



A Scoping Study of (Who, What, When, Where) Semantic Tagging Services

Document details

Authors:	Anna Gerber, Lianli Gao, Jane Hunter eResearch Lab, The University of Queensland http://itee.uq.edu.au/~eresearch/
Version/Date:	<ul style="list-style-type: none">• v 1.0, September 30, 2010• v 2.0, November 3, 2010• v 3.0. November 23, 2010• Public release. February 22, 2011

Summary

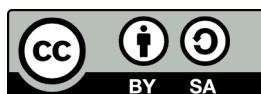
This document is a report on a Scoping Study of Semantic Tagging Services for the Australian academic research sector. The study identifies technologies and technical infrastructure to enable the semantic mining, analysis and linking of knowledge contained within distributed collections of resources. It focuses on tagging digital text resources with the names of historically or culturally significant *people*, *places*, *dates*, *events* and *topics/concepts*. The report discusses:

- What semantic tagging technologies/services currently exist;
- The maturity and desirability of these technologies;
- The optimum infrastructure that would be necessary to provide such a service;
- The optimum architecture and integrated set of services;
- Possible service providers;
- Recommendations for next steps.

The eResearch Lab at the University of Queensland prepared this report on behalf of the Australian National Data Service (ANDS).

Rights and Acknowledgements

This work copyright The University of Queensland, 2010 – 2011.



Licensed under Creative Commons Attribution-Share Alike 3.0 Australia. <<http://creativecommons.org/licenses/by-sa/3.0/au>>.



This project is supported by the Australian National Data Service (ANDS). ANDS is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy Program and the Education Investment Fund (EIF) Super Science Initiative.

Contents

1	Introduction and Terms of Reference	1
1.1	Background	1
2	Aims and Objectives	1
3	Methodology	2
4	Community Drivers and Types of Collections	2
5	Technological Survey	3
5.1	Overview	3
5.2	Open Source Standalone Applications	4
5.3	Web Services	6
5.4	Commercial Systems	9
5.5	Bio-medical Semantic Tagging Tools.....	10
5.6	Scientific and Chemistry Semantic Tagging Tools.....	10
5.7	Research - Semantic Tagging of Texts.....	11
5.8	Research - Semantic Tagging of Multimedia	12
6	Anticipated Future Trends	13
7	Conclusions and Recommendations	14
7.1	Assessment Results.....	14
7.2	Recommended Next Steps	17
7.3	Recommended Vocabularies	18
7.4	Infrastructure and Architecture	18
7.5	Service Providers	19
7.6	Conclusions.....	20
8	Author contact details	20
9	References	20
10	Appendix: assessment criteria	23

1 Introduction and Terms of Reference

1.1 Background

A series of discussions and workshops were held in 2008 and 2009 between the Australian Humanities and Social Sciences (HASS) community, the National eResearch Architecture taskforce (NeAT) and the Australian National Data Service (ANDS). These discussions identified the need for a service that enables textual documents (newspapers, historical manuscripts, theses) and other types of digital resources (images, video, audio, maps) to be processed and tagged with unique identifiers and controlled terms that identify the names of historically or culturally significant *people, places, dates, events* and *topics/concepts*. Moreover, if the tags are drawn from a controlled vocabulary and are represented in a machine-processable format (e.g., RDF¹, OWL²) then they provide the foundation for richer analytical and inferencing services that can uncover previously-unknown relationships between resources in disparate collections. The combined use of URIs (to identify the textual resources, segments and tags) and RDF (to record the tags/annotations), represents the Linked Data³ approach to connecting distributed documents over the Web. This is the recommended best practise for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web.

2 Aims and Objectives

The aim of this study is to identify the most effective technologies and the optimum technical infrastructure to enable the semantic mining, analysis and linking of knowledge contained within distributed collections of digital textual resources. Automated tagging techniques greatly reduce the time and effort required to generate fine-grained metadata – which in turn will facilitate the sharing and re-use of data and knowledge across historical, cultural and scientific disciplines. Alternative manual techniques such as crowd-sourcing of tags will also be investigated.

The study has involved discussions with individuals from the following projects and organizations: AustLit, School of History, Philosophy Religion and Classics (University of Qld), SETIS (University of Sydney), eScholarship Research Centre (University of Melbourne), Australian Scholarly Editions Centre (UNSW, ADFA), National Library of Australia.

The specific aims of the scoping study are to identify:

- What semantic tagging technologies/services (capable of analysing textual documents and tagging *who, what, when, where*) currently exist;
- An assessment of the maturity and desirability of these technologies;
- An assessment of the optimum infrastructure that would be necessary to provide such a service;
- A design for the optimum architecture and integrated set of services;
- An assessment of possible service providers;
- Recommendations for next steps.

¹ <http://www.w3.org/TR/PR-rdf-syntax/>

² <http://www.w3.org/TR/owl-ref/>

³ <http://linkeddata.org/>

3 Methodology

An eight-step approach was adopted in undertaking this scoping study. The specific steps are:

1. Identify and document the communities and applications that are driving the demand for this service;
2. Identify the collections of documents (+ their format, structure, genre, discipline etc.) that would generate most value if they were tagged (e.g., newspapers, theses, manuscripts, photos, maps, audio/video recordings);
3. Identify currently available tools for automated tagging. Evaluate these tools based on a set of criteria including: open source, standards-based, maturity/robustness, platform-independent, efficiency, scalability, interoperability, flexibility, tailorability, tag representation (SKOS, RDF, RDFa etc.).
4. Identify existing tools and approaches to streamline high quality manual tagging and evaluate them based on a set of criteria (see above). For example, crowd sourcing of tags.
5. Identify existing controlled vocabularies for generating and validating tags (e.g., Australian People names, Australian Place Names Gazetteers, ISO 8601 dates/times etc)
6. Identify methods and systems for managing controlled vocabularies (and their versions) e.g., thesauri, controlled vocabulary and ontology registries.
7. Design the optimum “tag generation, storage, re-use and management architecture” which integrates the set of services specified in the previous steps
8. Write the Final Report and Recommendations for next steps

4 Community Drivers and Types of Collections

Examples of the communities and applications in Australia that are demanding tools and services to automatically tag and annotate documents include:

- Historians – frequently want to apply text analysis to historical documents such as diaries, letters, memoirs, transcriptions of interviews etc. to identify and retrieve references to people, places, concepts, events etc. Examples of historical documents and collections that are of specific interest to Australian historians include: “Reports of the Cambridge Anthropological Expedition to the Torres Strait”, “The Endeavour Journal of Sir Joseph Banks”, the Australian War Memorial’s war diaries⁴ and the NLA’s collection of digitized newspapers⁵. An example of a major international project being undertaken in this domain is the Criminal Intent Project <http://criminalintent.org/> - this project is using GATE, Zotero and TAPOR to perform manual semantic markup and textual data mining on the Old Bailey Proceedings, a collection of 120 million words of structured text that documents court records of more than 197,000 individual trials held over 240 years in Great Britain.
- Literary Scholars – literature researchers (such as the AustLit community) frequently want to analyse texts to identify and analyse recurrences and patterns of words, phrases or topics. There are a vast array of text analysis tools available that enable users to determine the frequency with which words or phrases are used, create concordances, view words in context, and study patterns in texts⁶. However there are relatively few tools available that will automatically identify and annotate named entities (people, places, dates, concepts) within literary texts. Such tools are highly useful particularly for inferring relationships between literary texts, authors, ideas and places.
- Linguists – the Australian linguistic community is currently promoting the development of an Australian National Corpus (a massive online database of spoken and written language in Australia) to support scholars studying the Australian version of the English language and historical trends in Australian English. Assuming the establishment of a large scale Australian National Corpus, linguists will require information technologies to enable the semi-automated tagging, annotation and transcription of textual, audio and video

⁴ http://www.awm.gov.au/collection/war_diaries/

⁵ <http://newspapers.nla.gov.au/>

⁶ <http://digitalresearchtools.pbworks.com/Text+Analysis+Tools>

documents. Apart from the Australian National Corpus initiative, the other major linguistic archive in Australia is the Paradisec (Pacific and Regional Archive for Digital Sources in Endangered Cultures) project. Users of Paradisec require similar tools to the Australian National Corpus but also require tagging and annotation tools for multiple languages (especially endangered Indigenous languages). Linguists most commonly require part-of-speech tagging i.e., automatic identification of nouns, verbs, articles, adjectives, prepositions, pronouns, adverbs, conjunctions and interjections. Such tools are out of scope of this study. However the linguistic community in Australia also require “named entity tagging” tools that identify people, places, dates/times and concepts.

