Artequakt: Generating Tailored Biographies with Automatically Annotated Fragments from the Web

Sanghee Kim and Harith Alani and Wendy Hall and Paul H. Lewis and David E. Millard
Nigel R. Shadbolt and Mark J. Weal[1]

Abstract. The Artequakt project is working towards automatically generating narrative biographies of artists from knowledge that has been extracted from the Web and maintained in a knowledge base. An overview of the system architecture is presented here and the three key components of that architecture are explained in detail, namely knowledge extraction, information management and biography construction. Conclusions are drawn from the initial experiences of the project and future plans are described.

1 INTRODUCTION

The growth of the World Wide Web (Web) and the corpus of documents that it covers increased the demand for content to be annotated. Such annotation facilitates systematic search and discovery of knowledge and intelligent information processing. Annotating existing Web documents forms one of the basic barriers towards realising the Semantic Web ([11], [25]).

Annotations can be roughly classified into two types. The first is concerned with identifying textual entities in documents that match information *already existing* in a knowledge base, e.g. the word 'Rembrandt' in the document is matched to a painter's name in the knowledge base. Such annotations are normally restricted to the type and amount of information held in the knowledge base. The other type of annotation is involved in locating *new* factual information in documents based on a given domain classification structure, e.g. 'Rembrandt' in the document is the 'name' of a 'Painter', where *Painter* is a class in the ontology with the relation *name*. This new fact can be asserted in the knowledge base. This second type is the main approach taken to annotation in the Artequakt project.

Previous work on annotation has demonstrated the value of coupling Natural Language Processing (NLP) with ontologies ([13], [23]). The ontology can guide the annotation task by restricting it to a specific domain and, unlike "rigid templates", can provide it with knowledge inference and conceptual browsing facilities [13]. An ontology-based approach for annotation needs to deal with the issues of duplicate information across documents, managing ontology change, and redundant annotations [22].

Annotation can exist in different forms and be used in a variety of ways. One interesting possibility is to use it to restructure the original source material in new ways, producing a dynamic presentation tailored to the users needs.

Previous work on the production of dynamic presentations has highlighted the difficulties of maintaining a rhetorical structure across a dynamically assembled sequence [20], as a consequence there has been a focus on dynamic presentation decisions as opposed to narrative ones [14]. Where dynamic narrative is present it has been based around robust story-schema such as the format of a news program (a sequence of atomic bulletins) [12].

It is our belief that by building a story-schema layer on top of an ontology we can create dynamic stories within a certain domain. By populating the ontolgy through automatic annotation software we could allow those stories to be constructed from the vast wealth of information that exists on the World Wide Web.

1.1 The Artequakt Project

The Artequakt project aims to implement such a system around the domain of artists and their paintings, automatically producing tailored biographies of artists from fragments of information extracted from the Web. This is *not* an attempt to out-perform hand-crafted biographies, but rather to gather information from a wide variety of sources and target it specifically at the interests of a particular reader. The first stage of this project consists of developing an ontology for the domain of artists and paintings. A selection of information extraction tools and techniques are being developed and applied that attempt to automatically generate annotated content from online documents based on the project's ontology and WordNet lexicons. The annotations are stored in a knowledge base and will be analysed for duplications. In the second stage, narrative construction tools are being developed to query the knowledge base through an ontology-server to search and retrieve relevant facts or textual paragraphs and generate a specific biography. The automatic generation of tailored biographies is concerned with two areas of focus. Firstly, providing biographies for artists where there is sparse information available, distributed across the web. This may mean constructing text from basic factual information gleaned, or combining text from a number of sources with differing interests in the artist. Secondly, the project aims to provide biographies that are tailored to the particular interests and requirements of a given reader. These might range from rough stereotyping such as "A biography suitable for a child" to specific reader interests such as "I'm interested in the artist's use of colour in their oil paintings".

The expertise and experience of three separate projects are drawn together under the umbrella of the Artequakt project. These are:

The Artiste project - A European project working on a distributed database of art images in collaboration with partners that include the Louvre, the Uffizzi Gallery, the National Gallery and the Victoria and Albert Museum.

