# Publishing Historical Texts on the Semantic Web —A Case Study

Eeva Ahonen and Eero Hyvönen
Semantic Computing Research Group (SeCo)
Helsinki University of Technology (TKK) and University of Helsinki
http://www.seco.tkk.fi/
firstname.lastname@tkk.fi

***Abstract*—Historical texts are an important component of cultural heritage, and are being digitized and published on the web in various portals for the researchers and the public. However, searching and linking them with related contents is challenging due to the non-structured text form, digitization errors, and the differences and variations between old and modern language, including historical names (e.g. places), used for querying. This paper addresses these issues by presenting an approach and a system for publishing old texts on the semantic web. As a case study, an existing historical newspaper archive on the web is considered. In our model, semantic metadata is added to the text using automated concept extraction methods. Search is implemented with semantic techniques, by creating a multi-faceted search interface for the text materials. Problems due to OCR errors and spelling variants are addressed with a fuzzy string matching algorithm trying to guess corresponding words in a lexicon, and giving suggestions for corrected word forms. References between texts in the library as well as links between the library and external knowledge sources are formed by using shared ontologies for semantic annotations.**

***Keywords*-historical newspapers; automatic semantic annotation; multi-faceted search;**

## I. Introduction

Historical newspapers archived in libraries all over the world document history in an exhaustive way that has an undeniable value to researchers as well as other professionals and the public in general. Newspaper libraries have been digitalized in massive efforts by micro filming, and to enable full-text searches the filmed pages have been transformed into text either automatically using optical character recognition (OCR) techniques [1], [2] or by manual work.

As a result, libraries of historical materials exist relatively widely on the world wide web, featuring string search interfaces and limited, mostly structural metadata.

The typically very large size and heterogeneous nature of the historical text materials imply that the usability of the digital libraries could be greatly improved by using semantic techniques [3]. It seems, however, that so far historical text materials haven't been commonly adapted for publication in semantic web portals for cultural heritage [4]. The focus has been on publishing collections of art, photos, artifacts etc.

One major challenge in the way of adopting historical text materials to the semantic web is the difficulty of the automatic processing of old language containing a lot of OCR errors, variance, old expressions and grammatical forms, and obsolete names. As is stated in [5], the research of automatic text processing tends to focus on contemporary materials, that typically are structurally homogenous and by content cover only the recent past.

Automatic methods are required to annotate large text corpuses semantically. With historical content, this is especially challenging: Firstly, old language isn't necessary understood by standard natural language processing (NLP) tools. Secondly, there is variance caused by the processing of the material, such as noise and errors caused by the OCR process. Thirdly, there is the challenge of the absence of knowledge sources to be utilized in processing the historical content that describes resources, such as people and places, that don't necessarily exist in contemporary knowledge sources [5].

In this paper, these issues are addressed by a case study, where a large historical newspaper archive is adopted into the semantic web by automated methods. Our main research questions are:

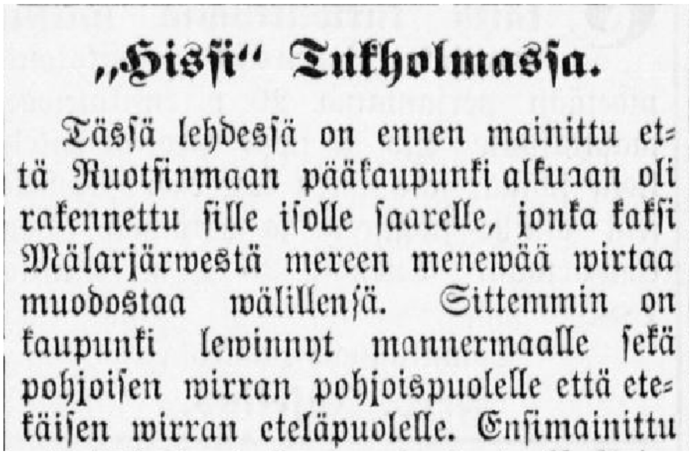
- *How to do linguistic analysis and semantic annotation on old text containing variance?*
- *How to link historical content semantically to existing ontologies?*
- *How to search the annotated material semantically?*

To create an intelligent user interface for searching the historical newspaper articles, semantic metadata is added to the material. OCR errors and spelling variants are tackled with an approximate matching algorithm trying to find corrected word forms in a lexicon. The ontology based semantic annotations form references within articles and text parts in the material, as well as between the archive and external information sources. For the end-user interface, a semantic multi-faceted search interface has been implemented that provides a useful alternative for the traditional full text search.

