Project Document Cover Sheet

Project Information			
Project Acronym	ASSIST		
Project Title	ASSIST – Education Evidence Portal and Media Frame Analysis		
Start Date	1st April 2008	**End Date**	30th September 2009
Lead Institution	University of Manchester		
Project Director	Sophia Ananiadou		
Project Manager & contact details	Brian Rea NaCTeM, Manchester Interdisciplinary Biocentre, 131 Princess Street, M1 7DN, Manchester. +44 (0) 161 306 3096		
Partner Institutions	EPPI and NCeSS		
Project Web URL	http://www.nactem.ac.uk/assist/		
Programme Name (and number)	e-Infrastructure		
Programme Manager	James Farnhill		

Document Name			
Document Title	Project Progress Report		
Reporting Period	June – November 2008		
Author(s) & project role	Davy Weissenbacher – Developer, Research Associate Brian Rea – Project Manager Sophia Ananiadou – Primary Investigator		
Date	30/11/08	**Filename**	ASSIST_Overall_Progress1_ShortV3.doc
URL	-		
Access	√ Project and JISC internal		☐ General dissemination

Document History		
Version	**Date**	**Comments**
1.0	30/11/08	Original submission document

JISC

Project Name	*ASSIST, NaCTeM University of Manchester*
Project Website	http://www.nactem.ac.uk/assist/
Report compiled by	*Sophia Ananiadou, Brian Rea, Davy Weissenbacher*
Reporting period	*June – November 2008*

Section One: Summary

The project has seen good progress over this period. We have constructed an alternative corpus from the LexisNexis database, large enough to guarantee valid results from our statistics-based Text Mining (TM) tools. We have encountered a well-known problem, missing or inconsistent metadata, and propose two solutions to tackle it. Consistent with the timeline, a demonstrator expanding on the ASSERT system and integrating a Named Entity Recognizer has been adapted to our corpus and has been deployed. In collaboration with our partner, we currently are in the process of qualitative evaluation of this demonstrator. This evaluation will lead us to enrich its design and improve access to document content for the sociologist. This evaluation is also an opportunity for us to define in collaboration with our stakeholder the hierarchy of named entities needed. The learning stage of the extra named entities will start in December.

Section Two: Activities and Progress

The initial objectives for the first 4 months were to create a corpus and process it with a baseline system (WP4).

The corpus provided by our partner (NceSS) as WP3 deliverable contained too few documents to allow correct behaviour of the text mining tools based on statistical information. We decided to extend the corpus according to the user requirements (T4.1). We have queried the LexisNexis newspaper database to collect a set of 4889 documents with a query defined by our partner (the lexical terms and the span time have been given to us). The metadata needed for the indexing of our documents has been computed automatically with the help of the document format produced by the database.

To process the corpus we have adapted the existing system ASSERT (T4.2). This system was initially designed to process abstract-length articles from publications. The features used to access the content of these documents are not suitable for the sociologist. Where the user of ASSERT would generally search for a specific piece of information in a restricted set of documents, the sociologists in ASSIST are more interested in an intelligible overview of a large collection of documents to discover the frames. The system has been adapted to process a new set of metadata (new date format, a quick summary of the content, a list of keywords, etc.). These metadata have been indexed to allow the user to query these pieces of information.

As specified by the work package (T4.3) we have designed a new dynamic web site as a prototype system to demonstrate it to our partner. We have deployed a beta version of our system to start a discussion with our stakeholders in order to define a suitable interface for frame analysis and to qualitatively evaluate the contribution of each piece of information for frame analysis. The web site has been secured and restricted access given to our partner. Access for other stakeholders will be made available upon request. The internal report ASSIST-D3 details the installation instructions and describes how to use it in a tutorial way..The internal report ASSIST-D1 provides a technical presentation of the system and its Text Mining components.

