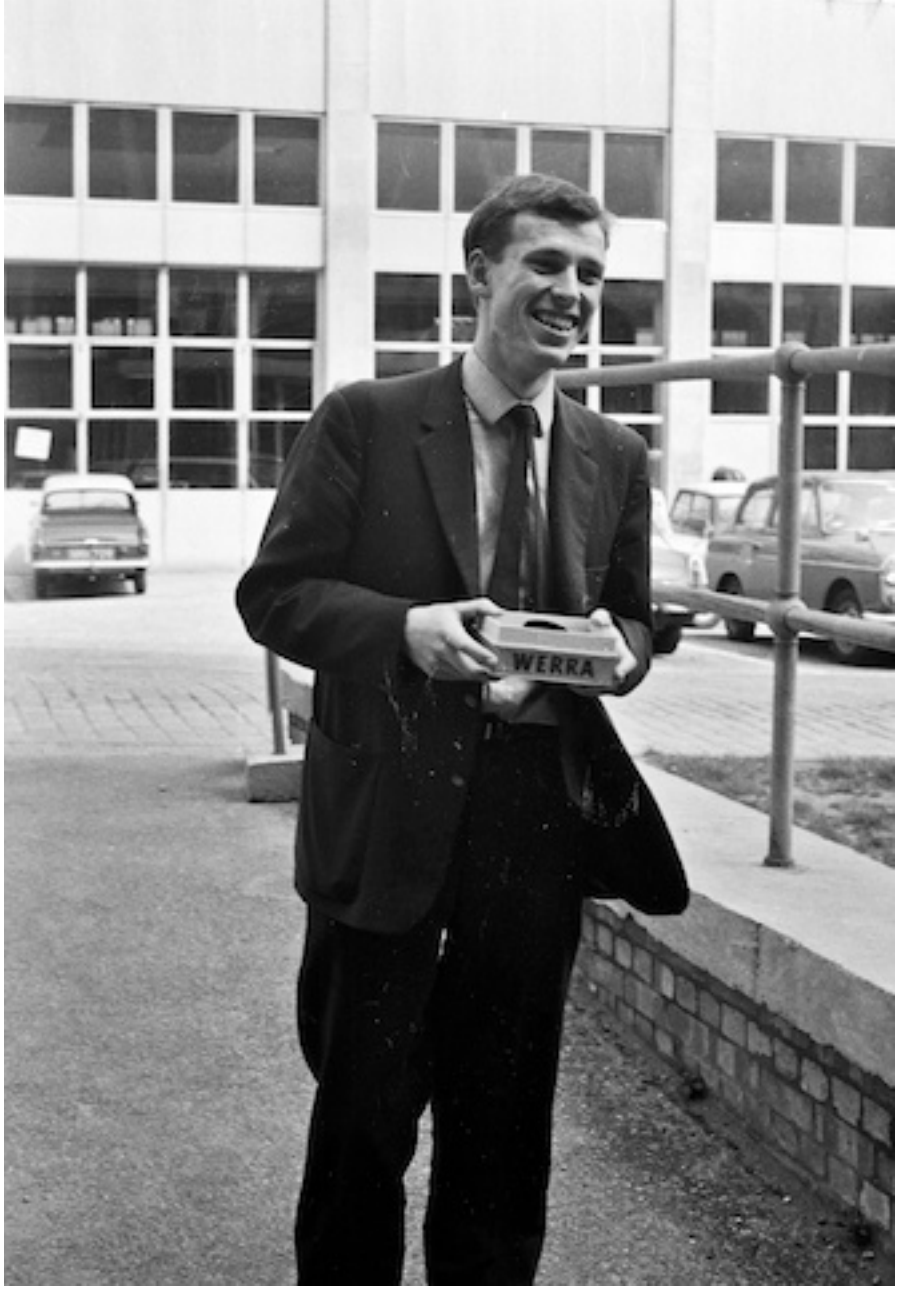
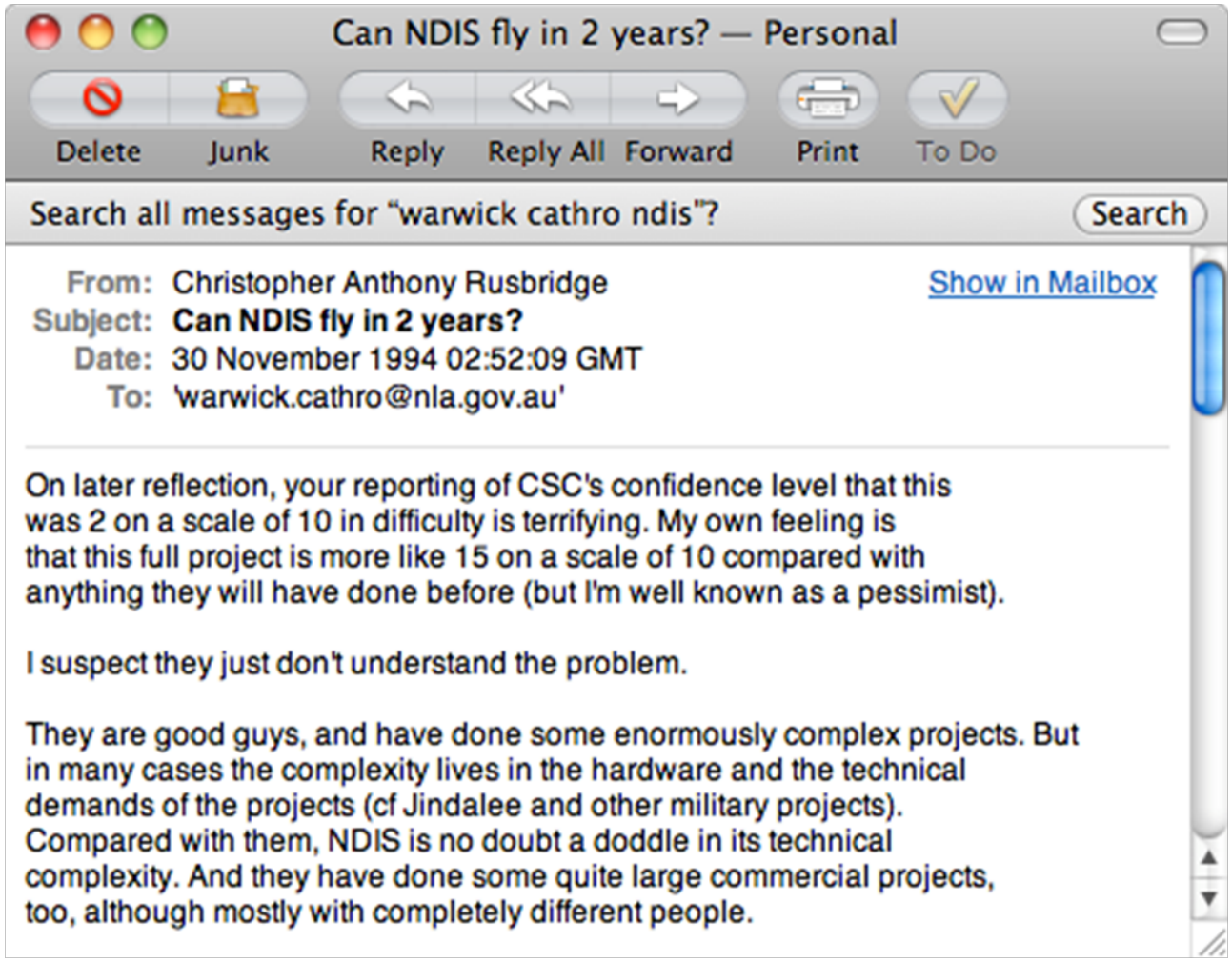
