Multifaceted, Collaborative Access to Cultural Heritage

Project IST-034984

D2.1 Conceptual Alignment of Multiple, Distributed Content Sources

Deliverable

Editor:	Nahum Korda
Contributors:	Straight Technology, Technion, Idea, JAFI
Dissemination Level:	Public
Version:	1.3
Date:	28-Apr-07

Executive Summary

Document Objective

Detailed exposition of the efforts undertaken in the task *T2.1 Distributed ontology alignment* of the work package *WP2 Semantics and Ontology Engineering*.

Targeted Audience

MOSAICA technological partners

Prerequisite

IST-034984 MOSAICA Technical Annex
D5.1 MOSAICA System & Application Specifications

Summary

This deliverable is a report on the accomplishments in the task *T2.1 Distributed ontology alignment* of the work package *WP2 Semantics and ontology engineering*. Due to the situation encountered at the onset of the project where no ontologies in the sense of Gruber's definition of ontology were available, the originally planned methodological approach was abandoned, and the original title of this deliverable *D2.1 Methodology for Alignment of Multiple, Distributed Ontologies* was modified into *D2.1 Conceptual Alignment of Multiple, Distributed Content Sources*. As the methodological approach suitable to the situation in the project we adopted the theoretical framework for alignment proposed in [Kent 2000]. In order to satisfy both principles assumed in this methodological approach, the MOSAICA Core ontology was developed as a common conceptual framework used for the conceptual alignment of the specific conceptualizations implied in the content sources. The content sources selected as the initial content for MOSAICA were selected by the consortium partners are the Jewish Encyclopedia Online and the eJewish.info directory. The methodology developed in MOSAICA by adapting the original methodological proposal in [Kent 2000] consists of two phases: 1) identification of instances, and 2) consolidation. Identification of instances is actually a methodological "shortcut" merging two procedural steps envisioned in the original proposal – "lifting" and "fusion" – into a single step. During the implementation several utility applications were developed.

© 2007 MOSAICA

This document contains proprietary information of MOSAICA and shall not be reproduced or transferred to other documents, disclosed to others, or used for any purpose other than that for which it is furnished, without the prior written consent of MOSAICA.

The trademark and service marks of MOSAICA, including the MOSAICA mark and logo, are the exclusive property of MOSAICA, and may not be used without permission. All other marks mentioned in this material are the property of their respective owners

Contents

1. Introduction 4
   1.1. Preliminary Remarks 4
   1.2. Structure of the Deliverable 7
   1.3. Motivation 8
   1.4. Terminological Clarifications 12
2. State of the Art 16
   2.1.1. Further Terminological Clarifications 17
   2.1.2. Theoretical Framework 18
   2.1.3. Implementations 19
   2.1.4. Methodological Conclusions 22
3. MOSAICA Methodology 23
   3.1. General Approach 23
   3.2. Development of MOSAICA Core Ontology 26
      3.2.1. Extraction of Distinctive Features 30
      3.2.2. Formalization 35
      3.2.3. Outcome: MOSAICA Core Ontology 39
   3.3. Identification of Instances 39
      3.3.1. Extraction of Instances from the eJewish.info Keywords 41
      3.3.2. Conversion and Alignment of the Jewish Encyclopedia Online 43
      3.3.3. Outcome: MOSAICA eJewish and MOSAICA JE ontologies 46
   3.4. Consolidation 46
      3.4.1. Consolidation Following Alignment of eJewish.info Keywords 47
      3.4.2. Consolidation of the Alignment of the Jewish Encyclopedia Online 49
      3.4.3. Outcome: MOSAICA Domain Ontologies 49
4. MOSAICA Implementation 51
   4.1. eJewish.info Keywords Parser and Converter 51
   4.2. SANDRA Search AND Retrieval Application) 52
   4.3. MOSAICA Crawler and Extractor 53
   4.4. Simple OWL Editor (SOE) 56
5. Conclusions 58
6. References 62

1. INTRODUCTION

1.1. Preliminary Remarks

This deliverable was planned as a report on the accomplishments in the task *T2.1 Distributed ontology alignment* of the work package *WP2 Semantics and ontology engineering*. When planning this task, the original expectation was that multiple ontologies on the Jewish cultural heritage exist somewhere "out there" on the WWW, and that the underlying semantics of the project could be based on these external resources by aligning them into a single, unifying ontology. – The factually encountered situation was rather different, as it became evident during the first months of the project.

One of the chief obstacles in the ontology engineering is currently the so called "knowledge acquisition bottleneck" ([Buchanan et al. 1993]), i.e. lack of domain ontologies that would enable innovative application of the already available semantic technologies. This is due to the fact that domain modeling is tedious and complex, and requires in-depth understanding of formal logic, and expertise in Web standards such as XML, RDF and OWL. However, domain experts – such as experts in cultural heritage – typically lack the required enthusiasm to develop such specific – additional – expertise in computer science, just in order to model their knowledge domain, so that it could be used by sophisticated computer software. Accordingly, the situation encountered at the onset of the project was that there was no ontology whatsoever on the Jewish cultural heritage on which the system developed in the project could base its semantics.

Hereby, a terminological clarification concerning the use of the term "ontology" is necessary. Guarino and Giaretta point out in [Guarino and Giaretta 1995] that the term "ontology" is often indiscriminately used either as "a particular conceptual framework at the semantic level", or as "a concrete artifact at the syntactic level, to be used for a given purpose". To the former use of the term "ontology" they refer as "conceptualization", and are reluctant to identify it with what they call "intensional" meaning of "ontology". Therefore, they adopt (and slightly adapt[1]) the definition of ontology by Gruber (which is probably the most quoted definition of ontology) in [Gruber 1993]:

[1] Guarino and Giaretta advocate an intensional approach to conceptualization inspired by the Montague semantics. According to that approach, relations used for conceptualization must be formally defined as constraints in respect to possible worlds to which the conceptualization intends to apply. Since such a set of constraints can be expressed in multiple models (which they name "ontological commitment"), no single ontology can be considered to be a **complete** account of a conceptualization. Therefore, they modify the Gruber's definition of ontology into: "a logical theory which gives an explicit, partial account of a conceptualization" (basically, inserting the specification "partial" into the original definition).

[1] "An ontology is an explicit specification of a conceptualization."[2]

This definition and the discussion around it are rather technical in character, and often ignored or misunderstood by domain experts. Based on our experience, it seems that the term "ontology" is often identified among the domain experts with conceptualization in general (at the "semantic level"), rather than with "explicit specification of conceptualization" (at the "syntactic level"). Accordingly, ontology engineers often get surprised when discovering what the actual level of conceptualization that they will need to handle in the application domain is.

Nevertheless, the domain of Jewish cultural heritage is well conceptualized, as evident when inspecting valuable cultural heritage resources, such as the Jewish Encyclopedia Online[3] and the eJewish.info[4] directory – despite the fact that these resources do not provide an "**explicit** specification" of their underlying conceptualizations. Accordingly, the efforts in the task *T2.1 Distributed ontology alignment* had to be redirected towards developing Jewish heritage ontology in the sense of the Gruber's definition – based on the formalization and alignment of the available content sources.

Hereby, an additional terminological clarification concerning the use of the term "alignment" is also necessary.

When discussing alignment, typically two formal conceptual models are envisioned as participating in the procedure. In this scenario, elements of one model need to be mapped into the other in such way that some kind of correspondence[5] between them is established. To that purpose a series of technologies have been developed as utility applications[6], as detailed in the section 2 State of the Art below.

However, According to Gruber's definition, the conceptual models encountered in the content sources such as the Jewish Encyclopedia Online or the eJewish.info directory could not be considered ontologies, since they are not "explicit specifications of conceptualization". As currently available technologies presume such explicit specification of conceptualization as a precondition for ontology alignment, not only that

[2] This is the compact version of the original definition. The full version is the following: "A specification of a representational vocabulary for a shared domain of discourse — definitions of classes, relations, functions, and other objects — is called an ontology."
[3] www.jewishencyclopedia.com
[4] http://www.ejewish.info
[5] In [Kiryakov et Al. 2001] there are two sets of primitives defined, called InterOntologyRel and IntraOntologyRel, each of which has a number of relations that aim to capture the correspondence of concepts originating from different ontologies (e.g., equivalent, more specific, meta-concept, etc.). A typology of these relations is given in the form of a hierarchy.
[6] Such as, SMART, PROMPT, Chimaera, or GLUE.

these technologies turned out to be unsuitable for MOSAICA, but evidently an altogether different methodological approach had to be sought.

This realization had a critical implication for the research efforts in the task *T2.1 Distributed ontology alignment*: a methodological approach was required that does not merely map one formalized conceptual model into another, but that rather aligns implicit conceptual models through their systematic explication. Indeed, outlines of required methodological approaches were found in the previous research efforts ([Kent 2000], [Kalfoglou and Schorlemmer 2002], [Stumme and Maedche 2001]), as reported in the section 2 State of the Art below, but it is worth noticing here that the situation encountered in MOSAICA is not completely unique in the research literature.

Gal et al. in [Gal et al. 2005] present a scenario in which query forms from various websites need to be aligned, in order to enable automatic submission of a query to multiple websites (for example, airline reservations). This scenario is relevant here not only because it describes a situation similar to that encountered in MOSAICA, but primarily because it demonstrates that conceptual alignment must be handled in practice in a broader context than merely as correspondence between two formal conceptual models. Gal et al. coined the term "application semantics" to describe the semantics that is embedded, for example, in constituents of a query form webpage, such as input fields, their corresponding labels, etc., or in core business rules that guide the layout of such webpage, such as the fact that pickup location always precedes drop-off location in car rental input forms. Although such application semantics cannot be considered ontology in the sense of the Gruber's definition, there is doubtlessly an underlying conceptual model behind it. Accordingly, although not "ontology alignment" in the strict sense of the term, the effort to establish correspondence between heterogeneous "application semantics" is certainly "conceptual" in its character, and certainly based on the underlying semantics.

Lacher and Groh in [Lacher and Groh 2001] present CAIMAN, a system, which uses machine-learning for ontology mapping. Although this system is of little interest to MOSAICA, the authors elaborate on a scenario where members of a community would like to keep their own perspective on a community repository. – They continue by arguing that each member in a community of interest organizes her documents according to her own categorization scheme – which they consider to be "ontology". This rather weak account of ontology justifies, to a certain extent, the use of a user's bookmark folder as a "personal" ontology. The mapping task is then to align this "personal" ontology with the directory structure of CiteSeer[7] (also known as ResearchIndex).

Although such notion of "personal" ontology can be used only metaphorically, CAIMAN is an important and well cited research effort in ontology alignment, and must

[7] Available at: http://citeseer.ist.psu.edu/.

be taken into account when considering the state of the art. It clearly demonstrates that in practice formalized conceptual models (i.e. "explicit specifications of conceptualization") may be an exception, rather than the rule – situation for which the "knowledge acquisition bottleneck" ([Buchanan et al. 1993]) was blamed above.

Therefore, rather than discussing "ontology alignment", which seems too narrow for the purpose of the discussion in this deliverable, we adopted the broader term "conceptual alignment", which seems more suitable to the context of MOSAICA.

Accordingly, the original title of this deliverable *D2.1 Methodology for Alignment of Multiple, Distributed Ontologies* was modified into *D2.1 Conceptual Alignment of Multiple, Distributed Content Sources*.

1.2. Structure of the Deliverable

Being a report on the accomplishments in the task *T2.1 Ontology alignment*, the intention of this deliverable is therefore to:

1. Clearly describe the problems encountered in our attempts to meet the user requirements that were defined in the project (and which were documented in the deliverable *D5.1 MOSAICA System & Application Specification*), and
2. Systematically present the solutions that we devised in order to overcome these problems.

Accordingly, in the remainder of this section we shall present the motivation for our efforts in this task, and how these efforts are related to the user requirements defined in the project. In addition, in this chapter we shall also provide some further terminological clarifications, in order to avoid possible confusion that might emerge as the result of inconsistent and inexact use of terminology.

In the section 2 State of the Art we shall briefly present the current situation in the ontology engineering related to the alignment of conceptual models. This chapter is not intended to serve as a survey of the past and recent efforts related to this subject, but rather as an overview of the difficulties encountered in the field by other researchers, as well as of various proposals and attempts to overcome these difficulties.

In the section 3 MOSAICA Methodology we shall introduce the specific methodology for alignment of heterogeneous content sources and their corresponding conceptual models, which was developed within the project. This methodology was designed primarily to meet the specific requirements of the system that is being developed in the project, and was intended to provide solutions that can be immediately implemented in a working system. Nevertheless, our ambition was also to devise a generic methodological

approach that could be applied to other domains of application, and even to other systems beyond MOSAICA.

In the section 4 MOSAICA Implementation we shall describe the implementation of this methodology in several (utility) components of the MOSAICA system, which were factually employed to provide the basic ontologies used today by all components of the MOSAICA system. This description will not go into unnecessary technical details regarding the implementation, but will rather focus on the architectural design of the implementation, as well as on the requirements that these components had to meet.

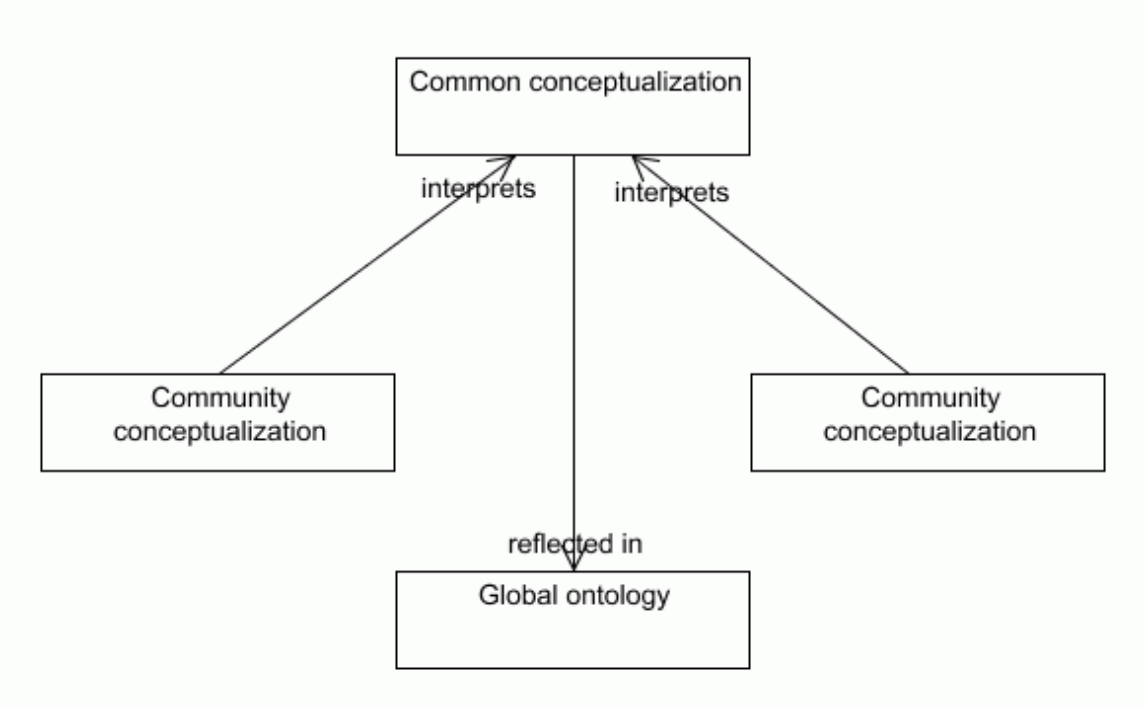
Finally, in the section 5 Conclusions we shall present our conclusions regarding how MOSAICA methodology for conceptual alignment relates to the current state of the art in the research field.

1.3. Motivation

MOSAICA is a complex system integrating multiple external content sources into a homogeneous display framework. In order to make this possible, a unifying conceptual framework is required, which could enable heterogeneous content to be uniformly mapped into a common conceptual model. The specific difficulty encountered in MOSAICA is in the fact that not only MOSAICA's **content** is distributed and heterogeneous, but MOSAICA moreover utilizes heterogeneous **conceptual models**. This is necessary, since MOSAICA unifies content sources such as digital libraries and document repositories that have their own, internal conceptual models and data schemas. In order to unify the heterogeneous content, MOSAICA must therefore initially unify heterogeneous conceptualizations of that content.

This situation is well conceptualized by Kent in [Kent 2000], and we shall outline it here in slightly modified manner, which will be justified in the section 3 MOSAICA Methodology below. The conceptual model of MOSAICA is schematically depicted in Figure 1.

Figure 1 MOSAICA conceptual model



There are multiple "community ontologies", or in the case of MOSAICA community conceptualizations. Such community conceptualizations are, for example, the Jewish Encyclopedia Online and the eJewish.info directory. Nevertheless, a "common ontology" – or in case of MOSAICA: a common conceptualization – should be also assumed. This common conceptualization would be in MOSAICA the conceptual model of the domain of the cultural heritage in general, and of the Jewish cultural heritage in particular. Various community conceptualizations could be then viewed as specific interpretations of the common conceptualization, potentially with different coverage of the domain of discourse. – For example, the Jewish Encyclopedia Online covers the Jewish cultural heritage only until the beginning of the 20th Century.

