Humanities and the Grid: Early exemplars and potential projects

Paul S. Ell
Director
Centre for Data Digitisation & Analysis
Queen’s Belfast

ISGC 2008
Summary

- The CDDA perspective
- e-Science in the arts and humanities
- Electronic Cultural Atlas Initiative Exemplar
- Irish Studies Exemplar – as a problem to be solved
- The way forward – or how to make the Grid relevant in the humanities
CDDA’s objectives

- To develop strategic humanities e-resources
- To use these resources in its own research and publish scholarly books and journal articles
- To develop methodologies that assist in the management and interrogation of the source materials to produce new perspectives and scholarship
Data Outputs

- Historical census data for Britain
- Welsh historical statistics
- Mortality statistics
- Hearth Tax Data
- Statistics on Religion
- Scottish National Dictionary
- Dictionary of the Older Scottish Tongue
- British Parliamentary Papers with BOPCRIS
- Database of Irish Historical Statistics
- Irish texts
- Key holdings from QUB Library Special collections
- British Parliamentary Papers referring to Ireland
- Act of Union Virtual Library including images and some OCR work
- Image scans of Latin texts for Ireland
- Stormont papers
- Historical diaries relating to China
- Convict database for Down County Museum, Living Linen
- JSTOR Irish Studies Library
- Total funded work = TW$350,000,000
Scholarly outputs

- Methodological work concerned with the application of new techniques: CUP GIS in Historical Geography, papers in Historical Methods, the Journal of the Royal Statistical Society and the International Journal of GIS.

- The use of e-resources in traditional scholarship: CUP book on Victorian Religion, Historical Geography etc.

- Books and papers combining new methodological approaches and e-resources: Counting Heads, Historical Atlas of Warwickshire, Mapping the Famine etc.
Grid technologies

The three aspects of e-Science are likely to have varying impacts in the humanities and arts

- **Access Grid**: Is this really distance learning with a better internet connection? Are humanities scholars going to change the fundamental way they do research?
- **Computation Grid**: Do humanities and arts scholars need high-powered computing power?
- **Data Grid**: The key technology that will fundamentally change scholarship in the humanities and arts.

This is reflected in an upcoming article in the *International Journal of Humanities and Arts Computing* by David Robey, Head of ICT with AHRC. He states:

There should be no doubt about the potentially transforming impact of e-Science on the A&H. The most obvious application, though not as we shall see the only one, is in data grid and related applications. The ‘data deluge’ in the A&H may be less of a problem than in the social sciences, but it is a real problem nonetheless.
Unique challenges in the humanities and arts: The Data Grid

- In the humanities the data grid is not as concerned with moving large amounts of data as in the sciences (although image databases can be large)
- It is more concerned with heterogeneous, fragmented, partial, disparate e-resources which are often small
- Information overload - the digital deluge
- Resource discovery problems
- Interface and data harvesting problems
- Integratory difficulties
- Data in ever more complex multimedia formats - not just text but numbers, images, objects, video, sound files
- How to organise data - by subject, by chronology, by location - or all three...
- But there are exemplars...
CDDA - a microcosm of the issues

- Significant funding invested in developing e-resources - from JISC, AHRC, BA, ESRC, and internal investment
- e-resources based on outstanding analogue sources and key research interests at QUB
- Issues relating to under use, maintenance, sustainability, access and interlinking
- Lack of investment in infrastructure in the UK - AHRC decision not to continue funding AHDS
The proof: early exemplars

The Electronic Cultural Atlas Initiative

- UC Berkeley-based project with almost 1,000 humanities and arts academic affiliates from around the world holding spatially referenced e-resources
- Metadata that allows registered distributed datasets to be retrieved on the fly at object level
- Software – TimeMap – which allows retrieved data to be selected and visualised and exported
ECAI Technology

IT Infrastructure Components

- The ECAI Metadata Clearinghouse
- TimeMap™ Software
- ECAI Servers
- ECAI User Support System and Knowledgebase

IT Architecture Goals

- Use distributed institutional resources
- Follow best practices for distributed internet computing
- Encourage interoperability
- Create custom solutions to link and enhance technologies when needed
- Create services and tools to support users
- Use a team structure to organize contributions of ECAI affiliates

Technical Methodology

- The technical architecture and technology design is developed collaboratively by participants who are scholars and technical specialists.
- The goal of allowing access to primary data and information through a spatio-temporal interface allows access to information that is multilingual, cross disciplinary, and from multiple institutions. The technology needed to accomplish this goal crosses many technical fields and paradigms.
- To allow access to distributed primary resources, ECAI is using a library model where resources are cataloged, but not centralized. Scholars, data and resources are distributed.

Technical Teams

ECAITech is the main group of individuals interested in the technical design and development of ECAI. In addition, there are several technical subcommittees focusing on aspects of the technology. The leaders of these groups comprise the ECAITech steering committee. ECAITech has a listserv for communication and meets regularly at ECAI conferences. Work of the technical groups is posted in the members' area of this site.
ECAI Metadata Clearinghouse

The Metadata Clearinghouse plays a central role in ECAI. One of the primary goals of ECAI is to be able to drill down through time and space and choose from a rich array of primary materials from multiple scholars and sources. Therefore, one of the primary functions of ECAI is to create a catalog of a wide range of cultural data and make it accessible in a single interface.

