(Digital)
Cultural Heritage: From Preservation to Access to Promotion to Where?

Alan F. Smeaton
Dublin City University
Presentation Overview

• Apologies to Thomas C. Südhof, “Mechanisms of Neurotransmitter Release”, Nobel Prize winner Physiology/Medicine 2013

• Digital Cultural Heritage … 20 years, to now

• Capturing intangible culture

• Preservation – Accessibility – Access – Promotion

• Looming hurdle in promoting digital CH … content-based access

• Content access for CH digital imagery
Digital Cultural Heritage Started

- Digitisation became cost-effective 20 years ago
- Clamor for CH digitisation, national, EU, philanthropic
ISOS

- **Irish Script On Screen**
  - DIAS project with DCU assistance
  - Make digital images of ancient Irish manuscripts available in an online Digital Library
    - Small (150-500K) JPEG images
    - Large (1.5-5MB) JPEG images
  - [http://www.isos.dias.ie/](http://www.isos.dias.ie/)
  - Metadata includes manuscript details and page numbering

- The collection contains manuscript pages from manuscripts from collections from universities, museums and Abbeys

- Helps to make ancient Irish manuscripts available to general public as well as to Celtic scholars
ISOS Important Features

• **Strict Editorial control of the objects in the library**
  - Not an upload process
  - An expensive digitisation process means that careful planning is required before any manuscript is included.

• **No need for people to travel to access the library**
  and the ISOS digital library is available a

• **Anyone can access the collection**
  - Though large JPEG images are reserved for registered library users

• **ISOS helps to digitally preserve degrading manuscripts**
ISOS Manuscripts
ISOS Purpose

- So the aims were PRESERVATION and ACCESSIBILITY (not access or promotion!)

  www.isos.dias.ie

- Content-based Access non-existent though there was some work on using word shapes as a basis for retrieval

- Group similarly-shaped characters using a reduced alphabet, based on considerations of appearance of letters
Word Shape Mapping

A ← A-Z b d f h k l t
x ← a c e m n o r s u v w x z
g ← g p q y
j ← j

Dublin is beautiful

• ... maps to “AxgxxxAA ix AxxxAiAxA”
Word Shape Mapping

Based on a lexicon of 318,636 correct word forms, many share the same WST

\[
\text{xxxAxxx} = \text{numbers, nuclear, murders, and 281 others}
\]
\[
\text{xxxA} = \text{card, rush, cost, and 204 others}
\]

but

\[
\text{xxAxixxxxAAx} = \text{retrievable, untrimmable}
\]
\[
\text{xxAxixxx} = \text{retrieve, cadmiums, calcines, entwines, matrices, retainer, sobrinus}
\]

while the following are all unique

\[
\text{automobiles, Environmental, contribute, criminals, military, ignored, origin, vitamin}
\]
Provided a form of search and navigation based on actual content

While typed English has some uniqueness, hand-written Old Irish was ... impossible
Cultural Heritage

- Much early digital CH was Preservation and Accessibility .. what about the internet (not the same as social media) ?

- The Internet Archive ... archive.org ... non-profit digital library, free public access ... 200 employees scanning books, $10m per annum,

- 10 petabytes 3 years ago = $10,000,000,000,000,000 B = $10^{16}$ bytes = 400 Billion web pages

- Slow to retrieve because roboticised storage!
<table>
<thead>
<tr>
<th>Name</th>
<th>Website</th>
<th>Wayback Machine Records</th>
<th>Total links</th>
<th>Rank(Global &amp; Ireland)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agriland</td>
<td>agriland.ie</td>
<td>40 (2012-2015)</td>
<td>377125</td>
<td>292,928/2,061</td>
</tr>
<tr>
<td>Labour</td>
<td>labour.ie</td>
<td>504 (1998-2014)</td>
<td>1063988</td>
<td>1,060,560/6,273</td>
</tr>
<tr>
<td>Fianna Fail</td>
<td>fiannafail.ie</td>
<td>406 (1998-2014)</td>
<td>735891</td>
<td>1,408,305/6,663</td>
</tr>
<tr>
<td>Fine Gael</td>
<td>finegael.ie</td>
<td>293 (1998-2014)</td>
<td>882846</td>
<td>807,786/4,330</td>
</tr>
<tr>
<td>Sinn Fein</td>
<td>sinnfein.ie</td>
<td>577 (1998-2012)</td>
<td>1884479</td>
<td>615,420/7,449</td>
</tr>
<tr>
<td>Green Party</td>
<td>greenparty.ie</td>
<td>388 (1999-2014)</td>
<td>517009</td>
<td>4,830,354/none</td>
</tr>
<tr>
<td>Anti Austerity Alliance</td>
<td>antiausterityalliance.ie</td>
<td>9 (2014)</td>
<td>16317</td>
<td>3,199,124/none</td>
</tr>
<tr>
<td>Teagasc</td>
<td>teagasc.ie</td>
<td>360 (1998-2014)</td>
<td>1260273</td>
<td>303,114/3,330</td>
</tr>
<tr>
<td>Public Policy</td>
<td>publicpolicy.ie</td>
<td>22 (2007-2015)</td>
<td>144127</td>
<td>1,175,738/8,030</td>
</tr>
<tr>
<td>NESC</td>
<td>nesc.ie</td>
<td>92 (1999-2014)</td>
<td>16003</td>
<td>4,254,207/none</td>
</tr>
</tbody>
</table>
## Sample Websites

