# TAO Methodology: Transition of Legacy Systems to Semantic Enabled Application

Hai H. Wang[1] Terry Payneinst[1] Nick Gibbins[1]
Danica Damljanovic[2] Kalina Bontcheva[2] Ahmed Saleh[1]

[1] University of Southampton, {hw, trp, nmg, amms}@ecs.soton.ac.uk
[2] University of Sheffield, {D.Damljanovic, bontcheva}@dcs.shef.ac.uk

**Abstract.** Despite expectations being high, the industrial take-up of Semantic Web technologies in developing services and applications has been slower than expected. One of the main reasons is that many systems have been developed without considering the potential of the web in integrating services and sharing resources. Without a systematic methodology and proper tool support, the migration from legacy systems to Semantic Web Service-based systems can be a very tedious and expensive process, which carries a definite risk of failure. There is an urgent need to provide strategies which allow the migration of legacy systems to Semantic Web Services platforms, and also tools to support such a strategy. In this paper we propose a methodology for transitioning these applications to Semantic Web Services by taking the advantage of rigorous mathematical methods. Our methodology allows users to migrate their applications to Semantic Web Services platform automatically or semi-automatically.

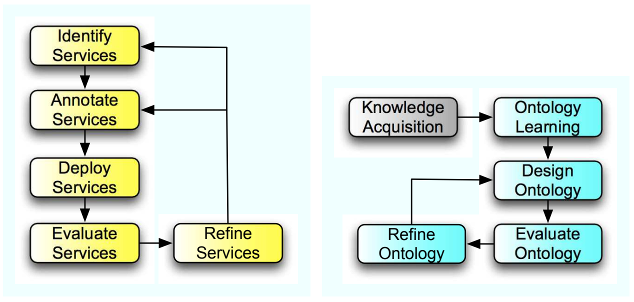
## 1 Introduction

Semantic Web Services combines the Web Services and Semantic Web enabling technologies. By semantically annotating the relevant aspects of declarative Web Service descriptions in a machine-readable format that can facilitate logical reasoning, such service descriptions become interpretable based on their meanings, rather than simply on a symbolic representation. The advantage of this is that many of the tasks involved in using Web Services can be (semi-)automated, for example: discovery, selection, composition, mediation, execution, monitoring, etc. Thus, Semantic Web Service Research [15] has been recognized as one of the most promising technologies to emerge, exhibiting huge commercial potential, and attracting significant attention from both industry and the research community. Despite its great prospect of success, the industrial take-up of Semantic Web Services technologies has been slower than expected. This is mainly due to the fact that many legacy systems have been developed without considering the potential of the Web for integrating services and sharing resources. The migration of legacy systems into semantically web-enabled environments involves many recursive operations that have to be executed with rigor due to the magnitude of the investment in systems, and the technical complexity inherent in such projects. In this context, there are three main issues to be considered, namely: 1) Web Accessibility dealing with the transformation of components of the legacy system that are exposed as Web services,

2) Service Transformation where the exposed Web services are mapped to the corresponding Semantic Web Service representations and 3) Semantic Annotation where the Semantic Web Service is annotated using the relevant domain ontology. Without a systematic methodology and proper tool support, the migration from legacy systems to semantically enabled applications could be a very tedious and expensive process, which carries a definite risk of failure. There is an urgent need to therefore provide strategies that support the construction of ontologies which facility the migration of legacy systems to Semantic Web Services platforms, and also tools to support such a strategy.

This paper proposes a new methodology for addressing the above issues in particular Web Accessibility and Componentization, which in turn could lead to an automatic Platform Transformation. The main idea of the methodology is to identify the components/steps for creating web services to represent system functionality and semantically annotate such services through domain ontologies elicited from system documentations. Typically, the step for creating web services will be merged with the procedures of learning ontologies (from system documentation) to facilitate future ontology pruning and refinement of web service descriptions which eventually leads to bridging the gab of interoperability and hence moving the system closer to SOAs. This methodology is part of Transitioning Applications to Ontologies (TAO) project[1]. TAO is a project in the European Sixth Framework Program. The goal of the TAO project is to define methods and tools for transition of legacy information systems to semantic enabled services, enabling semantic interoperability between heterogeneous data resources and distributed applications.

