DAICA - Digital Assistant Investigating Cultural Assets

Lothar Hotz[1], Dan Cristea[2], Justyna Pietrzak[3], Martin Povazay[4], Brigitte Rauter[4], and Daniela Buleandra[5]

[1] HITeC e.V. c/o University of Hamburg, Germany,
hotz@informatik.uni-hamburg.de
[2] University Alexandru Ioan Cuza and Romanian Academy, Romania,
dcristea@info.uaic.ro
[3] Eleka Ingeniaritza Linguistikoa S.L., Spain, justyna@eleka.net
[4] P.Solutions Informationstechnologien GmbH, Austria,
{martin.povazay,brigitte.rauter}@psolutions.at
[5] SIVECO Romania SA, Romania, Daniela.Buleandra@siveco.ro

Abstract. Besides web pages, the web offers access to an immense variety of digitized source material, inventories and catalogues hosted by libraries and archives relevant for humanities and social sciences (HSS) studies. In practice, a remote access to HSS information is considerably hampered by several barriers: Researchers interested in a specific topic do not know which institution harbors information related to a specific topic; Data collections are equipped with unique user interfaces and offer different data structures; Language barriers impede information exploitation; Retrieval mechanisms do not provide intelligent access to semantically related information. In this paper, we describe an Digital Assistant Investigating Cultural Assets (DAICA) for research and information procurement in HSS, guided by the vision of a digital information space of cultures. The DAICA will support HSS studies by autonomously identifying appropriate resources and presenting topical investigation results. In particular, the DAICA will integrate technology and provide a solution for analysing historical digitized documents, performing semantical search in deep data structures, automatic translation, extending a search by meaningful relations, creating summaries of identified resources, and providing user interactions for complex search results.

Keywords: Semantic search, machine translation, summarization, optical character recognition, cultural heritage

1 Introduction

The web offers an immense variety of digital or digitized source material, inventories and catalogues relevant for studies in humanities and social sciences (HSS). Institutions such as libraries, museums, archives, local and regional authorities, parliamentary and media documentation services provide a wealth of information, organized on local websites and in principle accessible for inquiries and research from anywhere in the Internet, however not by ordinary search engines - they reside in the "deep web" and require special access methods.

Resources exist in a multitude of media types: unstructured and structured text, with and without metadata, full text accessible or not, images, and videos. Media comprise historical or cultural texts, biographical information, newspapers articles, maps of cities, regions, and countries, paintings, photographs, and much more. In addition to varying data formats, the user interfaces to these resources differ widely, both regarding language and functionality. Hence, research is difficult, and the outcomes are all too often incomplete and insufficient. Typically, tedious manual investigations are needed to study a specific topic: one has to get into contact with many possible sources, get acquainted with access modalities, retrieve data, commonly by language-specific key-word search, determine relevance across language barriers, and summarise the resulting material. In HSS studies, the task is further aggravated because of diverse cultural regions, each with its own history and tradition, and hence, often with a different understanding of seemingly identical terms.

This situation is faced by scholars, students, journalists, public, and media if they want to investigate abroad data sources, not only data of their local libraries. In principle the same applies to investigations by business and commercial aggregators. There is clearly the need to pave the way for barrier-free access, i.e., support in accessing distributed repositories, in translation services, in fast interpretation of new unknown resources, and collaboration with other interested users about this matter to cultural data.

Fortunately, basic technology for multilingual and semantically enhanced search in multimedia databases is available. But in order to effectively support HSS studies, techniques have to be adapted and integrated. For example, relevance criteria including temporal and geographical proximity, cultural vicinity, or taxonomical distance of terms must be taken into account for a semantic search to be effective. Optical character recognition (OCR) must be invoked for searching scanned documents; Machine translation must be used for identifying, linking and presenting relevant multilingual data, summarisation and visualization and user-device interaction facilities must be provided for complex and multifaceted data. Altogether, these methods can allow effective retrieval of historical and cultural data from the deep web.

As a technological innovation, this paper presents a concept of an intelligent HSS research assistant called *Digital Assistant Investigating Cultural Assets, DAICA*, which will be used by scholars of the humanities and by the interested public or the industry having the need for investigation a specific topic. Summarisations will be returned of semantically related documents, articles, texts not only of web-pages but also of other data sources.

This paper describes the use cases of a DAICA (Section 2), a new search procedure in Section 3 integrating semantic search, machine translation, OCR, entity discorvery, and summarization. These technologies are integrated in a configurable framework (Section 4 and Section 5).

2 A Use Case of DAICA

Figure 1 presents a use case of a digital assistant presented in this paper with user and background interactions processed by DAICA. Main features are the proactive discovering of the user's interest topics ("A1"), the background actions based on localization if the users is moving ("A2") and on OCR for interpreting original ancient documents. By further observing the user's writing activities, DAICA computes related sources in the web and, hence, supports the user by his investigation activities ("A3").

