Enriching Audiovisual Collection Metadata with event-related concepts

Thesis Design
Master Information Sciences

Vrije Universiteit Amsterdam
Faculty of Science

Name:	Evangelos Trantos
Student Number:	2503700
e-mail:	trantos@student.vu.nl

Supervisors: Marieke van Erp
Roxane Segers

Table of Contents

1. Introduction

The rapid development of the Web, together with its ever growing infant, the Semantic Web, has undoubtedly opened a plethora of new research paths. Simultaneously, knowledge and information have entered the era and realm of digitization, i.e., the transformation-conversion of analog information into digital information. As a result, cultural heritage institutions offer their vast collection of exhibits to the public through their websites. These exhibits – we will refer to them as *objects* in the remainder of the document – come in the form of artworks (e.g. paintings), artefacts, audiovisual archives (e.g. collections of videos, stills, texts and audio) and even weapons.

The ‘Agora’ Project[1] utilizes these objects and tries to link them with events. It constitutes a collaboration between the Computer Science and History department of the Vrije Universiteit Amsterdam, the Netherlands Institute of Sound and Vision[2] (Beeld en Geluid) and the Rijksmuseum Amsterdam[3]. The objective of this collaboration is the development of a *“platform for interactive exploration of heterogeneous heritage collections, in which museum objects can be placed into an explicit (art)historic context. Through this context, objects from highly diverse museum collections can be related, resulting in a more complete and illustrated description of historical events”* [2]. This creates a pioneer illustration of object collections, since the users will be able to browse the collections based on the event that each object relates to.

This study will be undertaken within the Agora project. Specifically, utilized in this study will be objects from the EUscreen[4] and OpenImages[5] collections of the Netherlands Institute of Sound and Vision. Both of these collections contain audiovisual material with events from the beginnings of the past century until today. When exploring the objects by browsing through their websites, additional information appears for each of them. This information is also called *metadata.* It exists to acquaint the user with all the elements relating to the object’s creation and with background information concerning the object. In other words, it places the object in a historical context, i.e., a historical chronology, where a sequence of events takes place. For instance, for every object in the collection, the EUscreen website provides additional metadata such as the title of the object, the broadcast date, the production year, the publisher of the object, its genre (e.g., news), etc. Finally, a summary and description of the object assist in placing it within a historical context. Thus, by exploring the metadata, and particularly when reading the description, one can establish a correlation between the object and a historical event. However, this correlation is not always obvious. The objective of this study is to utilize the existing metadata and link each object to an event or series of events. This way, the websites will be event-based browsable, generating a learning perspective for the users. The approach to be used is called metadata enrichment, i.e., creation of enhanced information that results in more accurate, complete and consistent metadata, and it will be discussed in the following sections.

[1] http://www.agora.cs.vu.nl
[2] http://portal.beeldengeluid.nl/
[3] http://www.rijksmuseum.nl
[4] http://www.euscreen.eu/
[5] http://www.openimages.eu/

The remainder of the document is structured as follows. Section 2 serves as a continuation of the introduction, while it also provides an overview of related work. Section 3 consists of placing the problem of this research into a context and of posing the appropriate research questions to tackle it. Section 4 constitutes the methodology section, where the phases of realizing the goal are described. In section 5, there is an illustration of the necessary planning, whilst, as the document draws to its close, the expected results are discussed described.

2. Theoretical Background

In this section, we will provide additional and explanatory information concerning the metadata schema of EUscreen and OpenImages. Furthermore, we will mention studies and approaches, which display goals similar to ours and perform metadata enrichment in order to reach them.

2.1 Metadata schema

The metadata of the objects constitute a momentous factor in linking the objects to events, for they embrace all the necessary information regarding the event they are associated with. Consequently, the more information about the objects, the easier it becomes to link them to events. The quality of this metadata, however, varies a great deal, depending on the object. In some cases, too little information might be provided, due to the fact that there is limited information about the object in the first place. Ordinarily, the standardized metadata of an object in a collection encompasses the following schema [1]: (i) the object ID, (ii) the object type, (iii) the date of the object creation, as well as more textually-intensive fields, such as (iv) the object title, (v) the creator of the object, and (vi) a textual description. The latter, as mentioned in the introduction, can provide additional insights, such as an object-event correlation, which is indispensable when creating an event-based browsable website.