The remainder of this paper is organized as follows. Section 2 provides a high level explanation of the methodology. Section 3 presents cookbook-style guidelines on how to adopt the methodology using the tools developed by TAO. Section 4 discusses the evaluation of the methodology and tools. Finally, Section 5 and 6 present the related work, conclusions of this paper and future work.



(a) SOA design lifecycle (b) Ontology design lifecycle

**Fig. 1.** SOA and ontology design lifecycle

[1] `http://www.tao-project.eu/`

## 2 High-level methodology

The transitioning methodology outlined in this section is presented as a high-level composite lifecycle, which highlights the interactions between existing methodologies for developing Service-Oriented Architectures (which include Web Services and Enterprise Architectures), and for ontology design. In this paper, we develop this abstracted lifecycle sketches of SOAs and building domain ontologies from text, and demonstrate how and where these should be linked. The design lifecycle for a service-oriented system is largely divorced from both the specific methodology used to create the system, and from the lifecycle of the individual services within such a system. Figure 1(a) shows a sketch of the design lifecycle for an SOA system.

**Service Identification:** This process refers both to the identification of existing services which can be repackaged within an SOA system, and also to the identification of required functionality (from a business process modeling exercise, for example) that does not currently exist in operational form, and the subsequent implementation of such functionality as services.

**Service Annotation:** In order to facilitate loose coupling of component services through brokerage and matchmaking, it is necessary to describe the services in an SOA system in sufficient detail that a service requester can find an appropriate service that meets their needs.

**Service Deployment:** Here we refer to the deployment of services within a service execution environment. This may include publishing/advertising services through public service registries, discovering services where by clients can identify candidate services that may fulfill their requirements, selecting services where they chose the most appropriate service(s), composing services where clients integrate several independent services to achieve an overall goal, and finally executing services either through direct invocation or through workflow management systems.

**Service Evaluation:** This process refers to the ongoing monitoring of an SOA system to determine whether it meets its design goals.

**Service Refinement:** The refinement of an SOA system typically takes one of three forms: the introduction of new functionalities through the creation of new services; the refactorisation of existing service functionality (through aggregation or further decomposition, for example); or the refinement of the service descriptions to better facilitate service matchmaking and brokerage.

As the design lifecycle for an SOA system is largely independent of the particular methodology used, the design lifecycle for a domain ontology (Figure 1(b)) can be separated from the specific knowledge acquisition and modeling methodology used. Domain ontologies usually describe the conceptualization of entities, relations between them, instantiations and the axioms related to a specific domain (such as wines or cars). These domain ontologies can either specialise concepts introduced in some other top-level ontologies (which describe very general concepts like space, time, event, which are independent of a particular problem or domain), such as $Dolce$[2] and $OpenCyc$[3] or

[2] `http://www.loa-cnr.it/DOLCE.html`

[3] `http://www.opencyc.org`

they can be created from scratch for a particular domain. A common process to create domain ontologies from scratch can contain several steps:

**Ontology learning:** *Ontology learning* refers to the use of techniques for automatically or semi-automatically extracting ontologies from existing document corpora. As such, the output from an ontology learning process should not be considered as the finished product, but as a first cut that is solidly grounded in the available documentation, and which will inform the later design of a more polished ontology for production use.

**Ontology Design:** The ontology design process refers to the process of formally codifying the knowledge that has either been manually acquired from a domain expert, or (semi-)automatically extracted from a document corpus. This process may also encompass the identification and reuse of appropriate components within pre-existing ontologies, the alignments of the designed ontology with pre-existing ontologies, or the modularisation of the ontology to facilitate such alignment in future.

**Ontology Evaluation:** The ontology evaluation process assesses whether or not the designed ontology is fit for purpose.

**Ontology Refinement:** This process refers to the refactorisation of the designed ontology to better represent the problem domain.

The TAO transitioning methodology provides a logical approach for connecting the above lifecycles (i.e. SOA and Ontology design) through the following three main points.