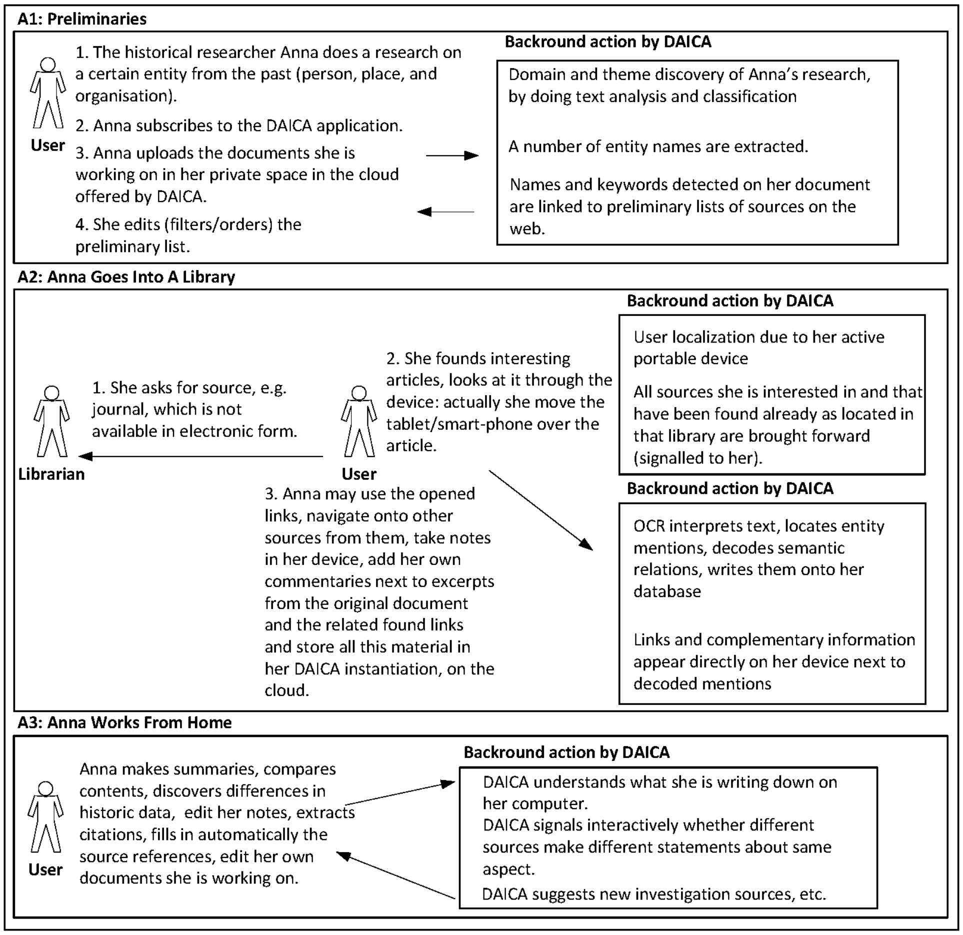
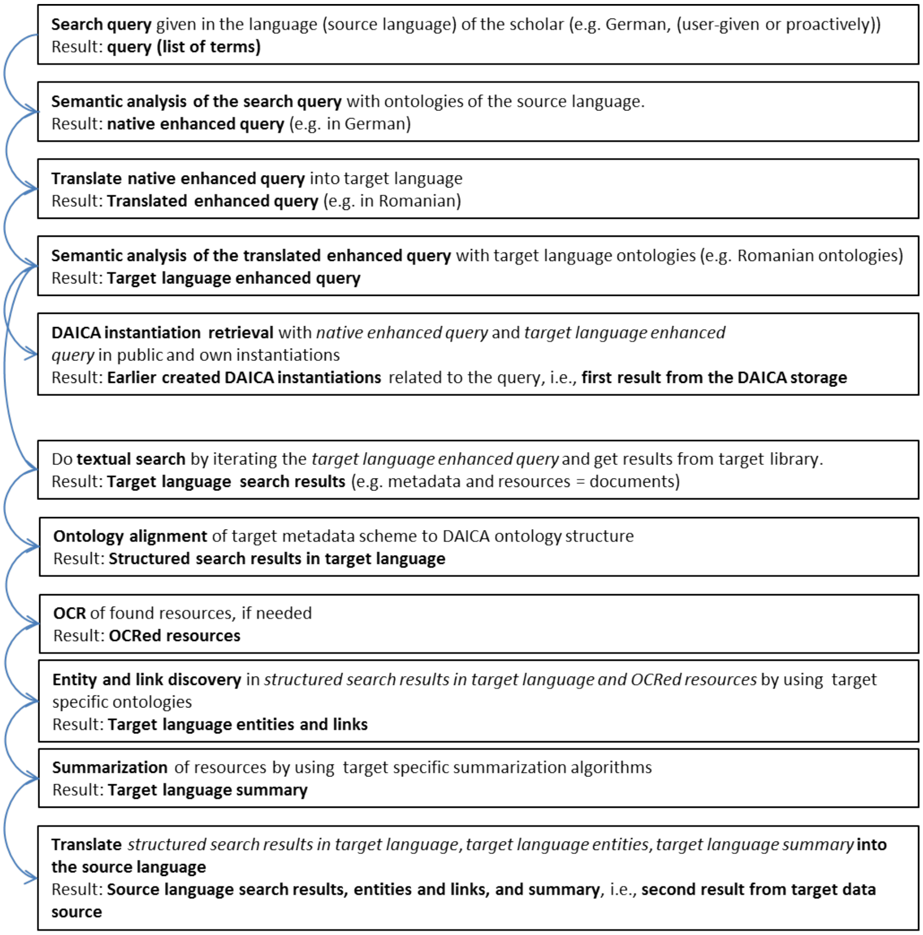

