



## JISC Project Plan

### *Overview of Project*

#### 1. Background

As part of its £22m Digitisation Programme, JISC has funded the creation of a range of well used cultural and scholarly content in research, learning and teaching. The purpose of this project is to leverage this investment in those resources by using geography as a means to cross reference and unify. In this respect, geographical referencing has the capacity to provide an entry point into a wealth of other JISC digitised content in a similar way that Google is using geography as an 'organising principle' for its resources.

Much of this content is rich in geographical information (names of places, regions, and countries, plus other geographical information such as rivers, mountains etc), whether this information be embedded in the metadata (such as terms that describe audio items in the British Library's' Archival Sound Recordings) or within the digitised texts itself (such as in BOPRCRIS' Eighteenth Century Parliamentary Papers).

Traditionally, it has been difficult to exploit the richness of this geographical information. However, recent developments in natural language processing (which allows for digital identification of geographical information) and a developing infrastructure for delivering such information via the web has allowed for further work in this area.

#### 2. Aims and Objectives

The aim of this project is to demonstrate the value of geographically enriching metadata for the purpose of augmenting resource search and discovery methods.

The objectives are:

- To ingest diverse metadata (including the textual content of a resource) and to semi-automatically enhance this with explicit geographical referencing
- To build show case demonstrators illustrating the potential afforded by the enrichment process
- To evaluate the efficacy and accuracy of the methodology and its applicability and potential for enhancing other JISC collections

#### 3. Overall Approach

This particular project tasks the JISC Data Centre at EDINA, in association with the University of Edinburgh's Language Technology Group (LTG), with enriching the geographical information held by three of the extant JISC's digitisation projects. These projects jointly contain millions of words of text and implicit reference to geography via placenames. Moreover it is only via the use of specialised processing techniques such as proposed here, that the potential power afforded by explicit geographical referencing can be unlocked. These projects are:

- **Histpop (History Data Service,HDS) - The Online Historical Population Reports – <http://www.histpop.org>**
- **BOPCRIS 18th Century Parliamentary Publications - [www.parl18c.soton.ac.uk/](http://www.parl18c.soton.ac.uk/)**
- **Archival Sound Recordings - <http://sounds.bl.uk/>**

Project Acronym:  
Version:  
Contact:  
Date:

Each of the three above projects will provide EDINA with their relevant meta/data including existing geographical information. EDINA will then make use of the experimental GeoParser tool (developed by the University of Edinburgh's Language Technology Group), to enrich the data by identifying place-names and adding explicit georeferences (co-ordinates) to the metadata.

The enriched data for two of the projects (BOPCRIS and Histpop) will then be embedded in an experimental geographical service to be hosted by EDINA. This will allow users to search resource collections via a map-based interface and provide links back to the citation of the place-name in the original resource.

The third enriched data set (Archival Sounds) will be returned to the BL and a geographical resource discovery service built alongside the BL's existing ASR web interface.

## 4. Project Outputs

The following will form the key outputs from the project:

1. Enriched data sets for the three digitisation projects cited above:
  - These enriched data sets will have identified place names explicitly georeferenced i.e. geographical co-ordinates added to the metadata.
  - Each source of enriched data (excepting the APR data which by default is publicly accessible and falls outside of the ac.uk domain for which Ordnance Survey licensed data is constrained) will have two versions – one utilising Ordnance Survey co-ordinates and one using IPR free coordinates (taken from the geonames.org CC licensed data source). The use of the latter, while not as extensive as OS co-ordinates, will allow for the data to be exploited in open access scenarios (the use of OS co-ordinates is currently restricted by rights restrictions).
  - Optionally, if other aspects of data can be identified (e.g. currencies, persons) they may be added to the enriched metadata
2. An end-of-project evaluation and summary report making suggestions for further work if appropriate,
3. Two versions of an experimental map oriented service providing joint geographical access to BOPCRIS and Histpop (to be delivered via EDINA, but pointing back to the original resources). The OS version will be protected via Shibboleth; the other will be open access.
4. An experimental service providing geographical access to Archival Sounds (to be delivered by the BL as part of its Archival Sounds resource) and based upon IPR free enriched metadata.

