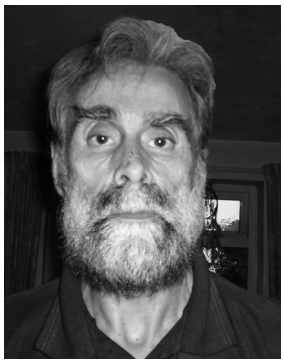
Pros and cons of online archive data for academic research

Online archive data offers great potential for easing the data discovery and integration process during research. With a focus on spatial information and two specific York projects, two particular developments are identified as necessary to facilitate this process. The first, and relatively simple to deliver, concerns the inclusion of spatially explicit metadata with the archive catalogues. The second concerns the way archive material is digitized and delivered to the end user, and difficulties resulting from the prevailing approach of delivering digital image reproductions from the archives are explored. An additional service within the data delivery process to assist the end user in ingesting the material is also discussed.



PETER J HALLS
GIS Advisor
University of York

Introduction

Online archive data offer a potentially rich resource for academic research and the potential to save researchers a significant amount of work in terms of data discovery and preparation. Most researchers expend a great deal of effort in locating required data and transforming them into a form suitable for use. Research builds on existing knowledge and, especially in social science disciplines, such knowledge is often expressed in the outputs of previous work. As online archives grow and deepen – especially as data are increasingly included as addenda to published research papers – researchers should in theory find it easier to discover and obtain the data that supports the research findings described in the papers. This is an ideal: however, attaining this ideal may require further development of those online archives. As my work primarily concerns the study of the way phenomena are distributed and interact over geographic space, I shall focus on the spatial aspect of archive data.

Spatial information

Most information contains a 'spatial' component, if only the place of publication. When I talk of a spatial component I am concerned with much more than maps and plans: cartography, in terms of the modern maps and plans with which we are all familiar, is a comparatively young discipline – little more than a couple of hundred years – although diagrams and maps in other forms have been around for three, maybe four millennia. Place names, cited in news bulletins, for example, set a spatial context for that particular item. There is a great deal of geographical information that can be mapped, but which itself comprises descriptive data, such as demographic facts and figures indexed by postal code, for example. Spatial information really is ubiquitous.

However, spatial information does have a number of peculiarities that must be addressed if it is to be linked usefully to other data. These peculiarities require rather more specialist treatment than does text – indeed, one peculiarity, especially with more recent material, may be that locations may be defined using numeric, co-ordinate systems rather than textual names. The tools used by end users facilitate the integration of information by means of these co-ordinate references. Frequently, spatial metadata facilitates rapid and easy integration with other materials and, particularly for visualization purposes, with map data.

Place as spatial information

The term 'place' can be remarkably imprecise. 'London', for example, refers to an area of some 600 or so square miles – not a very helpful designation for one trying to zero in on a specific

location within the metropolis. The term 'place' has a long pedigree, but has no truly consistent meaning: it emerges from human experience and refers to any location that matters to the human using it. Culturally, a 'sense of place' may refer explicitly to a human experience and the sharing of that experience with others. Consider its application in literature: Dickens frequently describes real places – parts of London in 'The Old Curiosity Shop' and 'Little Dorrit', for example. It might be an interesting project to trace the activities of a character in such a novel, using the descriptive text against contemporary mapping. Non-fiction may employ similar, though perhaps less detailed, descriptions. The recently launched Old Bailey Online service (http://www.oldbaileyonline.org/) links to digitized case records, some of which contain beautifully detailed descriptions of parts of London. If linked to census records, for example, such data might give the historian an opportunity to explore the activities of both victim and offender.

One of the problems as yet unsolved by the Old Bailey Online project, but which is a constant in archive records, is that of variant spellings and alternate names for the same location. Standards have been developed in recent years, primarily to support automation in the postal delivery service and efficiency for the emergency services. However, online resources drawing on digitized sources now provide access to material created over several centuries, much of it created well prior to the standardization of place names. Where these archive materials are now searchable electronically, whole new opportunities for the re-use of this information are now opening up, although that is not to say that the problems of such re-use (including the problems posed by the much greater exposure of non-standard place names) are all now solved!

Working online archives

Let me explore two examples: one from a recent MA research project at York and the other rather more speculative.

A student project

Within the University Libraries and Archives at the University of York is the Borthwick Archive, a repository of historically important collections, some dating from the fourteenth century or earlier. Recently, a student undertook to investigate the extent to which Benjamin Seebohm Rowntree's investigation into poverty in York (*Poverty, A Study of Town Life*, 1901) could be supported from the archive records. The student chose to use offences related to drunkenness as a test case and to try to map the incidents against the historic Ordnance Survey mapping available from the Digimap historic data service. In theory this would be a straightforward project: all the relevant information had been digitized. However, there are various ways of digitizing and delivering digital information: both the map data and the court records have been digitized by scanning the originals and delivering the (digital) copies as images – which are not themselves searchable. The first task for the student was to transcribe the selected court records, so that the information could be processed as text, from the archive images. Since the map data was also delivered in the form of image files, the student had to work out the numeric co-ordinates of all the addresses mentioned in the court records before it was possible to discover the spatial spread of incidents and so to explore how this pattern was distributed across the city. The amount of work needed to achieve this probably exceeded the original digitization effort by an order of magnitude, or more: there was no realistic means of automating the process. Alternative digitization processes for both text and map resources do exist – character recognition techniques for text are readily available, for example – but it seems that they are rarely applied.