During October and November, we have selected and integrated into the ASSERT system an existing Named Entity Recognizer (NER) called BaLIE[1], this way completing the deiverable T6.1. BaLIE is open source java software released under GNU GPL licence. This makes it easily integrated in our system. This NER uses a semi-supervised learning algorithm to extend the gazetteer it applies to a document to recognize NEs. This reduces the number of examples needed to adapt the NER for new categories of NE and consequently the time spend to annotate these examples. The author claims similar performances of its system compared with a full supervised learning based system. Currently we use the default gazetteers designed for general NE and the base-line version of the system. Next month, at the request of our partner we will extend the task of the deliverable T6.1. Taking advantage of the stakeholders evaluation of our system we will perform a NE learning stage to extend the hierarchy of the NEs recognized by the NER to match up the specific needs of our partner. We will lead a full evaluation of the extended NER on our corpus.

Through our close involvement with the National Centre for e-Social Science (NCeSS) during this project and the previous ASSERT project, we have co-organised a workshop on text mining in the social sciences as part of the 4th International Conference on e-Social Science. We used this as a core dissemination activity for the project.

Section Three: Institutional & Project Partner Issues

[1] Website: http://balie.sourceforge.net

Discussions have taken place with Elisa Pieri to anticipate the further semantic metadata enrichment (WP6).

The first enhancement is the integration of the Named Entities module (T6.1). Text mining tools usually use a simple hierarchy of Named Entities which contains: person names, association or company names, and dates or amounts of money. But we can also extend this hierarchy to include names which refer to other objects (e.g. names of buildings, names of military planes...) or set in detail the categories of existing named entities (e.g. governmental association, name of a fiction person...). Because the adaptation of the named entity recognizer is time consuming, we must define, in advance, the most suitable hierarchy according to our partner's requirements. The discussion is still in progress.

The second enhancement is the annotation of the NaCTeM corpus with semantic frame metadata (T8.1). In order to define these annotations, we have collected a sample of the current codes applied by Dr. Pieri to discover the ID frames in the NCeSS corpus (T7.1). A first result is the large variation in the type of annotations added to the document. Some codes are simple and can be automatically annotated (e.g. synonyms of the word 'public' or variation of the named entity 'Sir Crosby'). Others are complex and can only be defined and added by the sociologist (like passages describing the information that can be stored on an ID). At this time, the set of annotations that we can automate is unknown. We continue our analysis in collaboration with our partner to define these annotations.

Section Four: Outputs and Deliverables

Presentations on the ASSIST Project were given in the "Text Mining and the Social Sciences" Workshop *in conjunction with the 4th International Conference on e-Social Science. This presentation gives an overview of the system architecture and discusses how this system can improve the research work for the sociology community. Available at:* http://www.nactem.ac.uk/assist/slides.htm/

Base System prototype (D4.3): A web demonstrator showing how the ASSERT based system classified mass-media documents. This uses the NaCTeM alternative corpus described in Section Two. Available at: http://nactem3.mc.man.ac.uk:8080/ASSIST/

Section Five: Outcomes and Lessons Learned

It has been noted that the license conditions on the LexisNexis documents are still an issue we have to deal with for the creation and the demonstration of the results of our tools on this corpus. This issue has to be addressed as early as possible to not delay progress on the rest of the project. It should be advised that stakeholders provide an alternative corpus from the beginning of the project.This was addressed in our initial risk analysis, but the multiple resources we should have in place feature similar license conditions. Each must be handled individually.

To perform a qualitative analysis of written documents the sociological community has acquired its own tools. These tools are gathered under the generic name CAQDAS tools. As an annotation tool, a CAQDAS tool can be thought of as a direct competitor of the Text Mining based system implemented during this project. However, a preliminary study shows that the annotation added during the analysis is complex and manually applied to the document. The automatic calculation of these annotations is currently impossible with existing Text Mining tools. Consequently not only are CAQDAS tools not a competitor of our system but they can, in a future work, be embedded as a complementary tool within it.