This article describes the main goals and challenges of the project, reports the current status and first results of the work, and outlines next steps on the way to creating a semantic archive of a large collection of historical newspapers. In the following, the underlying archive used in our study is first described.

## II. Historical Newspaper Archive of the National Library of Finland

The National Library of Finland has a digital archive of historical newspapers (http://digi.lib.helsinki.fi/sanomalehti/secure/main.html), that consists of over 900,000 pages of old newspaper material. The archive includes every newspaper published in Finland during the years 1771–1890. The newspapers have been microfilmed and then turned into text using OCR methods specially designed for the old character set, the German 'Fraktur", used in the newspapers in the 18th and 19th century. Figure 1 shows an example of the text. The article's title is "Hissi Tukholmassa" which means "Elevator in Stockholm". In the original paper the article is completed with a painted picture of the elevator.



„Hisſi“ Tukholmasſa.

Täsſä lehdesſä on ennen mainittu että Ruotſinmaan pääkaupunki alkuaan oli rakennettu ſille iſolle ſaarelle, jonka kakſi Mälarjärwestä mereen menewää wirtaa muodostaa wälillenſä. Sittemmin on kaupunki lewinnyt mannermaalle ſekä pohjoiſen wirran pohjoispuolelle että etekäiſen wirran eteläpuolelle. Enſimainittu

Fig. 1. A part of an article in the newspaper Turun Lehti published on the 14th of March, 1885

There is also an index for a part of the material covering some 400,000 article titles. The index itself is a historical subject heading classification of newspaper contents based on the topic of the article. For example, there is a category ”Arts history” that contains subdirectories ”Graphical Arts”, ”Music”, ”Theatre” etc. Apart from the historical index, no other content-based indexing or metadata is available for the material.

The material used in this study for demonstration and testing purposes consists of the historical index containing some 400 000 article titles. The article index is stored in a relational database. The metadata is described in table I.

TABLE I
Metadata schema for the article index

| | |
|---|---|
| article_title | The title text of the article |
| article_id | Identifier for the article |
| directory_title | The title text of the directory |
| directory_id | Identifier for the directory |
| publication | Name of the publication |
| publication_date | Date of the publication |
| publication_volume | Volume number of the publication |
| issn | International serial number for the volume of the publication |
| volume_id | Identifier for the volume of the publication |

### A. Language Change and Variance

One of the most significant characteristics of the material is its varying language. The texts are written during the end of the 18th century and the 19th century, during which time the (written) Finnish language was still evolving. The spellings of words change significantly when moving from the 18th century old Finnish into the 19th century early contemporary Finnish. Also at that time the written form of Finnish language was not very settled, and depending on the writer, a lot of concurrent variants for the same words were used.

Another thing that has added complexity and variance to the language is the automated process of turning the micro filmed papers into text. Some of the papers may have been in a bad condition at the time of the micro filming, and therefore the text can't be read at all sections. An additional challenge for the OCR process is the old character set, that isn't always too easy to interpret without awareness of the context. For example, some characters, such as the lower case *w* and the lower case *m* look almost exactly the same. As a result, the text contains a lot of variance that should be somehow addressed in the searching or annotating process in order to provide better search results.

### B. Current Web Portal

The National Library has a web portal providing 1) a text-based search system for the material and 2) a topic-based index for the indexed part of the material. In the text-based search system an expanding algorithm can be applied for the search phrases to cover some of the different spellings. Date and the name of the publication can be applied to limit the search. In the topic-based index however, no further searches or restrictions apart from the selected directory can be done. This means that when a single category of the index can contain hundreds of articles, only manual search through them is possible.

### C. The Vision for the New Web Portal

The goal of our work is to provide a more intelligent and flexible approach to searching the archive and to help researchers and other users to find more meaningful information from the material. This is to be achieved using semantic techniques. As in some earlier systems [6], the material will be automatically annotated using ontologies that describe the concepts characterizing the texts. This way the metadata is essentially turned into small RDF (Resource Description Framework, http://www.w3.org/RDF/) graphs connected together by the shared ontologies.

In our first experiment described in this paper, facet analysis is used as a basis for the annotation. Recognizing and annotating different semantic roles in which the concepts appear enable multi-faceted (view-based) searching [7], [8] to be applied to the material.