The Equator IRC - An EPSRC funded Interdisciplinary Research Centre that, amongst many other activities, is investigating the use

[1] Intelligence, Agents, Multimedia Group, University of Southampton, SO17 1BJ, UK

of narrative techniques in information structuring and presentation.

The AKT IRC - An EPSRC funded Interdisciplinary Research Centre looking at all aspects of the knowledge lifecycle.

Although focussing on artists and their paintings, the techniques being developed could be applied to other domains.

This paper will examine the overall proposed Artequakt architecture, looking at the three main component parts, namely, knowledge extraction, knowledge representation and storage and narrative generation.

2 ARCHITECTURE OVERVIEW

Figure 1 illustrates the systems architecture used for the initial Artequakt demonstrator. Three key areas can be identified.

The first concerns the knowledge extraction tools. These are to be used to extract factual information items together with sentences and paragraphs from web documents that might be manually selected or obtained automatically using appropriate search engine technology. The fragments of information are passed to the ontology server along with metadata derived from the vocabulary of the ontology.

The second key area is the information management and storage. The information is being stored by the ontology server and consolidated into a knowledge base, focused on artists and paintings.

The final key area, is the narrative generation. The Artequakt servlet will take requests from a reader via a simple web interface. The reader request will usually include an artist for whom to generate a biography in a particular style (chronology, through the paintings etc.) and also any user information; for example, the narrative might be generated specifically for a child or an art historian. The server then uses story templates to render a narrative from the information stored in the knowledge base. The rest of this paper will examine these three areas in more detail.

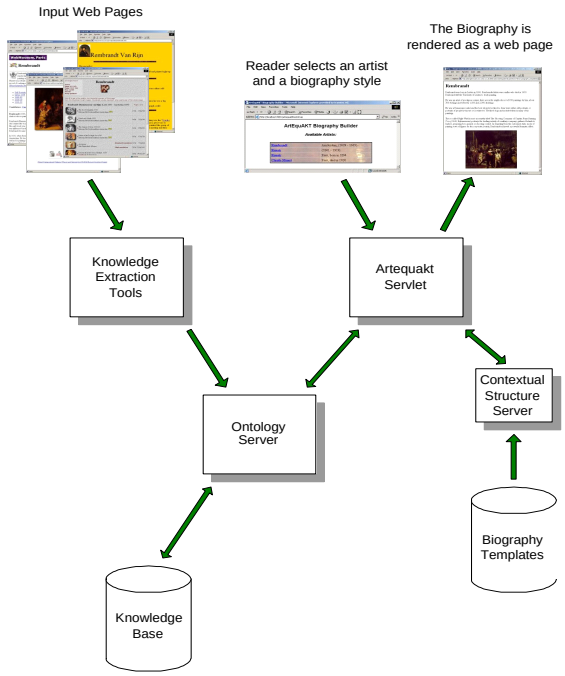

Figure 1. The Artequakt Architecture

3 KNOWLEDGE EXTRACTION

The aim of the knowledge extraction section is to extract and identify factual information from the Web-based documents and to structure it appropriately for entry into the knowledge base. Much of the information from the Web is in the form of natural language documents. One of the promising approaches to providing easy access to such documents is centred on information extraction that reduces them into tabular structures from which the fragments of documents can be retrieved as answers to queries. However, the effort and time needed for annotating a large number of texts and the prerequisite of acquiring background knowledge that stipulates which types of information are extractable, are major challenges toward exploiting such extraction techniques for practical purposes([25]). Work such as ([4],[13]) investigated the application of machine learning techniques in order to automatically identify patterns from annotated example texts.

In particular, whilst many attempts have been made to extract information from the Web by using manually annotated texts, no robust and reliable methodologies are yet available. Documents from the Web use limitless vocabularies, structures and/or composition styles for defining approximately the same content, implying that it is of little use to make efforts to locate recurrent syntactic patterns. For example, although content similarity between two biographic documents might be expected, expressions used for both sources may vary dramatically.