Section Six: Evaluation

Aside from ongoing evaluation activities, with the collaboration of our stakeholders, we have started a qualitative evaluation of the baseline system. In the form of an open discussion, we have asked our partner about the choice of metadata and pieces of information displayed and the ergonomics of the web site. The following list presents the questions:

- *The query interface*
    - *user guide (bubble help, list of metadata available for the query (Named Entity, title, author...), commented example of query (available operators and wildcard characters))*
- *The list of documents*
    - *metadata of interest, document snippet, membership of a cluster (using coloured highlighting or numerical references)*
    - *pieces of information accentuated (colour, font, disposition...), default value (number of terms in the list, threshold for the weight of the terms)*
    - *management of missing metadata (a default missing value, removing incomplete documents from the list, interface for user correction)*
- *The list of clusters and their visualisation*
    - *pieces of information associated with a cluster and their default values (use of the 'all documents' cluster and the 'other documents' cluster,...)*
- *The document interface*
    - *pieces of information highlighted (Named Entities, terms above a certain threshold, terms of the query,...)*
    - *addition of snippet or metadata in the list of related documents*
    - *mention the membership of the document in a set of clusters*

The result of this questionnaire will emphasize the important information and lead us to redesign the demonstrator interface. We assume that this new design will improve the sociologist's analysis by facilitating access to the document content.

Section Seven: Dissemination

On 18th June 2008 NaCTeM and NCeSS ran a workshop on 'Text Mining in the Social Sciences' as part of the 4th International Conference on e-Social Science. Attendance was good with a number of presentations from different groups around the UK using text mining methods in social science research. A full programme can be found at http://www.ncess.ac.uk/events/conference/programme/workshop2/ *and copies of the presentations can be found at* http://www.nactem.ac.uk/assist/slides.htm*. To allow for wider dissemination videos of the full presentations will also be available through the ReDReSS project (*http://redress.lancs.ac.uk/*)*

Communications have been addressed to different conferences organized by the sociological community. During these communications we intend to advertise the demonstrator to the community. We emphasize on the complementarities between the CAQDAS tools used by the sociological community for the media qualitative analysis and our Text Mining based demonstrator. Showing the improvement in their daily work expected from the use of Text Mining software we are collating a list of interested users to target future dissemination activities and to widen the community actively engaged with text mining in the social sciences.

In collaboration with our stakeholder NCeSS we have submitted 5 shorts papers for the Media, Communication and Cultural Studies Association (MeCCSA) conference (14-16 January 2009, Bradford)[2]. The papers have been accepted after reviewing and a panel is now available for our communication.

A short paper has been submitted for a 30-minute presentation in the fourth International Conference on Interdisciplinary Social Science (8-11 July 2009). We currently await a reviewing decision.

Section Eight: Risks, Issues and Challenges

During the collection of the alternative NaCTeM corpus, we have encountered a well-known problem in the community. The documents in our corpus are extracted from the LexisNexis database and present different formats to access the metadata. Some formats are rare and have not been processed by our general interface. As a consequence, some metadata (like the title or the date of publication) are missing for a subset of documents in the index. These pieces of information cannot be queried, and these documents are ignored in the result, even if they are pertinent. A solution to tackle this issue is to implement a specific interface for each format. However, this is a partial solution. For each new format we have to extend the implementation of the preprocessing interface. We propose a second solution, more pragmatic. We intend to integrate in the next version of the demonstrator a new interface allowing the user to correct on-line the missing or noisy pieces of information.

The building of the alternative corpus from LexisNexis raises a licensing issue on the documents. According to their license we are allowed to download our corpus, but we cannot keep it on a hard disk more than 90 days, and we cannot publicly demonstrate the document processed by our system. Our solution to avoid the time constraint for the use of the document is to renew the corpus each 90 days (the query to collect the documents can be saved and the pre-processing stage is automated). To deal with the demonstration constraint, NCeSS is currently in discussion with the 'Guardian' newspapers to buy the data which could be publicly demonstrated.

