The Role of the Web in the Provision of National Data and Information Services: *the MIDAS Experience*

by Julia Chruszcz, Keith Cole and Anne McCombe *

MIDAS is a JISC designated national data centre for the UK higher education community providing on-line access and support for a range of large and complex datasets, such as censuses, surveys, time series databanks, bibliographic and full text databases. In this context, MIDAS is part of the developing JISC funded National Distributed Electronic Resource which is seeking to promote and extend access to electronic information and services to the entire UK higher education community.

The expectations of users have changed considerably and we have had to rethink how we deliver data and information to the researcher's desktop. It is not enough to promote awareness of the data resources and their potential applications in teaching and research. We also have to convince the users that their time is being used efficiently, that they can easily identify the data that they want, extract it and where appropriate put it into a suitable format for secondary analysis. For us this means creating appropriate interfaces for the data - simple enough for a once-off selection and versatile enough for more sophisticated use.

This paper addresses the influence of the Web and the expectations of its users on the services provided by MIDAS. It shall describe some the new interfaces to data and information which will be of particular interest in both research and teaching.

MIDAS overview

MIDAS (http://www.mimas.ac.uk/) is a JISC (http://www.jisc.ac.uk) designated national data centre for the UK higher education community providing on-line access and support for a range of data and information resources. Along with the other JISC funded data centres (BIDS and EDINA), data services and projects, MIDAS is part of the developing Distributed National Electronic Resource (DNER) which is seeking to promote and extend access to electronic information and services to the entire UK higher education community (http://www.jisc.ac.uk/cei/dner_colpol.html). MIDAS is based at Manchester Computing at the University of Manchester. Manchester Computing also hosts other projects, such as COPAC (http://copac.ac.uk/copac/) and the SuperJournal Project (http://www.superjournal.ac.uk/sj/).

The data and information resources available through MIDAS include on-line access to the UK Census of Population; government and other continuous surveys; national and international time series databanks; digital map data; satellite images; chemical information systems; bibliographic data resources and electronic journals. A key component of the service is the provision of a range of specialist support services, such as documentation, training and research support, relating to the data and information resources available on MIDAS. Many of the services are run in collaboration with other organisations, such as The Data Archive at the University of Essex.

MIDAS also provides access to a range of software packages and the large scale computing resources (i.e. memory, disk space and CPU time) required by users wishing to undertake complex data analysis.

MIDAS also provides facilities and support to projects wishing to exploit the internet to provide wider network access to data and/or information resources for teaching and research purposes. Two key data sharing and gateway services that have developed considerably over the past couple of years include NetEc (http://netec.mcc.ac.uk), which is a collection of projects which aim to improve the usefulness of electronic networks in Economics, and GENUKI (http://www.genuki.org.uk), which provides information relating to the study of genealogy and family history.

The service continues to grow both in terms of the range of services offered and numbers of registered and active users.

The history of MIDAS and the Web

Historically, Manchester Computing has always made efforts to ensure that service specific information is available on-line. Prior to the installation of a UNIX platform for the MIDAS service in 1993, the on-line help systems were developed using proprietary tools and access was restricted to users logged onto the system. With the advent of a UNIX based service, the on-line information system moved to gopher. This client-server approach to providing access to information about the various datasets, software and other services available via MIDAS proved

extremely flexible. It could be used by users actually logged onto MIDAS as well as users with local access to a gopher client.

Until relatively recently, the gopher server was used as the primary method of providing access to information. With the advent of the Web, the initial function of the MIDAS home page was to simply provide an alternative interface to the information held in the gopher. However, the MIDAS Web site (http://www.mimas.ac.uk/) soon developed into a much more sophisticated information resource which was more suited to meeting the diverse information needs of data set users. Service specific Web pages were established; documents, reports and newsletters were made available in an HTML format; experimental Web interfaces to datasets, and information were also developed together with the provision of links to related Web based resources. Compared to the graphical user interface offered by the Web, the hierarchical text based gopher system soon started to look antiquated. As use of the Web and access to Web browsers became more widespread, it seemed appropriate to transfer the MIDAS information server from gopher to Web. Increasingly, the service specific MIDAS Web pages are starting to become information gateways in their own right. A good example are the Web pages relating to the statistical packages on MIDAS (http://www.mimas.ac.uk/stats/).