The system developed for managing and cataloging data is the ECAI Metadata Clearinghouse. ECAI's Metadata Clearinghouse includes records which describe datasets or map layers. ECAI metadata records include bibliographic and descriptive metadata for resource discovery allowing identification of resources relevant to a particular place, time or thematic interest. They also contain connection metadata for each dataset which allows software such as TimeMap™ to connect to the data server across the Internet (or to local files), drill down into the data and integrate results from different datasets (heterogeneous distributed database query.)

In order to make the clearinghouse globally usable, we are conducting research to make the clearinghouse multi-lingual. Part of our efforts include a project coordinated by the Japan team to translate the metadata records into Japanese. Therefore, it is important to note that records contributed to the clearinghouse may be translated into Japanese or another language as this research continues. If you have any questions about the Japan team project, please contact Shoichiro Hara, National Institute of Japanese Literature, Japan, hara@riji.ac.jp.

The ECAI Metadata Clearinghouse can be used to find and access data.

- Search the ECAI Metadata Clearinghouse

It can also be used to register data for others to use.
Sydney TimeMap™ at the Museum of Sydney

Sydney TimeMap™ is a computer kiosk at the Museum of Sydney, on the site of first Government House, launched in November 2000. It is the first product of a research partnership between the Archaeological Computing Laboratory at the University of Sydney, ESRI Australia and the museum. The kiosk uses TimeMap™ software to deliver digital resources related to the history of Sydney and was created by Ian Johnson, Andrew Wilson and the team at the ACL.

The kiosk is located in the museum’s information centre. It uses a dual-screen display to deliver maps and associated resources.

Users can create a map with linked resources by choosing from themes or from a timeline.
But there are problems with the ECAI model

- TimeMap geo-data browser is not robustly supported
- Bespoke ECAI metadata must be applied to datasets at object level to make them fully functional
- No way to automate the application of metadata
- Few datasets have been registered
Clearinghouse Search

1. Draw search area with:
   - [ ] Dataset Density: ECAI
   - [ ] View Mapspaces
   - [ ] Dataset Density: Rumsey
   - [ ] View Datasets

2. Use the time bar to select a time frame for the search

3. Text search
   - [ ] Ireland

4. [Search]

Search is carried out within the spatial/temporal extent and collections shown to the right.

Search Extent

- Longitude: -180 to 180
- Latitude: 90 to -90
- Time: -250 to 2005

Search Settings

- Search collections:
  - ECAI clearinghouse
  - David Rumsey map collection

- Sort by:
  - [ ] Spatial relevance

Find text in:
- [ ] Title or subject
- [ ] Include element descendants

Historical map georeference:
- [ ] Passable fit or better

Find area of interest with zoom and drag tools

Click MapSpace bounding boxes for interactive map

Reset
ECAI Clearinghouse Search Results

Refine search

Longitude 180 to 180
Latitude 50 to 0
Time 250 to 2005

Find Ireland

Search

Search within results

No datasets match your search criteria.

Metadata clearinghouse maintained by Archaeological Computing Laboratory, University of Sydney.
An exemplar problem: Irish Studies

- Poorly defined subject area
- No cohesive e-resources currently exist
- But quite a lot of e-resources data are there - Database of Irish Historical Statistics, Act of Union Virtual Library, Historical Hansard, JSTOR journals
- Challenge to bring these together
Reality Check

- e-Science pre-supposes sophisticated levels of information literacy and information skills
- e-resources represent a high challenge environment for the majority working in the humanities
- Need to create a controlled environment to develop skills necessary for e-Science
Developing a roadmap

- Few real examples of the potential of the Data Grid being realised.
- Vital need for e-infrastructure before the Data Grid will meet potential – place name gazetteers; chronological gazetteers; subject indexes
- Need for a geo-temporal data browser - TimeMap?
- Need for detailed metadata - Enhanced metadata or context sensitive intelligent searching
Roadmap Project I: Infrastructure based

- Develop a comprehensive place name gazetteer using the English Place-Name Society analogue gazetteer
- Test gazetteer using AHDS/ECAI holdings
- Cost – TW$40,000,000
- Provide hooks to link e-resources together and develop exemplar projects - Domesday II
- Other e-infrastructure developments underway with partners
Redmile