[2] Website: http://www.meccsa.org.uk/

Section Nine: Collaboration and Support

More and more numerical documents are accessible via database on line (Medline[3], Citeseer[4], LexisNexis[5]...). This allows us to collect a huge corpus. However, the formats of the documents provided by these databases are heterogeneous and partial. The consequence is an unreliable pre-processing stage for these documents, i.e. some pieces of information are missing for some documents whereas they are expected (see the previous section). The point we would like to discuss with other projects is the existence of a plan to normalize the representation of the documents. Such standardization would make possible the implementation of a good quality API to collect documents and, as a consequence, improve the natural language processing of these documents.

Section Ten: Financial Statement

Formatted: Don't snap to grid

Section Eleven: Next Steps

- *integration of the qualitative evaluation by our partner to improve the design of the demonstrator*
- *extending the hierarchy of Named Entities according to the needs of our partner (T6.1 extension)*
- *integration of the anaphora resolution system in the pipeline (T6.2 – due to 01/01/09)*

[3] Website: http://www.ncbi.nlm.nih.gov/pubmed/
[4] Website: http://citeseer.ist.psu.edu/
[5] Website: http://www.lexisnexis.com/

Project Name	*ASSIST, NaCTeM University of Manchester*
Project Website	http://www.nactem.ac.uk/assist/
Report compiled by	*Sophia Ananiadou, Brian Rea and Davy Weissenbacher*
Reporting period	*June – September 2008*

Section One: Summary

The project has seen good progress over this period. We have prepared a corpus of 1300 heterogeneous documents from educational websites. We have encountered a well known problem, that of the missing or inconsistent metadata, and propose different solutions to tackle it. According to the time line, a demonstrator expanding on the ASSERT system has been adapted to our corpus and has been deployed. In collaboration with our partner, we are currently in the process of a qualitative evaluation of this demonstrator. This evaluation would lead us to enrich its design and improve the access of the document content for the user of the search engine.

Section Two: Activities and Progress

The initial objectives for these 4 months were to prepare a corpus for the development stage and process it with a baseline system (WP4).

The corpus provided by our partner (EPPI) as WP2 deliverable is composed of 1300 'PDF', 'Microsoft Word' and 'XHTML' documents. As Lucene, the underlying search engine for our system, indexes only raw text documents, we had to convert all of the documents into this format. We have tested different open source extractors (with the constraint that they have to be written in Java):

- *pdfbox has been selected to process 'PDF' documents.*
- *POI and textmining.org to process 'Word' documents. POI has been selected for its convenient representation of the Microsoft Word document.*
- *Jtidy, NekoHTML and HTMLCleaner to process XHTML documents (all are DOM based parsers but due to the ill-formed documents, it is difficult, not to say impossible, to apply a SAX based parser.) Jtidy has been selected because it cleans not ill-formed XHTML documents and it allows us to access all the HTML tags and textual content conveniently.*

Currently, we have successfully extracted the full body of these documents.

To process the corpus we have adapted the existing functionality of ASSERT (T4.2). This system was initially designed to process abstract length articles from publications. The features used to access the content of these documents are not suitable for the user of the EEP portal. Where the user of ASSERT would generally search for a specific piece of information in a restricted set of documents, the users of the EEP portal are more interested in an exhaustive research of pertinent documents and an intelligible overview of the large collection of documents returned. The system has been adapted to process a new set of metadata (new date format, a quick summary of the content (note that this summary is provided by the author and not computed automatically), a list of keywords, etc..). These metadata has been indexed to allow for the query of these pieces of information.

According to the work package (T4.3) we have designed a new dynamic web site as a prototype system to demonstrate it to our partner. We have deployed a beta version of our system to start a discussion with our stakeholders in order to define a suitable interface for the frame analysis and to qualitatively evaluate the contribution of each piece of information for the frame analysis. The web site has been secured and restricted access given to our partner. Access for other stakeholders will be made available upon request.