These observations have led us initially to use a natural language-based extraction approach for a comparatively deeper content understanding from which various clues concerning semantic and syntactic features can be obtained. The use of an ontology coupled with a general-purpose lexical database (WordNet [17]) as a guidance tool for creating interesting relations is another dimension of our initial approach aiming at minimising reliance on domain-specific extraction rules. Figure 2 shows extraction results based on the example of *'Rembrandt's father was a miller who died in 1630'*. Two biographic pieces of information about 'Rembrandt's father' (i.e. 'job_title (miller)' and 'date_of_death (1630)'), were captured as well as the fact that 'Rembrandt' is a person and he is the *son* of a dead miller.

3.1 Natural Language Information Extraction

The capability of recognising a named entity without the annotation effort of humans or without the need to create extraction rules is one of the objectives of our approach. The idea is to make use of general-purpose lexical databases and to exploit the knowledge from syntactical and semantic analysis to clarify the types and structures of given information. Although the proposed approach may not be as sophisticated as manually annotated definitions, its contribution lies in its extensibility and practical nature (acceptable performance). We use a paragraph as a unit of semantic analysis instead of a sentence, since much of the critical information used for interpreting text is scattered in different sentences (as observed in [3]). Downloaded documents from the Web are first divided into paragraphs, which are then broken down into a group of sentences. The paragraphs are analysed as follows:

1. Syntactical analysis: A sentence is decomposed into a set of grammatically related phrases (e.g. a verb-phrase, or a noun-phrase). We have used the Apple Pie Parser, which is a bottom-up probabilistic chart parser and is freely available [21].
2. Semantic analysis:

- *Identification of main components*: each compound sentence is decomposed into simplified structures, each of which contains one clause, i.e. a simple sentence. Each clause is clustered as one of three parts: subject, verb, and object. Temporal properties are inferred from a verb tense (e.g. 'past', 'present'), and associated with the sentence. A writing style (e.g. 'first-person', 'third-person') can be derived from the personal pronoun if it exists in the sentence's subject.
- *Recognition of named entity*: two resources are used for determining whether or not a given word denotes a person's name. The first is syntactical tags, which are obtained as the result of the syntactical analysis carried out by the Apple Pie Parser. The second is gazetteers of people names, which are available as part of the GATE (General Architecture for Text Engineering, [6]) package. GATE provides text files which contain person names associated with gender attributes. A name which is not defined in GATE's text files will still be extractable if it is tagged as a proper noun. Heuristics and grammar rules are applied in order to extract only proper personal nouns.
- *Resolution of pronoun references (anaphoric references)*: a personal pronoun refers to a specific person, and acts as a subject ('he' or 'she'), an object ('him' or 'her'), or a marker of possession defining who owns a particular thing ('his' or 'hers'). Currently we are using a simple resolution function that runs at reasonably fast speed obtaining the best-guessed referent. Three attributes (gender, number, and structural information) are considered in determining the right referent.
- *Adding a missing subject*: a clause can inherit a subject from a main clause, since it is syntactically dependent on the main clause.

In Figure 2, the given example *'Rembrandt's father was a miller who died in 1630'* is divided into two clauses. The same subject (i.e. 'Rembrandt's father') is assigned to both clauses since the second clause is dependent on the first one. At this stage, 'Rembrandt' was successfully recognised as a person's name. Gazetteers provided by GATE do not contain the name 'Rembrandt', whereas syntactic tags for this sentence mark it as a proper noun.

3.2 Relation Extraction

To create a binary relationship between two extracted individual facts, knowledge about the pre-defined semantic relations will be required. Consulting the ontology, which specifies various relationships among classes, will act as a basis for decisions concerning which relations to use. A query is submitted to the ontology server to obtain such knowledge.

In order to reduce the problem of linguistic variation between relations defined in the ontology and the extracted facts, we will use three lexical chains (synonyms, hypernyms, and hyponyms) as defined in WordNet. For example, the concept of 'depict' is matched with 'portray' (synonym) and 'represent' (hypernym). In order to reduce over- and under-generalisation, we will consider only one-level of hypernyms and hyponyms when a given word is a verb.

The types of information are identified by tracing the hierarchies of hypernyms. For example, as shown in Figure 2, 'miller' is extracted as the job of Rembrandt's father since the hypernyms map to 'worker'. Factual data, such as a date or a city name, are extracted by using a date parsing program coupled with a simple grammar and the hypernyms defined in WordNet. In cases, where there are multiple matches, all relations are represented in outputs.