Each of the community conceptualizations comes with its own community resources. – These are the articles form the Jewish Encyclopedia Online, and the websites from the eJewish.info directory. In order to unify the use of these multiple resources, the common conceptualization must be elevated into "global ontology". Hereby, this "global ontology" must be indeed provided as ontology in the sense of the Gruber's definition, and not merely as an implicit conceptualization, since the system will need to use it in order to provide its functionalities.

The underlying complexity of such effort will be demonstrated in this deliverable on the integration of two thematically highly valuable cultural heritage resources such as the

Jewish Encyclopedia Online and the eJewish.info directory[8]. However, it is our intention to integrate in the future more and more semantic sources into the MOSAICA conceptual framework.

Jewish Encyclopedia Online is organized into an alphabetic list of encyclopedic articles and graphic files. The encyclopedic articles are in plain text without a uniform document structure (except its division into HTML paragraphs), although certain lexico-syntactic uniformity exists within a particular type of articles (for example, within articles on geographic locations, within articles on people, etc.). These encyclopedic articles are linked with the accompanying graphic files, as well as between themselves by a limited set of annotation relations.

On the other hand, the eJewish.info directory provides a certain taxonomic model organizing links to various websites into a conceptual hierarchy similar to the WWW directories, such as Yahoo!. Each link is accompanied with a short textual description of the content that could be found on the corresponding website, and sometimes also with a set of keywords from a predefined controlled vocabulary.

Evidently, these two resources are characterized by radically different organization (i.e. alphabetic ordering and annotation relations vs. taxonomy and keywords). These differences – even at the superficial level of physical organization – demonstrate well the need for a unifying conceptual model. There is no way to reconcile alphabetic ordering and taxonomy at the organizational level, which is necessary in order to integrate their content into a unified display framework. Such reconciliation could be accomplished only at a higher level of their common conceptualization.

This inherent complexity in unification of heterogeneous content sources is stipulated by the requirements defined for the MOSAICA system in the deliverable *D5.1 MOSAICA System & Application Specification*. All of the use cases specified in that document presume that displayed content is external to the system, and that it originates from multiple distributed content sources. Hereby, MOSAICA cannot merely serve as a "federation" layer providing a single-point access to these sources, but must provide the semantics for the integration of the heterogeneous content into a single-point access frame. This need for conceptual alignment is reflected in all specific requirements for the *Semantic Layer*, *GIS/TIS*, *Intelligent Presentation of Neighboring Objects*, *Virtual Expeditions*, and partially even for the *Distributed Content (P2P) Layer*. In particular the user requirements depicted in Table 1 stipulate the semantic integration of the content sources.

[8] These sources were selected in the work package WP1 according to the criteria presented in the internal deliverable D1.1 Selection of Web-based resources in Jewish heritage.

Table 1: User requirements relevant to conceptual alignment

Section	Requirement
2.2.2 Advanced Semantic Search	4. Applicable ontology relations should be included into the advanced keyword querying interface. Selection of the applicable ontology relations should be performed dynamically, depending on the previous choices (possibly, by selection of topics or domains).
2.2.3 Semantic Browsing	1. Semantic browsing must provide a hierarchic presentation of entire ontology, or its subsections.
	4. Clicking on a node in the hierarchical structure should display additional ontology relations and ontology concepts applicable to the selected node (i.e. ontological assertions where selected concept is the subject, a relation is the predicate, and another concept or a value is the predicate). It is recommended that the display of the additional ontological elements is either textual, or as a spider diagram, or any similar diagram.
2.2.4. Display of Search Results	1. Display of search results for both semantic search and semantic browsing must include: (1) display of relevant associated resources, (2) display of matched ontological concepts, and (3) display of other ontological elements (i.e. relations and concepts) related to the matched ontological concepts (i.e. ontological assertions where selected concept is the subject, a relation is the predicate, and another concept or a value is the predicate).
2.2.5. Semantic Annotation	8. Semantic annotation should apply logical inference during the submission, in order to generate additional annotation associations (for example, by inferring a historical period from date and location).

What is common to all these requirements is the need to integrate multiple content sources via their underlying semantics, and thus enable various conceptual models to seamlessly interoperate within the system. This further implies that irrespectively of the specific conceptualization present in each specific content source, the entire system must provide a uniform and unifying semantic surface on which all system components can operate. Therefore, the alignment of the heterogeneous conceptual models had to be the first critical challenge that the project undertook.

We would like to emphasize the fact that MOSAICA is **not** merely an experimental system being developed for research purposes. MOSAICA is envisioned as a robust, industrial product with commercial application. This stipulates a hands-on approach in research, rather than sheer “theoreticizing”, and demands practical solutions, and research results that can be immediately implemented.

1.4. Terminological Clarifications

In this deliverable we are going to distinguish between the **intension** and the **extension** of conceptualization[9]. The distinction between the two originates from the early days of the formal logic ([Frege 1892]), and expresses two different aspects of the semantics: connotation and denotation. However, for the purpose of the discussion in this deliverable, we shall use the following definitions[10]:

[2] **Intension** of a concept is its formal definition, which includes all properties that are required of all possible instances of the concept.

[3] **Extension** of a concept is the set that includes all instances to which the concept applies.

Accordingly, the intension of the concept "book", for example, would be "physical object consisting of a number of pages bound together"[11], while its extension are all existing books.

By analogy, the database theory distinguishes between the intension of a database, which is its definition in form of a schema, and the extension of a database, which are the actual values in the database ([Batini et al. 1986]). We shall further extend this analogy to any conceptualization, and thus define the intension and the extension of conceptualization as follows:

[4] **Intension** of conceptualization is the theoretical construct (model) that represents some phenomenon, data or theory, with a set of variables and a set of logical and quantitative relationships between them.

[5] **Extension** of conceptualization is the set of all statements (assertions) in the domain of discourse to which it is applied.

Accordingly, the intension of conceptualization is actually the conceptual model or "explicit specification of conceptualization", and is consequently ontology in the sense of the Gruber's definition.

[9] Not to be confused with the intensional and extensional conceptualization introduced in [Guarino and Giaretta 1995] (see below).
[10] The following definitions are a gross simplification of the underlying philosophical problem that lead to the distinction between the intension and the extension. Actually, the following definitions are rather definitions of the *intensional definition* of a concept, as distinguished from its *extensional definition* (such as enumeration).
[11] WordNet 2.1

This distinction is relevant for the better understanding of the initial situation encountered in MOSAICA. Thematic content sources on the Jewish cultural heritage – i.e. community conceptualizations – could be viewed as the extension of the conceptualization of the Jewish cultural heritage – i.e. of the common conceptualization: these content sources contain numerous assertions about individual concepts declaring logical and quantitative relationships between them. Nevertheless, these sources do not provide the corresponding intension of that conceptualization, i.e. they do not explicitly specify the underlying conceptual model.

Actually, the organization that they **explicitly** provide is not specific for the domain of Jewish heritage, and cannot be considered to model that specific domain of discourse. The Jewish Encyclopedia Online organizes articles in alphabetic order. Alphabetic ordering is certainly not a useful model for the domain of Jewish heritage, since it is not unique to any particular domain of discourse. The question is also whether the taxonomy provided by the eJewish.info could be considered more useful, since at least the upper level of this taxonomy could be also applied across a wide range of domains of discourse. – As we shall explain in the section 3.2 Development of MOSAICA Core Ontology below– the answer to this question was negative.

Nevertheless, content sources such as encyclopedias and directories contain also an implicit conceptualization, at the level of the content that they provide. For example, encyclopedic articles could be roughly classified into several distinct categories, such as articles on personalities, geographic locations, historical periods, events, artifacts, etc. Each category typically follows a loose, but nevertheless recognizable lexico-syntactic pattern used for narration that could be depicted with appropriate tools, and used for the explication of the underlying semantics, as factually demonstrated by the methodology developed in the project. As we shall argue in the section 5 Conclusions below, the main effort in the conceptual alignment of conceptual models that are not "explicit specifications of conceptualization" must be invested into discovery, analysis and explication of this implicit, latent semantics[12].

There is a large body of literature devoted to the problem of the computer-assisted explication and formalization of latent semantics (often referred to as automatic "knowledge acquisition"), which we shall not survey in this deliverable. It is nevertheless necessary to point out several highly relevant research efforts, in order to completely clarify the distinction that we make between the intension and extension of conceptualization.

Early attempts to automatically extract structured semantic information from free text (among others [Amsler 1980], [Calzolari 1984], [Calzolari 1988], [Chodorow et al. 1985], [Markowitz et al. 1986]) targeted electronic dictionaries, rather than documents.

[12] Comparable to the "application semantics" in [Gal et al. 2005].

This is not surprising, since the syntactic structure of dictionaries and glossaries, is much more uniform when compared to the variety of other free texts, and guarantees efficient and precise machine processing. Furthermore, the lexical information in the electronic lexical resources is dense, and every sentence is highly relevant for the knowledge acquisition.

For example, Dolan et al. describe in [Dolan et al. 1993] an automated strategy, which exploits on-line dictionaries to construct a richly structured lexical knowledge base. In particular, it is demonstrated how LDOCE[13] can be used to build a directed graph which captures semantic associations between words. The result is a huge and highly interconnected network of words linked by arcs labeled with semantic relations such as Hypernym, Part_of, Location, and Purpose. If formalized in a Description Logic, such outcome would in fact provide ontology in the sense of the Gruber's definition.

However, with the emergence of the WWW, the interest of the researchers turned from the lexical resources to online documents, web pages, and other completely unstructured text (among others [Hearst 1992], [Klavans et al. 2002]). Several approaches to automatic extraction of semantic information from free text were examined. A rule based system named DEFINDER was introduced by Muresan et al. in [Muresan et al. 2003] used to extract glossaries embedded in Web pages in order to build a terminological database. Bootstrapping technique was introduced in [Morin 1999] and [Riloff and Jones 1999]. Bootstrapping technique was combined with statistical approach by Daille in [Daille 1996].Term extraction by collocation was examined by Smadja in [Smadja 1993].

Of particular interest is the system called SMART presented in [Dori et al. 2004] for two reasons. One is that SMART was designed to automatically extract a software system model from a corpus of requirements documentation, which is then formulized in the Object Process Model (OPM). This demonstrates that even free text with much less internal lexico-syntactic coherence than electronic dictionaries can be effectively used to formulate formal, "explicit specifications" of the underlying semantics, which is of direct relevance to the situation encountered in MOSAICA. The other reason for interest in this system is in the natural language processing platform called Search AND Retrieval Application (SANDRA), that was also integrated in the implementation of the MOSAICA methodology, as presented in the section 4.2 SANDRA Search AND Retrieval Application) below.

Extraction of semantic information from free text documentation is evidently a well-established research field with significant achievements. These efforts lead to a conclusion directly relevant to the situation encountered in MOSAICA: in an application domain the intension of conceptualization should not be taken for granted (or it may be

[13] http://www.longman.com/ldoce

partial, and require to be supplemented with additional modeling). – Nevertheless, the required conceptual model could be acquired starting with the extension of the underlying conceptualization by systematically acquiring elements required for its "explicit specification". In that respect, SMART – which operates on the lexico-syntactic surface similar to those encountered in the Jewish Encyclopedia Online and the eJewish.info, and automatically produces a formal semantic model from the implicit semantics – indeed served as a valuable road sign in the development of the MOSAICA methodology for conceptual alignment.

2. STATE OF THE ART

As already stated, the following section is not intended as a survey of the research efforts performed in the field of ontology alignment. – There are already many excellent surveys available in the relevant research areas, such as [Shvaiko and Euzenat 2005], [Doan and Halevy 2005], [Halevy 2005], [Noy 2004], [Rahm and Bernstein 2001], [Kashyap and Sheth 1996], [Larson et al. 1989], [Batini et al. 1986], and in particular [Kalfoglou and Schorlemmer 2003] – and there is indeed no need for yet another survey in this deliverable. Instead, our intention is to point out several closely related research efforts, and outline the "lessons learnt in the trenches", in order to justify, and better explain the methodology that was developed and applied in MOSAICA.

Hereby, it must be noted that the majority of relevant research was performed within the database community in the related areas of database schemata alignment and federated databases. Database schema matching or integration is regarded by many as closely related, and even similar to ontology mapping, which follows the ever-increasing belief that ontologies are similar to database schemata. Although this statement has many supporters (mainly in the databases community), there are substantial differences which should be taken into account. [Noy and Klein 2002] identified a number of areas where ontologies and database schemata are significantly different:

1. Database schema evolution aims to preserve the integrity of data itself, whereas ontology evolution is more complex since ontologies can be seen as data themselves, and a typical query on ontology could result in elements of the ontology itself.
2. Database schemata do not provide explicit semantics for their data, whereas ontologies are logical systems, and hence the intended semantics is explicitly and formally specified.
3. Database schemata are not sharable or reusable – usually they are defined over a specific database, whereas ontologies are by nature reusable and typically extend other ontologies.
4. Traditionally, database schema development and update is a centralized process, whereas ontology development is more decentralized and collaborative.
5. Database schema evolution should take into account the effects of each change operation on the data, like addition of a new class. In ontologies, however, the number of knowledge representation primitives is much higher and more complex: cardinality constraints, inverse properties, transitive properties, disjoint classes, definition of logical axioms, type-checking constraints.
6. Databases make a clear distinction between schema and instance data, whereas in rich knowledge representation languages used for ontology modeling it is difficult to distinguish where the ontology ends and the instances begin.

Accordingly, in this section and throughout this deliverable, we shall draw certain conclusions coming from the research area of the database schemata alignment, although we shall not analyze these research efforts in detail, and we shall not base the developed methodology on the approaches developed in that area. Instead, we shall focus on several research efforts (theoretical and practical) directly focusing on the ontology alignment, and analyze some of them in a greater detail.

Hereby, we shall distinguish in the existing research efforts between two types of approaches: deductive and inductive. It is not easy to draw a clear-cut line between these approaches, since in practice they are often combined. In general terms, they could be distinguished by their starting point. Deductive approaches start with two "explicit specifications of conceptualization" – i.e. with what we defined as **intension of conceptualization**, which as a rule must be provided as two ontologies formalized in a Description Logic – and attempt to match their elements utilizing either heuristics[14], or machine learning algorithms. Typical examples of such approach are SMART [Noy and Musen 1999], PROMPT [Noy and Musen 2000], PROMPTDIFF [Noy and Musen 2002], Chimaera [McGuinness et Al. 2000], or GLUE [Doan et Al. 2002]. These approaches that work "top-down" are evidently of less interest for MOSAICA, since the situation encountered in the project was radically different, as already explained in the section 1.1 Preliminary Remarks above.

In contrast, inductive approaches – even when there is an intension of conceptualization at hand – start with the so-called "lifting": explication of the extension of conceptualization, typically by populating existing ontologies with instances (and their corresponding assertions). Inductive approaches that work "bottom-up" are therefore highly-relevant to the situation encountered in MOSAICA, and will be therefore examined in more detail.

However, first some terminological clarifications are necessary.

2.1.1. Further Terminological Clarifications

We shall adopt the definition of ontology alignment proposed in [Kalfoglou and Schorlemmer 2003].

First, they define ontology as a pair $O = (S, A)$, where S is what they call after [Bench-Capon and Malcolm 1999] the "ontological signature", i.e. a core ontology consisting of the vocabulary, and A is a set of axioms, specifying the intended interpretation of the vocabulary in some domain of discourse[15].

[14] For more on heuristics typically employed in matching strategies see section 5 Conclusions.Conclusions
[15] It is not completely clear whether instances belong to S or to A.

Next, they use this definition of ontology, in order to provide two definitions of ontology mapping:

[6] **Total ontology mapping** from $O_1 = (S_1, A_1)$ to $O_2 = (S_2, A_2)$ is a morphism $F : S_1 \rightarrow S_2$ of ontological signatures, such that $A_2 \models f(A_1)$.

[7] **Partial ontology mapping** from $O_1 = (S_1, A_1)$ to $O_2 = (S_2, A_2)$ implies a sub-ontology $O_1' = (S_1', A_1')(S_1' \subseteq S_1' and A_1' \subseteq A_1)$ such that there is a total mapping from O_1' to O_2.

It is worth noticing that for the extensions of conceptualization the size of those encountered in MOSAICA, only the partial ontology mapping could serve as a realistic objective in practice.