A1: Preliminaries
User
1. The historical researcher Anna does a research on a certain entity from the past (person, place, and organisation).
2. Anna subscribes to the DAICA application.
3. Anna uploads the documents she is working on in her private space in the cloud offered by DAICA.
4. She edits (filters/orders) the preliminary list.
Backround action by DAICA
Domain and theme discovery of Anna's research, by doing text analysis and classification
A number of entity names are extracted.
Names and keywords detected on her document are linked to preliminary lists of sources on the web.
A2: Anna Goes Into A Library
Librarian
1. She asks for source, e.g. journal, which is not available in electronic form.
User
2. She founds interesting articles, looks at it through the device: actually she move the tablet/smart-phone over the article.
3. Anna may use the opened links, navigate onto other sources from them, take notes in her device, add her own commentaries next to excerpts from the original document and the related found links and store all this material in her DAICA instantiation, on the cloud.
Backround action by DAICA
User localization due to her active portable device
All sources she is interested in and that have been found already as located in that library are brought forward (signalled to her).
Backround action by DAICA
OCR interprets text, locates entity mentions, decodes semantic relations, writes them onto her database
Links and complementary information appear directly on her device next to decoded mentions
A3: Anna Works From Home
User
Anna makes summaries, compares contents, discovers differences in historic data, edit her notes, extracts citations, fills in automatically the source references, edit her own documents she is working on.
Backround action by DAICA
DAICA understands what she is writing down on her computer.
DAICA signals interactively whether different sources make different statements about same aspect.
DAICA suggests new investigation sources, etc.

Fig. 1. Use Case "New Type of Assisted HSS Study"

3 A New Integrated Search Procedure

A main task of DAICA is the transparent integration of semantic search, machine translation, and all other capabilities such as summarization, OCR, and entity discovery (see Figure 2 which illustrates the complete process). A user spells out implicitly or explicitly the search specification (query) in his/her language, the *source language*, in the example German. DAICA translates this query and its ontological enhancements into the *target language* (here Romanian) of a data source and starts the search. The results and their enhancements through links and summarisation are again translated back into the source language. In the following section, the details about the involved components are given.



Search query given in the language (source language) of the scholar (e.g. German, (user-given or proactively))
Result: **query (list of terms)**

Semantic analysis of the search query with ontologies of the source language.
Result: **native enhanced query** (e.g. in German)

Translate native enhanced query into target language
Result: **Translated enhanced query** (e.g. in Romanian)

Semantic analysis of the translated enhanced query with target language ontologies (e.g. Romanian ontologies)
Result: **Target language enhanced query**

DAICA instantiation retrieval with *native enhanced query* and *target language enhanced query* in public and own instantiations
Result: **Earlier created DAICA instantiations** related to the query, i.e., **first result from the DAICA storage**

Do **textual search** by iterating the *target language enhanced query* and get results from target library.
Result: **Target language search results** (e.g. metadata and resources = documents)

Ontology alignment of target metadata scheme to DAICA ontology structure
Result: **Structured search results in target language**

OCR of found resources, if needed
Result: **OCRed resources**

Entity and link discovery in *structured search results in target language and OCRed resources* by using target specific ontologies
Result: **Target language entities and links**

Summarization of resources by using target specific summarization algorithms
Result: **Target language summary**

Translate *structured search results in target language, target language entities, target language summary* **into the source language**
Result: **Source language search results, entities and links, and summary**, i.e., **second result from target data source**

Fig. 2. Process of a search session with ontological enhancement, DAICA instantiation (see next section) retrieval, and machine translation

4 The DAICA Framework

In order to obtain sustainable, reusable results the objective of DAICA is to build a general framework which will be used for the development of complex specialised architectures, each accommodating different capabilities, selected during configuration sessions. These capabilities realize the basic technologies for investigation tasks, i.e., optical character recognition (OCR) for enabling the translation of text images into words, search which takes the meaning of queries and documents into account, machine translation for interpreting documents in foreign languages, summarisation for getting a quick overview of a document or article, entity and link discovery for identifying important persons and subjects in a text. However, it is not evident which capabilities to use at which time, or, if capabilities come in variants, which version is best for a given investigation task.