**Learning ontologies from service descriptions:** In the ontology design lifecycle, the Ontology Learning process attempts to automatically or semi-automatically derive a knowledge model from a document corpus. In our Transitioning Methodology, we have refined this to reflect the contribution made by the structured (but not ontologically-informed) description of an existing body of services (for example, service APIs and developer documentation, SOA design documentation, and so on). We call this refinement *Service-Oriented Ontology Learning*. It is our expectation that the ontology resulting from an automated ontology learning process should be treated as a candidate ontology which will be subsequently be evaluated and refined in the Ontology Design process. Whilst the ontology extraction process may yield some conceptualization of the relevant domain, much of the implicit domain knowledge inherent in a service description will not be captured; thus the extraction of an ontology from structured sources such as those mentioned above may not obviate the need for further work on the domain ontology. However, such structured sources relating to existing services do provide sufficient initial information that suggests the creation of a domain ontology, which can be further evolved.

**Using domain ontologies to augment semantic content and service:** The Service Annotation process described in the SOA methodology refers to the description of services at the signature level in languages like WSDL[4]. While these allow rudimentary service matchmaking and brokerage on the basis of the types of the inputs

[4] The signature within WSMO [16] describes a functional or atomic interface with inputs and outputs (defined as messanges).

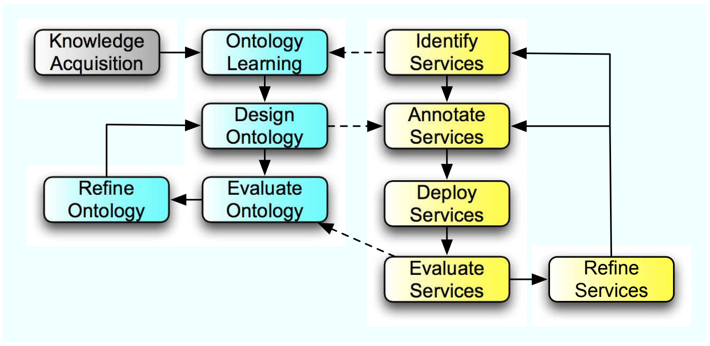


**Fig. 2.** High-level methodology.

and outputs of a service, these types are typically datatypes based on XML Schema, rather than richer knowledge-based types taken from an ontological characterisation of the domain. Thus, these interfaces need to be mapped to equivalent concepts within Semantic Web frameworks (such as OWL-S, WSMO or WS-WSDL) and annotated using the relevant domain ontology.

**Using feedback from service evaluation to refine ontologies:** Both the ontology design lifecycle and the SOA lifecycle contain evaluate-and-refine steps that represent a reflection on the performance of a system and the subsequent reengineering.

Figure 2 depicts an overview of the methodology and a representation of these steps. The interactions between the components of the two life cycles are effectively a refinement of processes within those lifecycles, and reflect the relationship between the products of each individual lifecycle. In particular, we note the task-oriented nature of a domain ontology which is defined with service annotation in mind when compared with a general-purpose ontology for the same domain.

## 3 Methodology cookbook, tool support and case study

The methodology presented in the previous section provides an abstract view about the important phases needed to be performed during the transition process. In order to support this methodology, TAO project has developed an open source infrastructure and a series of tools to aid the transitioning process. In this section, we present a cookbook style guideline about the usage of TAO tools. To better illustrate the idea, we use the transition of GATE system as a case study. GATE[5] is a leading open-source architecture and infrastructure for the building and deployment of *Human Language Technology* applications, used by thousands of users at hundreds of sites. GATE exhibits all the specific problems that large software architectures encounter, which enables us to evaluate the methodology and tools intensively . [8] presents the detailed view of the advantages and possibilities arising from building domain ontology and application of semantic enrichment of software artifacts in the GATE case study.

[5] `http://gate.ac.uk`

## 3.1 Transitioning cookbook

As mentioned before, TAO methodology has three main phases: the knowledge acquisition phase, the ontology learning phase and semantic content and the service augmentation phase. Each phase contains a set of tasks which may interact with each other. Figure 5 presents a UML diagram to illustrate the main transitioning process and we explain major activities in detail.

Given a legacy application, the domain engineers first check if there are some previously-developed ontologies for the application. Some public ontology search engines or public ontology libraries can be used for this task[6]. If such an ontology is found, it can be saved into the knowledge store developed by TAO for future usage, otherwise users have to derive the domain ontology from the legacy software. For the GATE case, we develop the ontology from scratch with the assitance of TAO tools.