The above schema rises from the Dublin Core standard [25], which is used to describe a wide range of web resources, such as audio, video and images. The Dublin Core metadata terms can also be used to provide interoperability for metadata vocabularies in the Linked Open Data cloud and semantic web implementations [26].

2.2 Related work

Hitherto, numerous studies that involved metadata enrichment have been conducted. One of them is, obviously, the research of the ‘Agora’ Project. More specifically, in [2], the metadata enrichment process is achieved by structuring the events in a historical event thesaurus. As a result, objects of different collections can be linked to each other. The data included were from the Dutch Wikipedia, the Rijksmuseum Amsterdam and the Netherlands Institute of Sound and Vision. The approach was concluded within four steps:

i. find the event names
ii. find event actors, locations and times,
iii. relate the event names, actors, locations and times, and
iv. link events to collections.

The reasoning of our methodology does not differentiate much from that. In particular, the first step of our study corresponds to steps (i) and (ii), for it involves acquiring the terms-concepts (e.g., names of people, locations, object ID) from the metadata schema. Moreover, after acquiring those concepts, they will be used to link the objects to events, which corresponds to step (iv) of [2].

Furthermore, Zervanou et al. conducted a study very similar to ours [4]. The goal of their study is the enrichment of existing cultural heritage metadata with "*automatically generated semantic content descriptors*". In detail, the overall process consists of six steps:

1. Text element extraction. The text snippet needed is found.
2. Language identification. Detection of the language used in the text snippets.
3. Term recognition. Identification of linguistic expressions which represent certain concepts.
4. Hierarchical Agglomerative Clustering. Approach to classify the extracted concepts.
5. Document classification. The documents are clustered based on semantics.
6. Evaluation process. Determines the accuracy of the term recognition process.

There is a resemblance between this study and our study, as it displays a similar goal, i.e., enriching the metadata of collections. Furthermore, our approach entails phases that correspond to both the Term Recognition and Hierarchical Agglomerative Clustering phases. These are the Concept Extraction and the Linking the Concepts phases, which will be further analyzed in the Methodology section. Since we also deal with large amounts of text snippets in our study, term recognition approaches seem appropriate in order to identify and extract the main concepts in the text. Moreover, while this study uses Agglomerative Clustering [36, 37] to classify the extracted concepts, our study will use the Simple Event Model (SEM) structure [7] to accomplish that. In general, we will adopt the two aforementioned phases of [4] in our study, adjusting them to our own data and methodology. For that reason, the study by Zervanou et al. can be considered as extremely useful, since it determines and describes a standard path for cultural heritage metadata enrichment. Consequently, we could say that this study, along with the one from 'Agora' [2], constitute the stepping stones of our study.

2.3 Techniques

In the previous subsection, we mentioned that in our study we will adopt the Term Recognition phase of [4]. However, we should also take into account numerous approaches, which specialize in recognizing and extracting terms-concepts from text, i.e., term extraction approaches. The objective of term extraction is to automatically extract relevant terms from a corpus. In classical terminology, a term is defined as the expression (or label, or representation) of a concept. For instance, in [4], the C-Value method [5] is utilized. C-Value constitutes a method which recognizes and extracts multiword terms from machine-readable special language corpora. Hence, similar approaches need to be found in order to extract terms in our study. Some examples of such approaches are the following:

- Kozakov et al. [27] report an approach of extracting domain specific glossaries from document collections used as a component of the IBM Textract system [28]. They mainly consider noun phrases and non-auxiliary verbs, including both single word and multi-word units. NLP tools such as a morphological analyser and a Part-of-speech (POS) tagger as well as a POS pattern filter are used to extract candidate terms

- Sclano and Velardi [29] designed another algorithm to extract domain terms. It consists of a linguistic processor and a set of filters. Given an input text, the linguistic processor is used to produce candidate terms by selecting typical terminological structures, such as compounds, adjective-nouns etc.

