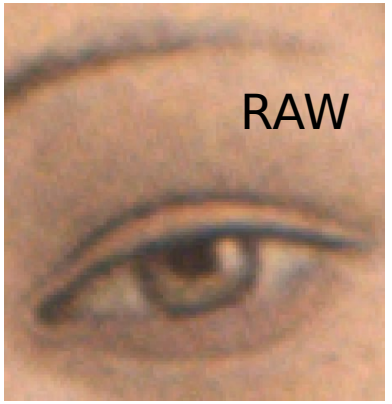
- JPEG is an open file format for digital pictures: ubiquitous, but slightly degraded, due to compression. Editing a JPEG file degrades it even more.
- Many digital cameras can save pictures in their *raw* format: a complete, untranslated copy to file of all the light signals (color, intensity etc...) which hit every single pixel of the camera sensor
  - Raw is the format which gives the highest possible image quality, much better than JPEG
  - Raw is to JPEG what word processing files are to PDF: if you want to obtain a different PDFversion without any degradation, you must have an OpenDocument original to modify, and generate a new PDF from that
  - Archiving pictures in raw format gives the highest guarantees to be able to process them in new ways in the future



Picture samples from www.oss-watch.ac.uk/resources/photo-files.xml

Formats in your life: look at my pictures! No, wait... (2)

- The risks of raw formats:
  - Hundreds of raw formats already exist
  - Some of them were at least partially encrypted
  - *"No one can predict how long a particular RAW file will be supported by a camera manufacturer (not even the company itself)"* (www.openraw.org/info/)



- The solution?
  - *"Open documentation of all RAW file formats by manufacturers is the quickest and most satisfactory way for OpenRAW's goals to be reached"*, (www.openraw.org/info/)
  - The Open Source dcraw format decoder at www.cybercom.net/~dcoffin/dcraw/
  - The Adobe Digital Negative standard (www.adobe.com/products/dng/)? Cfr "DNG is not the answer" at http://www.openraw.org/node/1482

Formats in your life: you got mail, can you also read it?

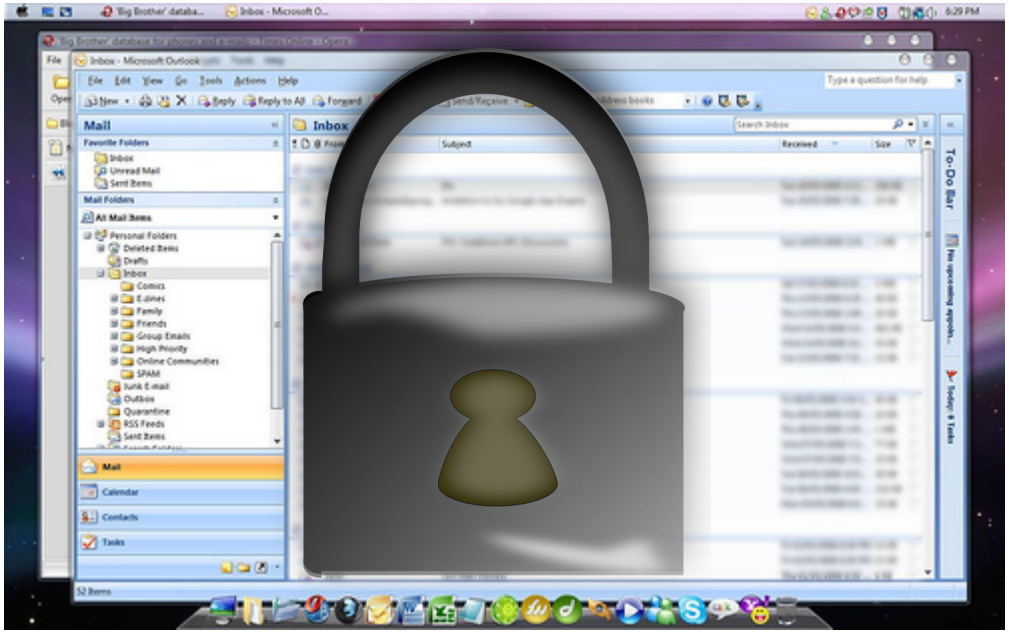
- The .pst (Personal Storage Table) file format used by Microsoft Outlook is proprietary.