Therefore, DAICA is defined as a framework, its infrastructure, and a suitable interface technology which can be used to interactively assemble architectures for selecting suitable components which implement the capabilities for a given investigation task. Various kinds of users, "aggregators", will use this kit to build **DAICA configurations** that best support their own needs or meet the investigation requirements of others.

DAICA instantiations will constitute another layer of outputs resulting from DAICA interactions. A DAICA instantiation (or instance) represents the combination between a DAICA configuration and a specific set of acquired resources from different data sources (usually referring to a specific topic a certain user has worked on). These resources will be accumulated by a user (or a community of users) during a series of work sessions with DAICA. Hence, all interactions typically with specific investigation goals and directed to specific resources, are stored, catalogued and offered for post-research re-use, for the benefit of their creators or future users. Hence, DAICA instantiations are collections of resources with respect to a topic.

Examples of DAICA configurations can be:

- DAICA-1: Capability to process contemporary German, Spanish, English and Romanian, with OCR, indexing, external linking of name entities, summarisation, and translation between these four languages;
- DAICA-2: Processing German and Romanian texts from 1850 to present date, OCR including the Gothic German and the transitional Cyrillic alphabet used in Romania in the middle of the XIX-th century, indexing and external linking of name entities, time expressions, summarisation, and translation between these two languages.

Examples of DAICA instantiations can be:

- Based on DAICA-2: Links to the bibliographical sources in the Library of Hamburg and Academy Library of Bucharest, knowledge-base with dated entries related to the migration in Germany and Romania in the XIX-th century;

- Based on DAICA-1: Links of German academic libraries with information from Basque archives, in relation to investigations accomplished by German scholars in the Basque Country in XIX century.

Hence, an instantiation summarizes all information about a specific topic, e.g., content identified in some libraries, notes made by the user. Furthermore, if made public, other users can make use of and refine such previously created instantiations through the DAICA instantiation retrieval mechanism. As such, the DAICA instantiations are the base for building a community discussing and further developing cultural topics.

DAICA uses a number of already existent technologies, which will be adapted to comply with the actual requirements. The DAICA capabilities include the following features:

- Specification and customization of the investigation task (proactive and triggered search specifications);
- Specification of data sources, their access, and their content in the form of metadata schemas, languages, or ontologies and used terminology, hence enabling access to foreign data and content without the need of manually traversing user interfaces or interpreting a library structure (by data source profiles);
- Analysis of ancient digitized documents (by pattern recognition and OCR, word spotting);
- Deep semantic search through data sources (by indexing and semantic search);
- Automatic translation of queries and resources (by machine translation);
- Linking of resources with expressive relationships on the basis of semantic entities (by entity identification, reference resolution, detection of temporal and spatial relations);
- Creating summaries of the identified resources (by automatic summarisation);
- Friendly end-user interfaces for the visualization of complex search results and dependencies in the Web for different types of devices: laptop, tablet, smart phone (by innovative visualization and user-device interaction);
- Easy configuration of new processing architectures to support a wide range of thematic investigations (by configuration facilities);
- Projects profiling for storing, retrieving and sharing of resources as DAICA instances (by instantiation facilities).

In summary, these features will be integrated in a generally applicable and customizable technological framework that will allow easy configuration of new architectures in order to help researchers and other categories of users to perform assisted cultural HSS investigations. Once installed in a DAICA platform, the framework can be used by aggregators (libraries, research institutes, administration) to configure new applications that will allow public users to get access to new data or to administer previously curated DAICA instantiations.

5 Technologies for DAICA

5.1 Existing investigation tools

The widely used Aleph integrated library system provides academic, research, and national libraries with the efficient, user-friendly tools and workflow support they need to meet the increasing requirements of the industry today and in the future. Built on an Oracle database, Aleph runs on a range of operating systems. Employing system-wide XML technology, Aleph offers third-party integration through an XML gateway. The product is based on industry standards, offering the ultimate in resource-sharing capabilities, full connectivity, and seamless interaction with other systems and databases.

Another solution used in libraries is DigiTool, which enables academic libraries and library consortia to manage and provide access to digital resources, both those that are created for use within the institution and those that are collected and maintained by the library for the benefit of the public.