5 Technological Survey

5.1 Overview

The aim of this section is to provide an overview of the technologies available for the automated tagging of “named entities” (e.g., persons, organizations, places, dates/times, quantities, concepts (e.g., chemicals, genes, proteins etc.)) within textual documents.

There exists a wide range of approaches to named entity tagging. One simple classification of Natural Language Processing (NLP) systems divides these into 4 types:

1. Statistical/machine learning approaches – these approaches require a large amount of manually annotated data (a training corpus) as training data
2. Linguistic grammar based approaches – these approaches are based on grammatical rules – they provide better precision but lower recall and are more time consuming than statistical approaches
3. Linguistic: Parts of speech (POS) parsers that identify and tag parts of speech (including nouns, verbs, articles, adjectives, prepositions, pronouns, adverbs, conjunctions and interjections).
4. Named Entity Recognition (NER) (based on ontologies/thesauri) and Disambiguation systems.

The most relevant to this report are NER systems – that locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, times/dates, quantities, monetary values etc. Most NER systems transform an unstructured block of text such the one below:

“Joseph Banks collected over 2000 plants in Australia in 1770”

into an annotated block of text:

<ENAMEX TYPE="PERSON">Joseph Banks</ENAMEX> collected over <NUMEX TYPE="QUANTITY">2000</NUMEX> plants in<ENAMEX TYPE="LOCATION">Australia</ENAMEX> in <TIMEX TYPE="DATE">1770</TIMEX>.

In this example, the annotations use the ENAMEX tags developed for the Message Understanding Conference (MUC) in 1990s.

Named Entity Recognition systems have been developed for specific types of entities (e.g., genes), for specific types of content (e.g., phone conversation transcripts, military reports, email), and for specific languages (English, Chinese, Japanese, Spanish, Dutch, Portuguese). For this report we are primarily interested in English NER systems.

The latest NER systems are quite brittle – they are primarily developed for and work well within a single domain – but don’t perform well when applied to other domains. However they do demonstrate near-human performance on English texts from the same domain as the training corpus. State of the art Java-based NER taggers such as the Stanford NER⁷ and the Illinois NER⁸ demonstrate scores of

⁷ <http://nlp.stanford.edu/software/CRF-NER.shtml>

⁸ <http://l2r.cs.uiuc.edu/~cogcomp/asoftware.php?key=FLBJNE>

93.39% at the Message Understanding Conference (MUC) compared with human annotators who scored 97.6%.

In the next sections below we identify currently available tools for automated tagging and provide a brief evaluation of each of them based on a set of criteria. The assessment criteria include the following:

- Open source or commercial – is the service free or licensed?
- Standards-based – is the system based on open standards?
- Maturity/robustness – is the system robust and mature?
- Platform-independence – will the system operate on different operating systems?
- Is the system available as a Web service or stand-alone application?
- Efficiency and performance
- Need for training and a training corpus of manually marked up texts
- Scalability – will it scale to massive online collections?
- Interoperability – will the system accept input and generate output to enable interoperability with other systems?
- Flexibility and tailorability - is the system designed so that it is relatively simple and easy to make changes and adapt it to a different discipline/ontology?
- Input formats (HTML, Word, PDF, TXT?) – does the system support a range of input formats? Does it support batch input of multiple files?
- Tag representation (SKOS, RDF, RDFa etc) – in what format are the tags generated?
- Availability of APIs - does the system include an API for developers?
- Ease of use and usability
- Specificity – does the system only identify people names or places or discipline-specific entities? If the service is designed for a specific discipline, and if so, which discipline?

5.2 Open Source Standalone Applications

In this section, we describe the most popular generic, stand-alone NER systems, the majority of which are written in Java. Many of them can be customised to identify specific entities (people, places, events etc). Many of them don't explicitly support semantic technologies like RDF, but they can be modified relatively easily, to generate this kind of output.

5.2.1 GATE (General Architecture for Text Engineering) <http://gate.ac.uk/>

- Open-source tool developed by University of Sheffield, can easily be embedded (Java jars) in other systems
- Includes *ANNIE*: - an information extraction and semantic tagger system which is extremely tailorable, supports multiple languages, customized gazetteers (based on flat list of terms or from an ontology)
- Extensive documentation on web site, however the system is relatively difficult to set up/configure – there is a set of training/certification modules.
- GATE cloud currently only available to GATE partners and in alpha phase.
- Current projects that use GATE include:
 - GATE/ETCSL - The project is building generic tools for linguistic annotation and Web based analysis of literary Sumerian
 - [EMILLE](#) (Enabling Minority Language Engineering) - Building a 63 million word electronic corpus of South Asian languages, especially those spoken in the UK
 - [OldBailey Online](#) - Named entity recognition on 17th century Old Bailey Court reports, using a combination of manual markup and GATE

5.2.2 YooName/Balie <http://yooname.com>

- Proof-of-concept built by PhD student based on semi-supervised learning
- Identifies 9 types of entities (100 sub-categories) including person, organisation, location, facility, product, event, natural object and unit.
- Evolved version of Balie (open source tool by same developer) <http://balie.sourceforge.net/>

- 5.2.3 Mallet (MAchine Learning for LanguagE Toolkit) <http://mallet.cs.umass.edu/>
- Open source (CPL) Java-based tool
 - Documentation aimed at people familiar with NLP – relatively difficult to get started
 - Sequence tagging features support Named Entity Recognition, using hidden markov models and linear chain conditional random fields (CRFs)
- 5.2.4 FreeLing <http://www.isi.upc.edu/~nlp/freeling/>
- Open source (GPL), APIs for both python and php
 - Supports multiple languages (including Portuguese, Italian, Spanish and English)
 - Recognises dates/times, quantities/ratios and named entities such as people.
 - Includes on-line demo <http://nlp.isi.upc.edu/freeling/demo/demo.php>
- 5.2.5 Illinois Named Entity Tagger http://cogcomp.cs.illinois.edu/page/software_view/4
- From University of Illinois at Urbana-Champaign
 - Tags *people, organisations, locations, miscellaneous*. Gazetteers are based on Wikipedia
 - Developed by L. Ratinov and D. Roth, Design Challenges and Misconceptions in Named Entity Recognition, CoNLL 2009
- 5.2.6 LingPipe <http://alias-i.com/lingpipe/>
- Java API with source code
 - Online demo - result is XML that labels the entities using ENAMEX tags identifying persons, organisations and locations
 - Can be trained to recognize entities from any domain or language based on regular expressions or dictionary
 - Free for research use, licenses available for commercial use
- 5.2.7 Open Pipeline <http://www.openpipeline.org/>
- Open source (Apache License 2.0) Java-based search pipeline platform
 - Includes wrappers for LingPipe and UIMA
 - Entity extraction via a commercial add-on
- 5.2.8 MinorThird <http://sourceforge.net/apps/trac/minorthird/wiki>
- A toolkit and collection of Java classes – provides machine learning methods for extracting entities, integrated with tools for manually and programmatically annotating text.
- open-source (BSD) Java libraries
 - annotation and visualisation system as well as entity recognition
 - Uses stand-off markup of textual documents stored in a database (TextBase)
 - Cohen, W. *MinorThird: Methods for Identifying Names and Ontological Relations in Text using Heuristics for Inducing Regularities from Data*, <http://minorthird.sourceforge.net>, 2004.
- 5.2.9 Stanford Named Entity Recognizer <http://nlp.stanford.edu/ner/index.shtml>
- Open source (GPL) Java-based tool (commercial license also available)
 - Needs to be trained to recognise entities e.g., person, location, organization
 - Additional tools available e.g., Perl module provides web service interface, and Apache UIMA annotator
 - Recent new release, active community with mailing lists for support
 - Output formats include XML, inlineXML and slashTags
- 5.2.10 TextPro/Typhoon <http://textpro.fbk.eu/>

TextPro/Typhoon is a classifier combination system for Named Entity Recognition (NER), in which two different classifiers are combined to exploit Data Redundancy and Patterns extracted from a large text corpus.

- Demo recognises persons, locations and organisations
- Works for both Italian and English
- Free for research/non-profit purposes
- Online demo available: <http://textpro.fbk.eu/demo.php>
- Typhoon also available as a web service (Italian only) <http://textpro.fbk.eu/typhoon>

5.3 Web Services

There is an increasing number of Web services available that perform named entity recognition on textual documents via a Web interface. The majority and the best of such services are not open source or free. There are some free web services (e.g., tagthe.net <http://www.tagthe.net/>) but they generally provide poor quality performance.