Furthermore, linking the annotations to common ontologies link the newspaper archive contents *internally* with each other and *externally* to it's broader context in the world of the semantic web or Web of Data. Internal links can point out e.g. articles about the same person or places in different decades, or the same event, such as inauguration of a king, in different newspapers or articles. External links provide recommendation gateways to other information sources, such as the Wikipedia, based on its semantically annotated version DBPedia, or other sources of the Linked Data Initiative (http://linkeddata.org). In our case of special interest is the RDF triple store of the CultureSampo portal (http://www.kulttuurisampo.fi/) [3]. This system contains 134,000 semantically annotated cultural objects (artifacts, photos, historical events, paintings, poems etc.), and nearly 300,000 additional resources such as persons and places. The content is annotated using the national semantic web infrastructure of FinnONTO ontologies [9] used for annotating the historical newspapers, too.

## III. Applying Methods of Semantic Web and Language Technology

This section describes the methods used to address the three main research problems stated in Section 1. So far the methods have been applied to a small portion of the material of 900,000 pages, to create a working application for demonstration and testing purposes. The material used in our study is the historical index of some 400,000 titles. Both the directory hierarchy and the article titles themselves have been adopted to the application.

### *A. Ontologies*

Several ontologies were used for annotating the subject matter of the articles. First, common annotation concepts were extracted from the text using the KOKO Ontology (http://www.yso.fi/onto/koko) [10] containing some 37,000 concepts. It is a collection of mutually mapped Finnish core ontologies including the Finnish General Upper Ontology YSO and some more specific ontologies that extend and refine the concepts in YSO.

Second, place names were extracted and annotated using geographical ontologies: a Finnish Geo-ontology SUO [11] that contains contemporary place types and place instances with coordinate information, and some additional metadata, and the Finnish Spatio-temporal Ontology SAPO (http://www.yso.fi/onto/sapo), that is a time series representation of the historical Finnish municipalities over time [12]. After the place names in the articles have been annotated with references to the Geo-ontology, thanks to the coordinate information, the news articles can even be placed and shown on a map.

### *B. Automatic Semantic Annotation*

Automatic semantic annotation is implemented using the information extraction tool Poka (http://www.seco.tkk.fi/tools/poka) [13]. Poka is a framework for automatic semantic annotation, that offers common tools to create an annotator for a specific need. Poka uses an open-source morphological analyzer for Finnish called OMorfi [14] to stem the words into their base forms. Stemmed words are compared to (stemmed) labels extracted from a given ontology, or any lexicon. When a label of an ontology term matches a stemmed word in the text, the respective URI is extracted from the ontology, and the subject matter of the text is annotated with this URI. In this way the text is turned into an RDF graph containing URI references to given ontologies.

Figure 2 shows the structure of the RDF graph at this stage of the implementation. The namespace Article is used for all resources that are local to this project. The Article:uriRef properties are references to the common Finnish KOKO-Ontology, i.e. their values are resources of the KOKO-Ontology. The hierarchical directory structure of the historical index is turned into rdfs:subClassOf -relations. The publication of the article and the date of the publication are made into object properties in order to be able to implement them as search facets. The other available metadata items, ISSN and volume number, are represented as literal properties of the articles themselves, as well as is the URL reference to the original source, that is derived from the article ID given in the metadata. The URL reference links the article title in question to the original newspaper shown in PDF form on the www pages of historical newspaper portal of the National Library of Finland.

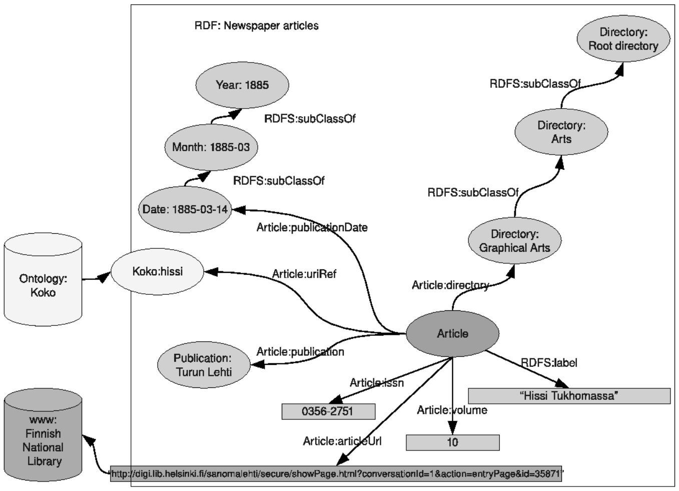


Fig. 2. A graphical representation of the structure of the metadata RDF graph of an article

### *C. Correction of Misspelled Words*

In the semantic annotation process, the problem dealing with variant nature of the old newspaper language is addressed using a fuzzy matching algorithm if no direct match is found for a word. The first phase of the annotation process is the stemming of inflected word forms into their base forms. This is done by OMorfi, that uses word lists for analyzing words.