Since many resources have a public exposure on the Web, other existing investigation tools or techniques which can be used for searching are the Web search engines and crawlers. Some open source or commercial tools (which can influence the solution) are: Datapark, ebhath, Eureka, Indri, ISearch, IXE, Lucene, Managing Gigabytes (MG), MG4J, mnoGoSearch, MPS Information Server, Namazu, Nutch, Omega, OmniFind IBM Yahoo! Ed., OpenFTS, PLWeb, SWISH-E, SWISH++, Terrier, WAIS/ freeWAIS, WebGlimpse, XML Query Engine, XMLSearch, Zebra, BBDBot and Zettair.

Besides search itself, one technique to be used when combining HSS data from multiple sources is data integration. State of the art approaches for data integration have adopted a schema-first (e.g., ETL, enterprise integration), a schema-never (e.g., search engines), or a schema-later (e.g., dataspaces) methodology.

Such tools provide the basic search and data access interfaces to library content. For DAICA, libraries operating those tools can and will be integrated through data source profiles. Furthermore, the provided search facilities of the tools will be used by the semantic search capability to perform keyword-based search.

A lot cultural assets are currently published through EUROPEANA. EUROPEANA bases its search functionalities on who, what, where, when and corresponding restrictions for media type, language, country, and provider. DAICA will base the search on semantic ontologies and multilingual access, thus, facilitating the document access for users. However, through the envisioned data source profiles, EUROPEANA can be integrated in the DAICA framework and, thus, be part of a DAICA investigation.

5.2 OCR, pattern recognition

For identification and retrieval of digitzed but not yet recognized documents, DAICA includes OCR (Optical Character Recognition) tools. The main challenges which have to be faced are the following:

- OCR must be based on a variety of historical fonts and spellings.
- Document images may have poor quality and may require image enhancement.
- Character and word recognition may be ambiguous due to noise.
- Word, sentence and semantic context must be exploited for disambiguation.

There exist several commercial and open source OCR tools, which perform high-quality OCR (up to 99%) for standard fonts and low-noise conditions [1]. On the other hand, character and word recognition results may be quite poor (below 80%) without prior knowledge of the font and without exploiting context information.

Hence, DAICA applies several innovative techniques to achieve high-quality OCR. First, OCR tasks are supported by their semantic context using meta-data and ontologies. Hence, ambiguities can be significantly reduced. For example, ambiguous readings can be refuted if the semantic distance (computed from an ontology such as WordNet) to the investigation topic exceeds a threshold. As a second innovative technique, applicable to manuscripts or unusual fonts, DAICA will allow word spotting based on patterns supplied by the investigator. This way, occurrences of similar patterns can be retrieved from a document. A third technique, mainly applicable to handwritten documents, will be the use of an advanced text-line finder which can cope with varying line orientations.

Thus, the approach for DAICA will be mainly based on existing OCR tools of the partners and open source tools, as well as low-level and context-supported computer vision and manuscript analysis [2,3,4].

5.3 Semantic search

A central goal of the DAICA is to provide support for studies of cultural heritage by extending keyword-based search to a much broader search based on semantic relations. A semantic search has the advantage of narrowing down ambiguous word meanings, especially across language barriers, and allowing proactive background search for related information. This goal can be achieved by a variety of techniques which try to take the intention of the user and the meaning of a query into account when searching in data sources.

There exist several approaches for semantic search as documented, for example, in the surveys, [5,6,7]. DAICA lays the focus on exploiting ontological information which are used in two fundamental ways: (i) to define, refine and expand the query topic; (ii) to find semantically related information in data sources.

Several publicly accessible implementations of semantic search approaches exist, including QWant, GoPubMed, Swoogle, and Google's Knowledge Graph, which deal with specific kinds of ontological representations. These techniques, however, do not meet the requirements for the intelligent agent conceived in this work: (i) DAICA will have to access a large number of heterogeneous content structures used in the archiving institutions for cultural heritage or similar data aggregations. Some may be supported by full-fledged Semantic Web ontologies, others by customized categorization schemes. In consequence it will be necessary to invoke ontology alignment in some form; (ii) Search will be multilingual, crossing language barriers between the user and information sources; (iii) In DAICA, the user can define a query by several kinds of topic descriptions, ranging from keywords, annotated images, graphic patterns, to coherent texts. Hence several heterogeneous measures for semantical distance will play a part, for example taxonomical distance, relatedness by names, time or geographical location, or chains of ontological structures; (iv) The user will be supported by proactive search, i.e., by autonomous background explorations through entity and link discovery in user's text writing; (v) Access to DAICA will be possible via mobile devices, and rendition of results will include summarisation.

The software for individual techniques is mostly available either as open source or detained by the authors. The main task for the DAICA is to conceive and integrate a tool combining the techniques in a user-friendly way.