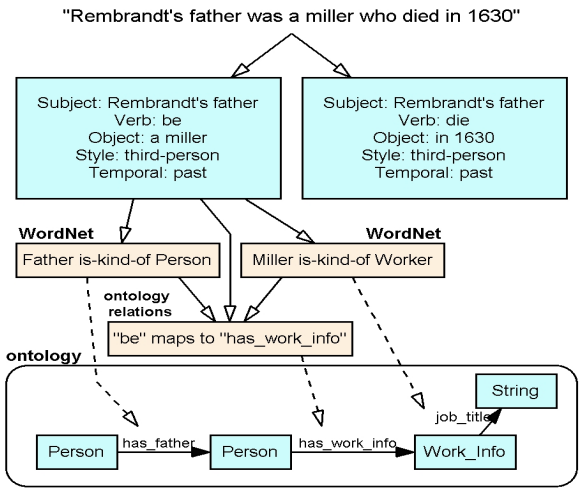

Figure 2. An example of knowledge extraction using ontology and WordNet

In Figure 2, relation extraction for both clauses is determined by the categorisation results of verbs (i.e. 'be' and 'die'). The 'be' verb poses a rather difficult case, since its semantic meaning is heavily dependent on other phrases, i.e. subject and object. According to WordNet definitions, one of its senses states 'work in a specific place, with a specific subject or in a specific function'. Since its synonyms (i.e. 'work' and 'follow') are matched with 'work', we exploit this relation to further examine whether or not it is related to 'job-information'.

In the second clause, since 'die' can be converted to the noun format 'death', the verb 'die' matches with two potential relations ('date_of_death' and 'place_of_death'). In this case 'date_of_death' was chosen since the '1630' was extracted from the same sentence and instantiated as date information.

The output from this section is an XML-formated representation of the facts, paragraphs, sentences and keywords identified in the knowledge extraction process. The XML files are sent to the ontology server to populate the knowledge base.

4 KNOWLEDGE REPRESENTATION AND STORAGE

4.1 Artequakt Ontology

An ontology is a conceptualisation of a domain into a machine readable format [7]. For Artequakt the requirement is to build an ontology to represent the domain of artists and artefacts. This ontology is being implemented in Protégé, which is a graphical ontology editing tool [18]. The main part of this ontology is being constructed from selected sections in the CIDOC Conceptual Reference Model (CRM - [5]) ontology. CRM was developed by ICOM/CIDOC[2] Documentation Standards Group to represent an ontology for cultural heritage information. It was built to facilitate the transformation of existing disparate museum and cultural heritage information sources into one coherent source.

The CRM ontology is designed to represent artefacts, their production, ownership, location, etc. This ontology was modified for Artequakt and is being enriched with additional classes and relationships to represent a variety of information related to artists, their personal

[2] http://www.cidoc.icom.org/

information, family relations, relations with other artists, details of their work, etc. The Artequakt ontology also allows the storage of textual paragraphs or sentences along with their source URLs so that at a later point they can be reorganised using the ontology as a guide.

4.2 Automatic Ontology Population

There is an increasing interest in building ontologies to provide a variety of knowledge services. Populating ontologies with knowledge is labour intensive and time consuming. Semi-automatic approaches to ontology population have been followed by for example [23] where relationships can be added automatically between instances if these instances already exist in the knowledge base, otherwise user intervention will be needed. OntoAnnotate [22] and OntoMat [8] are supporting tools of user-driven ontology-based annotations, where the produced annotations can be fed back to the ontology.

In this project we are investigating the possibility of moving towards a fully automatic approach of feeding the ontology with knowledge extracted from the web. As mentioned in section 3.2, this information is extracted with respect to the Artequakt ontology, and provided as XML files, one per document, using tags mapped directly from names of classes and relationships in the ontology. When a new XML file is produced (Figure 3(a)), it will be sent to the Artequakt ontology server which launches a program to parse the received file and populate the ontology with the newly provided knowledge (Figure 3(b)).

The ontology server is based on Java sockets and connected to the Artequakt knowledge base through the Protégé API. A limited inference engine is being built on this server to allow querying and the retrieval of specific information from the ontology, for example to get all paragraphs that mention the date of birth of a specific artist, get the artist of a painting, get all available facts about an artists, etc.

