Ed Bremner I

Ed Bremner I

Revision History
Revision 1 Wed, 08 Jun 2011 14:18:00 BST

Table of Contents

Introduction to Digitisation

Why digitise?

Digitisation is the process of creating an image or other digital representation of an original source. There are many types of digitisation and many reasons for a cultural institution to do it. This document will mainly treat instances where a digital image is created from an analogue document format through scanning or digital photography.

In a wider cultural context, most digitised items are created for online delivery. Digital publication helps to bring collections to a wider audience, provides positive publicity for an institution, and helps to satisfy the needs of an institution's stakeholders and the wider cultural market, who are increasingly coming to expect access to content online.

In the past, the focus of such digitisation was on providing digital images and enhanced bibliographic descriptions of cultural items. The challenge and expense of transcribing historical irregular typefaces has meant that searchable text functionality has not been routinely added. However, there are clear benefits to doing so. Even inaccurate capture of historical text greatly increases the number of access points to the material, while accurate capture opens up new avenues of research such as linguistic analysis. This is the area in which the IMPACT Project expects to make a particular contribution.

Background and current events

Digitisation of historical materials has traditionally focussed on the careful selection of particular valuable items or collections, often relying on private or research funding. These digitisation projects have typically had finite scope (e.g. they run for a set time, with a maximum budget agreed in advance), with decisions about image quality and resource functionality made to suit the purposes of the project. This type of digitisation still makes up a significant proportion of cultural institutions' digitisation effort.

While most often worthwhile, the main danger of such "boutique digitisation" has been that it encourages the production of digital material in distinct silos, outside any wider national or international context. This has resulted in a proliferation of technical standards at every stage of the digitisation process, needless duplication of effort across the global cultural sector, and in some drastic cases the technical obsolescence of the digital objects produced.

These dangers have been recognised for more than a decade, and digitising institutions have responded by forging partnerships with others engaged in similar work. Working together, cultural institutions can define

policies and standards corporately, and thereby better ensure that digital resources produced by museums, libraries and archives are sustainable indefinitely. The Europeana Project is a large scale outcome of this approach. Funded by the European Commission, Europeana combines the technical and managerial expertise of major cultural and research institutions across Europe to produce a sustainable portal for digital heritage resources.[1]

In addition to partnerships between digitising institutions, there is also a long history of partnerships between digitising institutions and commercial or research organisations. The most common form of partnership involves a cultural institution sub-contracting some aspects of the digitisation or publication process to a third party. In recent years, the private sector has increasingly been seen to lead this process, as evidenced by the Google Books Project. Begun in 2004, Google Books aims to digitise the entire printed content of several major research libraries, allowing access to and interoperability of digitised material on a scale never before seen.[2]

The IMPACT Project in context

The IMPACT Project is a research project funded under the Seventh Framework Programme of the European Commission[3]. The project partners are a mix of libraries, universities, research centres and commercial companies across Europe and beyond[4]. The project's name is a mnemonic contraction of Improving Access to Text, and it aims to influence international digitisation practice in a number of ways.

Firstly, it has researched tools that aim to improve the accuracy of optical character recognition (OCR) results. Secondly, it has developed tools to significantly advance knowledge in OCR technology and lead to further developments in the future. Thirdly, knowledge gathered and gained in these processes will be disseminated throughout the digitisation community using learning materials and training opportunities. Finally, IMPACT will develop into a sustainable Centre of Competence, which will continue to provide advice and support to European digitisation initiatives in the future.

As a result of this activity, a greater amount of Europe's cultural heritage will be digitised more efficiently, and will become more accessible through the provision of OCRed searchable text. Relationships between cultural institutions, the research community and industry will also have been fostered, leading to increased collaboration and more successful outcomes.

Why optical character recognition?

While not the only method for creating searchable text, optical character recognition (OCR) is the most cost effective method for large-scale projects where the primary aim is to increase the range of access to the material and the number of ways users may interact with it.

Once basic searchable text has been created with OCR, it not only allows users to discover relevant material which may otherwise have been impossible to find, the OCR results themselves can be reworked, refined and improved by research projects or indeed by a user community through cooperative correction initiatives.

Optical character recognition is treated in depth in later learning resources.

[1] *Europeana – About Us;* Europeana, 2008-2011;http://www.europeana.eu/portal/aboutus.html Retrieved 10.02.2011. For information on the contributing partners within Europeana, see: http://europeana.eu/portal/partners.html Retrieved 10.02.2011
[2] *About Google Books*: http://books.google.com/intl/en/googlebooks/history.html Retrieved 10.03.2011
[3] *Understand the Seventh Framework Programme*; Google Books; 2011: http://cordis.europa.eu/fp7/understand_en.html Retrieved 10.03.2011.
[4] IMPACT Partner Information; IMPACT website; 2008-2011: http://www.impact-project.eu/about-the-project/partner-information/ Retrieved 10.03.2011

Key Terms

Digitisation: the conversion of an object, image or signal into a discrete set of points known as binary digits. Text-based digitisation takes an image from a physical volume through scanning or photography and converts it into a code made of patterns of zeroes and ones. These patterns can be read by a digital computer and reconstructed as a representation of the original, allowing for easy transfer of information between computers or across a network such as the World Wide Web

Optical Character Recognition: the mechanical or electronic translation of images of handwritten, typewritten or printed text into machine-editable text. In a digitisation workflow, OCR is usually the last stage of document image processing and analysis, but in non-specialist use the term is often used to refer to the whole process of producing OCR text