Apart from those algorithms, Natural Language Processing (NLP), and, particularly, Information Extraction (IE) techniques can, also, be utilized in order to identify and extract terms in the text. For instance, for the task of extracting names of people from the text, Named Entity Recognition – or NER – (a subtask of IE) seems the most appropriate one. NER systems identify different types of proper names, such as person and company names, and sometimes special types of entities, such as dates and times, that can be easily identified using surface-level textual patterns. Since the 1990's, there has been a gigantic amount of approaches, from which we discriminate the following:

- Tjong Kim Sang and Meulder offered a standard experimental platform for NER. In [30] they describe the CoNLL-2003 shared task, which concerns language-independent NER and concentrates on extracting and recognizing persons, locations, organizations and names of entities that do not belong to the previous three groups. It is generally considered a benchmark and its dataset is drawn from the Reuters newswire.
- Turian et al. [31] go one step further, as their approach identifies several types that do not appear on CoNLL-2003, such as money, dates and numeric quantities. Its performance reaches the degree of 90.36%, whereas human annotators have a performance of approximately 97% [32].

Additionally, for the task of extracting terms other than names of people, numerous NLP tools, that are available over the Internet, seem competent. For instance, LingPipe is a tool kit for processing text using computational linguistics. LingPipe is used to do tasks like [33]:

- Find the names of people, organizations or locations in news,
- Automatically classify Twitter search results into categories,
- Suggest correct spellings of queries.

Other similar tools are:

1. *Mallet.* Mallet is a collection of tools in Java for statistical NLP: text classification, clustering and IE. MALLET is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text [34].
2. *Gate,* which is considered one of the leading toolkits for text mining and information extraction. One of the components it is distributed with is an IE component, ANNIE, which stands for "A Nearly-New IE system" [35].

Conclusively, we have mentioned and described studies, on which our own research will be based. Moreover, we enumerated possible algorithms, tools and approaches that might prove suitable for integrating and implementing our study.

3. Problem Statement

In this section, we will define the problem that triggered this research and analyze it extensively, so that it becomes clear what the exact aim of this research is. Since the overall study will be conducted within the ‘Agora’ Project, we will utilize objects from the collections of the Netherlands Institute for Sound and Vision, one of Agora’s project partners. More specifically, we will use the textual descriptions found in the Open Images and EUscreen objects. Open Images is an open media platform that offers a large collection of audiovisual archive material, which in its large majority consists of videos, even though image and audio files are also available. EUscreen offers a vast collection of video, text and audio files from audiovisual archives around Europe from the beginning of the twentieth century until today.

As mentioned in the Theoretical Background section, the quality of the metadata schema in a collection of objects varies a great deal. In other words, not all objects display the same amount of information about their background, let alone about the historical context where they belong. This is reasonable, for creating a collection of metadata records is bound to result in unreliable quality. The ultimate goal of this study is to enrich the metadata schema in such a way that the result will be to have consistent and reliable metadata. In that way, users will discover material more easily by browsing the collection in an event-based manner. However, for the website to be event-based browsable, the objects of the collections need to be linked to specific events.

Let us, first, focus on the metadata schema. The vast majority of the EUscreen and OpenImages objects display textual descriptions which contain information about the audio or video file. The problem is that, in some cases, the information could be:

a. *Missing.* Some objects do not display summaries or descriptions, a fact that causes an impotence to link the object to an event.
b. *Minimal.* In this case, the metadata schema is filled with gaps, i.e., there is no appropriate amount of information, because several fields of the metadata schema are empty.
c. *Excessive.* Here, the description is available, but too much redundant material exists, as well. This might cause an ineffective correlation between the object and events.
d. *Of low quality.* This constitutes the most common case. In fact, the level of detail among the descriptions is bound to differ. This causes an incapability to generate object-event correlations, as well.

This results in a significant impediment, as the objects cannot be placed within a historical context. The historical context of each object goes hand in hand with a specific chronology, i.e., a sequence of events. These events either inspired its creator or are reflected on the object itself. As a consequence, if the historical context does not exist or is not evident, no object-event correlation is possible. Hence, the problems to tackle are two. On the one hand we have the unreliable and non-detailed textual descriptions and on the other the inability to create correlations between objects and events.