Although the majority of services are commercial, some also have free components/versions with limited functionality/usage (e.g., 10,000 requests/day). Examples that apply this kind of restriction include:

- Evri
- OpenCalais
- AlchemyAPI

The most promising services also apply restrictions on the re-use of tags – for example, they don't provide a mechanism by which users can store the tags for re-use.

There are also many web services that to all intents and purposes are commercial because the amount of permitted free usage is very small: Meaningtool, Complexity Intelligenece, TextDigger.

Below is a survey of the most widely used, robust and best performing of the semantic tagging web services.

5.3.1 Evri <http://www.evri.com>

- Provides several APIs for NLP text analysis, content recommendations and relationships between semantic entities [43]
- The “*Get Entities Based on Text API*” extracts entities (people, places, things) from news articles, blog posts, twitter tweets and other web content. The full schema of entities is not published, but a zeitgeist of 1000 most popular is available. Entities include persons, locations, concepts, products, organisations and events as well as relations.
- Results are XML or JSON
- Evri entities are identified by Evri URIs (but no Linked Data URIs to other databases)
- Has a mobile application – filters and delivers personalized content via iPhone app
- Free, with no fixed limit for non-commercial use, however caching of results is not permitted – exemptions are possible (e.g., for academic use) by contacting the company. Commercial licenses available.

5.3.2 OpenCalais <http://www.opencalais.com/>

- OpenCalais is a product of Thomson Reuters that provides an open API that has been widely adopted by the open source community.
- Identifies specific entities, events and relations from the web and news domain (e.g., company merger, natural disaster, product recall, conviction etc). Also suggests social tags.
- A full list of available entities is available here: <http://www.opencalais.com/documentation/calais-web-service-api/api-metadata/entity-index-and-definitions>
- See also the online demo/web service: <http://viewer.opencalais.com/>
- User-defined vocabularies are planned for “some point in the future”)
- Many entities are identified using Calais URIs, some *sameAs* links to DBPedia and Freebase
- Supports disambiguation of companies, geographical locations and electronics products
- Results available as: RDF/XML, Microformats, custom XML (Simple Format), JSON:
- Provides character offsets that can be used to insert tags into content

- Free for up to 50,000 requests per day after registering for API key, subscription plans above that. Works on documents up to 100K.
- Supports English, French and Spanish
- Detailed documentation available on the website including RDF schema and demo
- It is the semantic tagging engine behind the OpenPublish platform (integrated with Drupal and WordPress)
- ClearForest <http://www.clearforest.com/> - also have a commercial product called OneCalais

5.3.3 Alchemy API <http://www.alchemyapi.com/>

- Automatically tags web pages, textual documents, scanned document images. Supports OCR to analyse scans of newspapers, documents etc.
- Supports multiple languages (English, Spanish, German, Russian, Italian + others)
- *Named Entity Extraction API* identifies specific entities including people, companies, organisations, cities, geographic features, anniversaries, awards, holidays etc.
- Entities identified by URIs from Linked Open Data (LOD) sources e.g. Freebase, UMBEL, CIA Factbook
- Disambiguation support (although seems to be missing disambiguated URIs for “person” entities) [43]
- Formats: XML, JSON, RDF, Microformats
- Requires an access key to access the API
- Free for up to 30,000 calls per day, can pay for commercial support.
- Detailed documentation available on website including RDF schema and online demo <http://www.alchemyapi.com/api/entity/>

5.3.4 Zemanta <http://www.zemanta.com/api/>

- Identifies the following entities: persons, books, music, movies, locations, stocks, companies (documentation does not mention events).
- Also returns related tags, categories, pictures and articles.
- Free for up to 10,000 API calls per day. Subscription plans above that.
- Returns RDF/XML, JSON, or custom XML
- Documentation says it supports custom taxonomies
- See a recent comparison with Open Calais: Linked Data Entity Extraction with Zemanta and OpenCalais <http://bnode.org/blog/2010/07/28/linked-data-entity-extraction-with-zemanta-and-opencalais>

5.3.5 OpenAmplify <http://www.openamplify.com/>

- Provides Natural Language Processing APIs, for use in commercial applications
- Analyses documents for topics (including named entities such as persons, organisations and locations), actions (i.e. events that can be identified by verbs such as give, learn, repair, request, say etc and when they have or will occur), style, demographics etc.
- Results available as custom XML and JSON formats
- Good documentation on website including code samples and tutorials.
- Free for up to 1,000 requests per day, commercial packages available beyond that.

5.3.6 Meaningtool <http://www.meaningtool.com/>

- Identifies entities (organisations, companies, locations, persons only), categories, keywords, language
- Supports English, Spanish and Portuguese texts
- Supports user-defined trees for categorisation
- Results available in JSON or custom XML
- Free for up to 1,000 requests per day (plans available above that)
- Good documentation and demo on website

- 5.3.7 Complexity Intelligence <http://www.complexityintelligence.com/>
- Free for 10,000 requests per month after registering
 - Identifies persons, companies, locations (perhaps more?)
 - Online demo available from web site
- 5.3.8 TextDigger <http://textdigger.com/>
- Semantic content tagger - free to tag 25 URLs per day (can purchase additional capacity)
 - Results are not sent automatically – must request page to be queued for tagging, and then retrieve the results via the web service.
 - Assigned tags are used to retrieve links to related web pages.
 - Results are returned as custom XML. Entities have numeric ids.
 - The results are stored in a database
- 5.3.9 Inform <http://www.informpublisherservices.com/>
- Commercial Web service
 - Not much information on their website – further information by enquiry only
- 5.3.10 mSpoke mSense <http://www.mspoke.com/mSense.html>
- Commercial Web service
 - Identifies named entities: people, places, organisations (also topics, categories)
 - mSense taxonomy based on Wikipedia, also allows customized taxonomy
 - mSense API available
 - Further information available through enquiry
- 5.3.11 Info(N)gen <http://www.infongen.com/>
- Commercial Web service
 - Default taxonomy includes entities such as company, industry, language, country, products – targeted at business, finance, pharma, energy, technology, consumer goods, retail, commercial services and media domains.
 - Customized taxonomies possible (must be created using the InfoNgen Taxonomy Wizard)
 - Results are RDF/XML or custom XML (via API or feed)
 - Further information available through enquiry
- 5.3.12 Alethes OpenEyes <http://www.alethes.it/openeyes.html>
- Commercial system
 - Website in Italian but can be applied to 8 languages including English
 - Example: <http://www.youtube.com/watch?v=VJdMM8Rhxdo>
 - Recognises people, organisations, places, quantities, dates, currency. Entities can be customized
 - Compatible with Apache UIMA
- 5.3.13 TagThe.net <http://tagthe.net>
- Returns custom XML containing tags identifying topics, locations, persons, (but not events). Also tags for title, size, content-type, author and language of the source document.
 - Does not markup content (or indicate location of entities within content).
 - Tags are text only (no identifiers or ontology)
 - Uses statistical approach (from FAQ). Analysis component is written in Java.
 - Free to use as-is. No limitations on use but also no service level guarantees.
 - Can invoke via HTTP requests

5.4 Commercial Systems

There are a wide range of commercial named entity recognition (NER) systems available. These systems typically use significant numbers of hand-coded rules, which enable them to achieve reasonable performance for limited numbers of entity types on well-circumscribed corpora, such as news articles. However they generally don't permit customization or tailoring for domains other than the one for which they were designed. Below we have described some of the more popular and widely used commercial systems for named entity tagging.