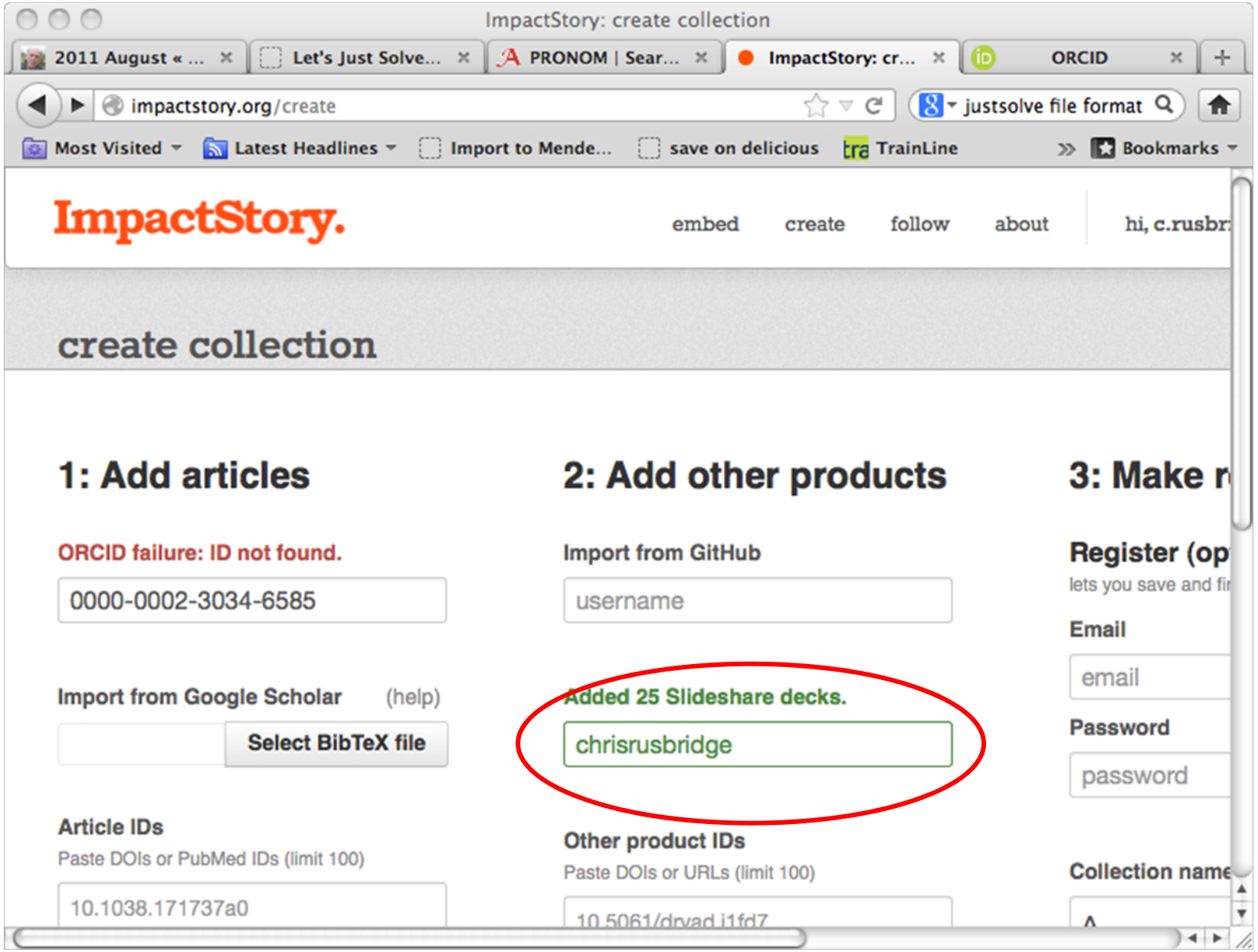
File Formats
What's the problem?
(To which crowd-sourcing is the answer!)