In case of non-existing descriptions, no real alternatives can be found using the existing metadata schema. In other words, the metadata of the object will have to be enriched with the assistance of an external source. Consequently, the problem exists when the descriptions are not missing from the schema. In that case, as mentioned in the Theoretical Background section, numerous concepts will be extracted from the available descriptions. These concepts will be utilized as links to other objects in order to breach the differences between the textual descriptions and create correlations between objects and events. The actual process will be described more extensively in the Methodology section.

Having identified the problems this study is going to tackle, the research questions arise automatically. As a result, following the above problem statement, the main research questions that arise from this research, but also drive it, are the following:

(1) To what extent are the concepts suitable to fill the gaps in the metadata schema?

The above research question raises additional sub-questions, since in order to complete the metadata enrichment we, first, have to link the concepts with particular events:

(1a) To what extent can we use these concepts to draw lines between objects and events?

(2) What are quality criteria for validation of the enriched metadata?

4. Methodology

This section is concerned with the approach and methodology behind this research. In order to illustrate that, the remainder of this section is structured as follows: we will analyze the steps, which are required to answer each research question. Consequently, what will be done is a distribution of the tasks regarding each research question.

4.1 Gathering the textual descriptions

Initially, the textual description part of the objects' metadata from EUscreen and Open Images, have to be collected. Here, we should keep in mind that all that is required is, in fact, just the metadata that accompany each object, not the object itself. As far as the EUscreen objects are concerned, the textual description we are seeking refers to the summary[6] of the object, whereas in the case of Open Images, it consists of the abstract[7]. As soon as the data collection is concluded, a statistical analysis will be performed. Specifically, the gathered data will be analyzed in order to examine aspects such as their volume and the quantity and quality of the information they provide. This step is illustrated in Figure 1, and constitutes the pre-enrichment phase.

[6] http://lod.euscreen.eu/resource/EUS_55F569268ACA42B186682960875F862B

[7] http://semanticweb.cs.vu.nl/pvb/browse/list_resource?r=__http://purl.org/collections/nl/openbeelden/ob_data.ttl968