## 5. Project Outcomes

The principle outcome will be an evaluation of the methodology employed for metadata enrichment and an assessment of the broader utility of georeferencing extant digitised resources. This will be used to communicate findings to a broader audience and to highlight relevance to other JISC projects and activities.

## 6. Stakeholder Analysis

Stakeholder	Interest / stake	Importance
Community user (researcher; student)	Ability to discover/locate resources	High
General member of the public	Ability to discover/locate resources	Medium
JISC	Assessment of geobrowsing as a resource discovery method Facility to enhance other	High

Project Acronym:  
Version:  
Contact:  
Date:

	extant resources and leverage investment Illustrate project synergies and interoperability via use of geography	
--	--	--

## 7. Risk Analysis

Risk	Probability (1-5)	Severity (1-5)	Score (P x S)	Action to Prevent/Manage Risk
Staffing	1	4	4	Staff contracts. Mitigated by short timescale of project
Organisational - Project partners fail to agree on work plan	1	4	4	Early consultation and consensual agreement to project plan
Technical – GeoParser performance inadequate to provide critical mass of correctly georeferenced material	3	4	12	Prior experience with BOPCRIS material and capacity to ‘tune’ the NLP software to nuances of particular data sources
External metadata suppliers fail to provide resource	2	4	8	Project buy-in by being project partner
Legal – IPR forbids use	1	5	5	Constrained within ac.uk and release of non-IPR enriched metadata for public consumption

## 8. Standards

Relevant web standards and those supported by majority browsers will be used for the demonstrators. The project will aspire to conform to JISC endorsed standards.

## 9. Technical Development

The principal technology deployed will be based on Natural Language Processing tools developed by the University of Edinburgh’s Language Technology Group.

Interface development for the demonstrators will be subject to technology assessment at build time but will aim to use best of breed open source technology solutions.

## 10. Intellectual Property Rights

As described in the JISC Generic Terms and Conditions of Grant, project partners recognise that the ownership of intellectual property rights made, discovered or created during the period of project funding will be indicated to them in the letter of grant. Communicating with partners and others involved in the project will ensure IPR issues are minimised. For project outputs including reports, JISC will be allowed to utilise, archive and disseminate the work in accordance with current JISC policy.

## *Project Resources*

## 11. Project Partners

Project Acronym:  
Version:  
Contact:  
Date:

*Partner:* EDINA  
*Role:* Consortia lead  
*Contact:* James Reid [James.reid@ed.ac.uk]

*Partner:* Language Technology Group  
*Role:* subcontracting NLP expertise  
*Contact:* Claire.Grover [grover@inf.ed.ac.uk]

*Partner:* HistPop, History Data Service (HDS)  
*Role:* Data provider, knowledge of historical geographical entities and disambiguation, empirical reviewer  
*Contact:* Matthew Woollard [matthew@essex.ac.uk]; Richard Deswarte [richardd@essex.ac.uk]

*Partner:* BOPCRIS  
*Role:* Primary Contact  
*Contact:* Julian Ball [J.H.Ball@soton.ac.uk]

*Partner:* British Library  
*Role:* Primary Contact  
*Contact:* Adrian Arthur [Adrian.Arthur@bl.uk]

NB. Given duration and scale of project it is unlikely a formal Consortia Agreement will be required.

## 12. Project Management

EDINA will act as the lead partner for the purposes of project administration and finance. Overall responsibility for the project will rest with senior staff at EDINA. The project will be co-ordinated by a Project Director, Dr David Medyckyj-Scott and a Project Manager, James Reid, based at EDINA. A work package focusing on project management will be produced as part of the project plan. A series of full project meetings will take place at project kick-off, mid-term and close. To limit the T&S outlay, the majority of communications will be conducted virtually by video/teleconferencing. Day-to-day communications amongst project members (both management and technical) shall be conducted by email and phone.