2.1.2. Theoretical Framework

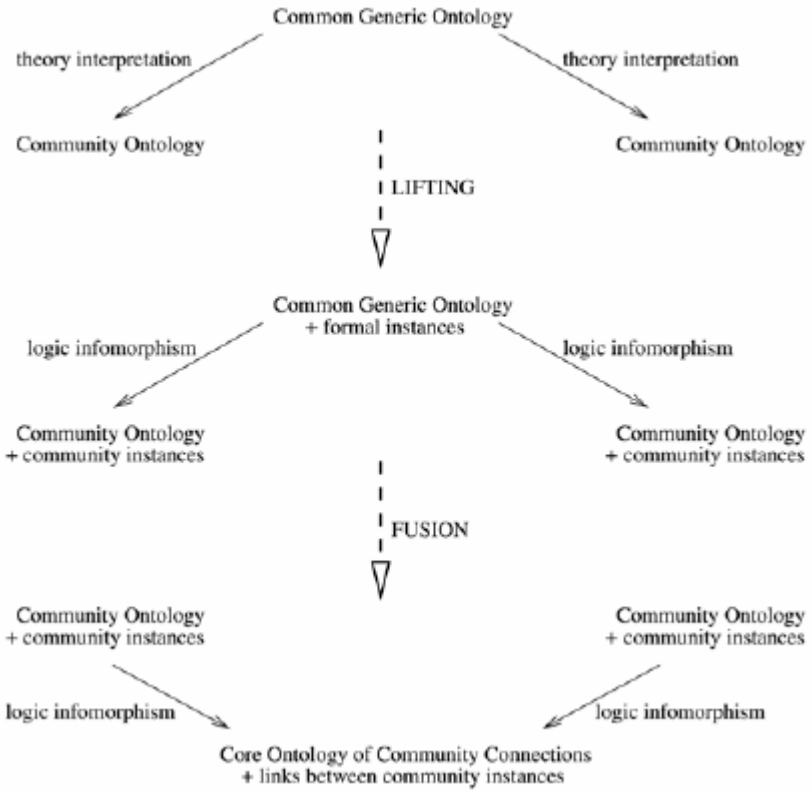
Kent in [Kent 2000] bases an idea for ontology alignment on the Information Flow Framework (IFF) introduced in [Barwise and Seligman 1997], and exploits the central distinction made in the channel theory between types (syntactic elements, like concepts and relations) and tokens (semantic elements, like particular instances and their corresponding assertions). Following the IFF, he assumes two basic principles:

1. A community with a well-defined conceptualization owns its collection of instances (controls updates to the collection, enforces soundness, controls access rights to the collection, etc.).
2. Instances of separate communities are linked through the concepts of a common generic ontology.

The former of these principles evidently holds for MOSAICA, since the content sources encountered at the onset of the project certainly provide an implicit conceptual model of the domain of discourse. However, the latter principle does not hold in MOSAICA, and as it will be demonstrated in the section 3.2 Development of MOSAICA Core Ontology below, a great deal of effort was invested, in order to comply with it.

Kent further describes a two-step process that determines the core ontology of community connections capturing the organization of conceptual knowledge across communities, as schematically depicted in Figure 2.

Figure 2 Two-step alignment process (adopted from [Kent 2000])



The process starts from the assumption that the common generic ontology is specified as a logical theory, and that the several participating community ontologies extend the common generic ontology according to theory interpretations. This process consists of the following steps:

1. **Lifting**. – This step incorporates instances into the ontologies (proper instances for the community ontologies, and the so-called "formal instances" for the generic ontology). This step could be viewed as **populating** ontologies.
2. **Fusion**. – In this step community ontologies are linked through the core ontology, which depends on how instances are linked through the concepts of the common generic ontology (see second principle above).

Kent's framework is purely theoretical, and no method for implementing this two-step process is given. Nevertheless, this framework seemed highly relevant and applicable to MOSAICA, given that the second principle above could be satisfied.

2.1.3. Implementations

Very close in spirit and in the mathematical foundations of IFF, Schorlemmer in [Schorlemmer 2002] studied the intrinsic duality of channel-theoretic constructions, and

gave a precise formalization to the notions of knowledge sharing scenario and knowledge sharing system. Drawing from the theoretical ideas of Kent's IFF and Schorlemmer's analysis of duality in knowledge sharing scenarios, Kalfoglou and Schorlemmer in [Kalfoglou and Schorlemmer 2002] propose the IF-Map methodology. To a certain extent, this proposal could be viewed as an attempt to practically implement Kent's theoretical framework, ad schematically depicted in Figure 3.

Figure 3 Ontology alignment proposed by [Kalfoglou and Schorlemmer 2002]

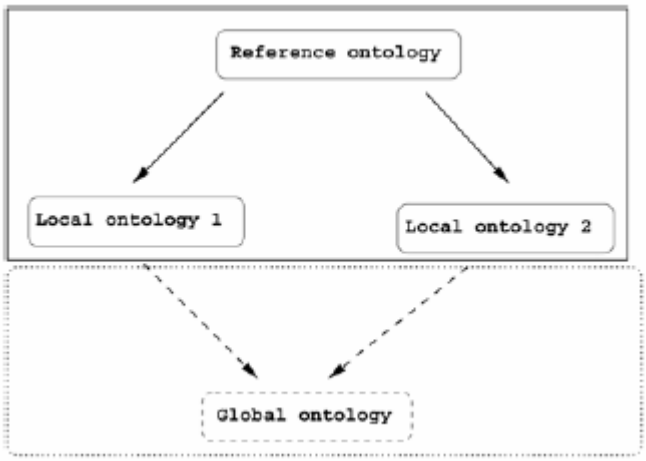


Figure 3 clearly resembles Kent's proposed two-step process for ontology sharing. The solid rectangular line surrounding *Reference ontology*, *Local ontology 1* and *Local ontology 2* denotes the existing ontologies. It is assumed that *Local ontology 1* and *Local ontology 2* are ontologies used by different communities, and populated with their instances, while *Reference ontology* is an agreed understanding that favors the sharing of knowledge, and is not supposed to be populated. The dashed rectangular line surrounding *Global ontology* denotes an ontology that does not exist yet, but will be constructed "on the fly" for the purpose of merging. – This ontology is similar to Kent's virtual ontology of community connections. The solid arrow lines linking *Reference ontology* with *Local ontology 1* and *Local ontology 2* denote information flowing between these ontologies and are formalized as logic infomorphisms. The dashed arrow lines denote the embedding from *Local ontology 1* and *Local ontology 2* into *Global ontology*. The latter is the sum of the local ontologies modulo *Reference ontology* and the generated logic infomorphisms.

It is worth noticing that the *Reference ontology* is defined merely as "an agreed understanding", rather than an "explicit specification of conceptualization", or "ontology signature". Accordingly, when implementing this model, it is not necessary to expect an intension of conceptualization ready for use, as long as there is a clear common understanding regarding the domain of discourse – which is certainly the case in MOSAICA. Accordingly, this implementation allows a starting point similar to that

encountered in MOSAICA, and provided both a solid theoretical framework, and the implementation directives.

Another highly-relevant and closely related research effort is presented in [Stumme and Maedche 2001]. It is based on work on Formal Concept Analysis (FCA) [Ganter and Wille 1999] and lattice exploration. The authors incorporate natural language techniques in a system called FCA-Merge to derive a lattice of concepts. The lattice is then explored manually by a knowledge engineer who builds the merged ontology with semi-automatic guidance from FCA-Merge.

Of particular interest is the fact that the input to the method is a set of documents from which concepts and the ontologies to be merged are extracted by utilizing natural language processing. These documents should be representative of the domain in question, cover all concepts from both ontologies as well as separate them well enough. As this method relies heavily on the availability of classified instances in the ontologies to be merged, the authors argue that this will not be the case in most ontologies and they opt to extract instances from documents.

In this respect, the first step of FCA-Merge could be viewed as an ontology population mechanism. Once the instances are extracted, and the concept lattice is derived, Stumme and Maedche use formal concept analysis techniques to generate the formal context for each ontology.

Next, the two formal contexts are merged to generate a pruned concept lattice. This step involves disambiguation (since the two contexts may contain the same concepts) by means of indexing. The computation of the pruned concept lattice is done by an algorithm, TITANIC, which computes formal contexts via their key sets (or minimal generators).

Having generated the pruned concept lattice, FCA-Merge enters its final phase, the non-automatic construction of the merged ontology, with human interaction. This construction is semi-automatic as it requires background knowledge about the domain.

Both of the presented approaches are similar and highly-relevant to MOSAICA in several points:

1. In contrast to the deductive approaches mentioned above, these methodologies do not stipulate intensions of conceptualization as their starting point. The latter even explicitly suggests starting with a document corpus, i.e. with what we refer to as the extension of conceptualization – arguing that this is the only way to achieve the required level of granularity of the ontological instances. Accordingly, the starting procedure in both of these approaches is ontology populating.

2. Both of these methodologies assume a common conceptual model ("reference ontology" in [Kalfoglou and Schorlemmer 2002], and lattice of concepts in [Stumme and Maedche 2001]) used for establishing the common ground between ontologies to be aligned. The "explicit specification" of this common conceptual model is hereby not a requirement.
3. Both of these methodologies perform alignment by classifying instances into the common conceptual model. Ontologies are considered aligned when both of them are mapped into that common model.
4. Both implementations combine automation and manual editing, resulting in semi-automatic systems. This is well-qualified in [Kalfoglou and Schorlemmer 2003] as a pay-off between "automation and high-quality mappings".

2.1.4. Methodological Conclusions

The analysis of the state of the art in the field of ontology alignment clearly pointed out that the situation encountered at the onset of the project is not completely unique, and that it could, and maybe even should be expected in many other domains – domains of cultural heritage, and broader. Actually, the deductive approaches seem applicable only in an almost "ideal" situation, such as alignment of two versions of the same ontology [Noy and Musen 2002]. In the "real world" it must be realistically expected that significantly more complex situations will need to be handled.

This analysis also provided very clear directives regarding the productive directions in which the research in MOSAICA needed to advance. It became evident that prior to the "fusion" a "lifting" ([Kent 2000]) is required as a necessary procedural step. It also became evident that a common conceptual framework needs to be established as the precondition for assuming the alignment procedure. As we shall present this in the section 3.2 Development of MOSAICA Core Ontology below, such common conceptual framework was developed in the project as the MOSAICA Core ontology.

Finally, this analysis also sets very clear boundaries for the expectations. It is evident that the implementation must opt for a semi-automatic system consisting of numerous components, including natural language processing platform and ontology editor.

3. MOSAICA METHODOLOGY

3.1. General Approach

According to [Smith et al. 2004], for aligning classes, individuals and properties in multiple ontologies OWL explicitly[16] provides three corresponding predicates: owl:equivalentClass, owl:sameAs, and owl:equivalentProperty.

The following explanation is provided for the predicate owl:equivalentClass: "The property owl:equivalentClass is used to indicate that two classes have precisely the same instances[17]" ([Smith et al. 2004]). Accordingly, when aligning two OWL-formatted ontologies, the equivalence between two classes could be in fact established only after determining that they both have identical instances. For example, let us consider the following situation[18]:

```
<owl:Class rdf:ID="WineColor">
  <owl:oneOf rdf:parseType="Collection">
    <owl:Thing rdf:about="#White" />
    <owl:Thing rdf:about="#Rose" />
    <owl:Thing rdf:about="#Red" />
  </owl:oneOf>
</owl:Class>
```

```
<owl:Class rdf:about="&alt;ColorOfWine">
  <owl:oneOf rdf:parseType="Collection">
    <owl:Thing rdf:about="&alt;White" />
    <owl:Thing rdf:about="&alt;Rose" />
    <owl:Thing rdf:about="&alt;Red" />
  </owl:oneOf>
</owl:Class>
```

Equivalence between the classes WineColor and ColorOfWine could be established through the following procedure:

1. All instances of the class WineColor would be identified with all instances of the class ColorOfWine using owl:sameAs predicate[19],

[16] We assume that other primitive predicates in OWL, such as owl:subClassOf or owl:disjointWith, could be used to the same purpose, although this is not explicitly suggested in the documentation.

[17] $\{A \equiv B : \forall x \in A, \forall y \in B : (x \in A \supset x \in B) \cap (y \in B \supset y \in A)\}$

[18] Based on the notorious Wine ontology available at: http://www.w3.org/TR/owl-guide/wine.rdf.

[19] It is not in the current scope of the discussion to establish whether the coincidence in the identical names of the individual concepts is the sufficient evidence for establishing identity of the individual concepts, or not.

2. It would be established that neither of the classes has additional instances that were not identified with instances of the other class, and
3. Class ColorOfWine would be formally defined as owl:equivalentClass of the class WineColor.

However, this strategy is at odds with all deductive alignment approaches presented in the section 2 State of the Art above: these approaches are deductive in the sense that they attempt to establish equivalence on the axiomatic base, and then deduce belonging of the instances of both classes to the same set. In contrast, the strategy implied in the description of the primitive owl:equivalentClass should be inductive, since it stipulates the class equivalence could be inferred only after identity of the corresponding instances sets is established.

In practice the inductive strategy may be applicable only to the enumeration classes that explicitly enumerate all of their constituents. Other classes could be semantically equivalent, although in each of the ontologies that are to be aligned we may encounter only a partial account of their constituents that may only partially overlap, or not overlap at all. Accordingly, even if there would be sufficient evidence to establish identity between instances of these classes in both ontologies, there may not be sufficient evidence that both classes have exactly the same instances.

Before proceeding, it is worth noticing that the deductive strategy is also inexact, and depends on the human judgment. The effectiveness of approximation in both cases depends in ultima linea on the quality and quantity of information about corresponding elements in both ontologies that are to be aligned.

The inductive strategy is of special interest in the context of MOSAICA. As presented in the section 1 Introduction above, MOSAICA faced the situation where two extensions of conceptualization were available, although none of them provided the corresponding intension of the underlying conceptualization. In such situation the inductive strategy seems to be the only applicable approach: this strategy allows starting with only conceptual extensions, and then building up the more generic relationships.

This is why we adopted the theoretical framework for alignment proposed in [Kent 2000]. In order to adopt this methodological approach, MOSAICA had first and foremost to comply with its basic assumptions:

1. A community with a well-defined conceptualization owns its collection of instances (controls updates to the collection, enforces soundness, controls access rights to the collection, etc.).
2. Instances of separate communities are linked through the concepts of a common generic ontology.

Regarding the former assumption, both the Jewish Encyclopedia Online and the eJewish.info directory could be certainly considered as required collections of instances with controlled updates to the collection, enforced soundness, controlled access rights to the collections, etc. These collections are indeed community collections with well-defined conceptualization of the domain of discourse.

Regarding the latter of the above assumptions, the problem was that no formal ontology in the domain of Jewish cultural heritage was available, although a common generic conceptualization could be certainly assumed in the domain of discourse. Accordingly, the precondition for the adoption of the alignment approach proposed in [Kent 2000] was development of a common generic ontology that could serve as the common conceptual framework for the fusion of the collections of instances. This ontology was envisioned as an upper-level ontology consisting of a limited number of generic classes and relations describing the domain of cultural heritage in general, and of Jewish cultural heritage in particular.

Hereby, the user requirements for the MOSAICA System had also to be taken into consideration. MOSAICA Core ontology is intended to be used as a navigational instrument – for example, providing menus in the GIS, or displayed as a navigational tree for semantic browsing. Accordingly, the usability issues had to be considered when designing the MOSAICA Core ontology – in particular, regarding its extent and organization.

An additional consideration was the actual coverage of the MOSAICA Core ontology. MOSAICA is envisioned as a system designed for a broad range of cultural heritages, and the Jewish cultural heritage is only the domain selected for the proof of concept. Accordingly, the MOSAICA Core must be actually a conceptual model of the cultural heritage in general. The upper-level elements specific for to Jewish cultural heritage – although necessary for the implementation – had to be separately modeled.

The efforts dedicated to the development of the MOSAICA Core are described in greater detail in the section 3.2 Development of MOSAICA Core Ontology below.

With the MOSAICA Core ontology both of the principles assumed in [Kent 2000] were satisfied, allowing the adoption of the proposed methodology. The proposed two-step methodological approach was then adopted as follows:

1. **Lifting**. – The first methodological step was devised as development of two separate ontologies consisting of only individual concepts (and corresponding assertions) extracted from the Jewish Encyclopedia Online and the eJewish.info directory. – The extraction procedures for both sources differ significantly, due to the differences in their internal organization, but similar heuristics was applied in both cases. In both cases extraction proceeded as populating MOSAICA Core ontology – because ontologies containing ontological schemas were not

developed for each content source separately. The extraction procedures are described in greater detail in the sections 3.3.1 Extraction of Instances from the eJewish.info Keywords and 3.3.2 Conversion and Alignment of the Jewish Encyclopedia Online below.

2. **Fusion**. – Fusion of ontologies was partly accomplished already through the extraction procedures, due to the design of the applied extraction procedures, as detailed in the section 3.3 Identification of Instances below. It was completed by cross-referencing the ontologies consisting of instances and the consolidation of the resulting ontology (the so-called "global ontology" [Kalfoglou and Schorlemmer 2002]), as described in the section 3.4 Consolidation below.

Accordingly, the entire alignment strategy developed in the task *T2.1 Ontology alignment* could be outlined as follows:

1. Development of the MOSAICA Core ontology (section 3.2 Development of MOSAICA Core Ontology),
2. Alignment of the Jewish Encyclopedia Online and the eKewish.info directory, which proceeded in two consecutive steps:
    a. Extraction of instances and their corresponding assertions (section 3.3 Identification of Instances), and
    b. Merging and consolidation (section 3.4 Consolidation).

3.2. Development of MOSAICA Core Ontology

The "lifting" ([Kent 2000]) presumes the existence of an ontology consisting of classes that are then populated by their corresponding instances. The practical problem with the adoption of the proposed approach was that the classes that formally describe such sets of instances were not explicitly provided. – Accordingly, an additional procedural step had to be included, providing grouping of the instances into meaningful classes.