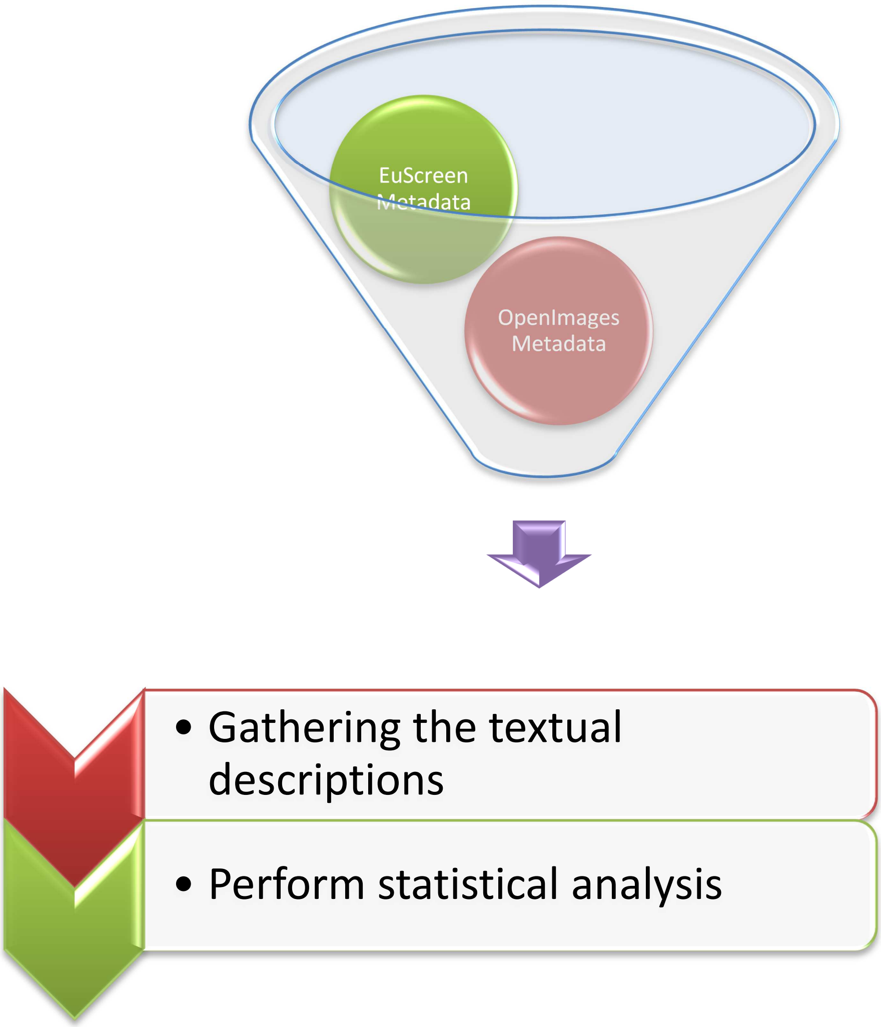

EuScreen Metadata
OpenImages Metadata
Gathering the textual descriptions
Perform statistical analysis

Figure 1. Phase 1: Data gathering phase

4.2 Concept recognition

In order to trigger the concept recognition process, we have to make sure that the metadata can be displayed in XML/RDF form. We chose XML due to the fact that the metadata of OpenImages can be displayed in XML/RDF format. In that way, the textual description regarding the objects of OpenImages can be found in the text contained below <ns5:abstract xml:lang="en">. These abstracts are found in ClioPatria[8] (the SWI-Prolog Semantic Web Server), where they can be obtained through a SparQL query. Similarly, the descriptions of the objects from EUScreen can be accessed through a SparQL endpoint[9]. Since the preferred metadata need to be in English, only English language abstracts have to be parsed.

The next step of this phase constitutes the concept extraction process, in which the XML/RDF files are parsed and the content is extracted. This brings us to the core of the phase, which is the recognition of the concepts. The objective of this phase is the

[8] http://semanticweb.cs.vu.nl/pvb/home
[9] http://oreo.image.ece.ntua.gr:10999/test/

identification of linguistic expressions, i.e., terms, which refer to defined concepts. These concepts mainly consist of the object name, its type and names of people.

For that purpose, several extraction tools will be utilized. In particular, what is needed is to apply term recognition approaches to perform concept mining, i.e., a technique, which extracts concepts from large text documents. Concepts are sequences of words that represent real or imaginary entities or ideas [10] and, thus, we will rely on Natural Language Processing (NLP) techniques [21, 22]. Since distinct NLP techniques are bound to display dissimilar results, more than one tool will be utilized. This approach will assist in identifying more concepts in the text. If, for example, a Named Entity Recognition tool is used, it is logical that its outcome will display names of people with more accuracy than other techniques.

4.3 Linking the concepts

After the process of identification and recognition of the concepts is completed, the concepts have to be classified in some way. This classification will create the bonds between concepts and objects. For instance, we will be able to identify objects that share the same concepts. Hence, these objects will refer to the same event. To achieve that, we will use the Simple Event Model (SEM) structure. The SEM consists of a Core class, which contains all entities that make up the context of an event, i.e., events, actors, places and times [7]. These entities constitute the core classes[10], where the extracted concepts will be classified. As a result, although an object might, initially, not provide enough information about the event it is related to, by classifying the concepts according to the SEM structure, we will probably gain some insight. To clarify, if an object displays the same amount of concepts with another object but its correlation with an event is not obvious, it is extremely likely that the second object correlates to an event. Consequently, an alignment between those two objects can be made, since it is more than sure that the objects refer to the same event.