(and so is the dbx format used by Outlook Express)

  - It is “a database file format”, recording also searches and sort procedures.

  - *“These design decisions reflect the core architecture of Outlook and Messaging Application Programming Interface (MAPI)”*.

(http://support.microsoft.com/kb/269520)



- Ways to use it with other software exist, but why go for unnecessary complications?

Formats in your life 2.0: Social Networking

- **Test:** all those who know what is the format in which Facebook stores **their own data about their own lives...** please raise your hands!

- *"Mark Zuckerberg has committed Facebook to opening up its data... But we have no timetable..."*

- In the meantime:
  - Join the group "Are there 100,000 people for open data in Facebook?"

    www.facebook.com/group.php?gid=7935290927
  - Take your Facebook Data with you! www.chrisfinke.com/2008/01/03/take-your-facebook-data-with-you/
  - Read about Social Network Data Portability

    http://mastersofmedia.hum.uva.nl/2008/12/06/facebook-connect-openid-the-format-war-for-your-identity/
  - Check out the status of any other social network or web service you use

The costs of not using digital documents

- June 2007, Italy:
- “the amount of costs and suboptimal procedures related to management of paper documents is around 3 and 5% of GNP, with an end impact on the national system which can be estimated from 42 to 70 billions of Euro”
- “adoption of digital documents makes it possible to save from 50 to 90%, depending on the service”
- Sources:
  www.cameradicommercio.it/cdc/id_pagina/26/id_tema/x/id_cp/11/id_ui/1432/id_prov/x/id_ateco/x/t_p/Osservatorio-permanente-sul-documento-digitale.htm
  www.infocamere.it/osservatorio.htm

The cost of *preserving* digital documents

- From the report “Comparing Preservation Strategies and Practices for Electronic Records, 2000” *(www.rlg.org/en/page.php?Page_ID=245)*
  - *DEFINITION: Preservation of digital information is not so much about protecting physical objects as about specifying the creation and maintenance of intangible electronic files whose intellectual integrity is their primary characteristic.*

- The study was conducted on behalf of the Preservation Task Force of the International Research on Permanent Authentic Records in Electronic Systems (InterPARES Project)
  - Goal: find the cost for institutions to preserve, maintain, and access electronic records
  - Only a few [organizations] were far enough along to have developed cost figures
  - The interviewees ranged from large national archives, to projects developing testbeds
  - The range of [total] costs for electronic record preservation is from $10,000 - $2.6 million per year [per organization...]

How long is a long term?

- How much time does an organization have to detect and fix any format-related problem that may cause data loss?
  - Until now, even if many people don't realize it, this time has been scaringly short:
    - **“What does 'long-term' mean in the context of Digital Preservation? 1. "five years or more"!!!!**

      (www.digitalpreservationeurope.eu/what-is-digital-preservation/index.php#07)

    - **“Digital information lasts forever—or five years, whichever comes first”**

      (Jeff Rothenberg, 1995, in “Ensuring the Longevity of Digital Information”,www.clir.org/pubs/archives/ensuring.pdf)
- This mainly happened because of ignorance among end users, starting from Public Administrations
- As the previous examples prove, we have only started to realize the true cost of such ignorance, and many data may be already lost forever
- Luckily, today we have both the consciousness and the tools to stop repeating these errors

Are open file formats against copyright?

- Long answer: NO!
  - Yes, today, formats (mixed with patents and bogus “encryption” schemes) are still used, in some cases, to restrict access and copying
  - but open formats are NOT an attack to content creators, or to copyright in general. This is a different problem!
- **Demonstration:** Have you noticed that all the first victims in all the Format Wars that we have just mentioned... are the copyright holders?
- Discussing copyright only makes sense if, initially, the content was under full control ol of its author, or whoever paid for its production. If your information can be fully accessed only with one program by one single vendor, it's not really yours...

...and any further discussion on who “holds” the “right to copy” is just meaningless!