5.4.1 SAS Text Miner <http://www.sas.com/text-analytics/text-miner/index.html>

- Mines text from PDFs, HTML, Word docs in multiple languages
- Identifies named entities, parts of speech and provides visualisation of concepts
- Support for many different entity types, including person and company names, locations, dates, addresses, measurements, and e-mail and URL addresses.
- Supports user customization of entity lists
- Commercial system, was previously known as Teragram

5.4.2 Leximancer <http://www.leximancer.com/>

- Commercial
- Standalone software or hosted solution
- Visualisations as well as named entity recognition

5.4.3 Megaputer PolyAnalyst <http://www.megaputer.com/polyanalyst.php>

- Commercial product
- Supports keyword and entity extraction as well as categorization and clustering of textual documents

5.4.4 Trifeed TRAILS <http://www.trifeed.com>

- Identifies entities including people, companies, places, events, books, movies, dates, currency with associated attributes (eg a person's position). Can also extract relations and quotes.
- Commercial
- Aimed at online news domain
- Demo available on website: <http://www.trifeed.com/new-demo.jsp>

5.4.5 Nogacom ClassLogic <http://www.nogacom.com/>

- Commercial entity extraction and classification based on Nogaclass data classification platform
- Focused on the business domain: entities include customers, suppliers, partners, products, competitors, locations etc – from their own business taxonomy
- Supports 32 languages
- Website contains mostly marketing material – not much technical information

5.4.6 NetOWL <http://www.sra.com/netowl/>

- Extractor recognises entities based on their own NameTag (people, organisations, places, addresses, dates etc), Link, Event (affiliation, transaction etc) and Cyber Security ontologies
- Supports multiple domains: Business, security, finance, life sciences, military, politics etc
- Supports multiple languages
- Output includes custom XML and OWL
- Provides term extraction and visualisation tool (Java-based)
- APIs for Java, C and can be run as a Web service

5.4.7 Basis Technology Rosette Entity Extractor (REX) <http://www.basistech.com/entity-extraction/>

- REX uses statistical modeling to learn patterns from large corpora of native language

- Identifies and tags people, organizations, locations, dates using gazetteers
- available for Chinese, Japanese, Korean, Arabic, Farsi, Urdu, Russian, Dutch, English, French, Italian, German and Span

5.5 Bio-medical Semantic Tagging Tools

The majority of discipline specific NER systems have been developed for text mining of biomedical literature and MEDLINE abstracts. Below are some of the most popular and robust tools in this area. They generally enable the identification and tagging of biomedical entities such as: protein, DNA, RNA, Cell Line and Cell Type.

5.5.1 BioNLP <http://bionlp.sourceforge.net/>

BioNLP is an initiative by the Center for Computational Pharmacology at the University of Colorado to create and distribute code, software, and data for applying natural language processing techniques to biomedical texts. It has generated a number of tools but the most relevant are:

- [Knowtator](#): a Protege plug-in for text annotation.
- MutationFinder: extracts biomedical entities from text

5.5.2 PennBioIE <http://bioie ldc.upenn.edu/>

The aim of the PennBioIE project was to develop better methods for information extraction, specifically from biomedical literature and are annotating texts in two domains of biomedical knowledge:

- inhibition of the cytochrome P450 family of enzymes (**CYP450** or **CYP** for short)
- molecular genetics of cancer (**oncology** or **onco**)

5.5.3 ABNER <http://pages.cs.wisc.edu/~bsettles/abner/>

ABNER (A Biomedical Named Entity Recognizer) is an open source (CPL) Java tool for molecular biology entity extraction. It recognizes proteins, DNA, RNA, cell line and cell type

5.5.4 POSBIOTM/W <http://isoft.postech.ac.kr/Research/Bio/bio.html#Requirements>

POSBIOTM/W is a workbench for machine-learning oriented biomedical text mining system. It is intended to assist biologist in mining useful information efficiently from biomedical text resources.

5.5.5 GENIA Tagger <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/>

The GENIA tagger analyzes English sentences and outputs the base forms, part-of-speech tags, chunk tags, and named entity tags. The tagger is specifically tuned for biomedical text such as MEDLINE abstracts and identifies proteins, DNA, RNA, cell_line and cell_type.

5.5.6 AIIAGMT <http://bcsp1.iis.sinica.edu.tw/aiiagmt/>

This NER system developed by the AIIALab at Academia Sinica in Taiwan, performs tagging of gene and gene products mentioned in textual documents.

5.5.7 DECA – Disease Extraction http://www.nactem.ac.uk/deca_details/start.cgi

DECA focuses mainly on disambiguation of model organisms commonly used in biological studies, such as E. coli, C. elegans, Drosophila, Homo sapiens. Given an article, DECA automatically identifies the species-indicating words (e.g., human) and biomedical named entities (e.g., protein P53) in the text, assigns a unique [NCBI Taxonomy](#) ID to each entity.

5.6 Scientific and Chemistry Semantic Tagging Tools

5.6.1 OSCAR3 (Open Source Chemistry Analysis Routines)

<http://sourceforge.net/projects/oscar3-chem/>

OSCAR3 is a set of software modules designed to enable semantic annotation of chemistry-related documents It provides two modules: OPSIN (a name to structure converter) and ChemTok (a tokeniser for chemical text) which are also available as standalone libraries. It also attempts to identify:

- Chemical names: singular nouns, plurals, verbs etc., also formulae and acronyms, some enzymes and reaction names.
- Ontology terms from the ChEBI ontology (<http://www.ebi.ac.uk/chebi/>)
- Chemical data: Spectra, melting/boiling point, yield etc. in experimental sections.

In addition, where possible the chemical names that are detected are annotated with structures, either via lookup or name-to-structure parsing ("OPSIN"), and with identifiers from the chemical ontology ChEBI.

5.6.2 SAPIENT – Semantic Annotation tool for Scientific Research Papers
<http://www.aber.ac.uk/compsci/Research/bio/art/sapient/>

A web application designed to take as input, full scientific papers that are represented in XML, and compliant with the SciXML schema. Supports the annotation of papers using topics/concepts in Physical Chemistry and Biochemistry taken from CISP (Core Information about Scientific Concepts). Examples of entities are: Background, Conclusion, Experiment, Goal, Hypothesis, Method, Model, Motivation, Object of Investigation, Observation, Result (based on the EXPO ontology). It provides both manual annotation and auto-annotation tools. The automatic annotation is performed by Oscar3, and generates colour-coded annotations.

5.7 Research - Semantic Tagging of Texts

Automatic semantic annotation requires training to carry out the annotation process autonomously. As such substantial human contribution is required to generate the training corpus and to maintain the corpus as the domain ontology evolves over time. For this reason significant research has been focused on semi-supervised approaches that don't require a large annotated corpus for training but may require some manual bootstrapping to start the learning process. A simple way to categorize semantic tagging systems is as follows:

- Machine-learning methods such as Amilcare that require an annotated training corpus;
- Rules-based systems – that rely on manually-created rules;
- Pattern-based systems – that require an initial set of seeds in order to discover patterns.

Armadillo [20] uses a pattern-based approach to annotation, based on the Amilcare information extraction system [21]. It is especially suitable for highly structured Web pages. The tool starts from a seed pattern and does not require human input initially - although the patterns for entity recognition have to be added manually.

The knowledge and information management (KIM) platform [22] consists of an ontology and knowledge base as well as an indexing and retrieval server. RDF data is stored in an RDF repository, whilst search is performed using LUCENE. KIM is based on an underlying ontology (KIMO or PROTON) that holds the knowledge required to semantically annotate documents, and on GATE to perform information extraction.

Magpie [23] is a suite of tools that supports the fully automatic annotation of Web pages, by mapping entities found in its internal knowledge base against those identified on Web pages. The quality of the results depends on the background ontology, which has to be manually modeled and populated.

MnM [24] is another tool that supports semi-automatic annotation based on the Amilcare system. It uses machine learning techniques and requires a training data set. The classical usage scenario MnM was designed for is the following: while browsing the Web, the user manually annotates selected Web pages in theMnM Web browser. While doing so, the system learns annotation rules, which are then tested against user feedback. The better the system does, the less user input is required.

The PANKOW algorithm [25] is a pattern-based approach to semantic annotation that makes use of the redundant nature of information on the Web. Based on an ontology, the system constructs patterns and combines entities into hypotheses that are validated manually.

S-Cream [26] is another approach to semi-automatic annotation that combines two tools: Ont-O-Mat, a manual annotation editor implementing the CREAM framework, and the Amilcare system. S-Cream can be trained for different domains provided the appropriate training data and proposes a set of heuristics for post-processing and mapping of information extraction results to an ontology. S-CREAM

uses the Amilcare machine-learning system together with a training corpus of a manually annotated set of documents, to automatically suggest appropriate tags for new documents.

ConAnnotator [27] uses Support Vector Machines (SVM) and Natural Language processing (NLP) approaches to facilitate the automated generation of annotations with the support of the domain ontology.

The SemTag system [28] is based on the TAP ontology (which is very similar to the KIM ontology). The system firstly annotates all occurrences of instances of the ontology. Secondly, it disambiguates the elements and assigns the correct ontological classes by analysing context.

More recently, the OntoNEO [29] system has been developed by Choi and Park to automatically semantically annotate named entities in texts. OntoNEO claims to have 18% better performance than the SemTag algorithm – by using a Hidden Markov Model (HMM) to represent the probabilistic model of named entities from a corpus of documents.

The SCORE system [30] for management of semantic metadata (and data extraction) also contains a component for resolving ambiguities. SCORE uses associations from a knowledgebase to determine the best match from candidate entities but detailed implementation details are not available.

In ESpotter, named entities are recognized using a lexicon and/or patterns [31]. Ambiguities are resolved by using the URI of the webpage to determine the most likely domain of the term (probabilities are computed using hit count of search-engine results).