Contents

- About me
- What problem?
- Crowd-sourcing
- What do we want to do about file formats?
- Jason Scott and “Just Solve the Problem!”
- My Powerpoint 4.0 adventure
- Open Letter to Microsoft
- Take-home message…


WERRA


Can NDIS fly in 2 years? — Personal
Delete
Junk
Reply
Reply All
Forward
Print
To Do
Search all messages for "warwick cathro ndis"?
Search
From: Christopher Anthony Rusbridge
Show in Mailbox
Subject: Can NDIS fly in 2 years?
Date: 30 November 1994 02:52:09 GMT
To: 'warwick.cathro@nla.gov.au'
On later reflection, your reporting of CSC's confidence level that this was 2 on a scale of 10 in difficulty is terrifying. My own feeling is that this full project is more like 15 on a scale of 10 compared with anything they will have done before (but I'm well known as a pessimist).
I suspect they just don't understand the problem.
They are good guys, and have done some enormously complex projects. But in many cases the complexity lives in the hardware and the technical demands of the projects (cf Jindalee and other military projects). Compared with them, NDIS is no doubt a doddle in its technical complexity. And they have done some quite large commercial projects, too, although mostly with completely different people.


ImpactStory: create collection
2011 August « ...
Let's Just Solve...
PRONOM | Sear...
ImpactStory: cr...
ORCID
impactstory.org/create
justsolve file format
Most Visited
Latest Headlines
Import to Mende...
save on delicious
TrainLine
Bookmarks
ImpactStory.
embed
create
follow
about
hi, c.rusbr
create collection
1: Add articles
ORCID failure: ID not found.
0000-0002-3034-6585
Import from Google Scholar
(help)
Select BibTeX file
Article IDs
Paste DOIs or PubMed IDs (limit 100)
10.1038.171737a0
2: Add other products
Import from GitHub
username
Added 25 Slideshare decks.
chrisrusbridge
Other product IDs
Paste DOIs or URLs (limit 100)
10.5061/dryad.j1fd7
3: Make r
Register (op
lets you save and fi
Email
email
Password
password
Collection name

- In short
  - Bit of a nerd
  - Bit OCD
  - Interested in the past
  - Interested in “contributing”
- I’m not alone!

File format problem?

- What can I do if (when) this file doesn't open?
- (Big archives have other problems at scale that are not my topic here!)

File format solutions?

- Lists of FFs with useful information
- FF specifications (and variants)
- Tools to identify FFs
- Tools to process old FFs in current environments
- Tools to migrate FFs to current environments
- Tools to emulate old environments so we can process old FFs

Crowd-sourcing?

- “Experts” only have some of the expertise
  - And they’re short of time
  - And have other priorities
  - And they often want to charge you for stuff
- Many people have small amounts of particular expertise- in depth!
- Many people want to contribute
- (And of course a few want to destroy…
  - But they will anyway)
- Put all this together?

Crowd-sourcing PRONOM

- "The online registry of technical information. PRONOM is a resource for anyone requiring impartial and definitive information about the file formats…"
- XLSX
  - Fmt/214
  - "outline record only"
  - "developed by- none"
  - "If you are able to help by supplying any additional information concerning this entry, please return to the main PRONOM page and select 'Add an Entry'."
- 2 August 2011 sent in suggestion
  - "get back to you in 10 days…"
- 25 January 2013 pretty much unchanged!

Summary

Name	Microsoft Excel for Windows
Version	2007 onwards
Other names	
Identifiers	PUID: fmt/214
Family	
Classification	Spreadsheet
Disclosure	
Description	This is an outline record only, and requires further details, research or authentication to provide information that will enable users to further understand the format and to assess digital preservation risks associated with it if appropriate. If you are able to help by supplying any additional information concerning this entry, please return to the main PRONOM page and select 'Add an Entry'.
Orientation	Text
Byte order	
Related file formats	None.
Technical Environment	
Released	
Supported until	
Format Risk	
Developed by	None.
Supported by	None.
Source	Digital Preservation Department / The National Archives
Source date	28 Oct 2009

Apologies to TNA…

- PRONOM is a Good Thing
- TNA has its own priorities (signatures)
- My beef is with the unfulfilled crowd-sourcing offer, rather than PRONOM!

- PRONOM could be (or could have been) a Truly Great Thing!

“Just Solve the File Format Problem”

- Jason Scott July 2012 blog post
- November 2012 to be the first “Just Solve” month
- File Formats to be the first Just Solve problem

QuickTime™ and a
decompressor
are needed to see this picture.

‘Everyone knows this problem. It’s why old novelists cry they can’t pull their first novel out of Wordperfect. It’s why someone who used U-matic tapes to record the first meetings of a famous protest group goes “oh well”. It’s why, in all things, someone looks at anything older than five years, and goes “bye”, figuring there’s nothing they can do.

‘And I’ve had to listen to the mewings about this problem for at least 20 years now, in various forms. A lot. And then the person lights up about maybe solving this problem, and then dims and says “well, we can’t really solve the problem”. Because they know – it’d take an army of people to do it.

‘Let’s make that goddamned army.’ Jason Scott

I’d like to be in that goddamned army!

Powerpoint 4.0 adventure

- Wanted more presentations on Slideshare
- Some would not open… older PPT 4 files
- Tried various approaches
- Blog and tweet feedback
- Iterating towards various solutions
  - First actual solution from Dave Clipsham!!! (not scaleable for me)
- Arrived at workable solution (for me): Zamzar
  - Initially promising, not quite there
  - Asked if they could do it… they said yes!!
  - Free for small quantities, subscription for premium service

Obsolete MS file formats

- Annoyed!
- MS does have specs of obsolete office formats
  - Only back to Office 97
  - Surely they can do better?
  - …and JUST Solve month coming up…
- Open Letter to Microsoft (Tony Hey) blogged
  - Over 100 commented in support
  - >600 peak views a day (normally ~10)
- Positive response from MS

Response from MS

“Chris

I have a reply from Jim Thatcher in the Office team:

1) We do not currently have specifications for these older file formats.