When is a file format standard really open?

"Minimal characteristics that a specification and its attendant documents must have in order to be considered an open standard:

- *The standard is **adopted and will be maintained by a not-for-profit organization**, and its ongoing development occurs on the basis of an open decision-making procedure available to all interested parties (consensus or majority decision etc.).*
- *The standard has been published and **the standard specification document is available either freely** or at a nominal charge. It must be permissible to all to copy, distribute and use it for no fee or at a nominal fee.*
- ***The intellectual property** – i.e. patents possibly present – of (parts of) the standard is made **irrevocably** available on a royalty-free basis.*
- ***There are no constraints on the re-use of the standard**"*

SOURCE: European Interoperability Framework for Pan-European eGovernment Services v. 1.0, November, 2004 (http://xml.coverpages.org/IDA-EIF-Final10.pdf)

What is XML, and why is it important?

- In binary formats, all information is stored in more or less arbitrary bit sequences
  - Minimum size, maximum reverse engineering difficulty

XML = eXtensible Markup Language

- Designed to make it easy to exchange information rather than locking it.
- XML files are in a plain (Unicode!) text format similar to HTML
- Takes more space and CPU time to be stored and process than binary data, but:
  - Compressing the final file recovers disc space and bandwidth
  - (Above all) has still huge advantages for developers and, ultimately, end users
- Reverse-engineering is much easier, compared with binary formats: data can be generated or processed with any existing text-processing tools, starting from those known and improved on since the '70s
- By itself, an XML is no more or less proprietary or open than binary formats.
- Its full benefits are only available when it's open in the sense previously explained

What is OpenDocument?

- An XML-based file format specification for texts, presentation and spreadsheets
- The solution to the problem created by Microsoft's .doc, .ppt and .xls formats (including OOXml)
- Developed to solve, in the office documents space, all the problems listed so far
- Formally approved as standard ISO/IEC 26300:2006

From the end users point of view, be they individual or whole nations, this guarantees that OpenDocument is not a one man, thoroughly undocumented hack that may disappear overnight.

- Already usable on all operating system with OpenOffice.org and many other programs

(cfr *www.opendocumentfellowship.com/applications*)

- Freely usable by anyone, without patents, royalties or other restrictions

(*www.oasis-open.org/committees/office/ipr.php)*

OpenDocument format internals

- Every OpenDocument file is simply a compressed Zip folder containing various elements
- Some of them are listed here:

content.xml

the actual textual content of the document.Complex XML markup, but still readable by humans

meta.xml

Metadata like Author name, Word count, Language, Date of last modification, etc

styles.xml

Style information like like font size, colour, page widthfor pages, characters, paragraphs...

Separate folders for binary objects

•Images
•Macros
•....

- More detailed overview: *http://opendocumentfellowship.org/Articles/IntroductionToTheFormatInternals*
- Official Specification: *www.oasis-open.org/committees/download.php/12572/OpenDocument-v1.0-os.pdf*