4.3 Consolidating the Knowledge-Base

When analysing web documents about selected artists, it will be inevitable that we extract duplicated information or even contradictory information. Handling such information is challenging for automatic ontology population approaches. Staab et al[22] stressed the problem of creating duplicate objects when extracting from different documents. They relied on manually assigned object-identifiers to avoid duplication. Our approach is attempting to identify and eliminate duplications automatically using a two-stage consolidation process.

The first stage is for the Artequakt ontology server to add all extracted information to the knowledge base regardless of what is already stored. This results in the creation of multiple instances of artists with possibly the same information (e.g. multiple instances of Rembrandt). The challenge is to identify which of these instances refer to the same artist, and which ones refer to genuinely different artists who happen to have the same name or information.

The second stage is to run a consolidation process to identify possible duplicate instances in the knowledge base, searching for clues in the rest of information available about these instances. This is why it is best to feed the new information to the knowledge base first (stage 1), which provide the consolidation process with more information to compare with.

The consolidation process involves applying a set of heuristics. Information extraction tools are sometimes only able to extract fragments of information about an artist, especially if the source document or paragraph is small or difficult to analyse. This results in the creation of new instances with only one or two facts associated with each, for example two artist instances with the name Rembrandt, but one instance has a location relationship to Holland, while the other has a date of birth relationship to 1609. One heuristic to apply here is to merge such shallow instances into one instance of Rembrandt with both location and date of birth relations, keeping the original source URLs of each fact.

Another heuristic is if two instances of same-name artists have equal values for their date and place of birth and death relationships, then these instances are likely to be duplicates, in which case they can be fused together as one instance, otherwise the two instances will stay separate. Such a heuristic helps to distinguish between same-name artists. The amount and type of information overlap between instances can be used to calculate a confidence value to indicate whether certain instances can be merged or left separate.

Another challenge in information consolidation is to identify exact matches. Identical information can exist in different versions. For example consider the sentences:

- Rembrandt was born in the 17th century in Leiden.
- Rembrandt was born in 1606 in the Netherlands.
- Rembrandt was born on July 15 1606 in Holland.

The sentences above provide similar information about an artist, written in different formats and specificity levels. To match the above sentences it will be necessary to enrich the current ontology with proper temporal and geographical representations. Some format varieties can be dealt with at the extraction level. For example the information extraction tools being used in this project can identify and extract dates in different formats, and provide it as day, month, year, decade, etc. This information could be fed to the temporal ontology and reasoned over to match between different time frames.

There has been much work on developing databases and gazetteers of place names, such as the Thesaurus of Geographic Names (TGN, [9]), Alexandria Digital Library (ADL, [10]), and WordNet which also provides some geo-information. Such sources can be integrated with the current ontology to provide knowledge on geographical hierarchies, place name variations, and other spatial information [1].

5 NARRATIVE GENERATION

While machines benefit from using structured ontologies to exchange information, human beings need a more intuitive interface. One of the most natural ways to do this is by story telling. There is a wealth of critical and philosophical thought concerning narrative that can be drawn on to assist in constructing a story (in this case a biography) from the raw information gathered. Figure 4 shows one way of viewing the layers that make up a narrative as proposed by Bal [2]. The raw facts and chronological collection of events in any particular tale is called the *Fabula*. For any given Fabula we could present the facts from different perspectives and in different sequences to produce a *Story*. We could then render any given Story into several different forms or *Narratives* (e.g. a film or novel).

In Artequakt the knowledge base can be thought of as our underlying fabula. To produce the eventual narrative (in our case pages of html) we need to first arrange sub-elements of the fabula into a sensible sequence and produce a story.

5.1 Biography Templates

The story structures we are using are human authored biography templates that contain queries into the knowledge base.