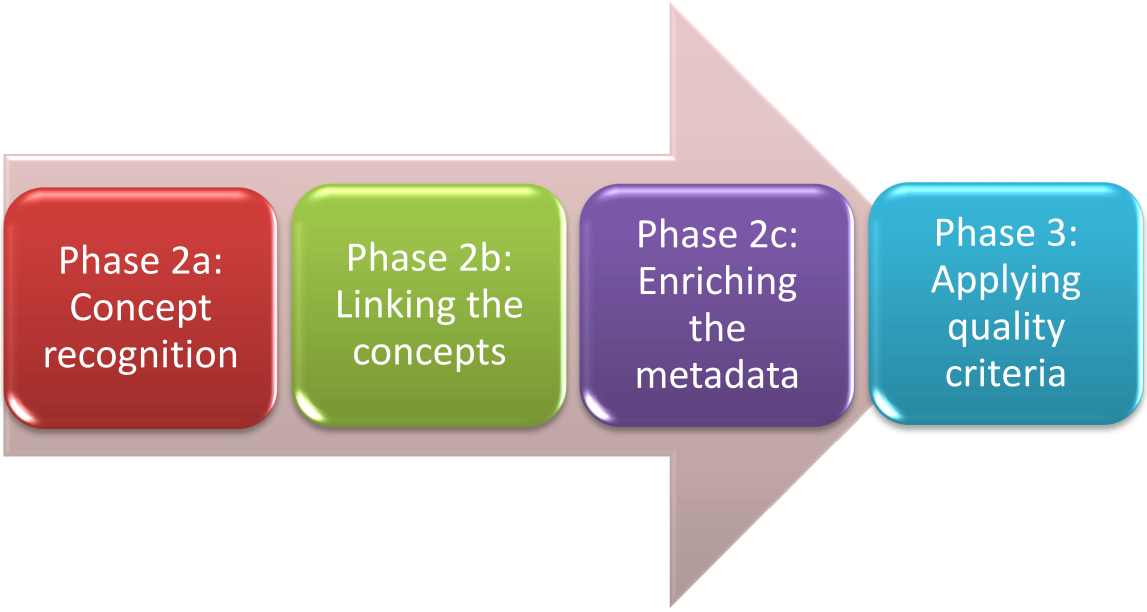
4.4 Enriching the metadata

After the above phase is completed, we can proceed to the enrichment of the objects’ metadata phase. In this phase, depending on the classification that was realized in the previous phase, we aim to enrich the metadata wherever they were lacking. For instance, if an object was not directly linked to an event and now is, this event will be mentioned in the textual description part of the metadata. The main objective of this phase is to improve the textual description, by adding the historical context and background behind its creation.

[10] http://semanticweb.cs.vu.nl/2009/11/sem/

4.5 Applying quality criteria

In this phase, we will analyze and validate the resulting enriched metadata of the objects that was realized in the previous phase. The goal of this phase is to understand what we are capable of doing now, in contrast to what we could do before the completion of the enrichment phase. For that purpose, experiments will be executed to test the validity of the metadata. These experiments will be in the form of queries, so that we can more effectively determine if the metadata are enriched in the desired way. The figure below illustrates the methodology phases that have to be executed. Phases 2a, 2b and 2c constitute the metadata enrichment stage, while phase 3 is a post-enrichment phase.



Phase 2a: Concept recognition

Phase 2b: Linking the concepts

Phase 2c: Enriching the metadata

Phase 3: Applying quality criteria

Figure 2. The enrichment and post-enrichment phases

5. Planning

The table below illustrates the tasks that need to be realized within this research and how they are scheduled:

	Tasks	Schedule
1.	Read literature	January 2012
2.	Form Theoretical background	January 2012
3.	Form research question	January 2012
4.	Literature review	February 2012
5.	Data gathering	March 2012
6.	Requirements for the metadata	April 2012
7.	Extraction from sources	April 2012
8.	Mapping and analysis	May 2012
9.	Analysis of enriched objects	May 2012
10.	Validation experiment	June 2012
11.	Reporting on results	June 2012

Table 1. Tasks and schedule

6. Expected results

In this section, we will deal with the results that we expect that this particular research will generate. In order to do so, we will start from the first phase of the approach that we introduced in the Methodology section. During the first phase, the textual descriptions are gathered. At this stage, we expect to gather all the English information regarding each object. While this phase is the starting point and, thus, is of major significance, we assume that all the available data will be gathered.

The second phase, i.e., the concept recognition phase, displays the first real difficulties of the research. We cannot be certain that the extracted concepts are the only ones that should be extracted from the available segment of text. For that purpose, a method that determines the validity of an extracted concept has to be utilized, so that it inspires a certain degree of assurance.