The Project Team will consist of:

EDINA (lead)  
Project Director, Dr David Medyckyj-Scott  
Project Manager, James Reid

Subcontracting staff at University of Edinburgh Language Technology Group (LTG)  
Dr Claire Grover  
Dr Richard Tobin

HistPop,HDS  
Project Lead Contact, Matthew Woollard, Associate Director, Head of Digital Preservation and Systems, UKDA  
Project Contact, Richard Deswarte, Social History Data Manager, UKDA  
Evaluation staff TBA.

BOPCRIS

British Library

Project Acronym:  
Version:  
Contact:  
Date:

### 13. Programme Support

No special arrangement required. Support will be via direct liaison with the JISC Digitisation Programme Manager.

### 14. Budget

See Appendix A.

## Detailed Project Planning

### 15. Workpackages

See Appendix B.

### 16. Evaluation Plan

Timing	Factor to Evaluate	Questions to Address	Method(s)	Measure of Success
End	Efficacy of georeferencing each data source	What is precision and recall performance? How well does disambiguation perform?	Statistical and qualitative	F score User assessment
End	Efficacy of georeferencing each data source	How well has the georeferencing performed?	Empirical	Comparison to sampled hand proofed resources at HDS

### 17. Quality Plan

Output Timing	Quality criteria	QA method(s)	Evidence of compliance	Quality responsibilities	Quality tools (if applicable)
Throughout	Efficacy of geoparsing	Recall and precision measurements	Test scores	LTG	NLP tools
Throughout	Interface	Usability assessment	Inhouse testing/consultant (BL)	EDINA	

### 18. Dissemination Plan

Timing	Dissemination Activity	Audience	Purpose	Key Message
End	reports	All stakeholders	Synthesis of project activity and results	Utility and value of georeferencing resources
End	Publicity to non-stakeholders	Technical and non-technical	Disseminate findings and activity more broadly beyond	Usefulness and benefits of georeferencing materials

Project Acronym:

Version:

Contact:

Date:

			confined stakeholder group.	
--	--	--	-----------------------------	--

## 19. Exit and Sustainability Plans

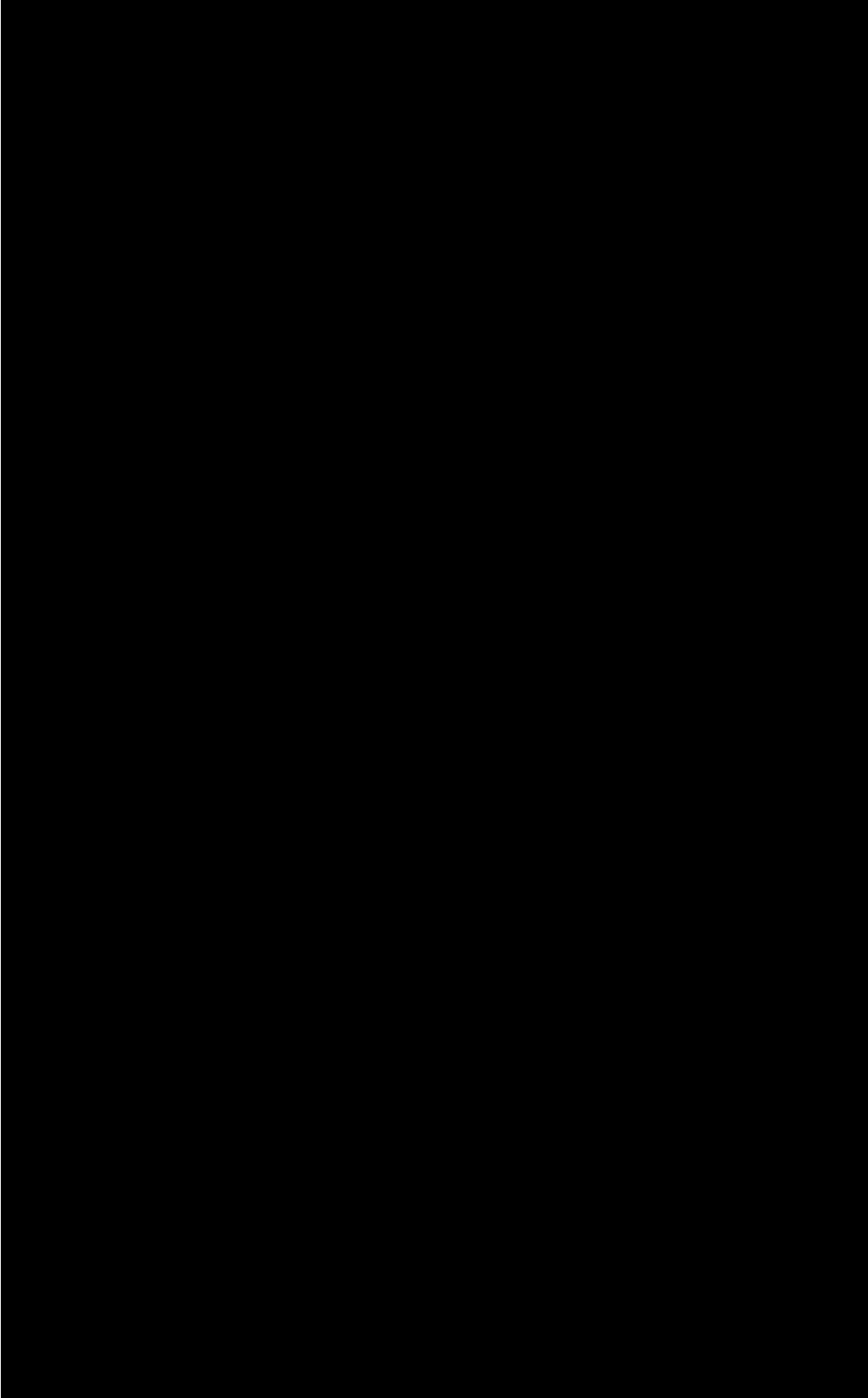
<b>Project Outputs</b>	<b>Action for Take-up &amp; Embedding</b>	<b>Action for Exit</b>
Demonstrators	Available for at least 12 months after project completion	
Reports	Available via partner sites and JISC	
Enriched metadata	Available to submitting partners for curation	