Hereby, a potential strategy could be to apply a clustering technique to the instances in both extensions. However, these extensions have only an **implicit**, latent underlying conceptualization. In contrast, clustering techniques requires common traits for grouping these instances to be known in advance, and assumes that they could be measured by some kind of distance metrics. Therefore, the clustering techniques can be of little help, and what is actually required here, is to explicate the underlying conceptualization by – first and foremost – establishing such common traits.

We shall refer to these common traits as "distinctive features", and formally define them in the section 3.2.1 Extraction of Distinctive Features below.

Initially, the strategy for acquiring and using distinctive features was planned as follows:

1. The processing of the eJewish.info was to be undertaken first. Although eJewish.info is much less lexico-syntactically homogeneous than the Jewish Encyclopedia Online, eJewish.info provides a controlled vocabulary used for annotation, and called eJewish.info Keywords. It was initially assumed that the hierarchical structure of these keywords could explicate these distinctive features by grouping keywords into logically defined subsets. These subsets could be then exploited in order to provide the required "explicit specification of conceptualization" in the domain of Jewish cultural heritage, and it was expected that they could be simply automatically converted into ontology.
2. The Jewish Encyclopedia Online articles were then to be aligned automatically with the eJewish.info Keywords by automated extraction of instances from it.

However, soon, numerous problems related to these expectations were detected, and this conversion proved not to be that trivial as initially expected. The following is the analysis of the problems encountered, intended not as a criticism of the eJewish.info approach to design and use of keywords, but rather as a detailed account of the situation regarding the semantic resources encountered in the project – situation that should be expected also in other domains of cultural heritage when applying the MOSAICA System to them.

1. **Design for a WWW directory, and not a knowledgebase**. – The hierarchical structure of the keywords provides a browsable layout typical of the WWW Directories, with the pronounced ambition to cover as much as possible of the potential interests of the prospective users. This structure is thus not designed to cover specifically the domain of cultural heritage, but rather to provide a broad range of topics satisfying potential curiosity in just any domain of life, as long as it is – somehow – related to Jewish communities and their day-to-day life. For example, topics such as "Entertainment", "Addictions" and "Human body" are also extensively covered. There are three immediate consequences of that approach:
    a. The hierarchical structure is both "too deep" (up to 10 levels) and "too broad" (65 keywords in the top level, but some levels have over one hundred keywords). Such structure is clearly inappropriate for the navigation based on dropdown lists and menus, and in most cases the impression was that the quantity was simply favored over the quality of information provided.
    b. Many of the keywords are too generic (for example, "Sciences" or "Computer Games"), and have nothing to do specifically with the cultural heritage in general, or with the Jewish cultural heritage in particular. Moreover, some keywords merge several well-distinguished concepts (for example, "Writers", "Screenwriters", "Novelists" and "Poets" are all assigned a single keyword), in order to provide a more generic category.

   c. The logical structure of the eJewish.info Keywords – as is – is clearly inadequate for conversion into ontology. For example, it was established that by such conversion of the eJewish.info Keywords the ultimate super-class of the concept "Concentration Camp" would be "Science and Technology"[20], leading to a logical assertion: "Concentration Camp is a kind of Science and Technology".

2. **Lack of internal logic**. – The division of keywords by levels often lacks any internal logic. For example, the top-level keywords contain "Science and technology", "Physical disabilities", and "Allegory". Besides that all these keyword could be of potential interest to a user, hardly any justification could be found for assigning them all to the same level. This situation is further worsened by the fact that eJewish.info Keywords are under a continuous revision, which is unfortunately undocumented so that some keywords are clearly not yet classified under an appropriate topic. For example, the top level contains also "Palestinian Authority", "Sharon (region, Israel)", and "Maloplskie (voivodship, Poland).
3. **Unsuitable content**. – The collection of keywords is often a mixture of topics from Jewish cultural heritage and the daily life in the State of Israel. Although the State of Israel is certainly an important topic in the Jewish cultural heritage, its day-to-day internal issues should be clearly distinguished from the more generic Jewish cultural heritage topics (for example, Israeli political parties that often appear and disappear from one elections to another could be hardly justified as of interest for the Jewish cultural heritage; similarly, Israeli state holidays and memorial days are mixed in the eJewish.info with the Jewish holidays and festivals). In addition, often the keywords exercise a strong Anglo-Saxon, even American bias, providing often, for example, English translations of the Hebrew terminology (e.g, "Passover"), or typical American customary spelling of Yiddish (e.g., "Shtreimel").
4. **Poor relationships**. – The relations used in the eJewish.info Keywords are extremely poor – only two: "parentKeyword" and "relatedKeyword ". Accordingly, the interpretation of these relations is left to the user's understanding of the topic, and they can imply: meronymy (for example, one geographical area that is part of another could be marked as either "parentKeyword" or "relatedKeyword "), hypernymy, hyponymy, or any other specific relation (such as, "adjacent to", "located in", "member of", etc.). Moreover, a single keyword could be related to numerous other keywords by a mixture of differing meanings of the same relation. For example, the keyword "Prayer and liturgy" is linked by the relation "parentKeyword" to "Haggadah" (which is a ceremonial text), "Yizkor" (which is indeed a prayer), and "Siddur" (which is simply the Hebrew term for a prayer book). Accordingly, most of the

[20] "Science and technology" is "parentKeyword" of "Science", which is "parentKeyword" of "Behavioral sciences", which is "parentKeyword" of "Sociology", which is "ParentKeyword" of "Management", which is "parentKeyword" of "Penal institutions", which is "parentKeyword" of "Concentration camps".

relations used in the eJewish.info are merely associative, and without any strict logical interpretation. For example, "North Africa" is "parentKeyword" of "Ancient Rome", which in turn is "parentKeyword" of "Roman Empire" – which may lead to an erroneous interpretation that the ancient city of Rome was actually located in North Africa.

5. **Structural inconsistency**. – There is a great deal of inconsistency in using relations for conceptual modeling. For example, geographical locations are related to each other by the relation "parentKeyword" (settlements are children of their corresponding countries or regions), but personalities are related to their professions by the relation "relatedKeyword ". This caused a significant problem when devising algorithms for automatic conversion, simply because these algorithms could not be applied to all keywords in a consistent manner, and required formation of specific subsets to which a particular algorithm could be applied. Moreover, the direction of relations was also inconsistent. For example, some personalities are "relatedKeyword " of a particular historical period, but also some historical periods are "relatedKeyword " of other personalities. Accordingly, both of the uses of the relation "relatedKeyword " had to be considered, in order to model a uniform logical relationship between historical periods and personalities.
6. **Incompleteness and redundancy**. – The use of the relations in the eJewish.info Keywords is both incomplete and redundant. For example, the Roman Empire is associated with North Africa, but not with Europe (or South Europe, or Italy) – which is evidently incomplete. On the other hand, the family Aaronsohn is associated with both "Ottoman period" and "20th Century", although the Ottoman period is specified to cover the 20th Century – which is evidently redundant.
7. **Mixture of class concepts and individual concepts**. – Another significant problem for devising algorithms for automated conversion was the fact that at the same level keywords could be converted either into class concepts or into instances of class concepts. For example, the keyword "Biographies" is "parentKeyword" of all personalities included in the eJewish.info Keywords, which would be considered as instances of the class concept *Personality*; however "Biographies" is "parentKeyword" also of the keyword "Autobiography", which is evidently another class concept that is not a subclass of the class *Personality*.
8. **Unsuitable labeling**. – The labels provided by the eJewish.info Keywords are also problematic considering their use in MOSAICA. Most of them are unsuitable for semantic indexing because they cannot be expected to occur in related texts phrased actually as they are in the eJewish.info Keywords (for example, "Aliyah, fifth, 1929-1939"). Some of them seem even unsuitable for display in MOSAICA, for example in dropdown lists and tree views.

Following this thorough analysis of the eJewish.info Keywords, it became soon evident that both extensions require to be modeled "from scratch".

At first, the intention was to undertake parallel efforts, and explicate two independent conceptualizations (what would correspond to the "community ontologies" in [Kent 2000]) – one for the Jewish Encyclopedia Online, and one for the eJewish.info. This attempt was soon abandoned, due to the following problems:

- It is difficult to completely independently model two closely related conceptualizations "from scratch" – even if it is done by two different persons. It became soon evident that the major differences in the emerging models are in the labels, for example: "Cultural Heritage Artifacts" vs. "Cultural Artifacts". Since labeling turned out to be arbitrary for both resources, this attempt to independently model the extensions seemed merely as an unnecessary "academic exercise".
- It was estimated that utilization of both resources in conjunction would be more beneficial, and that the resulting outcome would provide a higher quality if the content sources are cross-referenced during modeling.
- The emerging functional requirements from the entire MOSAICA System clearly pointed out that in addition to the internal organization of the content sources, additional aspects (related to the required functionalities) must be taken into consideration, and that consequently, an upper-level ontology will be required in order to satisfy these requirements.

Accordingly, rather than modeling two independent ontologies, the decision was taken to directly model that upper-level ontology, and the result was envisioned as **MOSAICA Core** ontology.

Accordingly, the methodology for development of the MOSAICA Core ontology proceeded through the following steps:

1. Extraction of the distinctive features (common traits) from both extensions, and
2. Formalization of these distinctive features in ontology.

In the following paragraphs we shall present each of these steps in detail.

3.2.1. Extraction of Distinctive Features

We define distinctive feature as follows:

[8] **Distinctive feature** is the characteristic shared among some individual items that clearly distinguish them as a group (class) from other items.

According to this definition a distinctive feature is actually class belonging. – This is not accidental, since by assuming existence of an underlying conceptual model in an extension we also assume that the elements is such extension are instances of classes formed by conceptualization.

Since these two extensions significantly vary in their appearance (one is a set of annotation keywords, and the other a set of online encyclopedic articles), the procedural approaches to them had also to vary significantly. These approaches are detailed in the following sections.

3.2.1.1. Conversion of eJewish.info Keywords into Classes and Relations

The procedural approach undertaken to convert eJewish.info Keywords into classes and relations could be described as proceeding through the following procedural steps:

1. **Development of the eJewish.info Keywords parser**. – From the thorough analysis of the eJewish.info Keywords it became evident that a great deal of manual effort will be required during the conversion. In order to facilitate this manual effort, and significantly minimize it, a small utility application was developed that can parse the XML formatted eJewish.info Keywords, and display them in a tree view for easier handling. This utility application is described in detail in the section 4.1 eJewish.info Keywords Parser and Converter below. During the following procedural steps this basic application was further developed, adding more and more functionalities to it – primarily those related to the algorithmics required to automate the conversion of relations between keywords into formal, logical assertions.
2. **Extraction of class concepts and removal of irrelevant keywords**. – The keywords that were manually identified as class concepts useful for the MOSAICA Core ontology were converted into formal classes. At the same time, keywords that either seemed too generic or unsuitable for the cultural heritage were removed.
   To that purpose the basic parsing utility was extended with a user interface enabling manual approval of the conversion of eJewish.info Keywords into classes.
   All classes were annotated using the proprietary annotation property[21] *see also* providing the reference to the corresponding keyword ID.
3. **Modeling of basic classes**. – The mapped class concepts were modeled into a hierarchical schema that strictly reflects the relationship of logical inclusion (expressed typically in natural language as “is a kind of”). To this purpose the utility application called Simple OWL Editor was provided by Straight Technology, and described in the section 4.4 Simple OWL Editor (SOE) below. A special attention was paid to the usability, ensuring that at leach level only a limited number of subclasses is modeled (in order not to exceed the “attention span” of an average user), and that a limited number of levels will be modeled (in order not to exceed user’s “tolerance” when requested to drill down into the

[21] That is to be distinguished from the native rdfs:seeAlso property, which is an rdfs:objectProperty requiring the referenced object to be formally defined, which is not the case with the eJewish.info Keywords.

hierarchical structure). – This was necessitated by – at the time – emerging user requirements.

4. **Modeling of basic relations**. – Next, a set of specific relations for each class, or group of classes was developed, in order to enable formulation of assertions about the future instances of these classes. As already mentioned, eJewish.info provides only two relations: "parentKeyword" and "relatedKeyword ". This is evidently insufficient if complex concepts in the cultural heritage domain need to be modeled. For example, in order to describe a class, such as *Historical Site*, at least the following relations are required (and were in fact developed): *is located in* (specifies the geographical location of the site in terms of country, region, settlement, etc., but not as coordinates on a map), *related to significant personality* (specifies an instance of the class Personality as significant for the site), *related to period* (specifies a historical period to which the site is relevant). Hereby, the eJewish.info Keywords were instrumental in sense that they provided a very good insight into the relationships that should be expected between various concepts in the domain.

3.2.1.2. Conversion of the Jewish Encyclopedia Online into Classes and Relations

The procedure for extracting distinctive features from a set of encyclopedic articles contained in the Jewish Encyclopedia Online in fact coincides with the methodological approaches for automatic acquiring of vocabulary from free-text documentation. Such vocabulary extraction is typically the first step in developing formal model "from scratch", and is an established research field, as pointed out in the section 1 Introduction above.

Accordingly, for handling the Jewish Encyclopedia Online natural language processing was applied, utilizing the platform named SANDRA, which was previously successfully applied to vocabulary extraction for conceptual modeling ([Dori et al. 2004]). SANDRA is presented in detail in the section 4.2 SANDRA Search AND Retrieval Application) below.

Hereby, circumstances significantly assisted our efforts: Jewish Encyclopedia Online is a well-organized lexical resource that, as mentioned above, maintains a recognizable lexico-syntactic pattern used for narration. Accordingly, the articles in it could be easily recognized as belonging to several distinct categories, such as articles on personalities, geographic locations, historical periods, events, artifacts, etc. It was already noted that the lexical information in the electronic lexical resources is dense, and that their lexico-syntactic structure is reasonably uniform, which guarantees efficient and precise machine processing[22].

[22] For example, it was noticed that the first sentence of each article defines the item that article is about, and therefore we applied NLP for vocabulary extraction only to the first sentence in every article.

The application of NLP to the Jewish Encyclopedia Online proceeded in an iterative manner, and was strongly influenced by the so-called "mutual bootstrapping" approach presented in [Riloff and Jones 1999]. The NLP approach to automatic acquiring of vocabulary from free-text documentation is based on the use of lexico-syntactic patterns that include empty slots. If the pattern is successfully matched to the document text, the empty slots are filled with the segments from the document text covered by them. Development of these lexico-syntactic patterns used for extraction is typically context specific (depends on the information that needs to be extracted), and requires a preliminary analysis of the underlying text. This is accomplished by using an initial, "bootstrapping" lexicon that extracts chunks of text that may be of interest for the design of the extraction patterns. The bootstrapping lexicon typically contains noun phrases that are then used to extract the verb phrases of interest. – These verb phrases are next transformed into extraction patterns. The idea of the mutual bootstrapping is to repeat this procedure in an iterative manner, integrate the extracted terminology into the bootstrapping lexicon, and then repeat the entire process, as long as satisfactory results are not accomplished.

Following this approach, we developed the following procedure for extraction of distinctive features from the Jewish Encyclopedia Online:

1. **Extraction of bootstrapping vocabulary**. – The bootstrapping lexicon was initially extracted by using SANDRA's categorization functionality. SANDRA provides a broad range of functionalities related to complex indexing for information retrieval. Detailed description of SANDRA is provided in the section 4.2 SANDRA Search AND Retrieval Application) below. Among other features, SANDRA provides the functionality of extracting common lexical patterns from a thematically uniform corpus of documents. To that purpose SANDRA does not require specific lexical resources, and it can provide the initial lexicon "from scratch".
2. **Formulation of lexico-syntactic patterns**. – The initial vocabulary was next used first to extract the lexico-syntactic context that may be relevant for design of extraction patterns. This procedure provided a very good insight into the relations that are required in order to describe the corresponding vocabulary. – This was in particular helpful for the interpretation of the only two relations provided in the eJewish.info Keywords, and assisted in their interpretation[23].
3. **Reiteration of the procedural steps**. – Hereby, an additional clarification is required, concerning the use of natural language processing in ontology engineering. The initial vocabulary was extracted by configuring SANDRA to detect commonly used noun phrases. This vocabulary was then used to extract the verb phrases as the context of this initial vocabulary. In conceptual modeling

[23] This is a good example of the cross-referencing effort mentioned above that was considered beneficial for the modeling of both extensions in conjunction, rather than independently.

the noun phrases coincide with the concepts, while the verb phrases coincide with the relations (properties). Accordingly, the entire procedure could be described as an iterative process of acquiring ontological classes, and then using them to acquire the related ontological properties. – By reiterating this procedure, the ontology is gradually constructed (and could be consequently formalized in OWL). These two procedural steps were, accordingly, repeated until the domain experts evaluated the resulting knowledgebase as sufficiently complete.

It is worth noticing that this approach is generic in the sense that it is neither language dependent, nor domain dependent. Since the procedure commences "from scratch", it could be applied to any natural language supported by the NLP platform. Moreover, the domain of application is also not restricted, since no specific lexical or semantic resources are required as a precondition. Accordingly, the methodology could be applied to additional domains of cultural heritage and in fact, to a broad range of domains of discourse in general.