The decision to transfer the information server from gopher to Web was not greeted with universal acclaim. Initial user feedback indicated that whilst the majority of MIDAS users had easy access to Web browsers there was still a sizeable minority of users with only telnet access and/or without access to a Web browser who would be inconvenienced if the gopher service were withdrawn. Consequently, it was necessary to install lynx – the non-graphical Web browser - on the MIDAS UNIX server to provide an alternative method of accessing information held as Web pages on MIDAS.

Responding to changing user expectations

There are currently a variety of different interfaces to the various data and information resources available on MIDAS. In part, this reflects the specialist software requirements of many of the datasets. For example, the majority of the government and other continuous surveys are supplied to MIDAS in SIR database format. Indeed, some surveys, such as the British Household Panel Study (BHPS), are actually supplied in multiple formats (e.g. SIR, SPSS, SAS and STATA). Although some of the services are entirely Web based (e.g. JSTOR, COPAC and ONS Databank) or use specific client-server software packages (eg Beilstein CrossFire). the majority still require the user to physically log onto a remote UNIX server – either via TELNET or X-Windows - and run one or more application packages.

There are a number of problems with this mode of working. The user is required to have a basic competence with the UNIX operating system, a reasonable prior knowledge of the structure of the data as well as expertise in one or more specialist data extraction and analysis packages. Whilst this might not pose a problem for computer literate and network aware researchers, it is does represent a significant barrier to less experienced users, such as undergraduate students, across a broad range of disciplines who might have extremely limited data requirements.

The expectations of users have changed dramatically over the last couple of years and we have had to rethink how we deliver data and information to the researcher's desktop. It is not enough to promote awareness of the data resources and their potential applications in teaching and research. We also have to convince the users that their time is being used efficiently, that they can easily identify the data that they want, extract it and put it into a suitable format for secondary analysis and that they can obtain informed advice on types of use. For us this means creating appropriate interfaces for the data – simple enough for a once-off selection and versatile enough for more sophisticated use.

Over the last couple of years MIDAS has endeavoured to respond to changing user expectations by trying to develop more Web based interfaces to many of the data and information resources. A number of the Web based interfaces to MIDAS services have been developed in house using CGI programming techniques. For example, a Web based interface to the ONS time series databank for the UK has been developed (http://www.mimas.ac.uk/ons/) in order to facilitate greater use of the data in teaching. This simple interface permits searching by keyword, browsing of series and saving of extracted series in different formats. It is important to note that many of these Web interfaces represent additional interfaces rather than replacements for existing interfaces requiring a UNIX login. The aim is to provide more appropriate interfaces to new categories of users rather force existing users into new modes of working.

Whilst it is relatively easy to develop simple search and retrieval interfaces to data and information resources – providing it does not require too much restructuring of the underlying datasets - it is much harder to develop more sophisticated interfaces. MIDAS is currently exploring the use of java tools to develop a Web enabled version of the cartographic data visualiser (cdv) which uses scientific visualisation techniques for exploratory spatial data analysis (http://www.mimas.ac.uk/JANUS/cdv/). However, building complex Web based interactive data extraction and visualisation systems is currently resource intensive and might not be worthwhile given the advent of commercially supported products. In addition, it is clear from the feedback from users that the ease of use of an

interface is often more important than its functionality.

One of the major brakes on developing more Web based interfaces is the current absence of Web enabled versions of many of the packages currently used on MIDAS for providing access to data. This is changing gradually but perhaps too slowly for most users. For example, the SAS/ IntrNet software product (http://www.sas.com/software/components/intrnet.html) can be used to develop Web-enabled SAS applications running on a central server. Similarly, ESRI Map Objects Internet Map Server (http://www.esri.com) provides a set of tools that can be used for serving maps over the Web. Both of these products will be evaluated for use on MIDAS.

Is accessibility the only constraint?