The third phase is the crucial one, for it consists of linking the concepts to events. The major concern that this phase displays is the classification of the extracted concepts according to the SEM structure. At this point, we cannot be sure that the concepts will be classified in an orderly fashion. For instance, several events are named after the territory where they took place, e.g., The Invasion of Normandy. As a result, the word Normandy constitutes both the territory where the event took place and part of the event name. In this case, depending on the approach that will be used, automatic classification will identify the word Normandy either as a place or as an event. Consequently, the approach that will be selected has to be able to discriminate such discrepancies. Moreover, there is also a possibility, where an object cannot be linked to any event. This is likely to happen in case the textual description is minimal and does not provide the appropriate amount of information to be linked to an event.

The next phase is where the metadata enrichment takes place. The results of this phase depend on the preceding one, since the whole enrichment process will be done according to the correlations generated before. On the other hand, during the final phase, the validity of the enriched metadata highly depends on the queries that will be carried out. Consequently, in a nutshell, the overall expected results seem to display a positive degree of confidence, for the amount of available data is gigantic, whereas the deficient textual descriptions are limited.

References

[1] C. Van Den Akker, S. Legêne, M. Van Erp, L. Aroyo, R. Segers, L. Van Der Meij, J. Van Ossenbruggen, G. Schreiber, B. Wielinga, J. Oomen, and G. Jacobs. Agora Digital Hermeneutics: Online Understanding of Cultural Heritage (2011). *In Proceedings of the 3rd International Conference on Web Science (WebSci'11).* Koblenz, Germany 14 – 17 June 2011. *Nominated for Best Paper award,* 2011.

[2] M. van Erp, J. Oomen, R. Segers, C. van den Akker, L. Aroyo, G. Jacobs, S. Legêne, L. van der Meij, J. van Ossenbruggen, and G. Schreiber (2011). Automatic Heritage Metadata Enrichment with Historic Events. *Museums and the Web 2011 (MW2011).* Philadelphia, PA, USA, April 6-9 2011.

[3] R. Segers, M.V. Erp, L.V.D. Meij, L. Aroyo, J.V. Ossenbruggen, G. Schreiber, B.J. Wielinga, J. Oomen, and G. Jacobs (2011). Hacking history via event extraction. *In Proceedings of K-CAP,* 2011, pp.161-162.

[4] K. Zervanou, I. Korkontzelos, A. V. D. Bosch, and S. Ananiadou. (2011). Enrichment and structuring of archival description metadata. In K. Zervanou and P. Lendvai (Eds.), *Proceedings of the Fifth ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (LaTeCH-2011), Portland, OR, pp. 44-53

[5] K. Frantzi, S. Ananiadou, and H. Mima (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries,* 3(2):115–130

[6] I. Korkontzelos, I. Klapaftis, and S. Manandhar (2008). Reviewing and evaluating automatic term recognition techniques. In Bengt Nordstrom and Aarne Ranta, editors, *In Proceedings of GoTAL '08, volume 5221 of LNCS,* pp. 248–259, Gothenburg, Sweden. Springer.

[7] W. R. V. Hage, V. Malaise, R. Segers, L. Hollink, and G. Schreiber (2011). Design and use of the simple event model (SEM). *Journal of Web Semantics* 9(2):128-136, July 2011

[8] R. Catizone, A. Dingli, H. Pinto, Y. Wilks (2008). Information Extraction Tools and Methods for Understanding Dialogue in a Companion. *In Proceedings of LREC, 2008*

[9] A. Chagnaa, C. Y. Ock, C. B. Lee, and P. Jaimai (2010). Feature Extraction of Concepts by Independent Component Analysis. *International Journal of Information Processing Systems,* Vol.3, No.1, June 2007

[10] A. Parameswaran, A. Rajaraman, and H. Garcia-Molina (2010). Towards The Web Of Concepts: Extracting Concepts from Large Datasets. *VLDB '10,* September 1317, 2010

[11] D. E. Appelt, and D. J. Israel. Introduction to Information Extraction tutorial. *Artificial Intelligence Center SRI International*

[12] B. Gelfand, M. Wulfekuhler, and W. F. Punch III (2000). Automated concept extraction from plain text. *In: AAAI 1998 Workshop on Text Categorization, Madison*, WI, pp. 13–17