3.2.1.3. Reconciliation of Extracted Vocabularies

The final results were two sets of vocabularies containing potential concepts for the future MOSAICA Core ontology. These vocabularies acquired from both extensions were manually cross-referenced, in order to determine overlaps, and resolve morpho-syntactic variations. This type of manual analysis must be performed by domain experts, since what is factually cross-referenced are two conceptualizations of the same domain of discourse, for which the vocabularies are only superficial references. Such analysis must accordingly establish the core set of classes meaningful and significant in the underlying domain of discourse.

An exemplifying excerpt of acquired vocabularies, and their final reconciliation is depicted in Table 2.

Table 2 Samples of vocabularies extracted from the Jewish Encyclopedia Online and eJewish.info

Jewish Encyclopedia Online	eJewish.info	Results of cross-reference
Rabbi	Biographies Rabbis	Personality Clergyman Rabbi
Festival	Rosh Ha-Shanah	Festival Rosh Ha-Shanah
Frank, Jacob, and the Frankists	Jewish Sects Jacob Frank	Movement Religious movement Messianic movement Frankists Personality Historical personality Heretic False Messiah Jacob Frank

3.2.2. Formalization

As evident from the examples above the cross-referencing included several typical strategies:

1. Identification of concepts: “rabbis” is simply an inflected form of “rabbi”.
2. Subsumption of one concept under another: “Rosh Ha-Shanah” is a “Festival” of Jewish New Year.
3. Subsumption of both concepts under a more generic category: “rabbis” and “rabbi” are both a subcategory in personalities.
4. Combined subsumption of concepts one under the other, and their subsumption under a more generic category: “Frankists” is a “social movement”, which is in turn a kind of a “Movement” (including political, religious, etc. movements).
5. Division of original concept: “Frankists” are a “Religious Movement”, while “Jacob Frank” is a “Personality”.

All of these strategies imply construction of a hierarchical structure of concepts, their properties and mutual relations. Accordingly, such cross-referencing is in fact modeling of the domain of discourse in a formal model. To that purpose OWL (Web Ontology Language) was selected as the preferred formalism, due to its world-wide acceptance as the W3C standard for semantic interoperability, and the extracted vocabulary and corresponding extraction patterns were formalized into initial MOSAICA Core ontology. Since this is the step in the presented methodology that requires direct involvement of the domain experts, Simple OWL Editor (SOE) – described in the section 4.4 Simple OWL Editor (SOE) below – was utilized as the authoring tool.

There are several obstacle encountered when conceptually modeling a domain of discourse using Description Logics, such as OWL:

1. OWL is an intricate formalism that requires familiarity with the field of ontology engineering, including the first and the second order logic, description logics, database design, XML and RDF/XML formats, URI technology, etc. However, it cannot be reasonably expected that domain experts who posses the knowledge required to develop ontologies, would in addition to their chief expertise also master the ontology engineering, just for the sake of undertaking ontology editing.
2. Currently available OWL editors are largely considered awkward and unnecessarily complex, so that domain experts are often greatly reluctant to learn how to properly operate them. Since these editors are designed for OWL, to use them assumes familiarity with OWL, which – as already stated – should not be expected in case of domain experts.
3. Currently available OWL editors largely neglect the multilingual aspect of ontology engineering. Although OWL provides mechanisms based on the RDF Schema and XML to support multilingual ontologies, the currently available OWL editors poorly support identification of ontology elements by their multilingual labels. Accordingly, these OWL editors necessitate inappropriate naming strategies, just to ensure human readability of ontologies.

The standard approach in this situation is to form “task forces” combined of both domain experts and ontology engineers, in order to jointly develop the required ontologies. In practice, this approach proved to be largely insufficient, and very few high-quality ontologies and software applications using them were developed.

An alternative approach is to provide ontology editor that would be sufficiently user friendly and simple, in order to enable the domain experts to develop independently high-quality ontologies with minimal effort, and without any specific expertise in computer science. SOE is an OWL editor based on the idea that the description logic behind RDF and OWL can be expressed in (almost) natural language, and thus enable domain experts to formulate logical assertions as relatively simple natural language clauses.

For the purpose of demonstration, the following is an example of a genus-differentia class definition formulated in SOE in natural language, and then automatically formalized in OWL:

Natural language:

White wine is wine that has specifically white wine color.

OWL:

```
<owl:Class rdf:ID="WhiteWine">
  <owl:intersectionOf rdf:parseType="Collection">
    <owl:Class rdf:about="#Wine" />
    <owl:Restriction>
      <owl:onProperty rdf:resource="#hasColor" />
      <owl:hasValue rdf:resource="#White" />
    </owl:Restriction>
  </owl:intersectionOf>
</owl:Class>
```

The procedure described in 3.2.1 Extraction of Distinctive Features above extracts distinctive features as a specific vocabulary used to describe various individual concepts.

This vocabulary, as we have presented above, falls into two categories:

1. Expression of belonging to a class, for example: "Ahimaaz Ben Paltiel: *Liturgical poet*" (stating that "Ahimaaz Ben Paltiel" is an instance of the class *Liturgical Poet*).
2. Expression of properties specifying this class, for example: "*author of* a family chronicle; *born in* Capua, Italy, 1017; *died about* 1060 *in* Oria" (hereby the construct *author of* is specific for the class *Writer*, which is a super-class of the class *Poet*, which in turn is the super-class of the class *Liturgical Poet*, and the constructs *born in*, *died about* and *died in* are properties of the class *Personality*, which is the super-class of the class *Writer*).

These expressions combined classify the individual *Ahimaaz Ben Paltiel* into the class *Liturgical Poet*, and are therefore distinctive features of that individual. However, these features imply a conceptual model depicted in Table 3 (assuming inheritance of properties by subclasses).

Table 3 Exemplary model of extracted distinctive features

Classes	Properties
Personality	born in died about died in
Writer	author of
Poet	
Liturgical Poet	

As demonstrated in this example, the vocabularies acquired in the previous procedural steps were systematically formalized in the OWL-formatted MOSAICA Core ontology.

The similarity between the natural language formulation of ontological assertions formulated in SOE and the expressions encountered in the actual documentation is striking already at the first glance, and it demonstrates well that when using SOE indeed

a minimal effort is required from the domain experts to formally model the corresponding domain of discourse.

The conceptual model depicted in Table 3 is then automatically converted into OWL by SOE as follows:

```
<owl:Class rdf:ID="personality">
  <rdfs:label xml:lang="EN">Personality</rdfs:label>
</owl:Class>
<owl:Class rdf:ID="writer">
  <rdfs:label xml:lang="EN">Writer</rdfs:label>
  <rdfs:subClassOf rdf:resource="#personality" />
</owl:Class>
<owl:Class rdf:ID="poet">
  <rdfs:label xml:lang="EN">Poet</rdfs:label>
  <rdfs:subClassOf rdf:resource="#writer" />
</owl:Class>
<owl:Class rdf:ID="liturgicPoet">
  <rdfs:label xml:lang="EN">Liturgic Poet</rdfs:label>
  <rdfs:subClassOf rdf:resource="#poet" />
</owl:Class>
<owl:ObjectProperty rdf:ID="bornIn">
  <rdfs:label xml:lang="EN">born in</rdfs:label>
  <rdfs:domain rdf:resource="#personality" />
  <rdfs:range rdf:resource="#location" />
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="diedIn">
  <rdfs:label xml:lang="EN">died in</rdfs:label>
  <rdfs:domain rdf:resource="#personality" />
  <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string" />
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="diedAbout">
  <rdfs:label xml:lang="EN">died about</rdfs:label>
  <rdfs:domain rdf:resource="#personality" />
  <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string" />
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="authorOf">
  <rdfs:label xml:lang="EN">author of</rdfs:label>
  <rdfs:domain rdf:resource="#writer" />
  <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#date" />
</owl:DatatypeProperty
```

It is worth noticing that the approach taken in MOSAICA was to utilize rdfs:label predicate (in conjunction with xml:lang attribute), in order to specify multilingual labels of all OWL elements. Accordingly, these labels are used for the display of the concepts and properties, rather than their actual names.

3.2.3. Outcome: MOSAICA Core Ontology

So far, the conceptual modeling focused on defining the distinctive features in the domain of discourse, and their formalization. Accordingly, a formal system of class concepts and their corresponding properties and mutual relations was developed as MOSAICA Core ontology. The current version of the MOSAICA Core is available at: http://www.straightech.com/OWL/MOSAICA/MOSAICA%20Core.xml.

The methodology applied to this purpose acquired this model from the extensions of the underlying conceptualization. Accordingly, it was not developed a priori and based solely on the knowledge of the domain experts, but rather systematically modeled the way in which the entities in the domain are actually described in the extensions.

Since this modeling was undertaken on the combination of multiple sources containing such descriptions (i.e. the Jewish Encyclopedia Online and the eJewish.info directory), the resulting model unifies the common conceptualization of these resources. Therefore entities described by all contributing resources can be directly associated with the elements of the model. – To that purpose the proprietary annotation property *see also* was deigned.

The MOSAICA Core ontology satisfies the second of the two principles assumed in [Kent 2000]. With that the initial conditions were met, in order to apply the two-step procedure proposed there.

3.3. Identification of Instances

Once the MOSAICA Core was in place, the next step was “lifting”: populating of the ontology with instances from the Jewish Encyclopedia Online and the eJewish Keywords. Generating formal instances is imminent for ontologies that are not populated with instances. This is common for upper-level ontologies, which are supposed to act as global ontologies that are sharable and agreed upon by different communities. Generating these instances is a core issue in [Kalfoglou and Schorlemmer 2002] and [Madhavan et Al. 2002]. Having a mechanism to populate ontologies with instances is an important aid for ontology mapping practitioners, as they can explore a different angle in mapping – to focus on the way local communities classify their instances.

The current procedural step – populating – could be also in a certain sense described as a standard phase in the ontology lifecycle. – This description is nevertheless misleading, and the real effect of this procedural step – in the context of the presented methodological approach – is however significantly different and far-fetching.

In a classical alignment scenario, we would face two independently developed ontologies – each independently populated with corresponding references to individual

entities. Consequently, alignment of classes in these two ontologies would result in merging of the class instances into a unifying set. However, this is not the case at hand. – Here the classes and the relations in the MOSAICA Core ontology were derived directly from the extensions, i.e. from the sets of references to individual entities. Accordingly, by assigning now the extensions to the corresponding classes defined in the unifying conceptual model (i.e. in the MOSAICA Core ontology) these extensions become aligned with that conceptual model. In addition, as soon as the instances from one of those extensions (in practice the eJewish.info Keywords were first processed), the additional extensions are matched to the existing extensions, extending the conceptual alignment to the alignment between the extensions. Accordingly, this strategy could be outlined as follows:

1. Design of the intension of conceptualization (i.e. MOSAICA Core ontology),
2. Alignment of an extension with the intension (identifies instances in the extension as the instances of the class concepts in the intension), and
3. Alignment of the extensions (identifies instances from multiple extensions as identical instances of a class concept in the intension).

Therefore, the current procedural step has in fact a much deeper methodological significance than merely populating ontology with class instances: it factually **aligns multiple resources into a unifying conceptual model**.

Accordingly, the methodological approach proposed in [Kent 2000] was not only adopted, but also adapted to a certain extent, in order to become more adequate to the situation encountered in MOSAICA. The fact that "lifting" is undertaken on the "common ontology" (rather than on the "community ontologies", which in MOSAICA do not exist), already implies the core aspect of "fusion". This is in particular emphasized by the fact that the additional extension (in practice the Jewish Encyclopedia Online) was used not only to populate this "common ontology", but that the instances extracted from it were also automatically identified with the already existing instances (extracted from the eJewish.info Keywords). We shall examine these modifications of the originally proposed methodological approach in greater detail in the section 5 Conclusions below.

The current procedural step was applied to each of the sources in a differing manner. This is due to the fact that the eJewish.info Keywords are a controlled vocabulary with an internal hierarchical structure, while the Jewish Encyclopedia Online is a website consisting of encyclopedic articles. The following is a detailed description of the particular procedures developed for each of the content sources.

3.3.1. Extraction of Instances from the eJewish.info Keywords

The following are procedural steps undertaken in order to extract instances from the eJewish.info Keywords, and correctly classify them under the MOSAICA Core classes:

1. **Mapping of keywords into MOSAICA Core classes**. – All keywords that were either adopted as classes in the MOSAICA Core, or that could be identified with a class in the MOSAICA Core were mapped utilizing the annotation relation *see also*. This procedural step merely extends the procedural step 2 in the section 3.2.1.1 Conversion of eJewish.info Keywords into Classes and Relations above.
2. **Extracting individual concepts from eJewish.info Keywords**. – It was assumed that in many cases the relation "parentKeyword" expresses the class/instance relationship between the keyword that was mapped as a class, and its children keywords. This assumption, however, often does not hold, since the relation "parentKeyword" was used in an associative, rather than in a logically strict and consistent manner, as mentioned earlier. For example, regions of the world were modeled as children of the continents, when in fact both should be considered individual concepts (e.g. "Africa" is not a class, but rather and instance of the class *Continent*). – The major difficulty in this procedural step was found in the inconsistent mixing of class concepts and individual concepts at the same level of conceptualization in the eJewish.info Keywords, as detailed in the point 7 in the section 3.2.1 Extraction of Distinctive Features above.
   In addition, often the relation "parentKeyword" was erroneously used instead of the relation "relatedKeyword". For example, "Ancient Rome" is defined as "parentKeyword" of "Roman Empire".
   Such situations could be solved only by manual inspection of the conversion. Accordingly, the parsing utility application described in detail in the section 4.1 eJewish.info Keywords Parser and Converter below was extended by a mechanism that requests confirmation (and possible correction) by a human user of the automatically performed conversion.
3. **Extraction of assertions from eJewish.info Keywords**. – In this step the relationships formed in the eJewish.info Keywords by using the over-generic relation "relatedKeyword " were systematically converted into their specific equivalents in the MOSAICA Core. The algorithms applied to that purpose used the domain specification of the class to which a particular individual concept was assigned as its instance by following this logic: if a keyword was mapped into a class in the MOSAICA Core, then relations that this keyword could possibly use to form assertions are only those for which its corresponding class is specified as their domain. For example, if the keyword "Poland" was mapped into the class *Country*, which is a subclass of the class *Location*, then all assertions about "Poland", could exclusively use either *is located in* or *is adjacent to* – because these are the only two relations in the MOSAICA Core for which the class *Location* is defined as their domain.

In addition, the range of the relations was also inspected for the following reason: although the number of options into which the over-generic relation "relatedKeyword" could be potentially converted was now restricted only to the relations for which a particular class was defined in the MOSAICA Core as their domain (i.e. to two – *is located in* and *is adjacent to* – in the previous example), this number of options could still be relatively high, and a mechanism was sought that could automatically select only a single – the most suitable – relation. For example, eJewish.info Keywords contain the following relationship: "Sarah Aaronsohn – 'relatedKeyword' – NILI". Hereby, in the previous step the keyword "Sarah Aaronsohn" was converted into an instance of the class *Personality* (more specifically, into an instance of the class *Spy*, which is a subclass of the class *Soldier*, which in turn is a subclass of the class *Personality*), and the keyword "NILI" was converted into an instance of the class *Organization* (more specifically, into an instance of the class *Military Organization*, which is a subclass of the class *Political Organization*, which in turn is a subclass of the class *Organization*). By inspecting the relations in the MOSAICA Core for which the class *Personality* was defined as their domain, only the following options for the conversion of "relatedKeyword" exist: *involved in* (with the range set to the class *Organization*), *is affiliated with* (with the range set to the class *Personality*), *is member of* (with the range set to the class *Family*), *is author of* (with the range set to the class *Source*), *is member of* (with the range set to the class *Community or Group*), *lived in* (with the range set to the class *Location*), *lived in* (with the range set to the class *Historical Period*), *participated in* (with the range set to the class *Historical Event*). However, since the keyword referred to by the relation "relatedKeyword " (i.e. "NILI") was already identified with an instance of the class *Organization* (i.e. *NILI*), then the only applicable relation from the MOSAICA Core could be: *is involved in* (since this is the only relation that has the class *Personality* defined as its domain and the class *Organization* defined as its range). Accordingly, the following assertion could be then automatically formulated: *Sarah Aaronsohn is involved in NILI.*

4. **Conversion of eJewish.info Keywords into MOSAICA Core labels**. – In order to support the multilingual approach MOSAICA Core elements are displayed by utilizing their multilingual labels (rdfs:label tag in conjunction with xml:lang attribute). However, eJewish.info Keywords support a different mechanism: the tag "keywordNonDescriptor" is used in order to provide alternate labels for each keyword. Since all eJewish.info Keywords are in English, the conversion into xml:lang attribute was trivial, but the problem encountered was that OWL supports only a single label (rdfs:label) per xml:lang attribute. Accordingly, only a single label had to be selected from eJewish.info alternative English labels.
To that purpose, additional feature was developed in the parsing utility application, which allowed selection of the preferred English label.
In addition, the parentheses often used in the eJewish.info labels were automatically removed, being considered redundant, and in order to improve the

readability (the information enclosed in parentheses was almost as rule implied in the inclusion relation, and therefore made evident in MOSAICA Core display).