1. REDMILE

Redmelde 1086 DB, Redmeld' 1343 Cl, 1388 Pat et passim to 1400, 1413 Rut et freq to 1486, 1533 ib
Redmild' a.1166 (e.15) BelCartB, 1208 FF, e.13 (e.14) BelCartA et passim to 1236 Rut, 1242 Cl et freq to 1300, 1308, 1319 Rut, 1427 Terrier, Redmilda Stephen Rut, l.12 (e.14) BelCartA et passim to 1222 (e.15) BelCartB, Redmilde 1230 RHug, 1240 Rut et passim to l.13 (e.15) BelCartB, 1363 Ipm, Redmyld' 1285 ib, 1289 Rut et passim to 1404 Wyg, 1410 Rut et freq to 1525 ib, 1535 VE, 1544 Rut, (~ in le Wall 1519 Wyg), Redmyld 1397, 1421 Cl et passim to 1483 Rut, 1518 Wyg
Redemild' 1254 Val, 1292 Rut et passim to 1343 (e.15) BelCartB, 1363 Ipm, Redemilde 1233 (e.14) BelCartA, m.13 (e.15) BelCartB et passim to 1308, 1312 Rut, Redemyl'd m.13 (e.15) BelCartB, 1274 Ass (p) et passim to 1370, 1411 Rut, Redemylde 1300 ib, 1321 (1449) WoCart et passim to 1428 FA, 1440 Cl, Redemylde' Edw 1 Rut, Redemild 1300 IpmR, Redemyl'd e.14 BelCartA, Reddymyl 1427 Terrier, Reddemeld 1486 Rut
Redamelna Hy 1 (e.15) BelCartB, Redmelina c.1155 Dugd, Redmeline e.13 Rut, Redmelne Hy 2 ib, 1230 RHug, Redmlyn c.1155 Rut
Redmilla Hy 1 (1333) Ch, Hy 2 (e.15), p.1250 (e.15) BelCartB, Redmill 1239 Cur, 1599 Wyg, Redmyll 1519 Fine, Redemilla Hy 1 (1333) Ch, Hy 2 (e.15), 12 (e.15) BelCartB, Edw 1 BelCartA
Radmilda Hy 2 (e.15) BelCartB, Radmild' m.13 Wyg, 1359 Pat, Radmilde 1288 Coram, Radmyld' 1371 Rut, 1482, 1489 Wyg, Radmylde 1409 PReP
Redemilde 1294 AD, 1343 Ipm, Rademylde c.1306 Wyg, 1343 Ipm, 1343 Fine, 1352 Ipm, Rademeld(e) 1350 Pat
Radnell 1524, 1525, 1531, 1532 Wyg, Radmyll 1518, 1530, 1531 ib, (~ in le Vale 1530 ib), Radmyle 1502, 1521 ib
Roadmap Project II: Content based on Irish Studies

- Using our Irish Studies holdings which are both varied and controlled
- Where existing work will gather multi-media resources and formats
- Which is professionally developed and maintained
- Initial funding with UC Berkeley from NEH to test searching and linking on bibliographic data
IRISH BOOKS

quarrel of Protestant and Catholic, which so far has prevented Ireland from realizing either its political or literary ideals. What a difference it would have made had some clergyman or parish priest transformed an Irish country town or village—Westport or Doneraile—into an Irish Concord! A thinker who would have unsealed the fountains of thought and claimed the full privileges of a human being on the soil of Ireland would have brought it far further toward the realization of its spiritual and political unity than ever did Parnell. But Ireland’s hour was not yet come, and just as, after Davis, politics became estranged from those interests which ensure the support of the wise, so literature estranged itself from those common interests which make it genuinely national. There was, first, the well-meaning Ferguson, sometimes really exalted in his personal poems and adaptations from the Irish, but whose more ambitious labours smell a little too much of the Record Office. There was also the attenuated Wordsworthianism of Aubrey De Vere. Lastly there has been the considerable literary movement initiated chiefly by Mr. Standish O’Grady, styled variously the “Irish Literary Revival,” the “Celtic Renascence,” &c., in which the chief factor has undoubtedly been the peculiar genius of Mr. W. B. Yeats.
has undoubtedly been the peculiar genius of Mr. W. B. Yeats. The meeting in modern Ireland of the modern with the ancient spirit is an important event, not only in the literary but in the spiritual history of Ireland, and perhaps the full significance of the work of Mr. Yeats and AE will only be apparent eventually. Mr. Yeats in particular understands the ancient Celtic spirit as Ronsard understood Græco-Roman antiquity, and is imbued with it in much the same way; and just as it was only when the modern world had learned to understand the ancient classics that it began to strike out in every direction on lines of its own, so perhaps the spirit of Ireland, through its self-recovery in this last poet of the line of Senchan Torpeist, is being made ready for new beginnings.

Meanwhile if we ask whether the voluminous literary
Named entities are linked to specific resources or dynamic searches over relevant databases.
Named entities not detected automatically can be added manually.
Conclusions

- The Data Grid will be the key area of e-Science activity in the humanities and arts
- Data Grid based e-Science in the humanities and arts is challenging
- Key infrastructure is required together with enhanced search capabilities
- Spatial organisation of data is poorly established in the humanities
- Chance to fully exploit the vast array of e-resources already available
- Opportunity for fundamental change in humanities and arts research
- Projects will be costly and the climate may not be right for infrastructural projects
Integrating e-resources through the data grid: statistics, maps, photographs, text, manuscripts, existing e-resources, websites, museum objects...