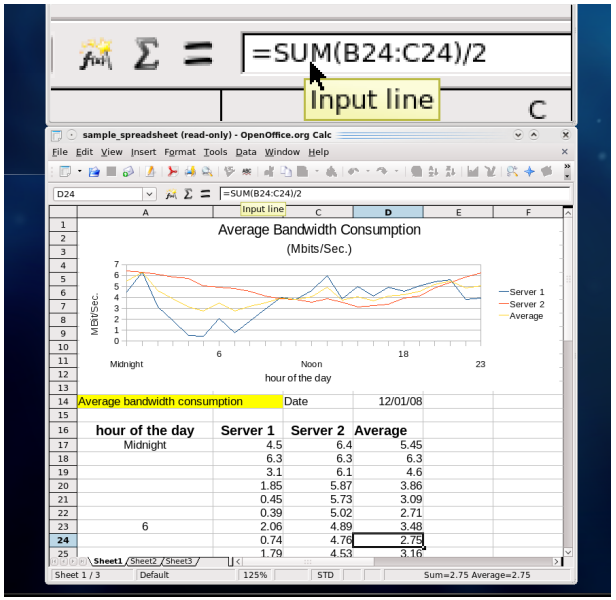
OpenDocument advantages

- Storing all content and metadata as plain (UNICODE) text guarantees full readability, to the point that files could be edited and restored manually. No locks or secrets!
- The ZIP archive structure guarantees complete and easy separation of:
  - Content
  - Visual presentation
  - User configuration
  - Binary objects
- Therefore, each of these elements can be automatically generated, modified, indexed or (for binary objects) scanned for viruses in a very efficient and clean way
- Existing standards reused whenever possible, to speed up development and adoption
- Above all: complete specification available and usable **without any restriction**

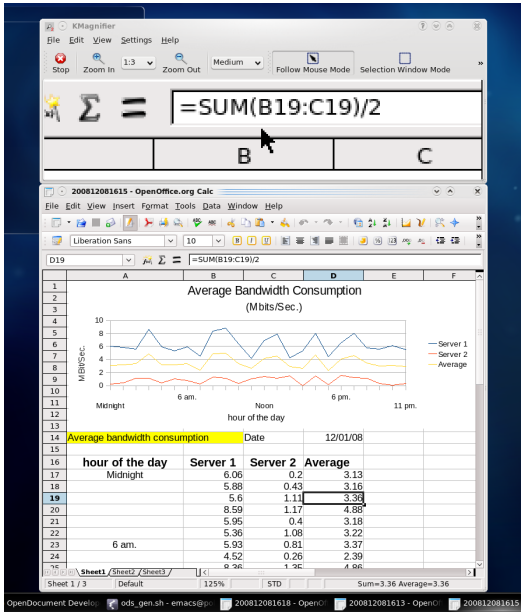
What is XML, and why is it important? (2)

- Thesis: *"Reverse-engineering is much easier, compared with binary formats: data can be generated or processed with any existing text-processing tools, starting from those known and improved on since the '70s"*

- Proof: creating different versions of the same, dynamic OpenDocument spreadsheet, with different data but automatically updated formulas and graphs, takes less than 100 lines of scripting, very little study of the format itself, no sw license fees.



Why update files manually, when the right file format lets your computer do it for you?



Marco Fioretti (marco@digifreedom.net)
http://mfioretti.net
http://digifreedom.net

Why not use PDF or RTF?

- **PDF**
  - It is a read-only picture of the printed version of a document
  - Made for presentation, not for structured, editable information
  - Many metadata (revision info, formulas...) are lost
- **RTF**
  - Belongs to Microsoft, that can change it (or its license) without notice or permission
  - While editable, it is not XML, so it cannot be processed and analyzed as efficiently as the latter

Could OpenDocument be used also by Microsoft?

- Long answer: yes, why not?
  - Remember: formats are (must be) independent from the software interfaces used to create and access information
  - The more ubiquitous OpenDocument becomes, the safer your data are!
- Can ODF be used in Microsoft Office today? Yes (I think):
  - May 2008: "Microsoft Expands List of Formats Supported in Microsoft Office" (www.microsoft.com/presspass/press/2008/may08/05-21ExpandedFormatsPR.mspx)
  - Sun ODF Plugin for Microsoft Office at www.sun.com/software/star/odf_plugin/
  - OpenXML/ODF Translator at http://odf-converter.sourceforge.net
- In practice, it depends on the type of file (text, presentations, spreadsheets) and how many other pieces of proprietary technology end up into an ODF file (macros?)

The direct competitor: Microsoft Office OpenXML

- ZIP-compressed archive of a number of files, describing content, metadata, styles, etc
- The separation into different files is not as clean as with OpenDocument
- Components are separated and linked according to Microsoft's Open Packaging Convention: another proprietary standard you must license in order to use the format
- OOXML Hyperlink example (the hyperlink value is stored in a completely separate file):
  - <w:hyperlink w:rel="rId1" w:history="1"><w:r><w:t>This is a hyperlink</w:t></w:r></w:hyperlink>
- OpenDocument equivalent (hyperlink is immediately accessible):
  - <text:a xlink:type="simple" xlink:href="http://example.com">This is a hyperlink</text:a>