5. **Conversion of remaining eJewish.info keyword labels**. – The labels that were not selected for the conversion into MOSAICA Core labels, where still preserved. The reason for it is explained in more detail in the following section 3.3.2 Conversion and Alignment of the Jewish Encyclopedia Online. This was accomplished by introducing the *expressed as* mechanism explained in more detail in the section 3.4 Consolidation below, which was defined as an annotation property of the type OWL data type property. Such property could be used with classes and instances, and its range is simply a string. In addition, it also supports the xml:lang attribute, which allows it to be used for multilingual purposes.

The conversion of eJewish.info Keywords covered the vast majority of the keywords: from the total of 3546 keywords included in the eJewish.info 2890 (81.5%) were either mapped into the MOSAICA Core ontology classes or extracted as instances of these classes, and only 656 (18.50%) keywords were rejected. Keywords that were extracted as instances are compiled in a separate ontology temporarily named "MOSAICA eJewish", and available at: http://www.straightech.com/OWL/MOSAICA/eJewish.xml. This ontology imports the MOSAICA Core ontology for reference.

3.3.2. Conversion and Alignment of the Jewish Encyclopedia Online

The following are procedural steps undertaken in order to align articles from the Jewish Encyclopedia Online with the instances of MOSAICA Core classes:

1. **Development of the utility application for processing of the Jewish Encyclopedia Online**. – Since the Jewish Encyclopedia Online was not available as an XML document, the initial effort was to develop a crawler capable of downloading and parsing the HTML pages from which the Jewish Encyclopedia Online consists. In addition, the crawler was also required to enable selection of a particular section, or even a single sentence or tag from the HTML document, since only such segment may be considered of a particular interest (for example, for the alignment purposes only the article's title, and the first sentence in the article's body text were considered relevant).
   This utility application was also expected to integrate the natural language processing platform called SANDRA. A detailed description of the developed utility application is provided in the section 4.3 MOSAICA Crawler below. SANDRA provides the Semantic Indexing functionality, which was utilized in the next procedural step.
   Unlike standard keyword indexing designed to generate inverted indices containing every word that occurs in a document, semantic indexing associates documents only with items from a predefined vocabulary ("controlled vocabulary"). Actually, the term "semantic" is justified only if such controlled

vocabulary is provided by ontology, i.e. as references to the formally defined concepts. In that case, such reference is merely a verbal representation of a semantic unit formally defined in the underlying ontology, and thus a technical means to associate (i.e. index) a document with that semantic unit. Rather than searching a document merely by its vocabulary, semantic indexing allows thus identification of a document by its meaning (hence the term "semantic"[24]).

2. **Alignment of instances from the Jewish Encyclopedia Online with the MOSAICA ontologies**. – As mentioned above, all class concepts in the MOSAICA Core ontology, as well as all individual concepts in the MOSAICA eJewish ontology were annotated using the *expressed as* annotation property, described in more detail in the section 3.4 Consolidation below. This property holds a string with various alternative expressions used to identify the corresponding concept. Some of the values of this annotation property were extracted automatically form the multiple labels provided with the eJewish.info Keywords, as explained above, and some were supplemented manually (for the additional class concepts defined in the MOSAICA Core ontology, or to improve the results in the current step).
   These annotations were – among other purposes – designed in order to enable the alignment of articles from the Jewish Encyclopedia Online with the corresponding concepts in the MOSAICA ontologies. The strategy was hereby as follows: individual concepts from the MOSAICA eJewish ontology were matched to the articles in the Jewish Encyclopedia Online utilizing the SANDRA's Semantic Indexing functionality. – Wherever a match was detected, an individual concept was created in a separate ontology (dedicated only to the Jewish Encyclopedia Online). An axiom using the relation owl:sameAs was formulated, binding the newly defined individual concept with the corresponding individual concept in the MOSAICA eJewish ontology.
   To that purpose, all annotations from both ontologies were first converted into a proprietary format used by SANDRA as the extraction patterns. These extraction patterns were next run on the entire body of the Jewish Encyclopedia Online. Hereby, the fact that the Jewish Encyclopedia Online is a well-organized lexical resource that, as mentioned above, maintains a recognizable lexico-syntactic pattern used for narration was exploited in order to define the most effective extraction patterns. This also allowed focusing only on the titles of the Jewish Encyclopedia Online articles for matching them to the instances from the MOSAICA eJewish ontology.
   Hereby, the major difficulty encountered was the morpho-syntactic variance, enhanced by alternative transliterations of Hebrew, Aramaic and Yiddish

[24] Alternative terminology for "semantic indexing" is "semantic annotation" that emphasizes the fact that through association with ontology a document text is enriched (annotated) with the semantic description of its content.

terminology into Latin alphabet. Summary of major types of this kind of difficulties is depicted by examples in Table 4.

Table 4 Morpho-syntactic diversity encountered in the resources

Type	Example
Transliteration of Hebrew into Latin alphabet	Rosh Ha-Shana, Rosh Ha-Shanna, Rosh Ha-Shannah
Translation	Rosh Ha-Shana , New Year
Ambiguity	born in (can be followed by place or year)
Under-specification	Movement (social or physical, etc.)
Over-specification	Sefer Ha-Zohar (usually referred to as only "Zohar")
Acronyms	Rashbi (Hebrew pronunciation of the acronym designating Rabbi Shimon Bar-Yochai)

In order to overcome these difficulties, the following mechanisms were devised:

a. **Transliteration**: the matching used Soundex value, rather than the original string. The Soundex algorithm applied was Daitch-Mokotoff Soundex[25] – a phonetic algorithm invented in 1985 by genealogist Gary Mokotoff, and later improved by Randy Daitch, both of the Jewish Genealogical Society. It is a refinement of the Russell and American Soundex algorithms designed to allow matching of Slavic and Yiddish surnames with similar pronunciation but differences in spelling.
b. **Translation**: translation was provided as annotation value where required.
c. **Ambiguity**: validation specification was provided in annotation value, signaling what kind of value is expected (for example, string or date).
d. **Under-specification**: Under-specified terms (typically used as upper-level classes) were not annotated, and consequently did not participate in the semantic indexing. Instead their more specific subclasses were annotated.
e. **Over-specification**: alternate expressions were provided as annotation values (similar to translations).
f. **Acronyms**: widely accepted acronyms were provided as annotation values (similar to translations and over-specifications).

There is an additional difficulty related to the alignment of individual concepts: the corresponding concept definitions may significantly differ in various conceptualizations. For example, the personality James Joseph Sylvester was associated with literature and law in the eJewish.info Keywords, but correctly[26]

[25] http://en.wikipedia.org/wiki/Daitch-Mokotoff_Soundex
[26] James Joseph Sylvester was a prominent British mathematician with significant contributions to the matrix theory. For a brief period of his life he studied law, and wrote a "mathematically flavored" treatise on poetry.

defined as a mathematician in the Jewish Encyclopedia Online. Wherever the concept definitions were conflicting, the conflicts were resolved manually resulting in the redefinition of the existing concept (for example, James Joseph Sylvester was correctly qualified as a mathematician).

3. **Extraction of additional instances from the Jewish Encyclopedia Online**. – Jewish Encyclopedia Online is a vast body of knowledge, containing over 15,000 articles. Since each article could be converted into at least[27] one instance of the MOSAICA ontologies, many of the articles are not covered by the keywords already converted into individual concepts. Accordingly, extraction patterns derived from the annotations using the relation *expressed as* from the MOSAICA Core ontology were applied, in order to extract additional individual concepts. Hereby, the fact that the Jewish Encyclopedia Online is a well-organized lexical resource that maintains a recognizable lexico-syntactic pattern used for narration was exploited again, allowing processing to focus only on the first sentences in the article's body text.

3.3.3. Outcome: MOSAICA eJewish and MOSAICA JE ontologies

The final result of the current procedural step were two domain ontologies:

1. MOSAICA eJewish ontology (available at: http://www.straughtech.com/OWL/MOSAICA/eJewish.xml), and
2. MOSAICA JE ontology (available at: http://www.straightech.com/OWL/MOSAICA/JE.xml).

These two domain ontologies are comparable to the so-called "community ontologies" in [Kent 2000], with one important distinction: both of these ontologies are populating the same "common ontology" (called "reference ontology" in [Kalfoglou and Schorlemmer 2002]). – In that sense, an important aspect of the "fusion" was accomplished already in this procedural step.

3.4. Consolidation

Consolidation was undertaken at two points in the procedure: 1) after conversion and alignment of the eJewish.info Keywords, and then again 2) after conversion and alignment of the Jewish Encyclopedia Online. In principle, consolidation is envisioned as a standard procedural step after every adoption of an additional semantic or content resource, and its alignment with existing semantic resources.

[27] Some articles are generic, and were split into several separate individual concepts. For example, "Frank, Jacob and the Frankists" was split into the individual concepts *Jacob Frank* (instance of the class *Personality*) and *Frankists* (instance of the class *Movement, Belief or Ideology*).

The major objectives of this procedural step are as follows:

1. Manual inspection of the results obtained through the automatic alignment, and their potential correction (including conflict resolution),
2. Establishing of the requirements for the potential modifications in the MOSAICA Core ontology that may be necessary in order to cater for the additional resources,
3. Verification of the factual correctness of the newly adopted resources,
4. Completing the uniformity of the results in terms of uniform transliteration standards, and uniform translation strategy (primarily translation of the original Hebrew, Aramaic, Ladino and Yiddish terminology),
5. Completion of annotations formed using *expressed as* relation, in order to provide improved capabilities for applying the Semantic Indexing to additional semantic and content resources, and
6. Supplementing ontologies with additional concepts of interest.

Although at both points in the procedure all of the above objectives were targeted, the emphasis was different in each.

3.4.1. Consolidation Following Alignment of eJewish.info Keywords

Since eJewish.info Keywords were the first external resource adopted into MOSAICA, the emphasis was on facilitating adoption of additional resources (primarily the Jewish Encyclopedia Online). Accordingly, the focus was set on the mechanisms designed for that adoption, and alignment of the additional resources.
The objective of the previous procedural step (i.e. identification of instances) when applied to an initial content source is simply to classify the references to individual entities into the correct classes (and to automatically formulate assertions about these entities using the corresponding properties and relations). However, when repeating this procedure on every additional content source, the extended objective becomes: to identify references to the same individual entity that originate from different content sources.

For example, “Moses Ben Maimon” is an article in the Jewish encyclopedia, and “Rabbi Moshe Ben Maimon” is a keyword in the eJewish.info Keywords. However, both refer to the same person (Egyptian philosopher, theologian and physician from the 13th Century). Accordingly, in the core ontology there should be only a single instance of, for example, class *Philosopher* referring to the famous Egyptian philosopher, and pointing to various electronic resources about him. Accordingly, it is necessary to identify between “Moses Ben Maimon” and “Rabbi Moshe Ben Maimon”, in order to avoid creating unnecessary multiple identical instances.

To this purpose the annotation mechanism using the annotation property *expressed as* was devised:

- An annotation property was added to the core ontology defined as:

```
<owl:DatatypeProperty rdf:ID="expressedAs">
  <rdf:type rdf:resource="&owl;AnnotationProperty" />
  <rdfs:label xml:lang="EN">expressed as</rdfs:label>
</owl:DatatypeProperty>
```

- This annotation property holds as its value the actual lexico-syntactic pattern that is used for text processing. Accordingly, rather than utilizing the label of an OWL element, semantic indexing can operate on these values instead.
- Annotation properties in OWL can be applied to both concepts and relations. Accordingly, this annotation property can be for both. As explained above, extraction of noun-phrases could provide concepts when conceptually modeling, while extraction of verb-phrases could provide relations.
- In addition, no restrictions could be set on annotation properties, so that the same annotation property can form multiple axioms with the annotated element. – This enables formulation of alternative terminology in the same language for a single element.

However, in order to apply this mechanism, it is necessary to annotate new instances in the ontology. Partly this was accomplished automatically, by converting labels contained in the eJewish.info Keywords, and partly it was supplemented during the consolidation procedural step (to this purpose SOE was utilized again).

The application of this mechanism undergoes the following steps:

1. **Identification of instances**. – For example, title of an article in the Jewish Encyclopedia Online is identified with a value of an axiom formed by using the relation *expressed as*.
2. **Creation of reference**. – Following such identification, an axiom is formed using the relation *see also*, referring the subject of the annotation to the URI of the article.
3. **Creation of additional annotation**. – In case that the match was not identical but only partial (for example, by using Soundex values, rather than actual strings), an additional annotation using the relation *expressed as* is formulated, using the new terminology (for example, the title of an article).

3.4.2. Consolidation of the Alignment of the Jewish Encyclopedia Online

At this point, the outcomes were the MOSAICA Core ontology, and two domain ontologies (eJewish and JE) aligned to the MOSIACA Core via instantiation mechanism[28], and to each other via referencing mechanism[29].

At this point, the emphasis was on the manual inspection of the alignment results (including the conflict resolution), as well as on the verification of the knowledgebase acquired from the process. Although a tedious manual effort, verification is an important objective of the consolidation, and it was stipulated by a decision of the MOSAICA Board of Editors as a condition for making the semantic resources publicly available. MOSAICA is envisioned as a knowledgebase that will be used as a reference source by many communities, including teachers and students. It is therefore of utmost importance to ensure that MOSAICA provides accurate, and reliable information.

3.4.3. Outcome: MOSAICA Domain Ontologies

During the consolidation, also reorganization of the all domain ontologies was undertaken. First MOSAICA Core ontology was split into two separate ontologies:

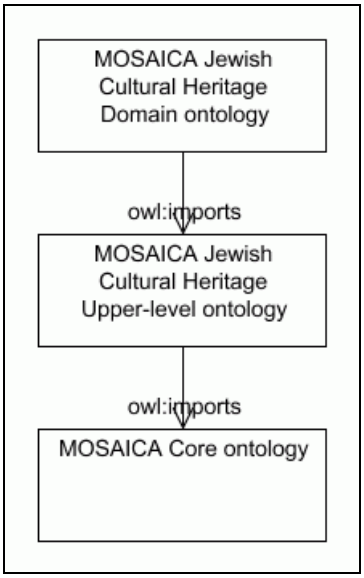
1. **MOSAICA Jewish Cultural Heritage Upper-level ontology**. – This ontology included only classes and concepts from the MOSAICA Core ontology that are specific for the Jewish cultural heritage. The reasoning behind the decision to extract these specific elements into a separate ontology that supplements the MOSAICA Core ontology was as follows: application of MOSAICA to additional areas of cultural heritage could benefit from an upper-level domain ontology that described the cultural heritage in general. However, such generic ontology should not include any of the specific elements. Since some classes and relations in the MOSAICA Core were modeled and are required specifically for the domain of the Jewish cultural heritage, these elements should be separated from the generic ontology into a separate ontology. This MOSAICA Jewish Cultural Heritage ontology is available at:
2. **MOSAICA Core ontology**. – The completely generic elements were left in the MOSAICA Core. It is envisioned that application of MOSAICA Core to additional areas of cultural heritage will result in other upper-level domain ontologies (like MOSAICA Jewish Cultural Heritage ontology) which supplement the MOSAICA Core. MOSAICA core is available at:

[28] OWL applies the rdf:type construct in order to define an individual concept as an instance of a class concept.
[29] Since both of the domain ontologies contain only individual concepts (and their corresponding assertions and annotations), only the owl:sameAs mechanism is applicable.

Next, the MOSAICA eJewish ontology, and the MOSAICA JE ontology were merged into a single ontology titled MOSAICA Jewish Cultural Heritage Domain ontology. The reasoning behind this decision was as follows: since the alignment identified some instances from the MOSAICA JE ontology with instances in the MOSAICA eJewish ontology and vice versa, MOSAICA JE ontology contains only the instances not covered by the MOSAICA eJewish ontology and vice versa. – Accordingly, there is no real justification to keep these ontologies separate, if both of them are incomplete without the other. Moreover, the housekeeping and maintenance of these ontologies is technically easier if they are merged into a single entity.

The combined usage of these ontologies facilitated via owl:imports mechanism is schematically depicted in Figure 4.