If a word is not found in the word list, no result can be retrieved with the basic method. This means that if a word is misspelled, or spelled in a historical way, that varies from the contemporary spelling, no morphological analysis can be retrieved for it using the standard analyzer. For these cases we use an additional part of the OMorfi, that tries to guess the morphological analysis and the base form of any given string of characters. In Finnish this is not a trivial task, since words inflect heavily and the inflected forms overlap so that without information about the original word or context it is almost always impossible to give just one option. For most cases the OMorfi guesser gives an extensive list of imaginary base forms and possible morphological analyzes for these.

This list is then compared to a lexicon retrieved either from the ontology used in annotation, or any other source that is considered exhaustive enough. A simple n-gram matching [15], [16] algorithm is applied to find the closest match between the guessed word forms and the lexicon:

$$sim(a,b) = \left( \frac{|A \cap B|}{|A| + |B|} \right)$$

where A is the set of n-grams formed from word a, and B is the set of n-grams formed from word b

A threshold is set, upon which a match can be considered successful. The match getting the highest score for closeness will be selected as the base form of the word. If none of the matched words get a score above the threshold, then the matching is considered unsuccessful and no base form was found for the word.

This simple method has produced promising initial results but it is also clear that several things remain to be done for improving the approximate matching. First of all, not all differences between the matched word forms need to be considered as errors or actual differences. For example spelling Finnish words with *w* (used in old Finnish) or with *v* (used in contemporary Finnish) shouldn't influence the similarity of the words, since the two are free variants that make no difference in meaning of the words. This kind of variations could therefore be ignored when calculating the word similarities. The same goes for certain context-dependent typical errors, in this case for example *w* and *m* could be considered free variants, since they are frequently confused by the OCR process. While these could also make a difference in meaning, it seems highly probable that the difference is caused by a mistake made by the OCR.

Secondly, the simple n-gram matching algorithm was chosen for testing purposes, but hasn't been proven to be the most efficient for this type of material. Approximate string matching has been the focus of extensive research, and many different algorithms have been specified for the purpose. In [15], [16], [17] quite a few are described that could be tested on the variant historical text.

The quality of results of the matching algorithm should also be evaluated in a more thorough manner than has been done so far. It should be made clear that the approximate matching and correction of spelling errors improves the search results more than it confuses the results with false corrections, ie. improves recall without deteriorating precision too much.

### *D. Multi-faceted Search*

After the text has been annotated with semantic information carrying different roles or types (such as people or places), a multi-faceted search interface can be created based on the RDF material.

For demonstration purposes, this can be done quickly and easily with Hako (http://www.seco.tkk.fi/tools/hako/), a lightweight tool for faceted semantic search engine interfaces for RDF materials. Hako takes an RDF file as input, and with some material-specific configurations creates a web application with a search interface for the RDF material. Any annotated properties can be configured as facets for the search.

Figure 3 demonstrates a sample of the newspaper archive shown in the Hako interface. On the left frame are the facets that can be used in the search. In this example they are: Topic-based historical newspaper directory ("Hakemisto"), Subject of the article based on common terms of the KOKO ontology ("Hakusana"), Time of publishing ("Julkaisu_pvm") and Name of the publication ("Paanimeke").