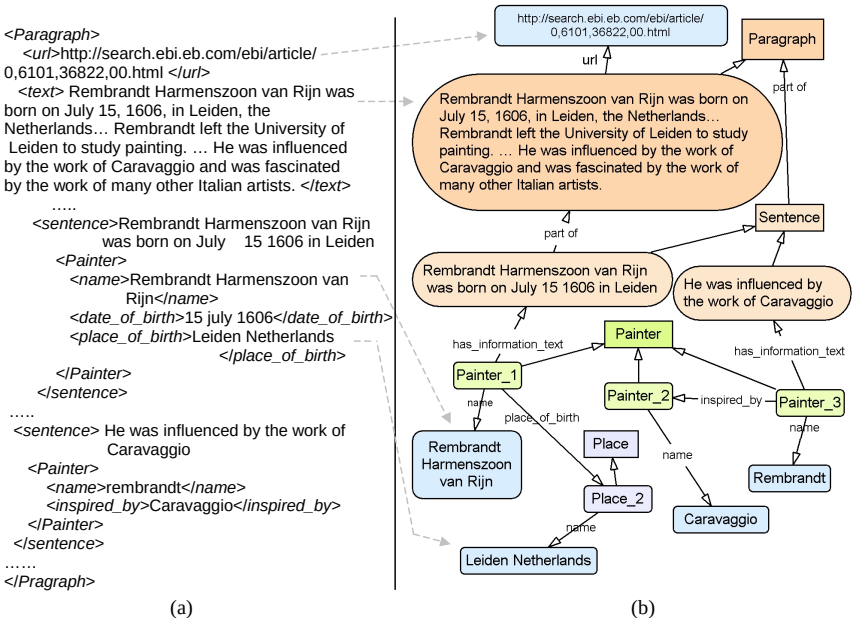

Figure 3. a) XML file of extracted information is sent to the ontology server, b) The server creates the relevant instances and relationships in the ontology.

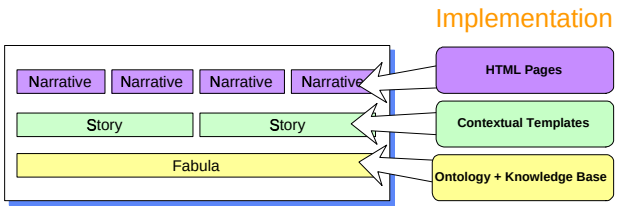

Figure 4. The Levels of Narrative

Previous work has stored queries into an ontological space as the destination of navigational links [24], by following the links the user causes the queries to be executed (and views the results). With Artequakt these basic links have evolved into more complex structures that arrange the queries into a sequence (a biography template).

The templates are written in XML using the Fundamental Open Hypermedia Model (FOHM) [16], which is capable of representing a variety of hypermedia structures including tours and links. The XML files are then loaded into the Auld Linky contextual structure server [15], which provides pattern matching facilities over the structures via HTTP.

Any given biography template may be constructed from several sub-structures. The basic structure used is the *Sequence*. This represents a list of queries that have to be instantiated from the knowledge base and inserted into the biography in order. These queries are authored using the vocabulary of terms defined within the ontology. Other structures allow more complex effects. A *Concept* structure contains several queries, any of which may be used at this point in the biography. A *Level of Detail* (LOD) structure is similar to a concept, but there is an ordering between the queries that corresponds to preference (i.e. preferably the highest numbered query should be used, if that's not possible the next highest, and so on). These structures may be nested (e.g. a sequence of concepts).

Some queries may retrieve paragraphs directly while others query the consolidated ontology for specific facts and construct sentences dynamically from the results. This can be useful for facts that have been inferred (and therefore there is no corresponding paragraph), or when there is no paragraph that fits the literary form of the rest of the biography (e.g. the biography is in third person, but all the available paragraphs are in first person).

The templates also contain contextual information on which parts of the biography structure are appropriate in different contexts (specified as a list of tag value pairs inside a context object). For example imagine that the user has specified that they do not have a good knowledge of artists. The template structure can specify that parts of the structure are only available to people with a good knowledge. Thus, when the user queries Linky for the template, the inappropriate parts that require this are pruned away.

Figure 5 shows an example structure being pruned. In this case a query into the ontology concerning artistic influences (here it would resolve into a sentence about Caravaggio) is removed because it would not make sense to a user who did not have a reasonable knowledge of artists. The resulting paragraph reads:

'Rembradt Harmenszoon van Rijn was born on July 15 1606 in Leiden. Rembradt's father was a miller who died in 1630. His early work was devoted to showing the lines, light and shade, and color of the people he saw about him.'