Figure 4 Importing mechanism for MOSAICA domain ontologies



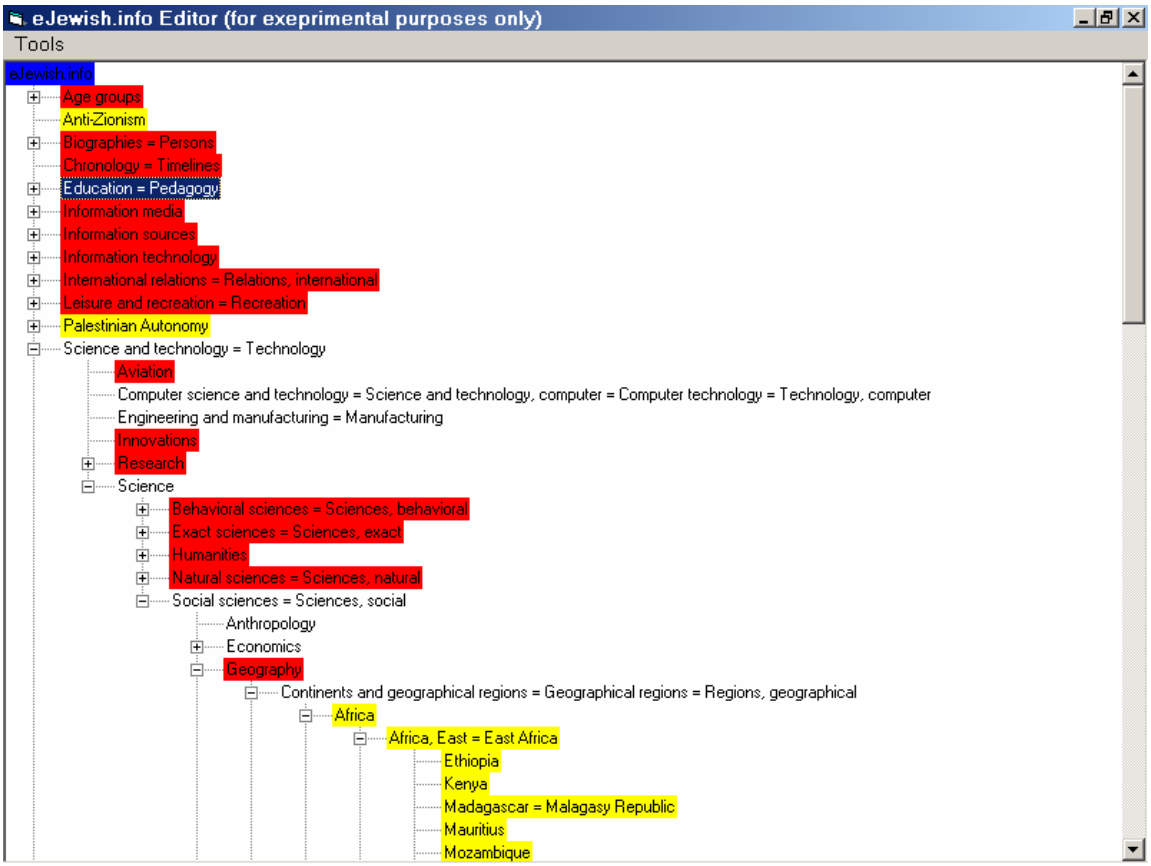
4. MOSAICA IMPLEMENTATION

The following chapters provide description of various utility applications utilized for the implementation of the alignment methodology.

4.1. eJewish.info Keywords Parser and Converter

eJewish.info Keywords Parser was coded as a desktop utility designed specifically to parse the XML-formatted file in which the keywords were provided and display it in a tree structure, as depicted in

Figure 5 eJewish.info Keywords Parser and Converter GUI



(The tree nodes marked by red background were rejected as irrelevant for the MOSAICA Core ontology, while the nodes marked by yellow background were identified with the elements from the MOSAICA Core vocabulary.)

Later, it was extended with a converter of eJewish.info relationships into OWL assertions, and more and more functionalities were gradually added to it.

The following are the basic functionalities that it supports:

1. Parsing of the specific eJewish.info XML schema format (this format was, however, not provided either as DTD or XML Schema),
2. Transformation of this format into an object model that supports ontological concept definitions and ontological relations,
3. Visual display of the object model in a browsable tree structure,
4. Integrated parsing utility for importing OWL-formatted ontologies,
5. Selection of the items in that object model by specifying its logical relationship to the elements from the imported ontology,
6. Extended heuristics for semi-automatic conversion of eJewish.info relationships into formal logical assertions,
7. Heuristics for extraction and selection of concept labels from the eJewish.info Keywords,
8. Generation of annotation axioms of two types: using the *expressed as* property (converting the eJewish.info labels), and using the *see also* property (associating the selected eJewish.info keywords with the concepts in the MOSAICA Core).

eJewish.info Keywords Parser and Converter is a utility application developed specifically for a particular purpose, and besides its use to immediately facilitate handling of the eJewish.info keywords it has no further use. Therefore, it was not released or published on the MOSAICA website.

4.2. SANDRA Search AND Retrieval Application)

SANDRA consists of two structurally independent components that were both used in different procedural steps in the methodology outlined above:

1. Content Engine provides the basic NLP functionalities combined with advanced text indexing functionalities,
2. Categorization Engine provides the functionality of clustering a corpus of documents by extracting their common vocabulary.

Content Engine is implemented as a finite-state transducer that operates on a text stream in four consecutive runs: (1) basic tokenization and basic tagging, (2) morpho-syntactic normalization and part of speech (POS) tagging, (3) chunk parsing and composite tagging, and (4) application of extraction patterns. – Each run operates on the outcome of the previous run.

Accordingly, Content Engine provides the following functionalities:

1. Keyword indexing (combining basic tokenization and morpho-syntactic normalization),
2. Semantic indexing (utilizing application of extraction patterns),
3. Named entity extraction (combining composite tagging and application of extraction patterns), and
4. Generation of the Normalized Document Format (NDF) utilized by the Categorization Engine (combining chunk parsing and composite tagging).

Categorization Engine is a suffix-trie parser and merger that operates on the NDF produced by the Content Engine. It follows the ideas of Zamir et al. in [Zamir et al. 1997] and [Zamir et al. 1999] to utilize suffix-trie for clustering of document corpora based on their common terminology.

Both engines are coded in Common LISP. Content Engine was used as a standalone executable (during the development of the MOSAICA Core ontology), and then also pipelined with the MOSAICA Crawler. Categorization Engine was used as a standalone executable only, since it was applied only in the initial stage, and not used later.

4.3. MOSAICA Crawler and Extractor

Following the exposition of the methodology in the previous chapter, the following functionalities were required from the implementation:

1. Download HTML pages from the WWW,
2. Parse HTML page for either embedded links to other resources, or for segments of text,
3. Provide semantic indexing of text segments utilizing MOSAICA Core ontology,
4. Append selected references to the MOSAICA Core ontology,
5. Provide user interface for manual editing of MOSAICA Core ontology.

The functionalities 3 and 5 above were provided respectively by two external systems that were utilized in MOSAICA:

1. Search AND Retrieval Application (SANDRA), which is a natural language processing platform designed for information retrieval of textual content, and
2. Simple OWL Editor (SOE), which is a desktop utility for display and editing of OWL-formatted ontologies.

The functionalities 1 and 2 above were provided by the MOSAICA Crawler and Extractor that was developed within the project, and that incorporates two subcomponents:

1. Download Manager designed to retrieve HTML pages from the WWW, and
2. HTML Parser designed to parse HTML pages retrieved from the WWW, and extract from them the relevant information.

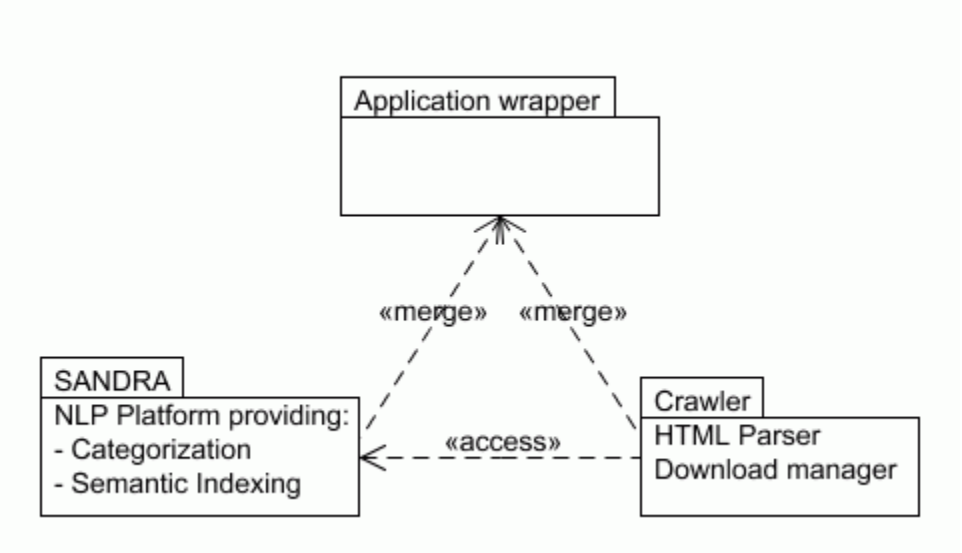
Moreover, SANDRA was wrapped together with that additional component into a utility application, which functions as an independent subsystem within the MOSAICA System. This application wrapper also embeds the functionality 4 above.

The entire system requires access to the MOSAICA Core ontology, which was provided in two different manners:

1. SANDRA originally required OWL-formatted ontologies to be converted into a proprietary format used by SANDRA's finite-state transducer. As a part of the efforts in WP2, a basic OWL parser was developed and integrated within SANDRA, in order to allow it a direct access to the MOSAICA Core ontology.
2. The wrapper application was pipelined with the external Ontology Server module, which provides all required editing functionalities.

Accordingly, the architectural design of this subsystem is schematically depicted in Figure 6.

Figure 6 UML deployment diagram of the MOSAICA conceptual alignment methodology implementation



Crawler binds together two interactive subcomponents:

1. Download Manager is a utility application implementing an HTTP client, and cache facility for management of the retrieving of HTML pages.
2. HTML Parser is an engine designed to parse HTML in order to: 1) extract links to other resources embedded in the page, and 2) extract particular segments of text, such as page title, specific paragraph or specific sentence in a paragraph.

The two subcomponents work together in such manner that Download Manager provides to the Parser HTML pages, Parser extracts links to additional resources of interest, and feeds them back into the Download Manager to retrieve them.

In addition, Parser extracts requested segments of text, passes them to SANDRA for further processing, and passes the retrieved processing results to the Application Wrapper.

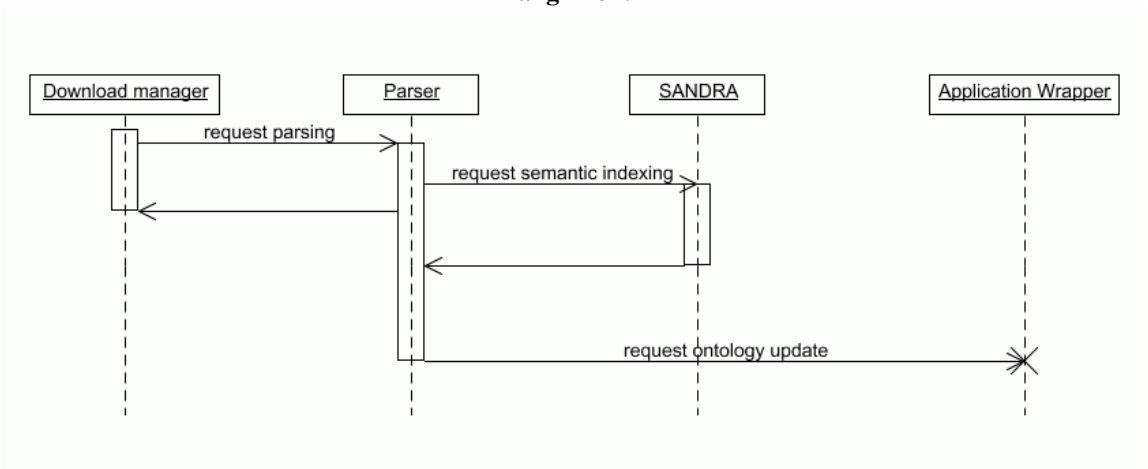
Crawler is coded in Java, and pipelined with SANDRA.

Application Wrapper binds all components into a single application, and provides programming interface for configuring all other components.

In addition, Application Wrapper interfaces the Ontology Server holding MOSAICA Core ontology, and uses it to edit it (append individual instances to the classes in it). Application Wrapper is coded in Java, and interfaces Java-based Ontology Server.

The workflows supported by the MOSAICA Crawler and Extractor are schematically depicted in Figure 7.

Figure 7 UML sequence diagram depicting internal workflows in the subsystem for conceptual alignment



Besides the initial procedural step in which the bootstrapping vocabulary is acquired by using only SANDRA as a standalone application, all additional procedural steps proceed in the following manner:

1. Download Manager retrieves the requested HTML page and passes it to Parser.
2. Parser extract requested information. If this information is links to additional resources, it is retrieved to the Download Manager. If this information is text segments, they are passed to SANDRA for processing.

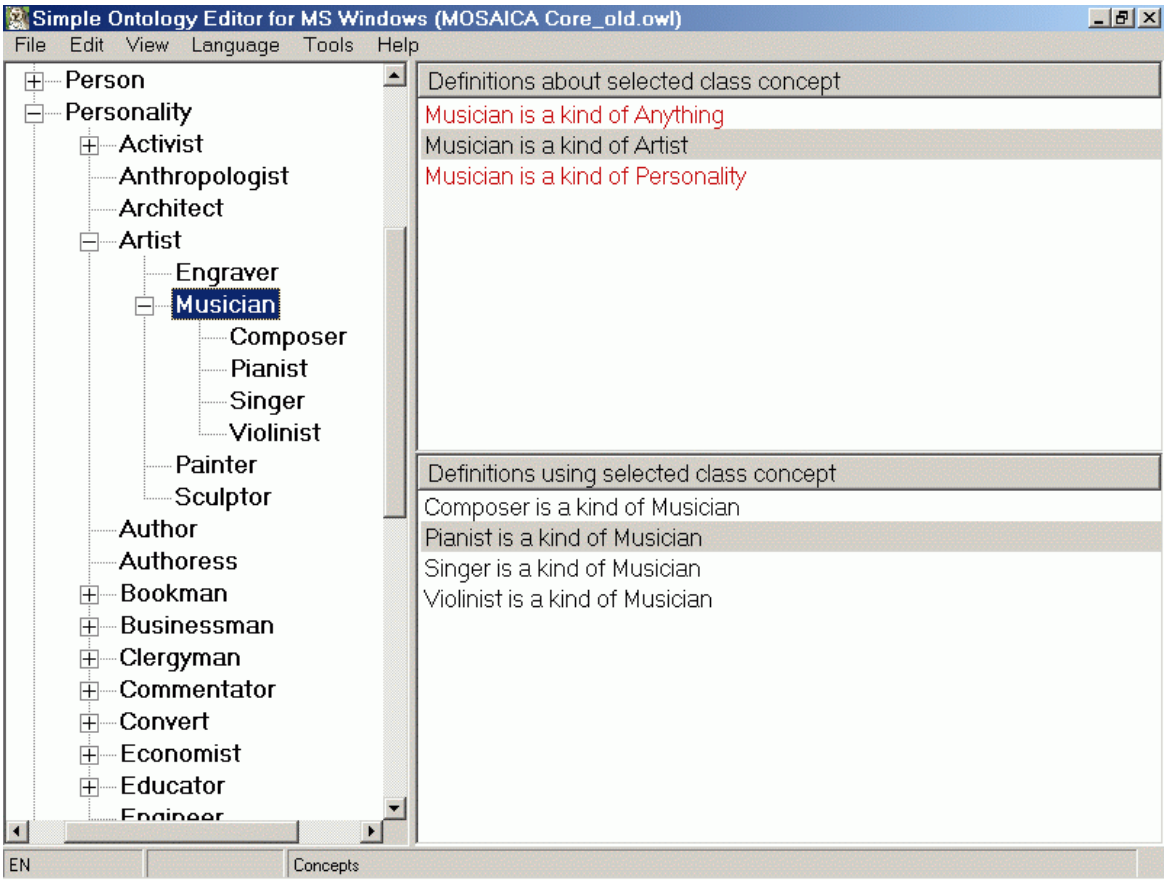
3. SANDRA returns processing results to the Parser, which passes them to the Application Wrapper.
4. Application Wrapper invokes Ontology Server to store the processing results.

4.4. Simple OWL Editor (SOE)

SOE provides a user-friendly graphic interface to OWL-formatted ontologies by converting logical assertions and other specifications formalized in OWL into (almost) natural language.

A screenshot of SOE displaying a part of MOSAICA Core ontology is depicted in Figure 8.

Figure 8 MOSAICA Core ontology (segment) displayed in SOE



Owl-formatted ontologies are converted into natural language by following these principles:

1. Instead of concept and property names their multilingual labels are used for representing them.
2. Logical operators used in OWL (e.g., set intersection and union, equivalence, inclusion, complementarity, etc.) are transformed into their natural language equivalents.
3. Complex OWL constructs (such as restrictions on property, enumerations, data ranges, etc.) are transformed into typical natural language definitions (such as genus-differentia definition, enumeration definition, etc.).
4. Editing is performed as formulation of natural language statements about the domain of discourse.
5. Validation is performed while editing, ensuring syntactic and semantic completeness and consistency of the developed ontologies.
6. Complete interoperability with other Semantic Web applications is ensured by providing only OWL input and output.

In order to provide extensive multilingual features, SOE embeds a translation utility allowing effortless translation of concept and property labels into multiple languages.

SOE is integrated with a proprietary logical reasoner that allows logical inference to be applied to the developed ontologies. The inferred logical assertions and specifications can be then integrated into the developed ontology, in order to ensure its completeness.

SOE is coded in VB6 providing a familiar MS Windows GUI. It was utilized as a standalone desktop application.

5. CONCLUSIONS

The methodology for ontology alignment developed and applied in MOSAICA was based on the theoretical framework proposed in [Kent 2000]. It also drew significantly from the implementation experiences provided in [Kalfoglou and Schorlemmer 2002] and [Stumme and Maedche 2001]. However, the implementation of the adopted theoretical framework was different, and we shall argue also more effective in particular circumstances.

The differences between the originally proposed methodological approach and its implementation in MOSAICA are summarized in Table 5.