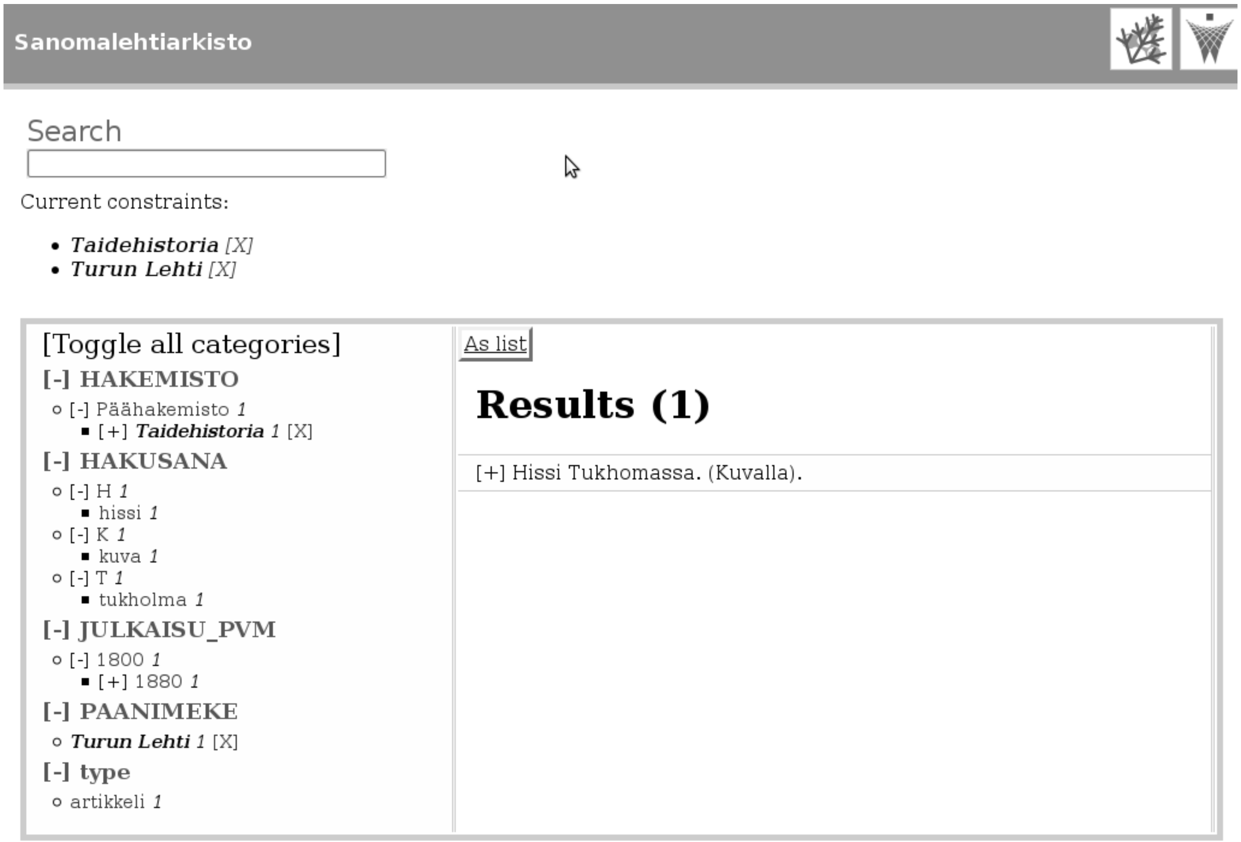
On the right frame is displayed the set of articles matching the current facet selections. In figure 3 a selection is made to show only the articles published in the paper "Turun Lehti" and belonging to the directory category "Graafinen taide" (Graphical Arts), leaving only one article that matches the two selections. Other existing properties for the current selection can be seen on the left frame.

A multi-faceted search interface allows for the user to filter the materials through selections on facet categories. After each selection the material in the result set is exposed along the facets, and the number of hits in each category can be seen. As a result, the user does not end up in "no hits" dead ends and doesn't have to guess what annotation values might exist in the remaining materials. At the same time, the facets expose for the user the relevant vocabulary by which queries can be formulated as facet category selections.

In end-user evaluations, faceted search has been found useful especially in situations where the user has difficulties in formulating the query but is rather willing to explore the materials [18]. If the user knows exactly what she is looking for, query word based Google-like search is often preferred. To support this, Hako also supports traditional text based querying. Since the properties used as facets are linked to ontologies, it is possible to make use of their ontological descriptions in for example query expansion: this would lead to for example "chairs" to be found when searching for "furniture". Hako also supports giving semantic recommendations based on the properties of the items. So when the user selects a particular article, recommendations can be made to other articles related to the same place or published in the same newspapers or otherwise related as specified.

## IV. Search Example

The advantages of the annotated material combined with the multi-faceted search are best described through an example. In



Sanomalehtiarkisto

Search

Current constraints:

- ***Taidehistoria*** *[X]*
- ***Turun Lehti*** *[X]*

[Toggle all categories]

**[-] HAKEMISTO**

- [-] Päähakemisto *1*
  - [+] ***Taidehistoria*** *1* [X]

**[-] HAKUSANA**

- [-] H *1*
  - hissi *1*
- [-] K *1*
  - kuva *1*
- [-] T *1*
  - tukholma *1*

**[-] JULKAISU_PVM**

- [-] 1800 *1*
  - [+] 1880 *1*

**[-] PAANIMEKE**

- ***Turun Lehti*** *1* [X]

**[-] type**

- artikkeli *1*

**Results (1)**

[+] Hissi Tukhomassa. (Kuvalla).