<b>Project Outputs</b>	<b>Why Sustainable</b>	<b>Scenarios for Taking Forward</b>	<b>Issues to Address</b>
Enriched metadata	Georeference persists with resource metadata (assuming IPR cleared)	Reingest into host systems and modify host interfaces to exploit georeferences	Compatibility of host systems New interfaces to exploit enriched metadata

Project Acronym:  
Version:  
Contact:  
Date:

## ***Appendixes***

### **Appendix A. Project Budget**

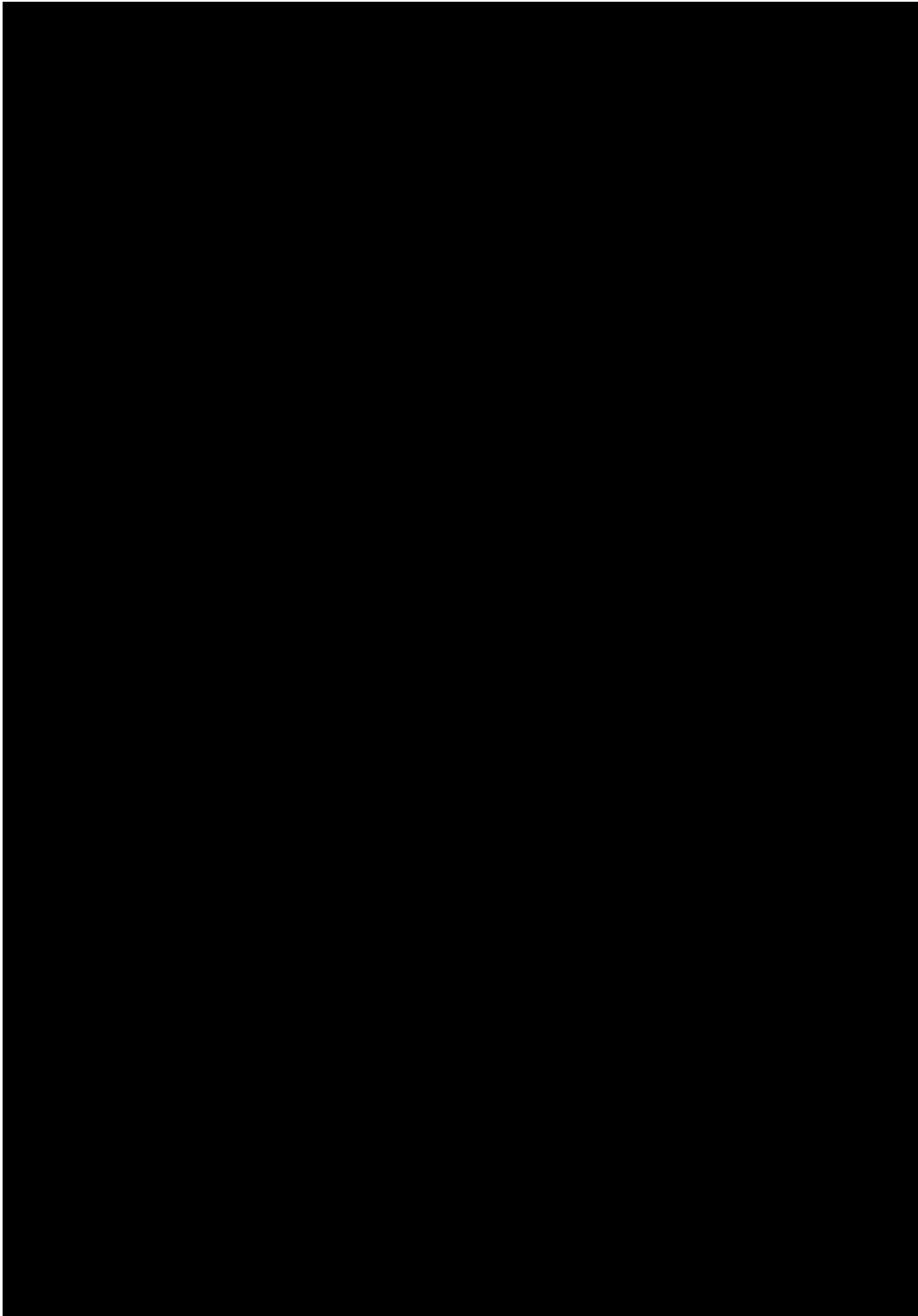


Project Acronym:

Version:

Contact:

Date:



Project Acronym:  
 Version:  
 Contact:  
 Date:

## Appendix B. Workpackages



<b>WORKPACKAGES</b>	<b>Month</b>	1	2	3	4	5	6
		Oct	Nov	Dec	Jan	Feb	Mar
1: Project Management		X	X	X	X	X	X
2: Data Assessment & Output Specification Development		X	X	X			
3: Data preparation and delivery		X					
4: GeoParser development		X	X	X	X	X	
5: Iterative Geoparsing / Georeferencing and assessment of outputs				X	X	X	
6: Final output run and data delivery						X	
7: Interface development EDINA BL				X	X	X	X
8: Testing & Formal Evaluation						X	X
9: Reporting							X

Project start date: *1<sup>st</sup> October 2008*

Project completion date: *31<sup>st</sup> March 2009*

Duration: *6 months*

				Milestone	Responsibility
<b>YEAR 1</b>					
<b><i>WORKPACKAGE 1: Project Management</i></b>	1/10/08	31/3/09			
<b><i>Objective: Ensure deliverables are met and that project remains on track</i></b>					
1. Create and agree project plan*	1/10/08	31/10/08	Agreed Project Plan	Yes	EDINA
2. Organise meetings and liaise with stakeholders*	1/10/08	31/10/08			
3. Oversee subcontracting and administrative arrangements	1/10/08	31/10/08			
4. Reporting (interim) and dissemination	1/1/09	31/1/09	Project status updates to stakeholders	Yes	EDINA
5. Progress monitoring	1/10/08	31/3/09			All
6. Communication	1/10/08	31/3/09			All
7. Reporting	1/2/09	31/3/09		Yes	EDINA/All
* Given timescales involved, these tasks will be performed prior to official project commencement (1 <sup>st</sup> October)					
<b><i>WORKPACKAGE 2: Data Assessment &amp; Output Specification Development</i></b>					
<b><i>Objective: To review existing metadata formats from data suppliers and to agree output specification</i></b>					