The Web has a major role to play in improving accessibility to electronic data and information resources. Whilst accessibility is a major problem that requires addressing, it is not the only constraint on more widespread and effective use of data and information resources. For certain datasets, increased availability and improved interfaces have not necessarily resulted in the expected growth in numbers of users that might have been expected. One factor that plays an important role is the general lack of awareness amongst the wider academic community about the availability, content, scope, structure of key data resources combined with a lack of understanding about their potential application in both teaching and research. Although some users may have a good awareness about individual datasets this may not be matched by an equivalent understanding of complementary data resources. In addition, the level and type of use of certain data sets can also be adversely affected by a lack of appropriate secondary analysis skills. For example, an absence of spatial data handling skills frequently acts as a major constraint on the of use of key spatial data resources, such as 1991 Census digitised boundaries.

The combined problems of accessibility, awareness and usability are being addressed by the KINDS (http://www.mimas.ac.uk/kinds/) project with respect to the spatial data resources available on MIDAS (http://www.mimas.ac.uk/maps/). An integrated set of Web based tools have been developed which enable relatively inexperienced users to search, browse and work directly with large and complex spatial data sets, such as the Bartholomew's digital map data for Great Britain. These tools include a spatial search engine; data set browsing; a help system which provides access to a knowledge base of geographical terms and concepts together with mapping/ data download facilities.

Feedback from users indicates that the complex registration process for many of the copyright/commercial data sets held on MIDAS does act as a major deterrent to use – particularly for teaching purposes. Although registration is frequently seen as part of the cost of obtaining free or low cost access to commercially valuable datasets there is a need to negotiate unrestricted academic access to more sample datasets to facilitate the development of teaching and learning materials. If the web is to be used to build virtual learning environments it is important that it is populated with meaningful sample datasets.

Developing sustainable data services

Use of data and information resources in academic teaching and research will inevitably generate a range of derived materials. These derived materials could include derived data sets; program code/algorithms; methodological notes; data quality reports or teaching materials. It is regrettable that large amounts of this value added material are frequently discarded and not recycled back to other users. Whilst funding bodies such as JISC are concerned about the long term preservation of electronic materials that constitute the DNER the issue of building dynamic and sustainable databases of derived material has not received much attention to date.

MIDAS has worked with a number of projects and users wishing to make derived datasets and/or software tools more widely available – particularly where this adds value to the services hosted by MIDAS. This has happened particularly with respect to the 1991 Census area and interaction statistics held on MIDAS. An example is the 1981 and 1991 Census population surfaces and associated access software (http://census.ac.uk/cdu/surpop/). The KINDS project is also currently exploring developing databases of derived material relating to the spatial data resources available on MIDAS. On the basis of experience to date, it is clear that the Web offers great potential for building collaborative virtual user communities around key data and information resources. However, there are a considerable number of technical, legal, organisational, quality assurance and cultural issues that require resolution before such dynamic web enabled databases of derived materials can be developed.

Exploiting emerging technologies

Keeping in contact with both existing and potential users of MIDAS services and alerting them to new developments is a major problem. Not all users subscribe to the relevant email distribution lists at mailbase (http://www.mailbase.ac.uk), there are problems with cross-posting and managing these lists can also be problematic as email addresses constantly change. Similarly, not all users visit the 'Latest News' section of the MIDAS Web site on a regular basis. One possible solution is the use of push channel technology as an automated method of keeping users notified about new developments. A variety of different products are now available for delivering customised news feeds to users. However, use of push channel technology is still relatively underdeveloped in the UK academic community despite its potential.

Another emerging technology is the use of plugin and helper applications which embed additional functionality within Web browsers and also enable users to access files in more specialised formats. For example, the Adobe Acrobat Reader plugin enables users to view, navigate and print documents, such as electronic journal articles, held in PDF format. From the MIDAS perspective, one of the key benefits of plugin technology is the extent to which it can significantly reduce interface development times. However, less technically competent users do express concerns about the concept of having to locate, download and install plugins – although in the future these tasks may be performed automatically by the browser.

MIDAS has also been experimenting with use of other types of plugin. As part of an ESRC funded project, MIDAS has been developing an entirely Web based interface to 1991 Census area statistics. A major component of this interface is the map based front end to the 1991 Census area statistics database. By downloading and installing the Autodesk MapGuide Viewer (http://www.mapguide.com/) the user is able to embed limited desktop mapping functionality within the Web browser. This enables the user to interact dynamically with a multi-layered map which incorporates digital map data from the 1:1,000,000 Bartholomew Europe dataset and digitised 1991 Census output area boundaries. By panning and zooming the user is able to identify and select areas for which 1991 Census area statistics are to be extracted.