5.4 Ontology management

In our approach, ontologies play an important role for obtaining meaningful search results in support of a user's investigation. All essential DAICA functionalities resort to ontologies, in particular semantic search, language translation, interpreted OCR, entity discovery, topical linking, and summarisation. Ontologies may provide concept names and definitions in terms of relations to other concepts, for example generalization, specialization, synonyms and antonyms. Standardized properties relate entities to important search criteria, such as location and time.

Due to the highly heterogeneous data sources of the cultural heritage and diverse evolved standards, investigations with DAICA have to cope with multiple ontologies in different languages, ranging from carefully designed OWL ontologies to simple databases characterized by metadata schemes. In order to determine the relevance of resources for a user query, DAICA must be able to align these ontologies with the semantics of the query. Several methods for query answering based on multiple and multilingual ontologies have been developed in the past decade, see [8,9,10] for surveys. Typically, there is a matching (or alignment) step where correspondences between heterogeneous ontologies are determined, and an interpretation step, where information relevant for a query is extracted.

In the DAICA infrastructure, ontology matching and interpretation will be performed for ontologies based on standards such as Dublin Core or Schema.org, on controlled vocabularies (WordNet and thesaurus vocabularies), and on existing biographical data standards and classifications. In several countries data sources are described by authority files of standardized metadata, in Germany: Gemeinsame Normdatei, beacon files, gazetteer data and links, files for *Common public corporation data (GKD, Gemeinsame Körperschaftsdatei)* with company and institution names, registers with personal names (PND), and *Common norm* *data file (SWD, Schlagwortnormdatei)* with commonly used tag words, categories, and subject headings. Mass data with named entity tagging and recognition data will enhance the scope of results and open up semantic relations and links to more resources.

5.5 DAICA instantiation retrieval

A DAICA instantiation represents a DAICA configuration and the resources acquired by a user or a community of users having close scientific interests for a specific topic using this configuration. As such, a DAICA instantiation represents a complete investigation case which is both, a useful documentation for the investigator and a valuable resource for similar investigations of other users. It is the objective of DAICA to support all users of the DAICA community by a case base of instantiations and case-based retrieval mechanisms.

Case-based information retrieval is a well-established technology, see [11,12] for surveys. While case-based retrieval has been originally conceived for feature-based object representations, applications to relational structures have proved quite successful [13]. More recently, case-based retrieval was further enhanced by ontology-based representations and corresponding similarity measures [14]. During the development of DAICA, a special theoretical attention will be given to an ontological organisation of the collections of DAICA instantiations. For example, issues of interest here are: demarcation strategies (when is it that two instances have to be considered as identical or distinct?), inheritance (is it that an instantiation A inherits parts of descriptions, sources, links, etc. from an instance B?).

5.6 Machine translation, multilingual processing in combination with semantic search and summarisation

Developing of efficient machine translation is a long-lasting and multi-level process. DAICA uses a mixture of the mature technologies of statistical, example-based and and rule-based machine translation (SMT and RBMT). As basic features, DAICA includes:

- Resource collection (semi-automatic parallel corpora extraction and dictionary building, with special emphasis on lesser-resourced languages and in-domain registers). These data will be used for training and tuning machine translation modules.
- Development of the query translation module. Previous experience and expertise of the partners will be used for adapting existing methods to language pairs of DAICA. SMT is language-independent, and the same toolkit can be used for any pair of languages provided specific single language texts and parallel texts for all language translation pairs exist. But the state of the development rule-based machine translation (RBMT) varies, depending on the language pair.

- DAICA will use the Apertium software [15]. Apertium is a classical shallow-transfer or transformer system, released under GNU Licence. Apertium includes dictionaries for language pairs involving Spanish. The Apertium MT engine consists of the pipelined modules for morphological anlaysis, part-of-speech tagger, and text generators as well as Statistical Machine Translation (SMT) based on the Moses toolkit [16].

Hence, our summariser is multilingual at the architectural level, meaning that it incorporates a pipeline of modules which has the same structure irrespective of the language of the processed document. However, initial elements of this chain (among which, the tokeniser, the POS-tagger, the lemmatiser, the NP-chunker, the clause splitter, the name entity recogniser, and the anaphora resolver) are strongly language dependent.

In the former ATLAS project[6], summarisers for Bulgarian, German, Greek, Polish and Romanian have been built, meaning that our general summarisation architecture has been adapted for all these languages by assembling basic-levels NLP modules supplied by partners. For DAICA, we will build summarisers for German, English, and Romanian, by re-using (and, where necessary, also enhancing) the German and Romanian basic components and including open-source modules for English.

The situation is somehow different when the language of the documents is old. Romanian or, for instance, has changed dramatically over time. Not only the lexica, grammar and syntax have evolved, but also the alphabet has changed from Old Cyrillic to Latin, with a mixture of the two, called the Cyrillic Transition Alphabet, used for a period in the middle of the XIX-th century. Based on previous work [17,18] we will study on diachronic Romanian morphology.