2) It is likely that those employees who had significant knowledge of these formats are no longer with Microsoft.

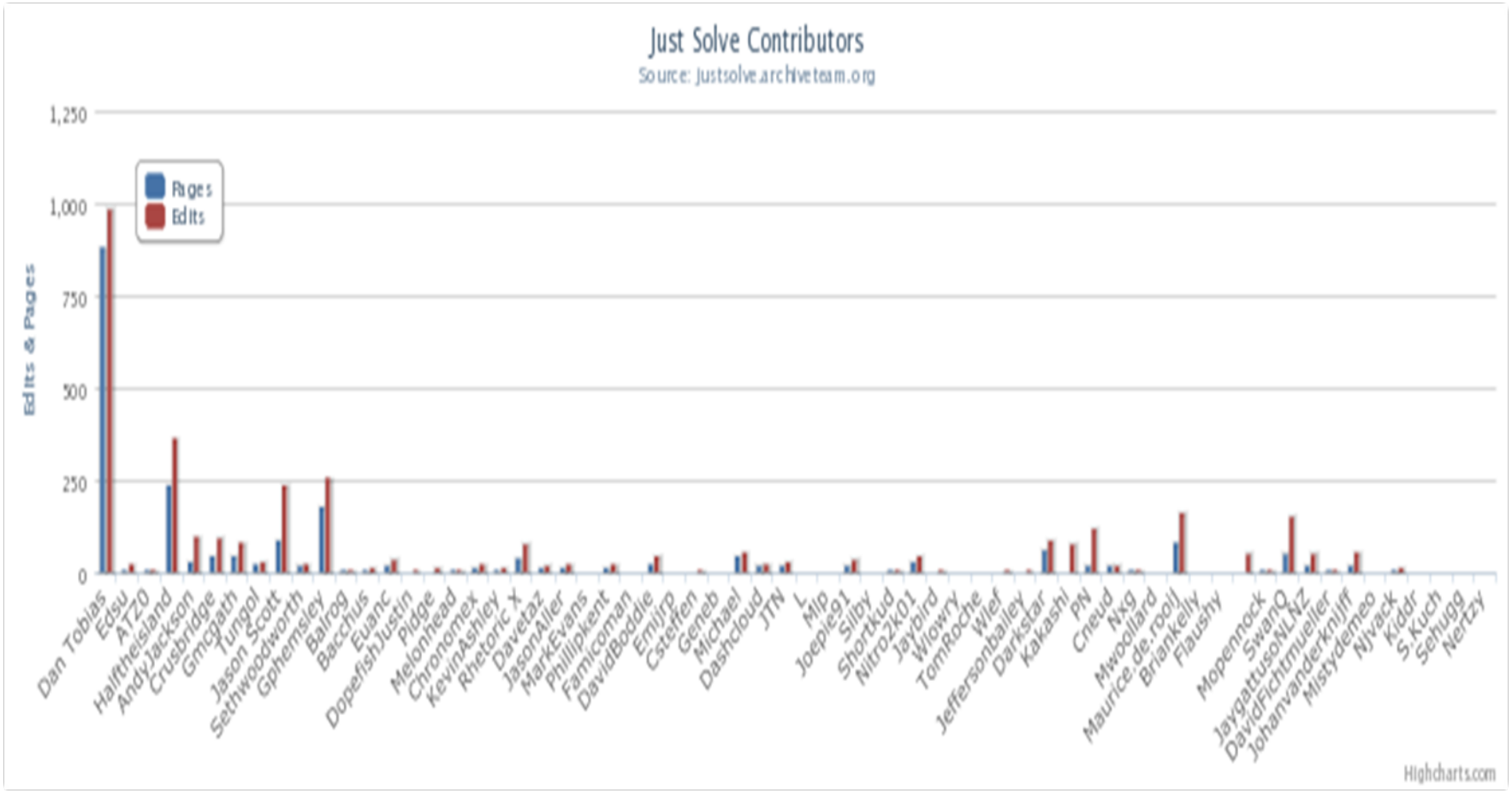
3) We can look into creating new licensing options including virtual machine images of older operating systems and old Office software images licensed for the sole purpose of rendering and/or converting legacy files.