Through our close involvement with the National Centre for e-Social Science (NCeSS) during this project and the previous ASSERT project we have co-organised a workshop on text mining in the social sciences as part of the 4th International Conference on e-Social Science. The Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) participated in the event that we used as a core dissemination activity for the whole project.

Section Three: Institutional & Project Partner Issues

Discussions have taken place with James Thomas to anticipate the integration of the Named Entity module (T6.1). Text mining tools usually involve a simple hierarchy of Named Entities which contains: person names, association or company names, and dates or amounts of money. But we can also extend this hierarchy to include names which refer to other objects (e.g. name of exams, awards...) or set out in detail the categories of existing named entity (e.g. name of school, name of a political person...). Because the adaptation of the named entity recognizer is time consuming, we must define in advance the most suitable hierarchy according to our partner's requirements. This discussion is still in progress.

During the preparation of the development corpus we have encountered a well known and difficult problem. Composed of documents coming various web sites, our corpus is a multi-source and a multi-format document corpus. As a consequence, metadata are missing for a subset of documents in the index, and noisy characters can be introduced in the textual content (see section eight for a detailed explanation). We have opened a dialogue with our partner to define simple heuristics to extract essential metadata (e.g. title) and remove the irrelevant characters.

Section Four: Outputs and Deliverables

Presentations on the ASSIST Project were given in the "Text Mining and the Social Sciences" Workshop *in conjunction with the 4th International Conference on e-Social Science. This presentation gives an overview of the system architecture and focuses on the improvement of the research work for the sociology community. Available at:* http://www.nactem.ac.uk/assist/slides.htm/

Base System prototype (D4.3): A web demonstrator showing how the ASSERT based system classified educational documents. This uses the development corpus provided by our partner described in Section Two. Available at: http://nactem3.mc.man.ac.uk:8080/ASSIST-EPPI/

Section Five: Outcomes and Lessons Learned

This project gives us the opportunity to compare two user constraints. Although our partners both work in social sciences, they express different needs which imply different adaptations for our system. It is clear with the task of Named Entity (NE) recognition: our partners are interested in distinct types of NE and multiples level of the NE hierarchy. We confirm a well established result in information extraction: to be efficient our systems have to be flexible according to the type and the domain of the corpus. This is a result that we have to take into account in the choice of our NE recognizer. We will favour a semi-supervised learning-based NE recognizer; this is the most flexible system currently available.

Section Six: Evaluation

Aside from ongoing evaluation activities, with the collaboration of our stakeholders, we have started a qualitative evaluation of the baseline system. In the form of an open discussion, we have asked our partner about the choice of metadata and pieces of information displayed and the ergonomics of the web site. The following list presents the questions:

- *The query interface*
  - *user guide (bubble help, list of metadata available for the query (Named Entity, title, author...), commented example of query (available operators and wildcard characters))*
- *The list of documents*
  - *metadata of interest, document snippet, membership of a cluster (using coloured highlighting or numerical references)*
  - *pieces of information accentuated (colour, font, disposition...), default value (number of terms in the list, threshold for the weight of the terms)*
  - *management of missing metadata (a default missing value, removing incomplete documents from the list, interface for user correction)*
- *The list of clusters and their visualisation*
  - *pieces of information associated with a cluster and their default values (use of the 'all documents' cluster and the 'other documents' cluster,...)*
- *The document interface*
  - *pieces of information highlighted (Named Entities, terms above a certain threshold, terms of the query,...)*
  - *addition of snippet or metadata in the list of relative documents*
  - *mention the membership of the document in a set of clusters*

The result of this questionnaire will emphasize the important information and lead us to redesign the demonstrator interface. We assume that this new design will improve the user's analysis by facilitating access to the document content.