Fig. 3. A sample of the newspaper material shown in the Hako semantic search interface with only one article matching the selection

the 14th of March 1885 Turun Lehti published an article with the title “Hissi Tukholmassa” (An Elevator in Stockholm), completed with a painted picture of an elevator. This article is part of the historical index, and belongs to a directory category called ”Graphical Arts”, presumably because of the picture.

When browsing through the index in the old search system, we find 17 pages under the ”Graphical Arts” directory with 20 article titles on each. No further searches can be made to the contents of a directory, so the only way to see what’s included in these 17 pages, is by browsing through them by hand.

In the Hako search interface, the multi-facet search allows for the user to limit the contents of the search result window by selecting (and unselecting) values facet by facet. After selecting the historical directory ”Graphical Arts”, for example, all other search facets now only show values that exist in the current selection. With other search facets we can further limit the search results by, for example, selecting the 19th century, or the year 1885, or March of 1885, or even the 14th of March 1885. We can limit the results to cover only the publication “Turun Lehti”, or select ontological indexing terms that have been added to the metadata in the automatic annotation process. For the article in the example these are “hissi” (=elevator), “Tukholma” (=Stockholm) and “kuva” (=picture).

Furthermore, the title in the historical index is actually misspelled as “Hissi Tukhomassa” (missing *l* as it should be “Tukholmassa”), but thanks to the automatically guessed correction of unknown words, it is nevertheless indexed with the term “Tukholma”.

## V. Discussion

This paper presented an approach to automatic linguistic analyzing of historical texts, and to the automatic semantic annotation of the variant text content. The goal was to achieve semantically annotated text, that is linked to commonly used ontologies. This allows any applications and services utilizing these ontologies to access the historical text content and vice versa.

The OMorfi morphological analyzer for Finnish was used to retrieve the base forms for the inflected word forms in the text. However, a standard tool such as OMorfi cannot always do morphological analysis due to either historical spelling variations or an error caused by OCR processing of the material. To solve the problem, a guesser was used to guess all the possible base forms and morphological analyzes for any given string of characters. Then an approximate string matching algorithm was devised to find the possible correct (or contemporary Finnish) word forms in a dictionary formed from an ontology describing common terms in Finnish. Where a match was found (either a direct match, or an approximate match after the guessing process), an URI reference was added to the subject annotation of the text. In this way the plain text corpus was turned into an RDF graph containing ontology links as well as other metadata already present in the material. A semantic search interface featuring a multi-facet search was then implemented on the material based on both the previously existing metadata and the automatically extracted ontology references as search facets.

It was shown, by an illustrating search example on a demonstrational implementation, how the research questions listed in the first section can be met through our approach:

- *How to do linguistic analysis and semantic annotation of old text containing variance?*
  The base forms were correctly derived and ontological references found for the search words of the title. One was misspelled, but the correct term was found regardless of that through the guessing and approximate matching procedure.
- *How to link historical content semantically to existing ontologies?*
  Through ontological matching and annotation of URI references, links between the text and the ontologies were formed. The articles are also individually linked to their original sources at the www page of the National Library of Finland.
- *How to search the annotated material semantically?*
  A multi-faceted semantic search interface was applied to the annotated material. This makes searching and browsing for information in the archive easier.

The system has been tested internally, and it looks very promising, as suggested by earlier end-user evaluations and systems using the faceted search paradigm [18], [19], [20]. However, formal evaluation of the semantic annotation scheme or the search system, e.g. by comparing the solution with the existing service on the web, has not been done yet. The new system is in any case more versatile than the existing service because it includes the old functionalities and adds on the new semantic features based on the underlying RDF model.

The existing historical newspaper service on the web is very popular among customers as an information source for professional as well as recreational users. A lesson learned in our case study is that the material itself is full of possibilities from a digital humanities point of view. However, without utilizing the underlying semantic it has definitely not been used to it's full potential. The first results reported in this paper show that the application of semantic techniques is likely to significantly improve the usability of a large historical archive. A fully non-utilized possibility of enriching the semantics of the archive is linking its content through interoperable ontologies with other semantic RDF repositories of cultural heritage, such as the national CultureSampo system.