4) One approach we could consider is for Microsoft to participate in a “crowd source” project working with archivists to create a public spec of these old file formats.

I think it would be sensible for you to talk directly with Jim – and Natasa in the UK – to see if there are some creative options that Microsoft could pursue with the archivist community.” Tony Hey

Just Solve 3

- Approach: first of my options
  - List plus short description
- Built a wiki
  - Not wikipedia… interesting why
- Listed hundreds of FFs
- Many many FFs briefly described


Just Solve Contributors
Source: Justsolve.archiveteam.org
1,250
1,000
750
500
250
0
Edits & Pages
Pages
Edits
Dan Tobias
Edsu
ATZ0
Halftheisland
AndyJackson
Crusbridge
Gmcgath
Tungol
Jason Scott
Sethwoodworth
Gphemsley
Balrog
Bacchus
Euanc
DopefishJustin
Pidge
Melonmead
Chronomex
KevinAshley
Rhetoric X
Davetaz
JasonAller
MarkEvans
Phillipkent
Famicoman
DavidBoddie
Emijrp
Csteffen
Geneb
Michael
Dashcloud
JTN
L
Mlp
Joepie91
Silby
ShortKud
Nitro2k01
Jaybird
Wlowry
TomRoche
Wlef
Jeffersonbailey
Darkstar
Kakashi
PN
Cneud
Nxg
Mwoallard
Maurice.de.rooij
Briankelly
Flaushy
Mopenmack
SwanQ
JaygattusoNLNZ
DavidFichtmueller
Johanvanderknijff
Mistydemeo
Njvack
Kiddr
S.Kuch
Sehugg
Nertzy
Highcharts.com

Navigation

Main page
File formats
Formats by extension
Still more extensions
Software
Glossary
Community portal
Recent changes
Random page

Toolbox

What links here
Related changes
Special pages
Printable version
Permanent link

XLSX

Office Open XML Spreadsheet (.XLSX) is the default file format for documents used by Microsoft Excel as of Excel 2007. In prior versions, the default version was XLS.

Contents [hide]

File Format

Name	XLSX
Ontology	• Electronic File Formats • Document • **XLSX**
Extension(s)	`.xlsx`

History

This (along with the other Office Open XML items DOCX and PPTX) was initially standardized as ECMA-376 in 2006. Three formats of this standard have been produced; the second version also corresponds to ISO/IEC 29500.

Incompatibility with earlier versions

Attempting to open XLSX files with earlier versions of Excel (pre-2007) results in garbage instead of a proper spreadsheet. A compatibility pack supposedly adds the ability to load the newer format into the older versions, but this doesn't necessarily work well in all cases. This can be a problem when people insist on e-mailing you files in the newest proprietary Microsoft formats when, most of the time, whatever they're sending could have been done fine in an entirely nonproprietary keep-it-simple-stupid format such as CSV or plain text. Open Office can open XLSX files, however.

Format

Like the other "Open XML" formats, this file format actually consists of various files (mostly XML) compressed into a ZIP archive, with this fact obscured from the end user by the use of a different file extension.

References

- ECMA-376 specification
- ISO/IEC 29500 specification
- How to open new file formats in earlier versions of Microsoft Office

So what do I want?

- More people involved in crowd-sourcing file format information
  - Especially the Just Solve wiki
  - Stuff on wikipedia where we can
  - Adding to PRONOM would be Excellent
- Someone to take leadership of the Microsoft obsolete formats issue
  - Grant to add older formats to Open/Libre Office? People to crowd-source the MS formats issue
  - An organisation that can seek grants, deploy people part time in their own interests, etc
  - Reverse engineering, adding stuff to Open/Libre Office
- People to contribute file examples

Resources

- My Pronom rant http://unsustainableideas.wordpress.com/2011/08/05/so-what-was-that-twitter-rant-against-pronom-all-about/
- Jason Scott's call http://ascii.textfiles.com/archives/3645
- Powerpoint 4 summary http://unsustainableideas.wordpress.com/2012/10/15/ppt-4-adventure-learning/ and related posts
- Zamzar http://www.zamzar.com/
- Open Letter to MS http://unsustainableideas.wordpress.com/2012/10/22/open-letter-ms-obsolete-formats/ and related posts

Take home message

- YOU can make a difference to the world!

QuickTime™ and a
decompressor
are needed to see this picture.

c.rusbridge@gmail.com