Section Seven: Dissemination

On 18th June 2008 NaCTeM and NCeSS ran a workshop on 'Text Mining in the Social Sciences' as part of the 4th International Conference on e-Social Science. Attendance was good with a number of presentations from different groups around the UK using text mining methods in social science research. A full programme can be found at http://www.ncess.ac.uk/events/conference/programme/workshop2/ *and copies of the presentations can be found at* http://www.nactem.ac.uk/assist/slides.htm*. To allow for wider dissemination videos of the full presentations will also be available through the ReDReSS project (*http://redress.lancs.ac.uk/*)*

Section Eight: Risks, Issues and Challenges

During the preparation of the development corpus we have encountered a well-known and difficult problem. Composed of documents from various web sites, our corpus is a multi-source and multi-format document corpus. Indexing such a corpus is known to be difficult.

After processing HTML documents with the parser, the logical structure of the file is lost. Title, subtitles, lists and references are flattened into a linear text. The consequence is the creation of ungrammatical sentences like "Key findings Use of formal tests Over 95 per cent of schools in the survey sample use the QCA optional tests in English and mathematics." To solve this problem we have adapted the output of the Jtidy parser to preserve the important parts of the logical structure of the XHTML documents using the XHTML tags.

With XHTML and Microsoft Word documents, we have seen the appearance of unexpected non-alphanumeric characters and words which are not in the original content of the document (e.g. the index, hyperlinks, footnotes...). To fix this problem we added a cleaning stage which deals with each type of unexpected characters and words independently. It should be noted that the proposed solutions are not perfect because of the diversity of the documents. Ungrammatical sentences and unexpected characters remain in indexed documents.

The problem of metadata extraction (namely: title, author, subject, keyword...) is more difficult. In the 'pdf' documents, the metadata can be found in a data cartridge but this cartridge is often left empty or filled with inconsistent information (e.g. in the user field, we can find the value 'user'). In the 'Word' documents a function serves to return the metadata of the document. Unfortunately, this function is imprecise for some metadata and returns unexpected information (e.g. the function does not return the title but the text in the logo placed before the title in the document). In the XHTML documents the metadata are clearly mentioned with meta-tags; however, some of them are missing (e.g. there is no meta-tag to mark the authors, and they are not emphasized with an HTML tag in the documents). To address the metadata problem there are at least two different strategies. The first is the application of heuristics to identify the metadata (for example the title is usually the first or second nominal phrase of the document with a large font). These heuristics are simple and easily implemented, but they return a lot of noise. The second strategy is to design a Text Mining tool to focus specifically on one type of metadata, for example a NER designed to extract the authors (a name of person in the first part or in the last part of the document is probably the author). If these tools are more reliable, it is still difficult and time consuming to design these tools for specific metadata. After a discussion with our partner, we decided as a start to apply simple heuristics to extract the essential metadata. The first qualitative evaluation by our partner should examine the value of this solution. If this solution is rejected, we will move to a TM tool based solution.

Section Nine: Collaboration and Support

More and more documents are accessible via on-line databases and repositories (Medline, Citeseer, LexisNexis...). This allows us to collect a huge corpus. However, the formats of the documents proposed by these databases are heterogeneous and partial. The consequence is an unreliable pre-processing stage, i.e. some expected pieces of information are missing for some documents (see the previous section). The point we would like to discuss with other projects is the existence for a plan to normalize the representation of the documents. Such standardization would make possible the implementation of a good quality API to collect documents and, as a consequence, improve the natural language processing of these documents. An alternative would be to investigate if other projects are looking at the discovery of this top level metadata from document content or format.

Section Ten: Financial Statement

Formatted: Don't snap to grid

Section Eleven: Next Steps

- *integration of the qualitative evaluation by our partner to improve the design of the demonstrator*
- *extending the hierarchy of Named Entities according to the needs of our partner (T6.1 extension)*
- *integration of the anaphora resolution system in the pipeline (T6.2 – due to 01/01/09)*