[13] A. C. N. Ngomo (2010). Low-Bias Extraction of Domain-Specific Concepts. *In: Posters of CICLING 2009*, Polibits (2009)

[14] E. Hovy, Z. Kozareva, and E. Riloff (2009). Toward Completeness in Concept Extraction and Classification. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 948–957*, Singapore

[15] B. Chen, W. Lam, I. Tsang, and T. L. Wong (2009). Extracting Discriminative Concepts for Domain Adaptation in Text Mining. *In Proceedings of KDD '09*, June 28–July 1, 2009, Paris, France.

[16] G. Heinrich (2005). Parameter estimation for text analysis. *Arbylon publications*

[17] W. R. V. Hage, V. Malaise, G. de Vries, G. Schreiber, and M. V. Someren (2009).Combining Ship Trajectories and Semantics with the Simple Event Model (SEM). *In Proceedings of EiMM'09*, October 23, 2009, Beijing, China

[18] S. Sarawagi (2007). Information Extraction. *Foundations and Trends in Databases*, Vol. 1, No. 3 (2007) 261–377

[19] A. Hotho, A. Nurnberger, G. Paass, and S. Augustin (2005). A brief survey of text mining. *Zeitschrift fuer Computerlinguistik und Sprachtechnologie (GLDV-Journal for Computational Linguistics and Language Technologie),* 20 (1) (2005), pp. 19–62

[20] J. V. Gemert (2008). Text mining tools on the Internet. *Isis technical report series,* Vol. 25

[21] M. T. Castellvi et. al. Automatic term detection: A review of current systems. *In Recent Adv. in Comp. Terminology '01*

[22] S. Loh et. al. Concept-based knowledge discovery in texts extracted from the web. *SIGKDD Explor. Newsl.,* 2(1), 2000

[23] http://www.blinkx.com/article/blinkx-introduces-enhanced-metadata-enrichment-online-video-automated~903

[24] J. Jung, R. Simon, B. Haslhofer (2011). YUMA - Crowd Sourced Metadata Enrichment for Online Collections. ERCIM News 2011(86)

[25] S. L. Weibel (2009). "Dublin Core Metadata Initiative: A Personal History." *Encyclopedia of Library and Information Science, 2nd ed., ed. Marcia J. Bates, Mary Niles Maack, and Miriam Drake*. London: Taylor and Francis

[26] http://dublincore.org/specifications/

[27] L. Kozakov, Y. Park, T. Fin, Y. Drissi, Y. Doganata, and T. Cofino (2004). Glossary extraction and utilization in the information search and delivery system for IBM technical support. *IBM Systems Journal*. 43, 3, 546–563

[28] M. Neff, R. Byrd, and B. Boguraev (2003). The Talent system: TEX-TRACT architecture and data model. *In Proceedings of HLT-NAACL Workshop on Software Engineering and Architectures of Language Technology Systems*, Edmonton, Alberta, Canada

[29] F. Sclano, and P. Velardi (2007). TermExtractor: a web application to learn the shared terminology of emergent web communities. *In Proceedings of the 3rd International Conference on Interoperability for Enterprise Software and Applications (I-ESA 2007)* (Funchal, Madeira Island, Portugal).

[30] E. T. Tjong Kim Sang, and F. De Meulder (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *In Proceedings of CoNLL-2003,* pages 142-147.

[31] J. Turian, L. Ratinov, and Y. Bengio (2010). Word representations: A simple and general method for semi-supervised learning. *In Proceedings of the ACL*, 2010

[32] E. Marsh, and D. Perzanowski, (1998). MUC-7 Evaluation of IE Technology: Overview of Results, 29 April 1998

[33] Alias-i. 2008. LingPipe 4.1.0. http://alias-i.com/lingpipe

[34] http://mallet.cs.umass.edu/index.php

[35] http://gate.ac.uk/sale/tao/splitch6.html#chap:annie

[36] http://nlp.stanford.edu/IR-book/html/htmledition/hierarchical-agglomerative-clustering-1.html

[37] http://www.speech.sri.com/people/stolcke/papers/sri-h4-lm/node7.html