### *A. Related Work*

There has been lots of efforts to digitalize historical newspaper archives, and as a result many digital libraries of historical newspapers exist on the web. However, the aim to find and develop automated processing techniques for the content of these libraries seems rather limited. In [5] an automated method of named entity recognition in a historical newspaper archive is presented. Here a learning material of semi-automatically tagged newspaper texts was used to teach the system to recognize the named entities belonging in ten different classes. A detailed report of the successes and failures of the task per class is presented. A pipeline for the automatic processing of historical newspapers is specified in [21]. The pipeline begins with OCR'd text in XML format, and includes article segmentation, genre and subject recognition, and finally event extraction. In [21] the results are reported for the first two steps, namely article segmentation and genre recognition. Their future plans include topic and event categorization from the material.

Metadata structures and some automatic or semi-automatic methods for digital libraries of historical newspaper material have been introduced also in for example [2] and [22].

However, none of the projects described are implementing ontological semantic annotation, or state any plans on implementing semantic web techniques on the materials.

In the field of adding historical and cultural content in the semantic web, the creation of the metadata and annotation of the content has usually been a manual process or in some cases semi-automatic with the help of semi-automatic annotation tools, for example. In addition to manual annotation of cultural items, some automatic annotation, such as the text descriptions of cultural items, has been implemented in for example [23] and [3].

## VI. Future Research

In future research the methods presented for automatic annotation will be developed further by a more detailed analysis and evaluation of the errors occurring in the process, as discussed earlier. We will also try to extend our automatic semantic annotation tools to recognizing and annotating historical events that recur in the material and should be of special interest to end-users of news materials. In [21] a plan for categorization of events in historical newspaper archives is made. In our own work, we are creating an ontology of historical events [24] whose first small version is already in use in CultureSampo. The plan is to apply this kind of ontology as a knowledge source for the historical event extraction task. In this way, references from the events occurring in the news material can be made to their respective ontological descriptions, and events can be linked with other related historical materials, such as photos, paintings, artifacts, persons and biographies, historical places, monuments etc. available in CultureSampo.

Ontology-based event detection in general seems to be a rather uncharted field of research. In the field of topic detection and tracking (TDT) [25], however, similar methods have been applied to recognize events described in several documents or text parts. Event patterns are created from news articles, and by matching these patterns news relating to the same event are found. Makkonen et al. [26] improved the results of their TDT task by applying semantic roles to patterns used for matching the events. In our own study we aim at creating the patterns from the historical event ontology, and match these to similar patterns created from the news material.

## Acknowledgements

This work is a part of the National Semantic Web 2.0 (FinnONTO 2.0) project 2008-2010, funded mainly by the National Technology and Innovation Agency (Tekes) and a consortium of 38 private, public and non-governmental organizations.

The work has been made possible through the help and cooperation of the people of the National Library of Finland: Majlis Bremer-Laamanen, Juha Hakala and Jukka Kervinen.

The Hako semantic search interface demonstrated in this paper has been developed by Joonas Laitio and Jussi Kurki in the FinnONTO2.0 project at the Semantic Computing Research Group in Helsinki University of Technology.

## References

[1] T. Kanungo and R. B. Allen, "Full-text access to historical newspapers," Laboratory for Language and Media Processing, University of Maryland, Tech. Rep., April 1999.

[2] R. B. Allen and J. Schalow, "Metadata and data structures for the historical newspaper digital library," in *CIKM '99: Proceedings of the eighth international conference on Information and knowledge management*. New York, NY, USA: ACM, 1999, pp. 147–153.

[3] E. Hyvönen, E. Mäkelä, T. Kauppinen, O. Alm, J. Kurki, T. Ruotsalo, K. Seppälä, J. Takala, K. Puputti, H. Kuittinen, K. Viljanen, J. Tuominen, T. Palonen, M. Frosterus, R. Sinkkilä, P. Paakkarinen, J. Laitio, and K. Nyberg, "CultureSampo—Finnish culture on the semantic web 2.0. thematic perspectives for the end-user," in *Proceedings, Museums and the Web 2009, Indianapolis, USA*, April 15–18 2009.

[4] E. Hyvönen. Springer–Verlag, 2009, ch. Semantic Portals for Cultural Heritage.

[5] G. Crane and A. Jones, "The challenge of virginia banks: an evaluation of named entity analysis in a 19th-century newspaper collection," in *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*. New York, NY, USA: ACM, 2006, pp. 31–40.

[6] K. Byrne, "Populating the semantic web—combining text and relational databases as rdf graphs," Ph.D. dissertation, University of Edinburgh, School of Informatics, 2008.

[7] M. Hearst, A. Elliott, J. English, R. Sinha, K. Swearingen, and K.-P. Lee, "Finding the flow in web site search," *Communications of the ACM*, vol. 45, no. 9, pp. 42–49, 2002.