5.7 Entity and link discovery

Recognition of entity mentions in texts (names of people, moments of time, countries, locations, events, organizations) and their correct interpretation in context is an issue of primary importance in DAICA. These mentions should open access gates to entries in the collection of accessible resources. Examples for points of interests are:

- Identify entity mentions in metadata field values and full texts and, if necessary, do their ontological interpretations, e.g., identify temporal entities and historical dates and events;
- Identify relevant relations between entities such as relations between instances: <person> is-in <place> (at <time>), <country> invades <country>, <person> signs <treaty>, etc.
- Identify collections of documents having contingent content such as <document> is-primary-source-for <event>, <document> in-relation-with <event>, <document> mentions <person> (at <time>);

[6] www.atlasproject.eu

One approach for entity discovery is the use of large repositories of entity names (gazetteers), such as person names, topics, locations, or temporal mentions. Ontologies and terminological databases can equally be used. This approach might look brute force, however, because of the existence of authority files and terminologies in library research, there is a huge amount of such entity storages and ontologies which can be used, similar to those mentioned in Section 5.4. Furthermore, we have recently built a large collection of regular expressions for the identification of geographical locations in free texts. As other means, larger contexts and syntactic analysis can be used to identify relations between entities. Once such relations are detected, the documents containing them can be tagged and indexed accordingly, providing information that can be used for intelligent retrieval.

5.8 Summarisation

Nowadays the quantity and diversity of data in the internet on whatever subject is extremely vast, which makes it more and more difficult to enquire on specific subjects. The needed information is usually hidden in an ocean of garbage data. This is one aspect of the well-known problem of information overload. One way to deal with it is to use summarisation techniques. In [19] a summary of a text is defined as a piece of text that conveys important information of the original one and that is not longer than half of the original length, usually significantly less than that. Summarisation is a hard problem in Natural Language Processing because, in order to do it properly, one has to really understand the point of a text. This requires semantic analysis, linking of mentioned entities (usually referred as anaphora resolution), discourse processing, and inferential interpretation.

Text summarisation methods can be classified into extractive and abstractive. An extractive summary includes sequences of words taken from the original document, which could be clauses, sentences or paragraphs. An abstractive summary does not reproduce sequences from the original document, but rather includes paraphrases of sections that mention important facts, events, and entities.

Many systems are known which perform automatic text summarisation, applying different techniques. Some use surface methods (involving no linguistic analysis but exploiting instead the format of the document), some take name entities from the original text as pivot elements and assume that the texts surrounding them is important and should stay in the summary (involving some kind of lexical analysis and classification methods); some relate significant features in the text and the summary and try to copy the ability to produce summaries from human-produced ones (involving learning and statistical methods); and some are based on discovering the discourse structure (involving processing at linguistic, syntactic and discourse level). The summarisation systems can also be considered from the point of view of the number of the processed texts, as single and multi-document, by the languages processed, as monolingual or multilingual, as well as by the genre of the processed texts.

The summarisation approach in DAICA is an extractive single-document process producing general or focussed summaries. It will enhance the approach described in [20], which is currently considered as one of the leading approaches in state-of-the-art automatic summarisation. It involves a long processing chain (including a tokeniser, a part-of-speech tagger, a noun phrase chunking module, a name entity recogniser, an anaphora resolution module, a clause splitter and a discourse parser). The improvements that we plan to realise in DAICA on the multilingual summarisation model, initially built in the ATLAS project, will concern a number of directions, including (i) the anaphora resolution engine - by adding rules that would allow coreference resolution on more finer criteria, (ii) the clause splitter module - by implementing and integrating in the calibration system of new machine learning algorithms, (iii) the discourse parser - by integrating the newly acquired enhancements of the Veins Theory [21], focussed towards reducing the search space in an incremental parsing process [22], or (iv) those using the recently proposed metrics of comparing tree structures [Mitocariu et al., 2013].

6 Summary

In this paper, we presented a concept for a digital assistant which integrates and combines semantic technologies such as interpreted OCR, semantic search, summarization, entity and link discovery, machine translation, and case-based retrieval for supporting users in investigation and research tasks. As a main focus, the assistant consideres resources of cultural heritage data sources such as libraries. However, the underlying technologies allow the application of the DAICA concept to arbitray Internet sources such as web pages or social media data. This paper represents a preliminary step of refining the conceptual and design principals before starting the actual development process of a new technology. However, the basic technologies which will be used for a complete DAICA system have been applied by us in similar approaches. We believe that the present day technologies, belonging to the domain of Artificial Intelligence, that have attained a theoretical and applicational maturity can be combined in DAICA in a very creative way.