Microsoft Office OpenXML (2)

- OOXML is so complex (+6000 pages) that it would take years to support it completely in any other application different from Microsoft Office:
  - *"OOXML... has been narrowly crafted to accommodate a single vendor's applications. Its extreme length... stems from it having detailed every wart of MS Office in an* ***inextensible, inflexible manner****. This is not a specification; this is a DNA sequence"*
  - *"The whole job [of creating an OOXML converter from scratch for Mac Word] is* ***roughly 20 man-years of development time****. That doesn't include testing, documentation, or localization. That would probably double the number of man-years, at least. But... is just for Word. We need additional teams for Excel and PowerPoint".*

Sources: www.robweir.com/blog/2006/01/how-to-hire-guillaume-portes.html

http://blogs.adobe.com/shebanation/2006/12/open_xml_one-way.html

Microsoft Office OpenXML (3)

- In 2008, even OOXML has been accepted as ISO standard (a mandatory requirement in government tenders):
  - Why approve two standards for the same thing?
  - The ratification process has raised objections and critiques never seen before in ISO
    - "**an unprecedented twenty countries have responded during the contradictions phase - most or all lodging formal contradiction**s with Joint Technical Committee 1 (JTC), the ISO/IEC body that is managing the Fast Track process under which OOXML (now Ecma 376) has been submitted. This may not only be the largest number of countries that have ever submitted contradictions in the ISO/IEC process, but nineteen responses is greater than the total number of national bodies that often bother to vote on a proposed standard at all."
    - "During the Ballot Resolution Meeting, most of ECMA's responses to problems in the specification were approved collectively without discussion".
    - "India, Brazil, South Africa and Venezuela filed for an appeal, citing insufficient review time and procedural irregularities".
- Sources:

www.consortiuminfo.org/standardsblog/article.php?story=20070206145620473

http://arstechnica.com/old/content/2008/05/ooxml-revolt-brewing-three-countries-appeal-iso-approval.ars

http://arstechnica.com/news.ars/post/20080302-xml-spec-editor-ooxml-iso-process-is-unadulterated-bs.html

http://arstechnica.com/old/content/2008/10/norwegian-standards-body-implodes-over-ooxml-controversy.ars

Microsoft Office OpenXML (4): why bother?

- Can all competitors use Microsoft's 6000 pages format? Maybe:
  - *Microsoft XML uses proprietary XML schemas*
  - *Only the license to use the software necessary to implement them is specified*
  - *Microsoft remains free to require a separate license to use the schemas.*
  - *License requirement might preclude any program that uses the file formats from being used in open-source software*
- *Such discrimination is unacceptable, against customer interest and the policies of many Public Administrations*
- *There is no commitment to delivering any future changes to the schemas or right to develop software implementing them under the same or more liberal license*

Sources: www.groklaw.net/staticpages/index.php?page=20050331183622861#A4, www.eweek.com/article2/0,1759,1829728,00.asp

- Above all, why bother? **"ODF has clearly won", says Stuart McKee, Microsoft Technology Officer,** www.infoworld.com/article/08/06/19/Red_Hat_Summit_panel_Who_won_OOXML_battle_1.html

When is OOXML acceptable?

- Today, given the present status of things:
  - there is no reason to ever save or distribute in MS Office 2007 or OOXML formats:
    - *completely new documents or*
    - *newer, modified versions of existing documents*
  - Please do **not** use such formats and demand that your schools and Public Administrations **never** use, require or accept them
  - Things may be different when it comes to preserving *already existing* documents that:
    - Are only available in older, undocumented Microsoft binary formats
    - **Must** remain **completely** (metadata, formulas, revision info...) accessible in the future, without any loss, even when no current software will support the original format anymore
    - Are only needed as reference, not for distribution
  - These are all cases where OOXML is, very likely, a better solution than ODF

Is it enough to say “OpenDocument”? Maybe not

- ODF and other “container” type formats (cfr MXF) have one “weakness”:
  - They can contain many different types of objects:
    - Macros, metadata, images and other multimedia content, digital signatures...
    - ...which may very well be proprietary!
  - A file format specification could never, and must not, put limits on the characteristics of any object that may be *embedded* in the main file
  - Especially when the discriminating factor is a *legal*, not technical one!
  - **The only possible solution is at another level, through “OpenFile” registered marks and/or policies**

To know more, please read http://robertogaloppini.net/2007/04/01/file-format-hidden-traps-in-opendocument-or-any-other-open-standard-and-how-to-avoid-them/

Format Wars 2007: meanwhile, China...