Table 1: A classification of approaches for semantically annotating texts.

System Name	Nature	Method
Armadillo	Automatic	Pre-defined ontology
	Semi-automatic	Adapted ontology
KIM	Automatic	Limited focus KIMO ontology
Magpie	Automatic	Pre-defined ontology
	Semi-automatic	Adapted ontology
MnM	Manual	Without training
	Semi-automatic	With training, KMi ontology
Pankow	Automatic	Limited focus
S-Cream	Manual	No training
	Semi-automatic	With training
SemTag	Automatic	Limited focus TAP ontology
OntoNEO	Automatic	Limited focus
SCORE	Automatic	Pre-defined ontology
ESpotter	Automatic	Weighted ontology

5.8 Research - Semantic Tagging of Multimedia

Because manual annotation of multimedia is so time-consuming, expensive and subjective, there has been significant research effort focused on automatic semantic annotation of multimedia. automatic low-level feature extraction tools are often employed to extract low level features (e.g., regions, colours, textures, shapes). The Semantic Gap refers to the difference between the low level features and the high-level semantic descriptions of the content (e.g., people, places, events, keywords) represented in discipline-specific ontologies. A range of approaches has been applied (with varying success) to bridge the Semantic Gap. Typically these approaches involve a combination of:

- manual annotation of corpuses of training content;
- interactively-defined inferencing rules (that specify rules for inferring high level descriptors from combinations of low level features);
- and neural networks or machine learning techniques

The most significant automatic/semi-automatic semantic annotation tools for multimedia are:

- **AktiveMedia** [32] – an ontology-based annotation system for images and text. It provides semi-automated annotation of JPG, GIF, BMP, PNG and TIFF images by suggesting tags interactively whilst the user is annotating.
- **Caliph and Emir** [51] - are MPEG-7-based Java tools which combine automatic extraction of low-level MPEG-7 descriptors with tools for manually annotating digital photos and images with semantic tags. The resulting metadata is stored as an MPEG-7 XML file which is used to enable content-based image retrieval.
- The **MPEG-7 SpokenContent Description Scheme Extractor** automatically recognizes speech, on which one can apply text-related annotation methods. The same applies for **Transcriber** [34].
- **M-OntoMat-Annotizer** [35] is a tool that allows the semantic annotation of images and videos for multimedia analysis and retrieval. It provides an interface for linking RDF(S) domain ontologies to automatically extracted low-level MPEG-7 visual descriptors.

Table 2: A classification of approaches for semantically annotating multimedia

System	Format Type	Nature	Technique
AktiveMedia	Images	Semi-automatic	Low-level semantics
Caliph	Images	Automatic	Low-level semantics
SWAD	Images	Automatic	Low and high-level semantics
MPEG-7 SCDSExtractor	Audio	Automatic	Speech
Transcriber	Audio	Automatic	Speech
4M	Video	Automatic Semi-automatic	Low-level semantics Machine-learning
M-OntoMat-Annotizer	Video	Automatic Semi-automatic	Low-level semantics Machine learning

6 Anticipated Future Trends

Named Entity Recognition has been a thriving field of research for almost 20 years. Over this time it has expanded to cover many different languages, domains, textual genres (email, news articles, blogs, tweets, web pages) multimedia and entity types. It has migrated from handcrafted rules-based approaches to machine learning approaches. Although supervised learning approaches (e.g., Hidden Markov Models) can achieve excellent results, they depend on the availability of a large corpus of annotated data. Although there are some such corpuses available, they are limited to specific domains and languages. Hence the research focus has shifted to semi-supervised (bootstrapping) and unsupervised learning techniques that don't require a large annotated corpus. It has also increasingly focussed on scalable approaches that work on large-scale collections of unstructured web documents [36-37].

The majority of research in this area is presented at the following annual conferences: CONLL (Computational Natural Language Learning), MUC (Message Understanding Conference), LREC (Language Resources and Evaluation), ACE (Automatic Content Extraction Program) and COLING (International Conference on Computational Linguistics). Monitoring of these conferences will continue to provide the most up-to-date information on the current state of the field. But increasingly, there are also relevant publications at the WWW (World Wide Web), ISWC (International Semantic Web) Bioinformatics and Digital Humanities (DH) conferences.

Of particular relevance to the HASS community is the AMICUS (Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts) project which aims to bring together researchers applying text processing to cultural heritage data, prominently narrative texts, such as folklore and scientific texts, to identify recurring motifs and patterns.

Also of relevance are the "Digging Into Data" projects [9] jointly funded by the NEH, NSF, JISC and SSHRC. It is a pity that Australia was not a partner in this multi-lateral e-humanities grant program between USA, Canada and the UK.

One significant emerging area is semantic publishing tools – tools that enable users to create and publish content with semantic markup already embedded. Some examples of such approaches include:

- OpenPublish – Thomson Reuters and Phase2 Technology recently released OpenPublish which combines Drupal with OpenCalais machine-assisted tagging and built-in RDFa formatting, to semantically tag textual documents as they are published <http://drupal.org/project/openpublish>
- Jiglu Insight and Jiglu Spaces – are commercial products that automatically tags content, finds hidden relationships to other content that’s been published and automatically creates links. <http://www.jiglu.com/>

Other highly topical and emerging areas of research that are relevant to this report include:

- Automatic semantic annotation of dynamic web documents, such as blogs, wikis and twitter/tweets i.e., unstructured textual resources that are constantly changing.
- Standardized interoperable annotation models such as the Open Annotation Collaboration [16], that promotes a common model based on Linked Open Data and URIs to ensure persistent and re-usable tags/annotations
- Hybrid semi-automatic semantic tagging systems – that combine machine-learning with rules-based approaches and crowd-sourcing to generate the training set and correct the results. For example, Finin et al, use Amazon’s Mechanical Turk to tag the named entities in twitter data [40].
- The application of cloud computing to high performance, large scale text analysis and named entity recognition e.g., Gate Cloud <http://gatecloud.net/>

7 Conclusions and Recommendations

7.1 Assessment Results

The above review of existing semantic tagging tools identified two specific candidates (OpenCalais and GATE) that warrant a more detailed evaluation. These two tools were chosen because of their relative stability, widespread adoption, previous applications, ease-of-use, flexibility and comprehensive documentation and open source community support. Table 3 below shows the outcomes of the detailed assessment of these two systems based on the criteria listed in the Appendix.

Table 3: Detailed Assessment Results for OpenCalais and GATE

Assessment Criteria	Open Calais	GATE
Open source or commercial? Free or licensed?	Commercial closed source. Web service is free to use for up to 50,000 requests per day, 4 transactions per second. Larger quotas can be purchased (e.g., Open Calais Professional: starting at US\$2,000 per month)	Free, Open Source (LGPL) No limitations on use.
Standards-based	Calais makes use of Semantic Web standards (RDF, OWL) and adheres to Linked Data principles for entity identifiers. The web service is built on top of standard HTTP and SOAP.	GATE developers are involved in the ISO technical committee (TC37) concerned with identifying, accessing and managing resources in language technology applications, and uses web standards such as XPointer and XML.
Maturity/robustness	Launched by Reuters in 2008, Open Calais is a stable, robust service used by a number of	Mature project led by the University of Sheffield with contributions from commercial

	<p>high profile online publishers. There is no Service Level Agreement (specifically no guaranteed uptime or response time) for free accounts; however, Open Calais claims 99.99% uptime.</p>	<p>partners and independent developers. GATE was developed in 1996, and the development team runs multi-platform regression tests on a continuous integration server to ensure robustness of the system.</p>
Platform independence	<p>Web service can be accessed from any platform.</p>	<p>Runs on platforms where Java 5.0 is available: Windows, Mac OS X, Solaris and Linux.</p>
Efficiency/Performance	<p>The speed with which Open Calais can process large collections of documents is limited by the API usage restrictions. Open Calais and their partners are continually training their system to provide accurate results for content from their application domain of online news; however this is an opaque process, and results cannot be tuned by end-users of the service.</p>	<p>GATE includes a benchmarking tool that tracks system performance across a corpus over time, using metrics such as Precision and Recall, allowing system performance to be finely tuned.</p>
Need for training corpus	<p>Not required, cannot be trained.</p>	<p>GATE can be configured to recognise entities using a rule-based grammar, and a gazetteer to identify named entities, which may be mapped from an OWL ontology. A training corpus is not required, however, <i>GATE Developer</i> optionally allows 'Gold Standard' data be used for evaluation and training machine learning algorithms. <i>GATE Teamware</i> supports the creation of training data from annotations created manually by a group of users.</p>
Scalability	<p>Open Calais is ideal for on-demand tagging of small documents. API usage restrictions make it less suitable for tagging large collections of long documents: free accounts are limited to 4 requests per second, professional accounts are limited to 20 per second, and all requests are limited to 100K input size. However an application built on top of Open Calais could implement caching to reduce the number of requests and aggregate results from multiple requests to overcome the input size restriction.</p>	<p>GATE is designed to be scalable, with a robust, modular architecture which supports load-on-demand from distributed data stores.</p> <p>The GATE Cloud Paralleliser (A3) was also recently released to support parallel execution of semantic tagging processes over large numbers of documents.</p>