**Knowledge acquisition** To derive the domain ontology from the legacy application using the TAO tools, users first need to collect the relevant resources about the legacy application.

***– Resources collection***

We identify some data sources which are commonly relevant to the description of a legacy application, such as application source codes, API, Java Doc etc. For more information about the potential data sources which may be related to the description of a legacy systems and their classification, please refer to [3]. For the GATE case study, first Java source code and JavaDoc files are collected. Those documents can be downloaded from `http://gate.ac.uk/download/index.html`.

***–Save the resource corpuses to TAO Repository***

After collecting all the related data sources, we store them in the repository. TAO project develops a heterogeneous knowledge store to store these data sources. The heterogeneous knowledge store is designed for efficient management of different types of knowledge: unstructured content (documents), structured data (databases), ontologies, and semantic annotations, which augment the content with links to machine-interpretable metadata. More information about this heterogeneous knowledge store can be found in [19].

**Ontology Learning** The purpose of ontology learning from software artifacts is essentially discovering concepts and relations from the source code, accompanying documentation, and external sources (such as the Web). Ontology learning is one of the most significant approaches proposed to date for developing ontologies. Previously, we have presented a detailed review of different ontology learning approaches [3]. In this paper, we show how to learn domain ontologies based on the TAO scenario; The final output of TAO will be a prototype that will include all pieces of software developed during the project and will be available as TAO Suite. LATINO[7], a part of the TAO Suite, is used for Ontology Learning. LATINO is a more-or-less general data-mining framework

---

[6] `http://swoogle.umbc.edu` or `http://swse.deri.org`

[7] `http://www.tao-project.eu/researchanddevelopment/demosanddownloads/ontology-learning-software.html`

that joins text mining and link analysis for the purpose of (semi-automated) ontology construction. The ontologies are constructed from the knowledge extracted from the data that accompany typical legacy applications. We introduce the term “application mining” which denotes the process of extracting this knowledge.

In the previous step, we collected a set of related data resources that describe the legacy application. To use LATINO to get ontologies from these resources, we need to first identify these resources' contents and structures.

***–Identify content and structure of software artifacts***

****Identify the text-mining instances***

****Assign textual document to instances***

****Determine the structure between instances***

Given a concrete TAO scenario, the first question that needs to be answered by a software engineer is – what are the *text-mining instances* ( which are used as graph vertices when dealing with the structure) in this particular case, i.e., the user need to study the data at hand and decide which data entities will play the role of instances in the transitioning process. It is impossible to answer this question in general  it depends on the available sources. Some potential choices include Java/C++ classes, methods, Database entities and the like. In the GATE case study, the instances are Java classes.

Next, we need to assign a textual document (description) to each text-mining instance. This step is not obligatory, and perhaps not even possible when the data is such that it does not contain any unstructured textual data. Again, there is not a universal standard for which text should be included, but it is important to include only those bits of text that are relevant and will not mislead the text-mining algorithms. Users should develop several (reasonable) rules for what to include and what to leave out, and evaluate each of them in the given setting, choosing the rule that will perform best. In general, for most legacy applications that have well-commented Java/C++ source code available, *class comment*, *class name*, *field names*, *field comments*, *method names* and *method comments* can be used.

The user may also identify the structural information, which is evident from the data. This step is also not obligatory, provided that textual documents have been attached to the instances. The user should consider any kind of relationships between the instances (e.g. links, references, computed similarities, and so on). Note that it is sometimes necessary to define the instances in a way that makes it possible to exploit the relationships between them. For Java/C++ classes, the potential links that can be extracted include inheritance and interface implementation graph, type reference graph, class, operation name similarity graph, and comment reference graph, etc. After this step, the data pre-processing phase is complete. More information about those types of links and the different calculations of link weight can be found in [12].