<table>
<thead>
<tr>
<th>Name</th>
<th>Website</th>
<th>Wayback Machine Records</th>
<th>Total links</th>
<th>Rank (Global &amp; Ireland)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politics</td>
<td>politics.ie</td>
<td>624 (2002-2015)</td>
<td>109415</td>
<td>73,104/361</td>
</tr>
<tr>
<td>The journal</td>
<td>thejournal.ie</td>
<td>1905 (2010-2015)</td>
<td>None</td>
<td>10,317/54</td>
</tr>
<tr>
<td>Ask About Ireland: Environment</td>
<td>askaboutireland.ie/enfo/irelands-environment</td>
<td>32 (2012-2014)</td>
<td>34159</td>
<td>216,744/14,165 .ie/Australia )</td>
</tr>
<tr>
<td>Trócaire</td>
<td>trocaire.org</td>
<td>621 (1996-2015)</td>
<td>635640</td>
<td>773,948/10,349</td>
</tr>
<tr>
<td>Irish Aid</td>
<td>irishaid.ie</td>
<td>38 (2009-2015)</td>
<td>33547</td>
<td>1,262,919/758,781 (US)</td>
</tr>
<tr>
<td>Concern</td>
<td>concern.net</td>
<td>543 (2001-2015)</td>
<td>774094</td>
<td>502,922/31,283 (Pakistan)</td>
</tr>
<tr>
<td>Christian Aid</td>
<td>christianaid.ie</td>
<td>60 (2004-2014)</td>
<td>36460</td>
<td>7,346,456/none</td>
</tr>
</tbody>
</table>
WayBackMachine

• WBM is accessed by URL and by date of snapshot

• WBM provides PRESERVATION and ACCESSIBILITY but not so good on access, or promotion … no search facility (which is not unreasonable)

Archive.org

• WBM is enabled by modern technology (storage) which has also enabled other forms of CH preservation
Sports

• **Sports / games are a form of intangible cultural heritage**
• **There are +3,000 traditional sports and games worldwide**
• **Rugby, Soccer, cricket, hockey are examples of the well-known, but …**
Longe Paume

- A form of tennis with virtual nets
- Played in Northern France
Tamburello

- A form of volleyball but with no nets, and played with tambourines
- Played in Japan
Jai-Alai / Pelota

• The fastest game in the world

• Like handball but with a thresher and played as a team!

• Played in Basque
GAA Football and Hurling ;-) 

- National game of Ireland
All examples of endangered sports!
Can Hollywood technology preserve minority sports?

27 October 2014 Last updated at 01:41 GMT

Performance capture is being used to record the movements of athletes belonging to several sports across Europe.

The technology involves placing a number of sensors on a subject, tracking their movements, and has previously been used in films like The Hobbit and Avengers: Age of Ultron.

The recording is part of an EU funded project that aims to digitally capture and preserve sports like hurling and Gaelic football for future generations.

BBC Click’s Nick Kwek talked to those involved in the process.

More at BBC.com/Click and @BBCClick.