In this way the biography structures will be tailored to the needs of each individual user. For our prototype we are concentrating on broad user classification (child/adult, etc) but it would also be possible to incorporate more sophisticated user modelling techniques (such as training sets [19]).

Once it has been retrieved from Linky the template has to be instantiated, by making each query in turn and then rendering the results into a html page for display.

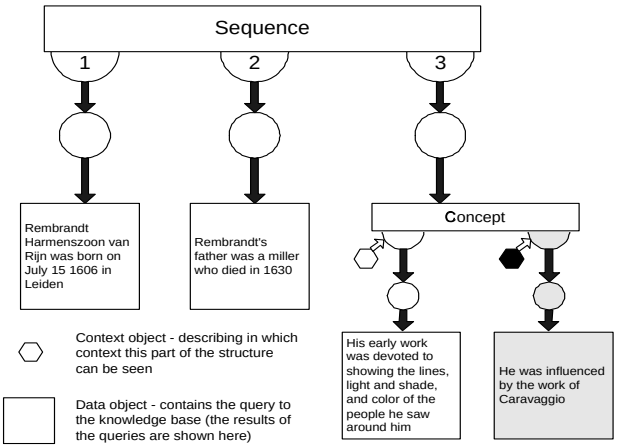

Figure 5. Template pruning: The black context (representing knowledge of artists) has failed, resulting in the shaded structure being pruned.

6 CONCLUSION & FUTURE WORK

In this paper we have described the basic architecture and initial work in the Artequakt project. Our aim is to be able to generate automatically tailored biographies from a knowledge base which has been automatically populated by annotating text fragments extracted from Web documents.

We are currently working on completing the initial prototype system by integrating the three main components identified in the architecture. We will then be able to assess the effectiveness over real data sources and begin the process of refining the constituant parts to improve the overall quality of the biographies served by the system.

Although some of the research issues in this process are particularly challenging, the final objective is to have an architecture in place which will allow us to explore some of the research issues that have arisen so far in more detail; for example, more comprehensive, automatic consolidation of knowledge bases, better techniques for knowledge extraction and more sophisticated narrative structuring of the knowledge fragments. To this end, progress has been made in the identification of an approach and the building of a prototype demonstrator for the project.

ACKNOWLEDGEMENTS

The work presented here is part of a larger project and we would particularly like to note the contributions of Hugh Glaser, Srinandan Dasmahapatra and David De Roure. This research is funded in part by EU Framework 5 IST project “Artiste” IST-1999-11978, EPSRC IRC project “Equator” GR/N15986/01 and EPSRC IRC project “AKT” GR/N15764/01.

REFERENCES

[1] H. Alani, *Spatial and Thematic Ontology in Cultural Heritage Information Systems,* Ph.D. dissertation, Computer Studies Department University of Glamorgan, U.K., 2001.

[2] M. Bal, *Narratology: Introduction to the Theory of Narrative,* University of Toronto Press, 1978. Trans. Christine van Boheemen. Torento. 1985.

[3] R. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, and V. Zue. Survey of the state of the art in human language technology, 1995.

[4] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery, ‘Learning to construct knowledge bases from the world wide web.’, *Artificial Intelligence,* (1-2), 69–113, (2000).

[5] N. Crofts, D.M. Dionissiadou, and M. Stiff, ‘Definition of the cidoc object-oriented conceptual reference model’, Technical report, International Organization for Standardization, (2000).

[6] H. Cunningham, K. Bontcheva, V. Tablan, C. Ursu, and M. Dimitrov, ‘Developing language processing components with gate (user’s guide)’, Technical report, University of Sheffield, U.K., (2002). available in http://www.gate.ac.uk/.

[7] N. Guarino and P. Giaretta, *Ontologies and Knowledge bases: towards a terminological clarification. Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing.*, IOS Press, 1995.

[8] S. Handschuh, S. Staab, and A. Maedche, ‘Cream - creating relational metadata with a component-based, ontology-driven annotation framework’, in *In Proceedings of the First International Conference on Knowledge Capture*, pp. 76–83, Canada, (2001).

