JPEG 2000 at the Wellcome Library

Christy Henshaw
Digitisation Programme Manager – Wellcome Library

JP2 Summit
12-13 May 2011
Library of Congress

wellcome trust
The Wellcome Trust

• A global charitable foundation

• Achieving extraordinary improvements in human and animal health

• Supporting the brightest minds in biomedical research and the medical humanities

• Exploring medicine in historical and cultural contexts
The Wellcome Library
The Wellcome Library

Collections of books, manuscripts, archives, films and pictures on the history of medicine from the earliest times to the present day.
The Wellcome Digital Library pilot, 2010-2013

Genetics and its Modern Foundations
A new online resource for everyone interested in the history of human and animal health.

Aims

• **build sustainable/expandable mechanism** – foundation stone for WDL
• **digitise key library holdings** - relating to a major Trust challenge area
• **digitise important third party content** – linked to theme
• **use innovative content and tools** – to encourage discovery and use
• **explore commercial partnerships** – enhance access to non-theme material
JPEG 2000 conversion – scope

- **Wellcome Images** – image library, legacy images, 300,000 images in the archive

- **Current projects** – pilot digitisation projects, 7m images 2010 - 2014

- **Long-term plans** – digitisation of *large proportion* of our collections (mainly special collections), 15m – 25m images 2014 and beyond
Type of content

• **Printed books** – early printed books, modern books (monographs), pamphlets, reports

• **Archives** – personal papers, institutional papers, unpublished works, mostly 20th century

• **Manuscripts** – unpublished, handwritten “manuscript books” and related materials, mostly 17th, 18th and 19th century, can be fragile

• **Artworks** – prints, paintings, posters, drawings, glass slides, etc.
The Francis Crick Archive
Books related to genetic research
Early printed books
Artworks, manuscripts
Decision to adopt JP2

JPEG 2000 was found to answer the following needs:

- **Storage costs** – 20/30m TIFFs stored on online, backed-up storage = multiple petabytes. Needed something cost-effective.
- **Quality** – needed a high-quality compressed format that would cover a wide range of content types.
- **Robustness** – needed a well-established image format with a high chance of long-term support.
- **Practical** – feasible to use in a Library digitisation workflow.
Finding our way

Working with JP2 opened up a whole new world – reading specifications, finding conversion software, so many choices.

Commissioned the report: JPEG 2000 as a Preservation and Access Format for the Wellcome Trust Library

Goal to find a single version of JPEG 2000 that would meet the needs of both long-term preservation and flexible delivery needs.
The result

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>File format</td>
<td>Part 1 (.jp2)</td>
</tr>
<tr>
<td>Compression</td>
<td>Lossy (6:1, 10:1)</td>
</tr>
<tr>
<td>Tiling</td>
<td>1024 x 1024</td>
</tr>
<tr>
<td>Progression order</td>
<td>RLCP</td>
</tr>
<tr>
<td>Decomp levels</td>
<td>5</td>
</tr>
<tr>
<td>Quality layers</td>
<td>8</td>
</tr>
<tr>
<td>Code block size</td>
<td>6, 64x64</td>
</tr>
<tr>
<td>Regions of interest</td>
<td>No</td>
</tr>
<tr>
<td>TLM markers</td>
<td>Yes</td>
</tr>
<tr>
<td>Bypass</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Embedding JP2

Chose [LuraWave](#) command line tool

- Some issues (bugs, or inconvenient implementations) arose, and all have been successfully addressed by LuraTech
- Created a firm consensus to use JP2 as the format for *all* still-image digital imaging (with one or two exceptions)
- No plans to use JP2 for digital video – but never say never
- Internal information sharing – digital archivists, systems administrators, IT department, programme board members
- External communication and networking
Current status, future plans

- Conversion of all new digital images is now carried out as standard.
- Nearing the final stages of a project to convert 450k image backlog to JP2 (reducing current footprint from 20 Tb to 5.5 Tb).
- Large projects use lossy JP2, legacy picture library uses lossless.
- Developed a strategy to determine compression levels.
- Currently using the GUI, but will use the command line interface with our new workflow system, streamlining conversion and QA.
- Medium term, will look at automating compression level selection.
Quality control for compression

- **Visual inspection**
- Color shifts, loss of detail, halo effects, pixelation, blurring, etc.
- Collection-based, representative sample
- Test range of compressions with intervals such as 2:1, 4:1, 6:1
- Once artefacts are discovered, step back to previous compression ratio
- Worst-performing image rules, for any particular collection
- Efficient for homogenous collections – less so for heterogenous collections with wide variety of content
- Archives particularly difficult – black and white compresses very well – colour drawings and photographs, not so well
Establishing the JP2K-UK group

• Unknown who in the UK were using JPEG 2000, or considering it
• Unknown who was even *interested* in JPEG 2000
• No one wants to work in a vacuum…
• Discovered a high level of interest: British Library, The National Archives, Oxford, King’s College London, Cambridge and Southampton Universities, Digital Preservation Coalition, commercial companies/consultants
• Loose affiliation of the like-minded – a user group
Remit of the JP2K-UK group

• Initial meeting in December 2009
• Everyone had a little knowledge – no one knew enough
• Agreed the need to approach JP2 implementation from practitioner’s point of view
• “Practitioner” meaning those who manage digital imaging strategies and implementation
• Agreed need to share information and collaborate
• Discussed ideas for a conference, and creating some guidelines for the user community
• Wellcome encouraged to write a blog about specific experiences working with JP2
Outputs

• JPEG 2000 Seminar, held in London in November 2010
  > 80 attendees
  > UK and European speakers and delegates
  > mostly non-technical audience

• Advocacy for practitioner’s needs
  > discussing and airing the needs and concerns of practitioners has influenced software developers, and even the JPEG Committee
  > JPEG 2000 at the Wellcome Library blog
    www.jpeg2000wellcomelibrary.blogspot.com
Future plans for JP2K-UK

• Guidance for practitioners
  > Human readable
  > Focus on practicalities
  > Enable practitioners to make informed choices
  > Advice on implementation

• Community building
  > Case studies
  > Lessons learned
  > Networking (nationally and internationally)