Share this story About sharing

[Image of a triangle]
Promoting CH

• Replay addresses preservation, accessibility?, one form of access and a fair bit of promotion by turning it into a game for public/children – outreach

• Promoting CH is difficult

• We like to think we like to engage in CH but it’s not easy to find interesting things, and we now like to find things

• Our attention span is reducing, especially younger generation, and we don’t like guided tours

• Information consumption has changed, live TV is becoming rarer, YouTube is user control, we find news

• We want control over what we consume
Promoting CH

• At odds with this trend, digital CH is accessed in ways where others, not the consumer, are in control, just like TV and newspaper editors

• Example ... TV archives

RTE.ie
Éamon de Valera (1882-1975)

With a political career spanning over 50 years, Éamon de Valera held the position of Taoiseach for three terms and was also President of Ireland from 1959 to 1973.
• **RTÉ Archive browsing is interesting, but could be enhanced …**
  - Locations could be built-in for visualisation, and navigation
  - Speech transcribed and searchable
  - Named entities (people, places, etc.) recognised and linked
  - … common characteristic is user control

• **User control for text navigation is easy … Google-like searching**

• **User control for visual navigation is not easy, we need to be able to automatically analyse imagery**
What does this mean?
Or this?
How good can a computer get at automatically annotating images without using any context at all?
MOUNTAIN VIEW, Calif. — Two groups of scientists, working independently, have created artificial intelligence software capable of recognizing and describing the content of photographs and videos with far greater accuracy than ever before, sometimes even mimicking human levels of understanding.

Until now, so-called computer vision has largely been limited to recognizing individual objects. The new software, described on Monday by researchers at Google and at Stanford University, teaches itself to identify entire scenes: a group of young men playing Frisbee, for example, or a herd of elephants marching on a grassy plain.

The software then writes a caption in English describing the picture. Compared with human observations, the researchers found, the computer-written descriptions are surprisingly accurate.

The advances may make it possible to better catalog and search for the billions of images and hours of video available online, which are often poorly described and archived. At the moment, search engines like Google rely largely on written language accompanying an image or video to ascertain what it contains.
Captioned by Human and by Google’s Experimental Program

**Human:** “A group of men playing Frisbee in the park.”
**Computer model:** “A group of young people playing a game of Frisbee.”
Captioned by Human and by Google’s Experimental Program

Human: “A young hockey player playing in the ice rink.”
Computer model: “Two hockey players are fighting over the puck.”
Captioned by Human and by Google’s Experimental Program

Human: “A green monster kite soaring in a sunny sky.”
Computer model: “A man flying through the air while riding a snowboard.”
Captioned by Human and by Google's Experimental Program

Human: “A person riding a dirt bike is covered in mud.”
Computer model: “A person riding a motorcycle on a dirt road.”
Captioned by Human and by Google’s Experimental Program

Human: “Three different types of pizza on top of a stove.”
Computer model: “A pizza sitting on top of a pan on top of a stove.”
Captioned by Human and by Google’s Experimental Program

Human: “Elephants of mixed ages standing in a muddy landscape.”

Model: “A herd of elephants walking across a dry grass field.”
How is captioning done?

Figure 1. Our model generates free-form natural language descriptions of image regions.
Deep Visual-Semantic Alignments for Generating Image Descriptions

Andrej Karpathy    Li Fei-Fei
Department of Computer Science, Stanford University
{karpathy, feifeili}@cs.stanford.edu

Figure 2. Overview of our approach. A dataset of images and their sentence descriptions is the input to our model (left). Our model first infers the correspondences (middle) and then learns to generate novel descriptions (right).
Figure 5. Example alignments predicted by our model. For every test image above, we retrieve the most compatible test sentence and visualize the highest-scoring region for each word (before MRF smoothing described in Section 3.1.4) and the associated scores $(u_t^T s_t)$. We hide the alignments of low-scoring words to reduce clutter. We assign each region an arbitrary color.
Image Captioning ... Solved?

- Computational processing needed for this is not scalable
- Domain knowledge data needed for this is not scalable
- Some major search engines are now introducing content-based access to imagery ... photos.google.com
- Based on pre-processing images for 00’s, 000’s, 0000’s of “concepts” ... is that scalable?
Aug 6
Searching visual libraries

- Scalable but only in a way
- These concept detectors are built by machine learning the differences between X and not X
- Requires much manual tagging and annotation
- An alternative to building these in advance is to build at query time … and where can we get lots of examples images of topic X … from Google of course!
Accessing Digital Cultural Heritage

• So we are getting content-access … and there will be trickle-down into digital libraries for cultural heritage

• Digital Cultural Heritage in 10 years …
  • New technologies, new media forms
  • New forms of Digital Cultural Heritage
  • New interaction modalities … tablets, augmented reality, brain sensing from wearables
  • New ways to navigate where we control
  • … but these are “just technology” changes and evolution

• Biggest change will be in people, expectations, demands because People want to be in control