[9] P. Harpring, ‘Proper words in proper places: The thesaurus of geographic names.’, *MDA Information,* (3), 5–12, (1997).

[10] L.L. Hill, J. Frew, and Q. Zheng, ‘Geographic names. the implementation of a gazetteer in a georeferenced digital library.’, *Digital Library*, (1), (1999).

[11] J. Kahan and M.-R. Koivunen, ‘Annotea: An open rdf infrastructure for shared web annotations’, in *In Proceedings of The Tenth International World Wide Web Conference, WWW10,* pp. 623–632, (2001).

[12] K. Lee, D. Luparello, and J. Roudaire, ‘Automatic Construction of Personalised TV News Programs’, in *In Proceedings of the Seventh ACM Conference on Multimedia, Orlando, Florida,* pp. 323–332, (1999).

[13] A. Maedche, G. Neumann, and S. Staab, *Bootstrapping an Ontology-Based Information Extraction System.*, Intelligent Exploration of the Web, Springer / Physica Verlag, 2002.

[14] C. Mancini, ‘From Cinematographic to Hypertext Narrative’, in *In Proceedings of the Eleventh ACM Conference on Hypertext and Hypermedia, San Antonio, Texas, USA*, pp. 236–237, (2000).

[15] D.T. Michaelides, D.E. Millard, M.J. Weal, and D. DeRoure, ‘Auld leaky: A contextual open hypermedia link server’, in *Hypermedia:Openness, Structural Awareness, and Adaptivity (Proceedings of OHS-7, SC-3 and AH-3), Published in Lecture Notes in Computer Science, (LNCS 2266), Springer Verlag, Heidelberg (ISSN 0302-9743)*, pp. 59–70, (2001).

[16] D.E. Millard, L. Moreau, H.C. Davis, and S. Reich, ‘FOHM: A Fundamental Open Hypertext Model for Investigating Interoperability Between Hypertext Domains’, in *HT00*, pp. 93–102, (2000).

[17] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, ‘Introduction to wordnet: An on-line lexical database’, Technical report, University of Princeton, U.S.A., (1993).

[18] M. A. Musen, R. W. Fergerson, W. E. Grosso, N. F. Noy, M. GrubeźY, and J. H. Gennari, ‘Component-based support for building knowledge-acquisition systems’, in *In Proceedings of the Conference on Intelligent Information Processing of the International Federation for Processing World Computer Congress, Beijing,* (2000).

[19] M. Pazzani and D. Billsus, ‘Learning and revising user profiles:the identification of interesting web sites’, *Machine Learning,* 313–331, (1997).

[20] L. Rutledge, B. Bailey, J. V. Ossenbruggen, L. Hardman, and J. Geurts, ‘Generating Presentation Constraints from Rhetorical Structure’, in *In Proceedings of the Eleventh ACM Conference on Hypertext and Hypermedia, San Antonio, Texas, USA*, pp. 19–28, (2000).

[21] S. Sekine and R. Grishman, ‘A corpus-based probabilistic grammar with only two non-terminals’, in *In Proceedings of the Fourth International Workshop on Parsing Technology,* pp. 216–223, (1995).

[22] S. Staab, A. Maedche, and S. Handschuh, ‘An annotation framework for the semantic web’, in *In Proceedings of the First International Workshop on MultiMedia Annotation,* Japan, (2001).

[23] M. Vargas-Vera, E. Motta, and J. Domingue, ‘Knowledge extraction by using an ontology-based annotation tool’, in *In Proceedings of the Workshop on Knowledge Markup and Semantic Annotation,* K-CAP’01,Canada, (2001).

[24] M.J. Weal, G.J. Hughes, D.E. Millard, and L. Moreau, ‘Open Hypermedia as a Navigational Interface to Ontological Information Spaces’, in *In Proceedings of the Twelth ACM Conference on Hypertext and Hypermedia, Arhus, Denmark*, pp. 227–236, (2001).

[25] R. Yangarber and R. Grishman, ‘Machine learning of extraction patterns from unannotated corpora: Position statement’, in *In Proceedings of Workshop on Machine Learning for Information Extraction,* pp. 76–83, ECAI,Berlin, (2001).