Interoperability with other systems	<p>OpenCalais provides the following extensions and plug-ins:</p> <ul style="list-style-type: none"> • Pipes service for integration with Yahoo! Pipes for RSS. • Integration module for Microsoft Office SharePoint Server 2007 (MOSS 2007). • Tagaroo plug-in for integration with WordPress blogs. • Gnosis Firefox extension <p>Third-party tools provide interoperability with other systems, for example, OpenPublish for Drupal CMS, Oracle Semantic Technologies platform.</p>	<p>GATE's Component model supports plug-ins providing interoperability with other systems e.g. for visualisation, machine learning and parsing and processing (LingPipe, OpenCalais, OpenNLP, UIMA). Supports JDBC connections to relational databases such as PostgreSQL and Oracle for input and output.</p> <p>Ontotext's MIMIR repository web application provides semantic indexing and search services over GATE.</p> <p><i>GATE Embedded</i> can be integrated into other Java-based systems.</p>
Tailorability – or discipline specific	<p>OpenCalais does not support custom vocabularies.</p> <p>The Calais ontology represents entities, events and facts related to the domain of online news (particularly political, business and entertainment news), and is available in OWL/XML format.</p>	<p>The rules and gazetteers used to recognize ontology entities are completely customisable.</p>
Range of input formats – batch input?	<p>Allows Text or HTML input. Input is limited to 100K (larger documents must be split and parts submitted separately). Primarily supports English (with limited entity recognition in French and Spanish). Calais does not support batching or aggregation of metadata across collections of documents.</p>	<p>Handles input as Text, HTML, SGML, XML, RTF, MS Word, PDF and email. Additional document formats can be added via custom Java classes via <i>GATE Embedded</i>.</p> <p>Can process input in many languages.</p> <p>Supports batch processing.</p> <p>Supports processing corpora.</p>
Tag representations	<p>RDF/XML; JSON; Microformats; SimpleFormat (XML)</p>	<p>Natively uses XML; XHTML; Java serialisation</p> <p>Export to other representations available via plug-ins.</p>
Availability of API	<p>Calais web service provides SOAP and REST-based APIs. Developers must sign up for a free API key in order to use the APIs.</p>	<p><i>GATE Embedded</i> library provides a Java API.</p>
Web service or application	<p>Web service</p>	<p>GATE is a standalone suite of tools:</p> <ul style="list-style-type: none"> • <i>GATE Embedded</i>: Java library • <i>GATE Developer</i>: graphical IDE • <i>GATE Teamware</i>: Workflow-driven web-app <p><i>GATE cloud</i> web service alpha is available to GATE partners</p>
Ease of use	<p>Easy to use: no configuration or set-up required prior to use. Web Service APIs are straight-</p>	<p>Requires specialist knowledge to configure to use custom ontologies and to tune for speed</p>

	forward to use and very well documented on the Open Calais web site. Low barrier to implementing applications using Open Calais: APIs can be accessed from many programming languages, including JavaScript	and accuracy. Extensive documentation and community support provided (via a mailing list). Commercial training modules are also available. Installing and integrating GATE with other systems requires knowledge of Java.
Extent of deployment (list projects)	Projects and companies using Calais include: Associated Newspapers, The British Library, CBS Interactive, CNET, The Huffington Post, Powerhouse Museum, VUE (Tufts), Al Jazeera, Associated Content	Around 35,000 downloads per year. Projects using GATE include: Greenstone (Waikato), Perseus (Tufts), CLARIN (Utrecht), UK National Archives, European Heritage On-Line (ECHO)
Other comments	Conditions of use include: Users must display the Calais logo, with a link to the Open Calais home page from the application or web site utilising the tags. Users must also incorporate the Calais-provided GUIDs when disseminating Calais-derived metadata.	

Although the OpenCalais web service is simple to use, robust and efficient when applied to general web articles and news items, it has a number of limitations in comparison to Gate. The primary limitations are: it is primarily aimed at extracting companies, products and events; it does not support custom vocabularies; and it is only free up to 50,000 requests per day.

The major advantages of GATE over OpenCalais are that: it is open source; highly customizable for specific domain applications; has a large community of developers who are regularly providing new tools⁹. Of particular interest is the recent development of GATE Mimir - a semantic repository (based on BigOWLIM) for storing, indexing and querying semantic text analysis output from GATE [44]. Also the recent development of the GATE Cloud Paralleliser enables fast, parallel execution of semantic annotation of large-scale documents over cloud computing (e.g., Amazon EC2).

7.2 Recommended Next Steps

We recommend that the next step should be the development of a pilot project in collaboration with the Australian humanities research community - to evaluate GATE (as a semi-automated semantic tagging service) in the context of an e-humanities project.

The aim of the pilot project would be to demonstrate the potential of text and data mining within a significant Australian historical context. It will also provide a working model for how a large corpus can be analysed online and semantically linked to other online collections to infer new knowledge. More specifically we recommend that:

- A set of digitized manuscripts/letters or texts, of significant historical and/or cultural significance should be selected as a test-bed - for processing and semantic tagging (e.g., Australian War Memorial War diaries, the Captain Cook Endeavour Journal, a sub-set of the NLA Newspaper Digitization collection). The selected digitized materials should facilitate interdisciplinary and collaborative research across the Australian university sector. For example, analysis of the Cook journals may add temporal depth to environmental datasets.

⁹ New GATE stuff, summer 2010 <http://gate.ac.uk/family/coming-soon/>

- Relevant entities and controlled vocabularies should be identified (see Section 7.3);
- The relevant GATE tools should be customized to specifically support the chosen entities, vocabularies and collection of texts. Then their performance evaluated based on the speed and accuracy of the results for the chosen testbed collection of texts.
- The extracted named entities/tags should be stored as RDF in an RDF triple store (i.e. the GATE Mimir (Multi-paradigm Information Management Index and Repository) semantic repository), with pointers to the precise textual segment/word (using XPointer or TEI) in both the text and scanned image of the manuscript.
- A Web interface should be developed that enables invited contributors (authenticated users) to view the automatically extracted semantic tags and correct/improve them as required
- The corrected semantic tags should be exposed to the Semantic Web and used to link the chosen collection of texts to other Linked Open Data (LOD) resources e.g., DBPedia, FreeBase, Geonames, UMBEL etc.
- RDF Graphs showing relationships between and across the textual documents should be generated and displayed as visualizations
- Semantic search, discovery and navigation services (using SPARQL over Mimir) should be developed over the semantically annotated collection – to enable browsing via social network graphs, genealogies, maps and timeline interfaces.

7.3 Recommended Vocabularies

Named entity recognition systems currently use the ENAMEX tags (*person, organization, location*) and TIMEX tags (*date, time*) expressed as embedded XML – it would be advisable to continue to use these tags. The following sub-categories are also available: *city, country, state/province, river*. Further useful sub-categories could be defined, specific to the discipline/application.

Rather than embed the markup in the document (as XML or RDFa), it is recommended that the markup be expressed as RDF and stored separately in an RDF triple store (such as Sesame) with pointers into the text (e.g., using Xpointer).

The Australian Name Authority File, as used by PeopleAustralia, would be the ideal source of people names to be used when identifying unique individual persons mentioned in texts (assuming the input documents are relevant to Australian history and culture).