***Creating feature vectors from contents and structures***

The text-mining algorithms employed by LATINO (and also by many other data-mining tools) work with feature vectors. Therefore, once the text-mining instances have been enriched with the textual documents we need to convert them into feature vectors. LATINO is able to compute the feature vectors from a document network. When this network is created based on the source code, such as in the GATE case study, it is common that a class has methods that return values of the type represented by another class. Also, comments in Java classes usually refer to other classes. For each of these cases, one graph would be created. In these graphs, vertices represent Java classes and edges represent references between these classes. After creating several such graphs they all have the same set of vertices. Next, different weights (ranging from 0 to 1) are assigned to each graph. In extreme case, 0 would be used to exclude the graph, and 1 to include it. Assigning weights is not a trivial process and requires lots of experimenting, experience, and intuition. Following the intuition, the user has to specify the weight setting and examine the results. If the results are not satisfying, the user has to change the settings and repeat the process again. To help the user set the parameters, OntoSight [13], an application that gives the user insight into document networks and semantic spaces through visualization and interaction, has been developed. For the usage of LATINO, we refer reader to [13].

These feature vectors are further used as an input for OntoGen[8] which is a semi-automatic data-driven ontology construction tool that creates suggestions for new concepts for the ontology automatically. OntoGen will be integrated with LATINO in later version.

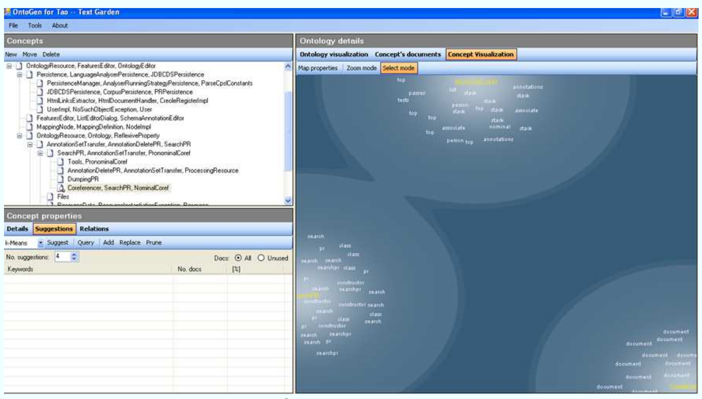
***Create domain ontology from feature vectors.***

The most important step of ontology development is identifying the concepts in a domain. Using OntoGen, this can be performed by using either a fully automated approach such as unsupervised learning (e.g. clustering), or a semi-automated supervised learning (e.g. classification) approach.

In the unsupervised approach, the system provides suggestions for possible sub-concepts of the selected concept and The supervised approach is based on Support vector machines (SVM) active learning method, which are a set of related supervised learning methods used for classification and regression. The user can start this method by submitting a query. After the user enters a query, the active learning system starts asking questions and labeling the instances. On each step the system asks if a particular instance belongs to the concept. The main advantage of unsupervised methods is that they require very little input from the user. The unsupervised methods provide well-balanced suggestions for sub-concepts based on the instances and are also good for exploring the data. The supervised method on the other hand requires more input. The user has first to figure out what the sub-concept is, then to describe the sub-concept trough a query and go through the sequence of questions to clarify the query. This is intended for the cases where the user has a clear idea of the sub-concept he wants to add to the ontology, although the unsupervised methods is not capable to discover it.

For the GATE case study, we have chosen the unsupervised approach, because we have little knowledge about the ontology. An example of automatically generated concepts visualised using OntoGen is shown on Figure 3. This figure depicts three concepts, namely *Nominal Coreferencer*, *Pronominal Coreferencer* and *SearchPR*. Each of these three concepts represent a separate Processing Resource (PR) in GATE. In OntoGen they are being clustered as belonging to the same group of concepts. If we decide to add this group to the ontology, one class will be created and three instances for each mentioned PR.

[8] http://ontogen.ijs.si/


OntoGen for Tao -- Text Garden
File Tools About
Concepts
New Move Delete
Ontology details
Ontology visualization Concept's documents Concept Visualization
Map properties Zoom mode Select mode
Concept properties
Details Suggestions Relations
k-Means Suggest Query Add Replace Prune
No. suggestions: 4
Docs: All Unused
Keywords No. docs [%]