[8] E. Hyvönen, S. Saarela, and K. Viljanen, "Application of ontology-based techniques to view-based semantic search and browsing," in *The semantic web: research and applications. First European Semantic Web Symposium, ESWS 2004, Heraklion, Greece*. Springer–Verlag, 2004, 92–106.

[9] E. Hyvönen, K. Viljanen, J. Tuominen, and K. Seppälä, "Building a national semantic web ontology and ontology service infrastructure—the FinnONTO approach," in *Proceedings of the ESWC 2008, Tenerife, Spain*. Springer–Verlag, 2008.

[10] K. Viljanen, J. Tuominen, and E. Hyvönen, "Ontology libraries for production use: The finnish ontology library service onki," in *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*, May 31 - June 4 2009, springer-Verlag.

[11] R. Henriksson, T. Kauppinen, and E. Hyvönen, "Core geographical concepts: Case finnish geo-ontology," in *Location and the Web (LocWeb) 2008 Workshop, 17th International World Wide Web Conference WWW 2008*. ACM International Conference Proceeding Series, Vol. 300, 2008, pp. 57–60.

[12] T. Kauppinen and E. Hyvönen, "Modeling and reasoning about changes in ontology time series," in *Ontologies in the Context of Information Systems*, R. Kishore, R. Ramesh, and R. Sharman, Eds. Springer–Verlag, 2007.

[13] O. Alm, "Tekstidokumenttien automaattinen ontologiaperustainen annotointi," Master's thesis, University of Helsinki, Department of Computer Science, September 2007.

[14] T. Pirinen, "Suomen kielen äärellistilainen morfologinen jäsennin avoimen lähdekoodin resurssein," Master's thesis, University of Helsinki, Department of General Linguistics, 2008.

[15] A. Pirkola, J. Toivonen, H. Keskustalo, K. Visala, and K. Järvelin, "Fuzzy translation of cross-lingual spelling variants," in *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. New York, NY, USA: ACM, 2003, pp. 345–352.

[16] U. Pfeifer, T. Poersch, N. Fuhr, and L. I. Vi, "Searching proper names in databases," in *Universitatsverlag Konstanz*, 1994, pp. 259–275.

[17] J. Euzenat and P. Shvaiko, *Ontology Matching*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007.

[18] J. English, M. Hearst, R. Sinha, K. Swearingen, and K.-P. Lee, "Flexible search and navigation using faceted metadata," University of Berkeley, School of Information Management and Systems, Tech. Rep., 2003.

[19] E. Hyvönen, E. Mäkelä, M. Salminen, A. Valo, K. Viljanen, S. Saarela, M. Junnila, and S. Kettula, "Museumfinland—finnish museums on the semantic web," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 3, no. 2–3, pp. 224–241, Oct 2005. [Online]. Available: http://www.cis.strath.ac.uk/ mdd/research/publications/

[20] G. Schreiber, A. Amin, M. van Assem, V. de Boer, L. Hardman, M. Hildebrand, L. Hollink, Z. Huang, J. van Kersen, M. de Niet, B. Omelayenko, J. van Ossenbruggen, R. Siebes, J. Taekema, J. Wielemaker, and B. J. Wielinga, "Multimedian e-culture demonstrator." in *The Semantic Web—Proceedings of the 5th International Semantic Web Conference2006*, November 5–9 2006, pp. 951–958.

[21] R. B. Allen, I. Waldstein, and Z. Weizhong, *Automated Processing of Digitized Historical Newspapers: Identification of Segments and Genres*. Springer Berlin / Heidelberg, 2008, vol. 5362/2008, pp. 379–386.

[22] D. Calvanese, T. Catarci, and G. Santucci, "Laurin: A distributed digital library of newspaper clippings," *World Wide Web*, vol. 4, no. 1-2, pp. 5–20, 2001.

[23] T. Ruotsalo, L. Aroyo, and G. Schreiber, "Knowledge-based linguistic annotation of digital cultural heritage collections," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 64–75, 2009.

[24] E. Hyvönen, O. Alm, and H. Kuittinen, "Using an ontology of historical events in semantic portals for cultural heritage," in *Proceedings of the Cultural Heritage on the Semantic Web Workshop at the 6th International Semantic Web Conference (ISWC 2007)*, 2007.

[25] J. Allan, Ed., *Topic detection and tracking: event-based information organization*. Norwell, MA, USA: Kluwer Academic Publishers, 2002.

[26] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi, "Simple semantics in topic detection and tracking," *Inf. Retr.*, vol. 7, no. 3-4, pp. 347–368, 2004.