Accessibility for all?

It has become increasingly apparent that whilst new digital telecommunications technologies can significantly widen access to data and information resources on a world wide basis they can also serve to reduce accessibility for other groups. It is in that context that many organisations such as the World Wide Web Consortium (W3C) are looking at ways in which access and usability can be improved for people with disabilities (http://www.w3.org/WAI/) or with access to text only browsers. It is clear from the ongoing debate on accessibility and usability issues as part of the design of HTML 4.0 (http://www.w3.org/WAI/References/HTML4-access) that it will be increasingly important for data and information service providers, such as MIDAS, to look at the way in which documents are structured and whether there is an over reliance on graphics for information presentation (e.g. frames), navigation aids and/or resource discovery (e.g. clickable image maps). Graphics and multimedia have an important role to play but this should not be at the expense of textual content and/or network efficiency. However, in an era of rapid technological development there will always be a tension between the desire to exploit leading edge technology in interface development and the requirements of low technology users and/or those with special needs.

Conclusion

The Web has already made a tremendous impact on the way in which data centres, such as MIDAS, deliver data and information to the desktop. In order to respond to changing user expectations it will be strategically important for MIDAS to continue to develop more Web based interfaces to its services – although these might not represent the only access method. However, for certain datasets this may not be achievable in the short term due to the absence of Web enabled versions of all the key data access/analysis packages used by the service. Therefore, for certain data resources it may be necessary for MIDAS to develop its own Web based interfaces – which may require a restructuring of the underlying data formats.

Irrespective of the access method, the Web does offer considerable scope for promoting more widespread and effective use of the unrivalled data and information resources that have been made available UK academic community. In this context, it is vitally important that the JISC funded data centres and data services continue to develop and enhance the knowledge base relating to particular data and information resources and provide access to more didactic materials. This will facilitate more effective resource discovery as well as contributing to the development of virtual learning environments and hopefully start to narrow the gap between actual and potential use of the DNER in teaching and research.

Acknowledgements

Keith Cole and Anne McCombe both of whom work for MIDAS have made significant contributions to this paper. Keith Cole as well as being a manager of the MIDAS services is also Director of the ESRC/JISC 1991 Census Dissemination Unit and the Assistant Director (Technical) of the Cathie Marsh Centre for Census and Survey Research (CCSR) at the University of Manchester. Keith has presented a similar paper specifically geared to Social Scientists at IRISS ’98 in the UK.

References

MIDAS (http://www.mimas.ac.uk/)

Joint Information Systems Committee (http://www.jisc.ac.uk)

Distributed National Electronic Resource (DNER) (http://www.jisc.ac.uk/cei/dner_colpol.html).

COPAC (http://irwell.mimas.ac.uk/sj/)

SuperJournal Project (http://www.superjournal.ac.uk/sj/).

NetEc (http://netec.mcc.ac.uk)

GENUKI (http://www.genuki.org.uk)

Office for National Statistics Databank (http://www.mimas.ac.uk/ons/)

SAS/IntrNet software product (http://www.sas.com/software/components/intrnet.html)

ESRI Map Objects Internet Map Server (http://www.esri.com)

KINDS (http://www.kinds.ac.uk/kinds/)

MIMAS: The Spatial Side (http://www.kinds.ac.uk/kinds)

Surpop V 2.0: Introduction (http://census.ac.uk/cdu/surpop)

Mailbase (http://www.mailbase.ac.uk)

Autodesk MapGuide (http://www.mapguide.com)

Web Accessibility Initiative (WAI) (http://www.w3.org/WAI)

WAI Resource : HTML 4.0 Accessibility Improvements (http://www.w3.org/WAI/References/HTML4-access.html)

* Paper presented at the 1998 IASSIST/CSS Conference, Yale University, New Haven, USA, May 19-22, 1998, Julia Chruszcz, Keith Cole and Anne McCombe, University of Manchester, Manchester Computing, Oxford Road, Manchester M13 9PL.