- OntologyResource, FeaturesEditor, OntologyEditor
  - Persistence, LanguageAnalyserPersistence, JDBCDSPersistence
    - PersistenceManager, AnalyserRunningStrategyPersistence, ParseCpslConstants
    - JDBCDSPersistence, CorpusPersistence, PRPersistence
    - HtmlLinksExtractor, HtmlDocumentHandler, CreoleRegisterImpl
    - UserImpl, NoSuchObjectException, User
  - FeaturesEditor, ListEditorDialog, SchemaAnnotationEditor
  - MappingNode, MappingDefinition, NodeImpl
  - OntologyResource, Ontology, ReflexiveProperty
    - AnnotationSetTransfer, AnnotationDeletePR, SearchPR
      - SearchPR, AnnotationSetTransfer, PronominalCoref
        - Tools, PronominalCoref
        - AnnotationDeletePR, AnnotationSetTransfer, ProcessingResource
        - DumpingPR
        - Coreferencer, SearchPR, NominalCoref
      - Files

**Fig. 3.** Concepts derived from GATE source code using OntoGen.

Apart from concept identification, OntoGen/LATINO also implicitly infers subsumption relations (i.e. subClassOf) between concepts (newer version will also be able to discover some other types of relations).The user can fully customize each of the concepts by defining its instances. The system helps here by detecting outliers both inside and outside the concept. If new data becomes available after the ontology is constructed, the system can help by automatically classifying new instances into appropriate concepts. For more detailed instructions about the usage of OntoGen/LATINO, please refer to [12].

***Design Ontology***

An important point to make is that the automated methods are not intended to extract the perfect ontology, they only offer support to domain experts in acquiring this knowledge. This help is especially useful in situations like ours when the knowledge is distributed in several documents. In fact no existing OL technique is completely unsupervised: a domain expert must be included somewhere in the knowledge acquisition loop. Therefore, the automatically acquired knowledge is post-edited, using an existing ontology editor, to remove irrelevant concepts and add missed ones. The changes mainly included deleting suggested concepts, as for the ontology it was not important to include too much details in certain cases, for instance distinguishing between the more than 30 types of Exceptions that could be thrown from the different Java classes.

After creating the domain ontology, we can save it into the TAO repository.

Now we are ready to augment the existing content of a legacy application (including the service definition) semantically. We present the details in the following subsections.

**Service and content augmentation** Content augmentation is a specific metadata generation task aiming to enable new information access methods. It enriches the text with semantic information, linked to a given ontology, thus enabling semantic-based search over the annotated content. In the case of legacy software applications, important parts are the service description, the software code and documentation. While there has been a significant body of research on semantic annotation of textual content (in the context of knowledge management applications), only limited attention has been paid to processing legacy software artefacts, and in general, to the problem of semantic-based software engineering. TAO has developed a tool named *Content Augmentation Tool (CAT)* to assist users to annotate heterogeneous software artifacts automatically (semi-automatically). In essence, *CAT* is capable of performing two tasks: semantic annotation – using Information Extraction some parts of the document content are marked and then linked to an ontology; and, persistent storage and lookup of augmented content, where document retrieval is based on relevance to a selected set of semantic annotations instead of relevance to words (like in keyword lookup). More information about *CAT* can be found at [7]. The main component of CAT is *Key Concept Identification Tool (KCIT)* which has been exposed as a Web service (CA service) and can be accessed at `http://gate.ac.uk/ca-service/services/CAService`. More information about CAT, KCIT and CA service can be found on `http://www.tao-project.eu/researchanddevelopment/demosanddownloads/content-augmentation-software.html` and also in [7] and [1].

To use *CAT*, we first need to identity which Web services users want to provide and also what kinds of other content need to be annotated.

***Identify services and other content to be annotated.***

Please note that normally the first step in creating a Web service is to design and implement the application that represents the Web service. This step includes the design and coding of the service implementation, and the testing to verify that all of its interfaces work correctly. After the Web service is developed, the service interface definition can be generated from the implementation of the service (i.e. the service interface can be derived from the application's Application Programming Interface (API)). Web services interfaces are usually developed in WSDL documents that define the interface and binding of the corresponding Web service implementations. In this paper, we assume that the Web services and the corresponding WSDL definitions for a legacy application have already been developed. Therefore, we focus on helping users to annotate the existing WSDL definitions to get SA-WSDL definitions. SA-WSDL [9] is one of the latest W3C recommendation for Semantic Web Service.