An administrative user

A second example of the difficulties researchers face in using digitized archival map data comes from the University of York's Centre for Medieval Studies. As an officer of the City of York Council, one of the duties of the City Archaeologist is to examine planning applications submitted to the Council for development proposals. Such applications are required to define the 'footprint' of the area affected by the application and many, perhaps most, of these analyses are now submitted electronically in a form which can be very easily integrated with the modern, digital, mapping of the city. However, for an historic city such as York, there is a constant likelihood of such proposals having an impact that would necessitate a formal archaeological investigation of the site. Such

investigations are costly, both in terms of the investigation itself and in terms of delays to the project. I have already mentioned the Borthwick Archives; York is fortunate in having a number of good archive resources, including the City Archive which offers an unbroken municipal record dating back to around the twelfth century. In the present case, the City Archaeologist requested a means by which he could use the planning application 'footprint' to drill down through these various archives in order to determine whether there is any recorded reference to the location in question. If such were found, it would probably mean that the application must be 'called in' to undergo the analyses required by statute. Quite apart from the pre-requisite that the archives be in digital form in order to enable such a process – and there are very good curatorial reasons why these materials should be digitized – many other potential problems become immediately apparent. For example, how to render the textual nature of the archive records so that those relevant to the search can be discovered by applying the site outline as search framework? The Old Bailey Online project encountered the issue of variant spellings; in York we have streets that have been renamed over the years, with the original names used for different streets at various points in time. To enlarge further on the nature and range of the issues involved in this instance would take a much longer article than I am allowed here.

Importance of location in accessible data

My purpose in citing these two examples is to illustrate the potential offered by archive resources and the need to be able to re-use material held in archive repositories and to use these materials in new ways, especially by exploiting the spatial content of the archive information so as to integrate it with information from other sources. Whilst these examples concern older historical materials, similar problems concern much more recent material. For example, a biochemical study may implicate a geographical factor hitherto unsuspected as an agent in the spread of some disease affecting more than one species. Those working on such a study may need to re-examine the work of previous researchers or analysts to test their theories; however, to do so would only be possible if the spatial pattern uncovered in the previous studies had been recorded in an accessible manner.

The future

What is required to facilitate such investigations? Two levels of action, I think.

A great deal can be accomplished if spatial details are included within the catalogue metadata. The commonly used Dublin Core element for 'coverage' makes it possible to define the spatial extent of the data described using placeName, numeric co-ordinates, or even complex polygon boundaries; unfortunately only the placeName option appears to be widely used. For the UK and Europe, a fairly recent EU Directive, INSPIRE (http://inspire.jrc.ec.europa.eu/), is setting standards for defining the spatial component of information in order to facilitate interoperability on the basis of location. Although the INSPIRE directive itself applies only to the holders of the 'authoritative source' of explicitly defined types of reference information, it seems quite likely that researchers will be seeking to use INSPIRE-compliant data as one of their information sources, even if our repositories do not fall under INSPIRE itself.

Secondly, while catalogue records support resource discovery, for the researcher or student resource discovery is only the very first part of the problem. Once potentially relevant data are located, these data must be obtained in a form suitable for integration into the project at hand. Using my City Archaeologist example again, what the end user may require is some sort of a service by which she/he can request that a resource be transformed during the delivery process into a form suited to her/his needs. From the perspective of the spatial component at least, the nature of such a transformation service is likely to be reasonably generic, though not necessarily simple.

Summary

Online archives are already making the discovery of research data much easier. However, there is scope for a greater inclusion of data and for recognition of the spatial content of the archived resources in terms of the metadata descriptions and in understanding of how such online sources may be used by end users. After all, easing the discovery of resources does little good if the format in which the resource is to be delivered is unsuitable or requires significant additional work for the end user. There is an opportunity to add

significant extra value to our online archive resources by making the digital resource as accessible and useful to the end user as the catalogue record.

■ **Peter Halls**
University of York GIS Advisor
Computing Service
Heslington
York YO10 5DD, UK
Tel: +44 (0)1904 323806
E-mail: P.Halls@york.ac.uk

To view the original copy of this article, published in *Serials*, click here:
http://serials.uksg.org/openurl.asp?genre=article&issn=0953-0460&volume=23&issue=3&spage=222
The DOI for this article is 10.1629/23126. Click here to access via DOI:
http://dx.doi.org/10.1629/23222
For a link to the full table of contents for the issue of *Serials* in which this article first appeared, click here:
http://serials.uksg.org/openurl.asp?genre=issue&issn=0953-0460&volume=23&issue=3