For locations, GeoScience Australia's Gazetteer of Australia 2008 contains 296,636 geographical names in Australia provided by members of the Committee for Geographic Names in Australasia: https://www.ga.gov.au/products/servlet/controller?event=GEOCAT_DETAILS&catno=65589

Regarding concepts, the Australian extension to the LCSH, that comprises additional Australian subject headings and references adopted for use in ABN, should be used: <http://www.nla.gov.au/librariesaustralia/cataloguing/auth/aust-lcsh.html>

7.4 Infrastructure and Architecture

Figure 1 shows the optimum infrastructural components/services and their integration within an overall system architecture. The optimum set of infrastructural components, that facilitates wide-spread access to the services and ensures interoperability between collections and tags, comprises the following components:

- Online access to collections of textual documents. Ideally these documents would be identified via unique persistent identifiers (URIs) and be encoded as HTML, XML or TEI (to simplify pointers to textual segments within the documents);
- An Automatic Semantic Tagging Web service that identifies and tags *people, places, date/time* and *concepts* that occur within specified textual documents. In addition to supporting the automatic creation of such tags, this service should ideally also support:
 - User corrections to the automatically generated tags;
 - Querying and browsing of semantic tags

- Visualization of tags and RDF graphs showing relationships both within and between documents
- A scalable RDF Triple Store (e.g., OWLIM¹⁰) – that stores the generated tags in a format that is conformant with the Open Annotation Collaboration specification [16]. Search and query of the tags in the triple store should be via a SPARQL querying interface.
- Optional output of the tags as either embedded RDFa or as ATOM feeds.
- A Controlled Vocabulary Registry – that stores controlled vocabularies including tag names (e.g., entity tags (person, organization, place, date, time) and domain-specific instances (people names, gazetteers, domain-specific concepts). This will be used to customize the Semantic Tagging service for a specific domain.

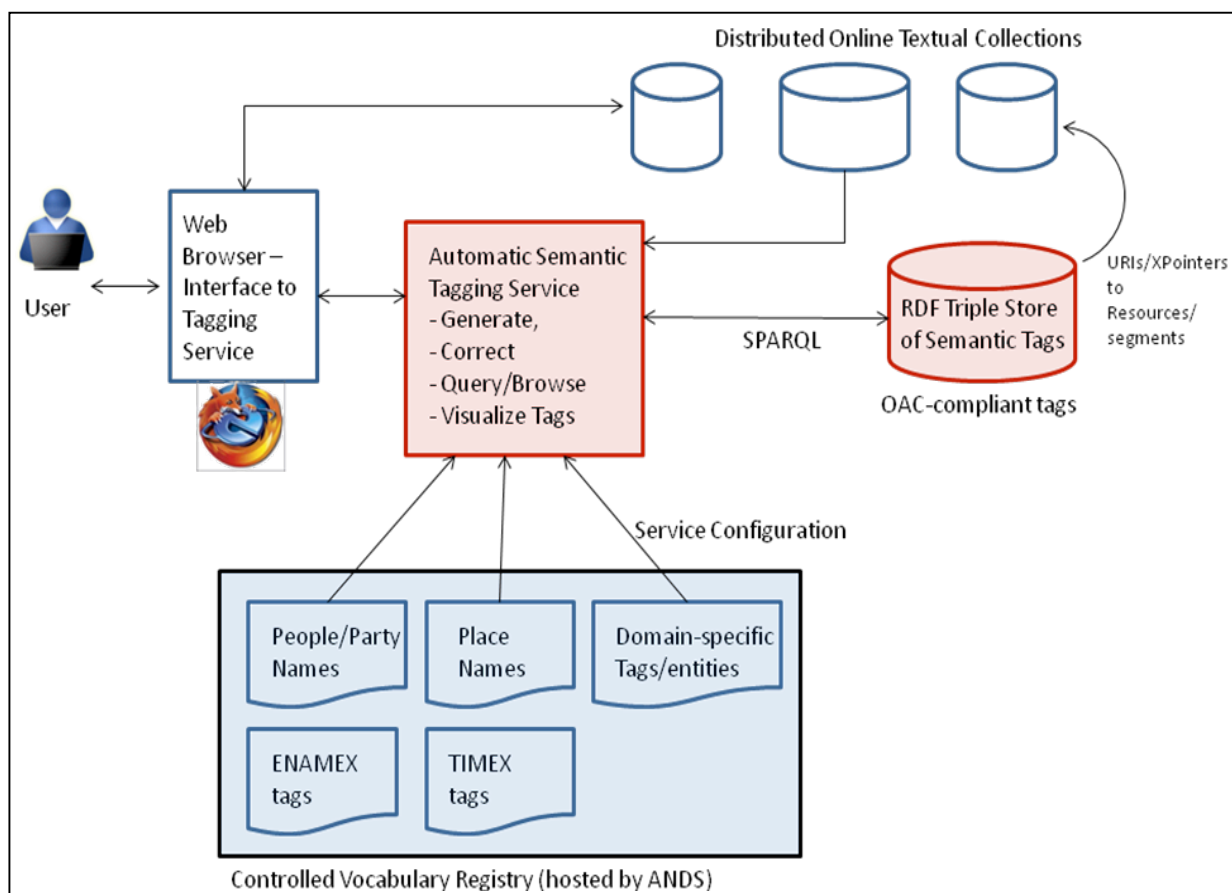


Figure 1: Integrated Set of Services and Optimum System Architecture

7.5 Service Providers

Further discussions are required before specific service providers can be identified. However it is envisaged that the resulting *Who/What/When/Where Semantic Tagging* Web service and the associated *Semantic Tag Triple Store* would fall within the scope of the National eResearch Collaboration Tools and Resources (NeCTAR) project's eResearch tools (RT) program. These two components are shown in the red in Figure 1.

In the longer term, an organization such as the National Library of Australia, may be interested in hosting the service, assuming that a suitable business model could be developed to cover the costs of support and maintenance.

¹⁰ <http://www.ontotext.com/owlim/>

The Controlled Vocabulary Registry (in blue in Figure 1) falls within the scope of the Australian National Data Service (ANDS) and hence it is anticipated that it would be provided, hosted and maintained by ANDS.

7.6 Conclusions

Humanities and Social Sciences covers a very broad range of sub-disciplines – including history, literature, linguistics, architecture, art history, ethnography, anthropology and archaeology. However language is central to the majority of humanities scholarship and hence textual processing is a common requirement across many digital humanities projects. Named entity recognition (NER) in particular, provides a way of making sense of large collections of unstructured text. It adds a semantic layer to the massive (manuscript, book, newspaper and theses) digitization projects currently being undertaken – exposing new relationships that might not otherwise be evident.

Although named entity recognition techniques are improving, they are still limited to a relatively small number of domains (news, sports, business). Domain-specific approaches are able to achieve reasonable performance for limited numbers of entity types on well-circumscribed corpora. However they don't easily permit customization or tailoring for domains other than the one for which they were designed. Although rules-based approaches are improving, machine-learning techniques that rely on large training corpuses still demonstrate the best performance within specific domains.

Our conclusion is that the optimum approach to named entity recognition is a combination of machine learning and user-input – both in the form of seeding rules and also corrections and feedback. This will require customization of one of the more robust, existing open source systems (GATE). Furthermore, by expressing the resulting high quality, automatically generated named entities in RDF (using controlled vocabularies) and storing them in the interoperable OAC-compliant format within a scalable RDF triple store (e.g., OWLIM), we open up the tagged textual collections to the Semantic Web and other Linked Open Data resources – enabling the development of richer search and browsing services and the inferencing of new relationships and knowledge.

8 Author contact details

Jane Hunter
Director eResearch Lab,
School of ITEE, The University of Queensland
St Lucia, Qld, Australia
Ph 617 33651092
Mob 0402 395797
Email: j.hunter@uq.edu.au

9 References

- [1] The Australian National Data Service (ANDS) <http://ands.org.au/>
- [2] NeAT National eResearch Architecture Taskforce Projects <http://ands.org.au/neat-projects.html>
- [3] Platforms for Collaboration Investment Plan <https://www.pfc.org.au/bin/view/Main/WebHome>
- [4] The Education Investment Fund
<http://www.innovation.gov.au/Section/AboutDIISR/FactSheets/Pages/EducationInvestmentFund.aspx>
- [5] The eResearch Lab, School of ITEE, The University of Queensland
<http://www.itee.uq.edu.au/~eresearch>
- [6] Text Mining for Scholarly Communications and Repositories Joint Workshop
<http://www.nactem.ac.uk/tm-ukoln.php>
- [7] TAPoR Text Analysis Portal for Research at the University of Alberta <http://tapor.ualberta.ca/>