- While Microsoft tries to fight OpenDocument with OOXML, China is developing its own home-grown open document format standard, called UOF (Unified Office Format):

- *“ China has embarked upon a very aggressive state-sponsored, and state funded, standards strategy...*
- *...China has evolved a two-pronged approach: use western standards with abandon, when they can be implemented for free, and develop its own standards in select, high volume areas, such as wireless technology, 3G telephones – and office suite formats - when they can't. UOF was adopted as a Chinese National Standard in May of this year, and implementers will need to obtain a license to multiple Chinese patents in order to implement it”.*

Source: http://consortiuminfo.org/standardsblog/article.php?story=20070817070419313

Open formats for Open States: some examples

- **Norway**: "Proprietary formats will no longer be acceptable in communication between citizens and government" (*www.vnunet.com/vnunet/news/2138935/norway-government-open-source)*
- **Extremadura, Spain**: “ODF will be the standard format in public administration and in schools”
- **Massachusetts, USA**: “As of January 2007, all software procured by the executive branch of the U.S. Commonwealth of Massachusetts must be able to read and write OpenDocument files. In addition, all internal documents will be stored in OpenDocument format”
- **European Union**: “Dr. Barbara Held, Enterprise and Industry Directorate-General of the European Commission Program... stated, "In the view of the European administrations and Member States, the ODF standard is at the very top of the pile by far from all other proposed open standards”
- For details and many, many more cases, please read:
  - www.odfalliance.org/resources/Annual-Report-ODF-2008.pdf
  - www.opendocumentfellowship.com/government/precedent

Open formats for Open States: Italy

- **2002:**
  - Sen. Fiorello Cortiana presents a law proposal mandating that all public administration only use Free software and publish their documents only in non proprietary formats (www.senato.it/leg/14/BGT/Schede/Ddliter/16976.htm)
  - On. Pietro Folena proposes similar regulations to promote IT pluralism and incentivate the diffusion of Free Software (www.senato.it/leg/14/BGT/Schede/Ddliter/17207.htm)
- **2003:**
  - Minister for Innovation and Technologies Stanca announces a directive stating that public Administrations shall privilege IT solutions which, among other things, can export data and document in at least one open format

    www.innovazione.gov.it/ita/comunicati/2003_10_29.shtm
- **2007:** ODF officially ratified by UNI (Italian Unification Organization)

Marco Fioretti (marco@digifreedom.net)
http://mfioretti.net
http://digifreedom.net

Open formats for Open States: Italy (2)

- **Emilia Romagna regional law n. 11, May 24th 2004:**

  Art. 5: "In order to guarantee to all citizens the greatest freedom to access public information, the Region actively promotes the usage of electronic file formats and databases in non proprietary formats"

- **Toscana regional law n. 1, January 26th 2004:**

  Art. 4: "[The principia and guidelines for all e-governments actions include] the usage, with respect to information which must be made publicly available, of open standards and file formats"

  (www.cesda.it/quadernidae/pdf/Pietrangelo_DAE2004.pdf)

- **Province of Pisa, November 7th, 2005:**

  "The Province of Pisa decided to implement an office automation system able to convert its documents to open formats, in order to allow its employees and also other Administrations and end users to convert their files in open formats"
  (www.salpa.pisa.it/salpa/cda/templates/detail_it.jsp?OTYPE_ID=2101&ID=220068)

- **Veneto Regional Law, December 2008:**

  "this law declares that no technical or legal (patents, licenses or trademarks) restriction on usage of digital archives admissible " (http://lucamenini.wordpress.com/2008/12/03/la-regione-veneto-approva-una-legge-per-il-pluralismo-informatico/)

Conclusions: first, some slogans

- Technology is legislation (uncertain source)
- Your own civil rights and the quality of your own life heavily depend on how software is used **around** you (M. Fioretti at http://digifreedom.net)
- Open formats make history - and maintain it

(G. Markham, http://business.timesonline.co.uk/article/0,,9075-1831039,00.html)

- If computer programs are pens, then think of file formats as alphabets... OpenDocument is the digital version of our alphabets.