Project Acronym:

Version:

Contact:

Date:

1. review sample of data supplied from each source	1/10/08	31/10/08			
2. agree amongst stakeholders exact specification for output	1/11/08	31/12/08			
3. Iteratively test and review	1/10/08	1/12/08			
4. Finalise and sign off agreed output specification	1/12/08	28/02/09	Agreed Output Specification	Yes	ALL
<b>WORKPACKAGE 3: Data preparation and delivery</b>					
<u>Objective:</u> To deliver data (all or significant representative sample) to EDINA/LTG for review and testing purposes					
1. Preliminary data delivery and assessment	1/10/08	31/10/08		Yes	All
2. Take delivery of data sample and review	20/10/08	31/10/08		Yes	All
3. Once Output Specification Agreed ,	20/10/08	28/11/08	Interim testing to agreed Output Specification	Yes	LTG/EDINA
4. Take receipt of complete data series from each data source					
<b>WORKPACKAGE 4: GeoParser development</b>					
<u>Objective:</u> To refine and tune the GeoParser to take account of source specific data quirks					
1. Run GP software tools over sample data and review performance	1/10/08	28/02/08			
2. Determine scale of tuning required per data source	1/10/08	31/12/08			
3. Refine software to accommodate issues	1/10/08	31/1/09			

Project Acronym:  
Version:  
Contact:  
Date:

identified above					
4. Iterative development and testing	1/10/08	28/02/09			LTG
<b>WORKPACKAGE 5: Iterative Geoparsing / Georeferencing and assessment of outputs</b>					
<u>Objective:</u>					
1. In conjunction with WP4 and WP2 determine optimum output formats	1/12/08	28/02/09			LTG/EDINA
2. Stakeholder review of outputs to inform iterative development	1/12/08	31/01/09			All
3. Formal assessment of GP performance (F-Scores, precision, recall)	1/02/09	28/02/09		Yes	LTG
<b>WORKPACKAGE 6: Final output run and data delivery</b>					
<u>Objective:</u> To produce ingest ready outputs to agreed Output Specification (WP2)					
1. Run GP over each data source in its entirety	1/02/09	28/02/09			
2. Deliver outputs to stakeholders in agreed format(s)	1/02/09	28/02/09	Final, full version output data delivered to stakeholders	Yes	LTG/EDINA
<b>WORKPACKAGE 7: Interface development</b> EDINA BL					
<u>Objective:</u> parallel development of map based					

Project Acronym:  
Version:  
Contact:  
Date:

search and discovery interfaces (BL/EDINA)					
1. Determine functional requirements for interface	1/12/08	31/12/08		Yes	EDINA/BL
2. Data receipt and verification. Ingest into host back end systems*	1/12/08	28/02/09		Yes	EDINA/BL
3. Agreement on technology options for interface requirements based on functional requirements	1/12/08	31/1/09		Yes	All
4. Build, test and deploy cycle	1/1/09	31/3/09			
5. Debug and review	1/1/09	31/3/09			
* may be based on sample data to agreed Output Specification to permit early design of interfaces					
<b>WORKPACKAGE 8: Testing &amp; Formal Evaluation (iterative)</b>					
<u>Objective:</u> To review and evaluate methodology and utility of georeferencing					
1. Empirical review of data sample to iteratively evaluate accuracy and utility of the geoparsing/georeferencing process	1/2/09	31/3/09	Formal report from HDS based on manual inspection of sample output	Yes	HDS
2. Qualitative assessment of geobrowsing interfaces for resource discovery	1/2/09	31/3/09			All
3. Formal summary of geoparser performance	1/2/09	31/3/09			LTG/HDS

Project Acronym:

Version:

Contact:

Date:

<b>WORKPACKAGE 9: Reporting and sign-off</b>  <u>Objective:</u> To synthesise findings and make recommendations for future work					
1. Produce final report	1/3/09	31/3/09	Final summative report	<b>Yes</b>	EDINA
2. Present findings to stakeholder group	10/3/09	31/3/09			All