Table 5 MOSAICA methodology and theoretical framework proposed in [Kent 2000]

<table>
<tr><th>MOSAICA</th><th>Two-step method [Kent 2000]</th></tr>
<tr><td>Identification of instances</td><td>Lifting</td></tr>
<tr><td rowspan="2">Consolidation</td><td rowspan="2">Fusion</td></tr>
</table>

These differences could be detailed as follows:

1. Lifting is envisioned as populating two “community ontologies” with “community instances”, as well as populating of the “common ontology” with the “formal instances”. Identification of instances, in contrast, is populating of the “common ontology” with the “community instances”.
2. Fusion is envisioned as merging of the “community ontologies” modulo “common ontology”. Identification of instances implied cross-referencing of the “community instances” and establishing of the identity relationship between corresponding “community instances” – which is an essential aspect of fusion.
3. Consolidation supplements the identification of instances, aiming at completing the fusion.

These modifications were primarily necessitated by the fact that the “community ontologies” did not exist in MOSAICA as intensions of conceptualization, but only as extensions. Consequently, the decision was taken to avoid the unnecessary effort to perform an “academic exercise” of creating two independent “community ontologies” “from scratch”, just for the sake of following the proposed methodological approach “to the letter”. Instead, both extensions of conceptualization were utilized to populate the “common ontology”, which automatically resulted in their fusion. It is worth noticing that every additional extension could be merged in the same way, extending the fusion to additional semantic resources.

In this sense, identification of instances is actually a methodological "shortcut" merging two procedural steps into a single step. This is – at least – not at odds with the original methodological proposal. Lifting is a necessary methodological step performed to "enrich" the information in the "community ontologies" to the level that enables their fusion. This is necessitated – as we argued at the opening of the section 3.1 General Approach above – by the typical definition of the class equivalence encountered in Description Logics, such as the one given in [Smith et al. 2004]: equivalence of classes implies their identical extensions (i.e. sets of instances). Accordingly, equivalence could be established only by inspecting and comparing the extensions of the classes, which requires them to be populated. However, inspection and comparison of class extensions is exactly what is undertaken by the identification of instances. Consequently, identification of instances merges lifting and fusion into a single procedural step, and provides a "shortcut" in the situation where no explicit "community ontologies" are available.

Another way of viewing identification of instances is merely as populating the "common ontology". "Common ontology" is a conceptual model common to a domain of interest shared by various communities who provide different interpretations of the shared domain. Accordingly, the effect of populating it would amount to exploration of different ways by which communities interpret and cover the domain of discourse. Nevertheless, the practical result would also be the alignment of these various interpretations in a common, unifying conceptual model – which is exactly what was aimed at by devising the identification of instances as a methodological procedure.

The most critical aspect of the identification of instances is certainly the heuristics employed in the identification. Halevy in [Halevy 2005] provides an extensive list of classes of heuristics that have been used for database schema matching – all of them applicable – mutatis mutandis – also for ontology alignment:

- **Schema element names**. - Element names (i.e. table and field names in databases, concept and property names in ontologies) sometimes carry information about their intended semantics (element names could be in ontologies extended to the corresponding labels, if they exist). The challenges involved in using this heuristics are that the use of synonyms, hypernyms, homonyms, abbreviations, acronyms and concatenations. In addition, in ontologies, which allow use of labels, the names are typically abstract (for example, GUIDs).
- **Data types**. - Schema elements that map to each other are likely to have compatible data types, but this is certainly not a rule. However, in many schemas the data types are underspecified (e.g., CDATA for XML). Moreovre, some data types could be decompositions of another data type (for example, in XML Schema either the data type Date could be used, or a date could be decomposed

using data types gMonth, gDay and gYear). In practice, considering data types is a useful heuristic for ruling out certain match candidates.

- **Data instances**. - Elements from two schemas that match each other often have similar data values (in databases), or similar instances (in ontologies). Similarities can arise in several ways: (1) values drawn from the same small domain, e.g., makes of cars or names of countries, (2) significant occurrences of the same values (e.g., superlatives describing houses for sale), or (3) patterns of values (e.g., phone numbers, price ranges). Data instances are extremely useful when available, but one cannot rely on their availability.
- **Schema structure**. - Matching elements in a schema are typically related to other related schema elements. For example, in an object-oriented hierarchy, it is often the case that if two classes match each other, then the children of these classes will also (at least partially) match. In XML DTDs proximity of attributes in a DTD (e.g., a Phone field next to Agent may suggest that it is the phone of the agent). However, relying on such a heuristic can be very brittle, and one of the main challenges is to find an initial match that drives the similarity of its neighbors.
- **Integrity constraints**. - Considering integrity constraints (for example, cardinality) on single attributes or across attributes can be useful for generating matches. For example, if two attributes are known to be keys in their respective schemas, then that provides additional evidence for their similarity.

However, not all of these strategies for schema matching are applicable to the identification of instances. Heuristics based on data types is meaningless, because instances are individual concepts, and do not have a data type specified. Heuristics based on the data instances is also meaningless, because what is matched in this step are instances, rather than classes from different schemas. Integrity constraints could be viewed in ontologies only as restrictions on property, which are set on classes, and makes the heuristics based on the integrity constraints also irrelevant for the identification of instances. In general, integrity constraints could be considered in ontologies as an aspect of the schema structure.

Accordingly, the following heuristics was employed for identification of instances in MOSAICA:

1. **Name and label matching**. – To this purpose the natural language processing was applied, including very complex procedure of text normalization (including application of a suitable Soundex algorithm), and application of extraction patterns. The fact that the matched instances were mostly either personal names or traditional terminology (transliterated from Hebrew, Aramaic, Ladino and Yiddish) removed significantly the problems related to synonyms, hypernyms, and homonyms, and thus ensured a relatively high accuracy.
2. **Schema structure**. – The domain and range specifications of relations were utilized in order to select the most adequate relation for a certain pair of

instances. – This strategy was used for extraction of assertions about instances, and it was applied after the extraction of instances.

The entire methodological procedure was designed to minimize the manual effort through automation. Ontology alignment has to be automatic in order to be practical. However, a high level of automation in text processing usually has the price tag of low-quality results. Our intention was to reconcile both requirements – automation and high-quality mappings – by adopting semi-automatic approaches. However, the non-automated part of the methodological procedure remains laborious and dependent on human experts (involving both ontology engineers and domain experts).

Nevertheless, a significant improvement was accomplished as advancing through the procedure: the most laborious step was certainly development of the MOSAICA Core ontology, where automation was pretty much marginal. Identification of instances from the first extension (eJewish.info Keywords) was already significantly less laborious, and mostly involved manual configuration of the extraction algorithms and inspection of the results. The identification of instances from the next extension (Jewish Encyclopedia Online) was already performed automatically by a component integrated into the MOSAICA System (i.e. not by a utility application developed to that purpose), and the manual effort was limited to the inspection and manual correction of the outcome.

We expect that the integration of additional semantic resources in the future will further minimize this manual involvement. These expectations are based on the following:

1. The quality of the extraction patterns used in the natural language processing increases with each additional source, which provides additional information on natural language variances that could be expected in the sources, such as transliteration, morpho-syntactic variances, synonymy and homonymy, etc.
2. With each additional semantic resource the MOSAICA Core ontology is enriched with additional relation definitions and their corresponding domain and range specifications. This schema structure enrichment provides an improved accuracy when applying the heuristics based on it.
3. Imprecision is higher for the classification of instances under the classes, than for identification of new instances with the existing instances. With each additional resource integrated that knowledgebase is increased, and fewer instances need to be classified, while more of them need to be merely identified with the existing, and already classified instances.

Following this reasoning based on the experience, we expect the semantic resources that will be integrated in the future to bring the entire procedure to the point where the only manual effort required is the approval of integrated ontology by the Board of Editors.

6. REFERENCES

[Amsler 1980] R. A. Amsler (1980). The structure of the Merriam Webster Pocket Dictionary, Ph.D. diss., University of Texas, Austin, Addison-Wesley Publishing Company, Inc.

[Barwise and Seligman 1997] Barwise, J and Seligman, J, 1997, Information Flow: The Logic of Distributed Systems Cambridge University Press.

[Batini et al. 1986] C. Batini, M. Lenzerini, S. B. Navathe (1986). A Comparative Analysis of Methodologies for Database Schema Integration. ACM Computing Surveys, 1986.

[Bechhofer et al. 2004] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, L. A. Stein, 2004. "OWL Web Ontology Language Reference - W3C Recommendation 10 February 2004". http://www.w3.org/TR/2004/REC-owl-ref-20040210/

[Bench-Capon and Malcolm 1999] Bench-Capon, T and Malcolm, G, 1999, "Formalising ontologies and their relations" Proceedings of the 16th International Conference on Database and Expert Systems Applications (DEXA'99) 250–259.

[Buchanan et al. 1993] B. G. Buchanan and D. C. Wilkins (ed.) (1993). Readings in Knowledge Acquisition and Learning: Automating the Construction and Improvement of Expert Systems, Morgan Kaufmann, San Mateo, CA., 1993.

[Calzolari 1984] N. Calzolari (1984). Detecting patterns in a lexical data base, Proceedings of COLING84, 170-173.

[Calzolari 1988] N. Calzolari (1988). Acquisition of semantic information from an on-line dictionary, Proceedings of COLING88, 87-92.

[Chodorow et al. 1985] M. S. Chodorow, R.J. Byrd, G.E. Heidorn (1985). Extracting semantic hierarchies from a large online dictionary, Proceedings of the 23rd Annual Meeting of the ACL, 299-304.

[Daille 1996] B. Daille (1996). Study and implementation of combined techniques for automatic extraction of terminology, in The Balancing Act: Combining Symbolic and Statistical Approaches to Language, eds., J. L. Klavans and P. Resnik, MIT Press, Cambridge, (1996).

[Doan and Halevy 2005] A. Doan and A. Halevy (2005). Semantic Integration Research in the Database Community: A Brief Survey. AI Magazine, Special Issue on Semantic Integration, 2005.

[Dolan et al. 1993] W. Dolan, L. Vanderwende and S. D. Richardson (1993). Automatically Deriving Structured Knowledge Bases from Online Dictionaries, PACLING 93, Pacific Assoc. for Computational Linguistics, 1993, pp.5-14.

[Doan et Al. 2002] Doan, A, Madhavan, J, Domingos, P and Halevy, A, 2002, "Learning to map between ontologies on the semantic web". In Proceedings of the 11th International World Wide Web Conference (WWW 2002).

[Dori et al. 2004] D. Dori, N. Korda, A. Soffer, S. Cohen (2004). SMART: System Model Acquisition from Requirements Text, in J. Desel, B. Pernici, and M. Weske (Eds.): BPM 2004, LNCS 3080, pp. 179 194, 2004. Springer Verlag Berlin Heidelberg 2004.

[Frege 1892] F.L.G. Frege (1892). Über Sinn und Bedeutung. Zeitschrift für Philosophie und philosophische Kritik 100, 25-50.

[Gal et al. 2005] A. Gal, G. A. Modica, H. M. Jamil, A. Eyal (2005). Automatic Ontology Matching using Application Semantics. AI Magazine, 26(1),21-32, 2005.

[Ganter and Wille 1999] Ganter, B and Wille, R, 1999, Formal Concept Analysis: Mathematical Foundations Springer.

[Gruber 1993] T. R. Gruber (1993). A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2), 199-220.

[Guarino and Giaretta 1995] N. Guarino, P. Giaretta (1995). Ontologies and Knowledge Bases: Towards a Terminological Clarification. N. J. I. Mars (ed.), Towards Very Large Knowledge Bases, IOS Press 1995, pp. 25-32.

[Halevy 2005] A. Halevy (2005). Why Your Data Don't Mix. ACM Queue, 2005.

[Hearst 1992] M. A. Hearst (1992). Automatic acquisition of hyponyms from large text corpora, Proceedings of COLING-92, Nantes, France, 1992.

[Kalfoglou and Schorlemmer 2002] Kalfoglou, Y and Schorlemmer, M, 2002, "Information-flow-based ontology mapping" in On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE Lecture Notes in Computer Science 2519, Springer. Pages 1132–1151.

[Kalfoglou and Schorlemmer 2003] Y. Kalfoglou and M. Schorlemmer: (2003). Ontology Mapping: The State of the Art. The Knowledge Engineering Review Journal, 2003.

[Kashyap and Sheth 1996] V. Kashyap, A. Sheth (1996). Semantic and schematic similarities between database objects: a context-based approach. The VLDB Journal, 1996.

[Kent 2000] Kent, R, 2000, "The information flow foundation for conceptual knowledge organization". In Proceedings of the 6th International Conference of the International Society for Knowledge Organization (ISKO).

[Kiryakov et Al. 2001] Kiryakov, A, Simov, K and Dimitrov, M, 2001 "OntoMap: portal for upper-level ontologies". Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS'01) 47–58.

[Klavans et al. 2002] J.L. Klavans, P. Davis, S. Popper (2002). Building Large Ontologies Using Web Crawling and Glossary Analysis Techniques, Proceedings of dg.o 2002.

[Lacher and Groh 2001] Lacher, M and Groh, G, 2001, "Facilitating the exchange of explicit knowledge through ontology mappings". In Proceedings of the 14th International FLAIRS Conference.

[Larson et al. 1989] J. A. Larson, S. B. Navathe, R. Elmasri (1989). A Theory of Attributed Equivalence in Databases with Application to Schema Integration. IEEE Transactions on Software Engineering, 1989.

[Madhavan et Al. 2002] Madhavan, J, Bernstein, PA, Domingos, P and Halevy, A, 2002, "Representing and reasoning about mappings between domain models" Proceedings of the 18th National Conference on Artificial Intelligence (AAAI'02).

[Markowitz et al. 1986] J. Markowitz, T. Ahlswede, M. Evans (1986). Semantically significant patterns in dictionary definitions, Proceedings of the 24th Annual Meeting of the ACL, 112-119.

[McGuinness et Al. 2000] McGuinness, D, Fikes, R, Rice, J and Wilder, S, 2000, "An environment for merging and testing large ontologies". In Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR-2000).

[Morin 1999] E. Morin (1999). Automatic acquisition of semantic relations between terms from technical corpora, Proc. of the Fifth International Congress on Terminology and Knowledge Engineering (TKE-99), Innsbruck, Austria, 1999.

[Muresan et al. 2003] S. Muresan, S. D. Popper, P. T. Davis, J. L. Klavans (2003). Building a Terminological Database from Heterogeneous Definitional Sources, National Conference on Digital Government Research (dg.o), 2003.

[Noy 2004] N. Noy (2004). Semantic Integration: A Survey of Ontology-based Approaches. Sigmod Record, Special Issue on Semantic Integration, 2004.

[Noy and Klein 2002] Noy, NF and Klein, M, 2002, "Ontology evolution: not the same as schema evolution" Also as: Smi-2002–0926, University of Stanford, Stanford Medical Informatics. Knowledge and Information Systems.

[Noy and Musen 1999] Noy, NF and Musen, M, 1999, "SMART: automated support for ontology merging and alignment" Proceedings of the 12th Workshop on Knowledge Acquisition, Modelling and Management (KAW'99).

[Noy and Musen 2000] Noy, NF and Musen, M, 2000, "PROMPT: algorithm and tool for automated ontology merging and alignment" Proceedings of the 17th National Conference on Artificial Intelligence (AAAI'00).

[Noy and Musen 2002] Noy, NF and Musen, M, 2002, "PROMPTDIFF: a fixed-point algorithm for comparing ontology versions" Proceedings of the 18th National Conference on Artificial Intelligence (AAAI'02) 744–751.

[Rahm and Bernstein 2001] E. Rahm and P. Bernstein (2001). A Survey of Approaches to Automatic Schema Matching. The VLDB Journal, 2001.

[Riloff and Jones 1999] E. Riloff and R. Jones (1999). Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping, Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99) , 1999, pp. 474-479.

[Schorlemmer 2002] Schorlemmer, M, 2002, "Duality in knowledge sharing" Proceedings of the Seventh International Symposium on Artificial Intelligence and Mathematics.

[Shvaiko and Euzenat 2005] P. Shvaiko, J. Euzenat (2005). A Survey of Schema-based Matching Approaches. Journal on Data Semantics, 2005.

[Smadja 1993] Smadja, F. (1993). Retrieving collocations from text: Xtract, Computational Linguistics, 19(1), 143-177.

[Smith et al. 2004] M. K. Smith, C. Welty, D. L. McGuinness (2004). OWL Web Ontology Language Guide, W3C Recommendation, 10 February 2004. http://www.w3.org/TR/owl-guide

[Stumme and Maedche 2001] Stumme, G and Maedche, A, 2001, "Ontology merging for federated ontologies on the semantic web" Proceedings of the International Workshop for Foundations of Models for Information Integration (FMII-2001).

[Zamir et al. 1997] O. Zamir, O. Etzioni, O. Madani, and R. M. Karp (1997). Fast and intuitive clustering of Web documents, Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, pp 287-290, 1997.

[Zamir et al. 1999] O. Zamir and O. Etzioni (1999). Grouper: a dynamic clustering interface to Web search results, Proceedings of the Eighth International World Wide Web Conference, Toronto, Canada, May 1999. Page 8.