(M. Fioretti, *Everybody's Guide to OpenDocument*, www.linuxjournal.com/article/8616*)*

Conclusions: what have we learned?

- Very often, data are the only reason to use software, not viceversa. Software comes **after** data and is at their service. The reverse can only give troubles
- Software producers may certainly claim rights on **their** programs, but there is no doubt that the data and metadata we produce or manage with those programs are only **ours**!
- Until now, file formats **have** been used by software producers to grant unfair advantages for themselves over their competitors and keep prices high without real justification
- There are at least three explicit examples of this practice in this seminar:
  - the Autocad advertising campaign
  - The two declarations (about MXF and ESRI files) that plain text formats and conversion filters were removed, or incompatible metadata were added, **only** when Free/Open Source Software had started to become a viable alternative
- Incidentally, what better proof than this behavior could you find that, in those field, FOSS may be a viable alternative today if it weren't for data already locked in proprietary formats?

Conclusions: what have we learned?

- The only way to guarantee that our data remain ours is to store them in file formats which are independent from any single software product
- In and by itself, Free/Open Source software is **not** a solution: many files in the examples above are lost not because of software licenses, but simply because:
    - Programmers didn't bother to leave any format documentation
    - End users didn't bother to demand it
- Only formats which are not only “Free as in Freedom” but also fully documented and officially maintained by a reliable, not-for-profit organization give real guarantees

Conclusions: what else have we learned?

- In many cases, the less formats there are for one specific task, the better.
  - Any conversion carries risks of data loss or degradation
  - This, too, is true regardless of software licenses: why should two programs which must not compete for market share represent the same data in different ways? The fact that you *can* write converters among Free Sw programs doesn't mean that it is a smart thing to create such a need (cfr http://digifreedom.net/node/56)
- Why is it so important to demand that file formats are managed properly?
  - Because it's a terribly serious problem, which has **already wasted huge amounts of public money** and damages public culture, services and education as well as our private lives
  - Because, unlike pension systems, health care, public education or transportation, it has a solution which is much quicker and cheaper to define and implement, so it makes much less sense to wait.

Conclusions: personal and public best practices

- Never, ever use a file format unless you are sure that you can reuse its content with other software
- Never buy goods and services which put your data in digital formats you can't fully reuse with any software outside of those goods or services
- Don't be fooled by shiny new software features, if they impact on formats: *"as long as innovation only impacts on software features, problems are unlikely, but saving **our** data in a new, scarcely documented format which only a few programs can read can be a real disaster!"* (Stefano Costa, Open Archeology, www.iosa.it)
- Demand that your Public Administrations do all of the above!

Useful Resources

- OpenDocument Fellowship
  www.opendocumentfellowship.com
- ODF vs OOXML: War of the Words
  www.consortiuminfo.org/standardsblog/index.php?topic=20071125145019553
- Digital Preservation Tutorial
  www.icpsr.umich.edu/dpm/dpm-eng/oldmedia/obsolescence1.html
- Sustainability of Digital Formats – Format Descriptions
  www.digitalpreservation.gov/formats/fdd/descriptions.shtml

Relevant initiatives

- Please participate to

  Document Freedom Day 2009!

  www.documentfreedom.org

- Join the Database of Digitally Free Schools at

  http://digifreedom.net/node/55

The end!

- I always welcome feedback, further material on these topics and occasions to discuss them!
- Please don't hesitate to come back with questions, suggestions, etc...
- For contact information, please see http://mfioretti.net
- Please go back home and convert all your files to open formats tonight!
- Thanks for your time!

Marco Fioretti