- [8] Voyeur Tools: See Through Your Texts <http://hermeneuti.ca/voyeur>
- [9] Digging into Data <http://www.diggingintodata.org/>
- [10] L. Ratnov and D. Roth, Design Challenges and Misconceptions in Named Entity Recognition, CoNLL 2009 <http://portal.acm.org/citation.cfm?id=1596374.1596399>
- [11] D.Nadeu, S.Sekine, "A survey of named entity recognition and classification", 2007 <http://nlp.cs.nyu.edu/sekine/papers/li07.pdf>
- [12] Chun-Nan Hsu, Yu-Ming Chang, Cheng-Ju Kuo, Yu-Shi Lin, Han-Shen Huang and I-Fang Chuang. Integrating High Dimensional Bi-directional Parsing Models for Gene Mention Tagging. *Bioinformatics* 24(13):i286-i294 <http://bioinformatics.oxfordjournals.org/cgi/reprint/24/13/i286>
- [13] Colin R. Batchelor and Peter T. Corbett, Semantic enrichment of journal articles using chemical named entity recognition Proceedings of the ACL 2007 Demo and Poster Sessions, pages 45-48, Prague, June 2007.
- [14] Peter Corbett, Colin Batchelor and Simone Teufel [*Annotation of Chemical Named Entities*](#). BioNLP 2007: Biological, translational, and clinical language processing, Prague, Czech Republic.
- [15] Semantic Annotation of Papers: Interface and Enrichment Tool (SAPIENT). Liakata M., Claire Q and Soldatova L. N. (2009) Proceedings of BioNLP 2009, p. 193--200, Boulder, Colorado.
- [16] Open Annotation Collaboration <http://www.openannotation.org/>
- [17] AMICUS Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts <http://ilk.uvt.nl/amicus/>
- [18] Lendvai, P., Declerck T., Daranyi S., Malec Sc., "Propp Revisited: Integration of Linguistic Markup into Structured Content Descriptors of Tales", Digital Humanities 2010, Kings College London, July 2010 <http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-753.html>
- [19] Council on Library and Information Resources (CLIR), "Working Together or Apart: Promoting the Next Generation of Digital Scholarship", Report of a Workshop co-sponsored by CLIR and NEH, March 2009 <http://www.clir.org/pubs/reports/pub145/pub145.pdf>
- [20] Dingli, A., Ciravegna, F. and Wilks, Y., Automatic Semantic Annotation using Unsupervised Information Extraction and Integration in *K-CAP 2003 Workshop on Knowledge Markup and Semantic Annotation*,(2003).
- [21] Ciravegna, F., Dingli, A., Wilks, Y., and Petrelli, D. 2002. Adaptive information extraction for document annotation in amilcare. In *Proceedings of the 25th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Tampere, Finland, August 11 - 15, 2002). SIGIR '02. ACM, New York, NY, 451-451.
- [22] Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., Goranov, M.: KIM – Semantic Annotation Platform. *Proc. of the 2nd International Semantic Web Conference*, Sanibel Island, Florida (2003)
- [23] Domingue, J., Dzbor, M., Motta, E.: Magpie: Supporting browsing and navigation on the semantic web. In: ACM Conference on Intelligent User Interfaces (IUI) (2004)
- [24] Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A., Ciravegna, F.: Mnm: ontology driven semi-automatic and automatic support for semantic markup. In: 13th International Conference on Knowledge Engineering and Management (EKAW2002) (2002)
- [25] Cimiano, P., Handschuh, S., Staab, S.: Towards the self-annotating web. In: Thirteenth International Conference on WorldWide Web (2004)
- [26] Handschuh, S., Staab, S., Ciravegna, F.: S-cream—semi-automatic creation of metadata. In: SAAKM 2002—Semantic Authoring, Annotation & Knowledge Markup (2002)
- [27] He Hu; Xiaoyong Du; , "ConAnnotator: Ontology-Aided Collaborative Annotation System," *Computer Supported Cooperative Work in Design, 2006. CSCWD '06. 10th International Conference on* , vol., no., pp.1-6, 3-5 May 2006

- [28] Dill, S., Gibson, N., Gruhl, D., Guha, R.V., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J.A., Zien, J.Y.: Semtag and seeker: bootstrapping the semantic web via automated semantic annotation. In: Twelfth International World Wide Web Conference (2003)
- [29] Choi, J. and Park, Y. 2007. Ontology-based automatic semantic annotation for named entity disambiguation. In *Proceedings of the 10th IASTED international Conference on intelligent Systems and Control* (Cambridge, Massachusetts, November 19 - 21, 2007).
- [30] Sheth, A., Bertram, C., Avant, D., Hammond, B., Kochut, K., Warke, Y.: Managing Semantic Content for the Web, *IEEE Internet Computing*, 6(4), (2002) 80-87
- [31] Zhu, J., Uren, V., Motta, E.: ESpotter: Adaptive Named Entity Recognition for Web Browsing, *Proc. of the 3rd Professional Knowledge Management Conference (WM2005)*, Kaiserslautern, Germany (2005)
- [32] Chakarvarthy, A., Ciravegna, F., Lanfranchi, V.: Cross-media document annotation and enrichment. In: 1st Semantic Authoring and Annotation Workshop (SAAW2006) (2006)
- [33] Lux, M., Becker, J., Krottmaier, H.: Caliph&Emir: semantic annotation and retrieval in personal digital photo libraries. In: CAISE Forum at the 15th Conference on Advanced Information Systems Engineering (2003)
- [34] Barras, C., Geoffrois, E., Wu, Z., Liberman, M.: Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication Special Issue on Speech Annotation and Corpus Tools* 33(1-2) (2000)
- [35] Petridis, K., Anastasopoulos, D., Saathoff, C., Timmermann, N., Kompatsiaris, I., Staab, S.: M-ontomat-annotizer: image annotation. Linking ontologies and multimedia low-level features. In: Engineered Applications of Semantic Web Session (SWEA) at the 10th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems (KES 2006)
- [36] Whitelaw, C., Kehlenbeck, A., Petrovic, N., and Ungar, L. 2008. Web-scale named entity recognition. In *Proceeding of the 17th ACM Conference on information and Knowledge Management* (Napa Valley, California, USA, October 26 - 30, 2008). CIKM '08. ACM, New York, NY, 123-132. DOI=<http://doi.acm.org/10.1145/1458082.1458102>
- [37] Downey, D., Broadhead, M. and Etzioni, O. Locating complex named entities in web text. In *IJCAI*, 2007. <http://turing.cs.washington.edu/papers/IJCAI-DowneyD1178.pdf>
- [38] Katharina Siorpaes and Elena Simperl, "Human Intelligence in the Process of Semantic Content Creation", *World Wide Web* Vol 13, No 1-2, 33-59, "Special Issue: Human-Centered Web Science; Guest Editors: Ernesto Damiani, Miltiadis Lytras and Philippe Cudre-Mauroux" <http://www.springerlink.com/content/r08076u01423023p/fulltext.pdf>
- [39] Reeve, L. and Han, H. 2005. Survey of semantic annotation platforms. In *Proceedings of the 2005 ACM Symposium on Applied Computing* (Santa Fe, New Mexico, March 13 - 17, 2005). L. M. Liebrock, Ed. SAC '05. ACM, New York, NY, 1634-1638
- [40] Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., and Dredze, M., "Annotating named entities in Twitter data with crowdsourcing" *Workshop on Creating Speech and Language Data With Mechanical Turk at NAACL-HLT*, 2010.
- [41] NeCTAR Consultation Paper, October 2010 http://nectar.unimelb.edu.au/docs/NeCTAR_Consultation_Paper_October_2010_FINAL.pdf
- [42] Nowack, B., "Linked Data Entity Extraction with Zemanta and OpenCalais", 28 July, 2010 <http://bnode.org/blog/2010/07/28/linked-data-entity-extraction-with-zemanta-and-opencalais>
- [43] DiCiuccio R., "Entity Extraction & Content API Evaluation", May 18, 2010 <http://blog.viewchange.org/2010/05/entity-extraction-content-api-evaluation/>
- [44] Ontotext, MIMIR Semantic Search Engine and Repository integrated with GATE <http://www.ontotext.com/kim/mimir.html>

10 Appendix: assessment criteria

The assessment criteria used for evaluating available tagging services listed in Section 5 include the following:

- Open source or commercial – is the service free or licensed?
- Standards-based – is the system based on open standards?
- Maturity/robustness – is the system robust and mature?
- Platform-independence – will the system operate on different operating systems?
- Efficiency and performance
- Need for training and a training corpus of manually marked up texts
- Scalability – will it scale to massive online collections?
- Interoperability – will the system accept input and generate output to enable interoperability with other systems?
- Flexibility and tailorability - is the system designed so that it is relatively simple and easy to make changes and adapt it to a different discipline/ontology?
- Input formats (HTML, Word, PDF, TXT?) – does the system support a range of input formats? Does it support batch input of multiple files?
- Tag representation(SKOS, RDF, RDFa etc) – in what format are the tags generated?
- Availability of APIs - does the system include an API for developers?
- Is the system available as a Web service or stand-alone application?
- Ease of use and usability
- Specificity – does the system only identify people names or places or discipline-specific entities? If the service is designed for a specific discipline, and if so, which discipline?