References

1. Singh, S.: Optical Character Recognition Techniques: A Survey. Journal of Emerging Trends in Computing and Information Sciences **4**(6) (June 2013) 545–550
2. Buhr, F., Neumann, B.: Evaluation of Retrieval Performance in Historical Newspaper Archives comparing Page-level and Article-level Granularity. Technical Report Technical Report FBI-HH-M-337/06, Universität Hamburg, Hamburg (2006)
3. Hotz, L., Neumann, B., Terzić, K.: High-Level Expectations for Low-Level Image Processing. In: Proceedings of the 31st Annual German Conference on Artificial Intelligence. Volume 5243 of Springer Lecture Notes in Computer Science., Kaiserslautern (September 2008) 87–94
4. Herzog, R., Neumann, B., Solth, A.: Computer-based Stroke Extraction in Historical Manuscripts. Manuscript Cultures, Newsletter (3) (2011) 14–24

5. Grimes, S.: Breakthrough Analysis: Two + Nine Types of Semantic Search. InformationWeek **2010** (2010) 1–21
6. Mangold, C.: A survey and classification of semantic search approaches. International Journal of Metadata, Semantics and Ontology **2**(1) (2007) 23–34
7. Mäkela, E.: Survey of Semantic Search Research. In: Proceedings of the Seminar on Knowledge Management on the Semantic Web, Department of Computer Science, Univ. Helsinki (2005)
8. Stock, K.: An approach to the management of multiple aligned multilingual ontologies for a geospatial earth observation system. In: Proc. 4th Int. Conf. on GeoSpatial Semantics, Springer (2011) 52–69
9. Rameshi, C., Gnanasekaran, A.: Methodology Based Survey on Ontology Management. International Journal of Computer Science & Engineering Survey (IJCSES) **1**(1) (2010) 1–12
10. Granitzer, M., Sabol, V., Onn, K.W., Lukose, D., Tochtermann, K.: Ontology Alignment - A Survey with Focus on Visually Supported Semi-Automatic Techniques. Future Internet **2** (2010) 238–258
11. Daniels, J.J., Rissland, E.L.: A case-based approach to intelligent information retrieval. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM (1995) 238–245
12. Mitra, M., Chaudhuri, B.: Information Retrieval from Documents: A Survey. Information Retrieval **2** (2000) 141–163
13. de Mantaras, R.L., McSherry, D., Bridge, D., Leake, D., Smyth, B., Craw, S., Faltings, B., Maher, M.L., Cox, M.T., Forbus, K., Keane, M., Aamodt, A., Watson, I.: Retrieval, reuse, revision, and retention in case-based reasoning. The Knowledge Engineering Review **20**(3) (2005) 215–240
14. Zidi, A., Bouhana, A., Mourad, A., Fekih, A.: An ontology-based personalized retrieval model using case base reasoning. In: Proc. 18th Int. Conf. on Knowledge-Based and Intelligent Information & Engineering Systems. (2014) 212–222
15. Forcada, M.L., Ginestí-Rosell, M., Nordfalk, J., O'Reagan, J., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Sánchez-Martínez, F., Ramírez-Sánchez, G., Tyers, F.M.: Apertium: a free/open-source platform for rule-based machine translation. Machine Translation **25**(2) (2011) 127–144
16. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: ACL 2007: proceedings of demo and poster sessions. (2007) 177–180
17. Simionescu, R., Cristea, D.: Automatic Morphologic Classification System for Romanian. In et al., L., ed.: BringITon! 2012 Catalog. Editura Univercity Al. I. Cuza, Iasi, Romania (May 2012) 52–53
18. Cristea, D., Simionescu, R., Haja, G.: Reconstructing the Diachronic Morphology of Romanian from Dictionary Citations. In: Proceedings of LREC-2012, Instanbul, Turkey (May 2012) 923–927
19. Radev, D.R., Hovy, E., McKeown, K.: Automatic Morphologic Classification System for Romanian. Computational Linguistics **28**(4) (2002) 399–408
20. Anechitei, D., Cristea, D., Dimosthenis, I., Ignat, E., Karagiozov, D., Koeva, S., Kope'c, M., Vertan, C.: Summarizing Short Texts Through a Discourse-Centered Approach in a Multilingual Context. In Neustein, A., Markowitz, J., eds.: Where Humans Meet Machines: Innovative Solutions to Knotty Natural Language Problems. Springer Verlag, Heidelberg/New York (May 2013) 109–136

21. Cristea, D., Ide, N., Romary, L.: Veins Theory. A Model of Global Discourse Cohesion and Coherence. In: Proceedings of 17th International Conference on Computational Linguistics - Coling '98, and the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - ACL '98, Montreal, Canada (August 1998) 281–285
22. Mitocariu, E.: Veins theory revisited. Dissertation, Univercity Al. I. Cuza, Iasi, Romania (2015 - in preparation)