***Annotate – automatically and manually***

As mentioned before, the main tool for performing Content Augmentation automatically is KCIT. KCIT identifies key concepts from software-related legacy content intelligently (more than exact text match, like many other existing approaches). It can also be configured to better adopt different use cases. For example, when preparing a document such as WSDL, it can be configured so that the tags' processing is enabled. Users then just click a button and KCIT goes through the WSDL file or other legacy content and automatically identifies the pieces of text or tag, which are related to concepts or relations defined in the domain ontology by using NLP techniques. After the process of automatic annotation is finished, users can validate results by visualising them e.g. in GATE GUI, correcting annotations if necessary, and adding new ones by manually selecting the text they want to link to the relevant concept from the ontology.

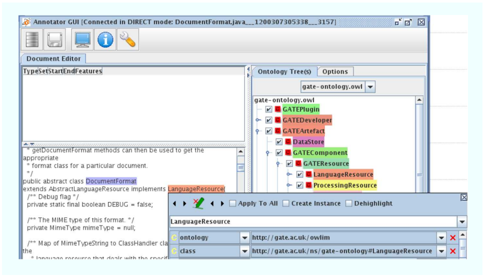
***View and revise annotations***

The results of the content augmentation process can be viewed graphically by the users as highlights over the original content inside the GATE GUI. For example, Figure 4(a) shows the results of processing the GATE class $DocumentFormat.java$ by the CA service . The highlights are the created semantic annotations, and the blue table at the bottom shows further details, in this case, of the annotation for $LanguageResource$ that refers to the class $LanguageResource$ from the GATE ontology.

The automatic annotating results could contain some flaws, and we need to ensure that these semantic metadata are correctly asserted. The annotation could be improper in several ways.

– Missed annotations. If the domain experts realize that there are some WSDL elements or texts in the legacy document, which should be annotated, but were missed by CA service, users can manually annotate them. If there is no proper concept within the existing ontology, new concept will be asserted into the ontology.
– Unnecessary annotations. It is possible that CA service has created some unnecessary annotations. Domain experts have to delete those annotations manually.
– Annotations with wrong concepts. If domain experts realize that the CA service has chosen the concept that is not the most suitable, they need to revise it.

Domain experts need to manually check the correctness of them. For example, as shown in Figure 4(b), the string *University of Sheffield* was not annotated as an *organisation*. To add this annotation to the document, first the text is selected/highlighted, and then the relevant ontology resource is chosen from the hierarchy on the right or from the drop-down list of resource names in the dialogue. In this particular examples we are annotating the string University of Sheffield as referring to the class Organisation.



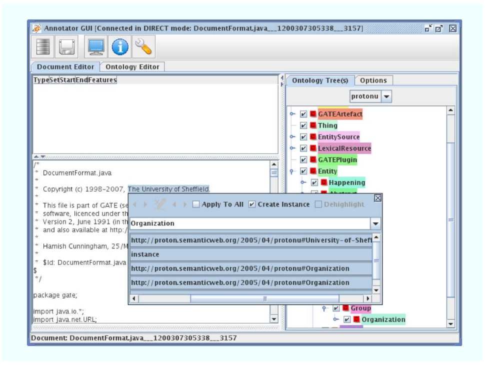
```
 * getDocumentFormat methods can then be used to get the
appropriate
 * format class for a particular document.
 */
public abstract class DocumentFormat
extends AbstractLanguageResource implements LanguageResource{
  /** Debug flag */
  private static final boolean DEBUG = false;

  /** The MIME type of this format. */
  private MimeType mimeType = null;

  /** Map of MimeTypeString to ClassHandler cla
the
```

gate-ontology.owl
- GATEPlugin
- GATEDeveloper
- GATEArtefact
  - DataStore
  - GATEComponent
    - GATEResource
      - LanguageResource
      - ProcessingResource

LanguageResource

| | |
|---|---|
| ontology | http://gate.ac.uk/owlim |
| class | http://gate.ac.uk/ns/gate-ontology#LanguageResource |

(a) Viewing results using Annotator GUI



```
/*
 *  DocumentFormat.java
 *
 *  Copyright (c) 1998-2007, The University of Sheffield.
 *
 *  This file is part of GATE (se
 *  software, licenced under th
 *  Version 2, June 1991 (in th
 *  and also available at http:/
 *
 *  Hamish Cunningham, 25/M
 *
 *  $Id: DocumentFormat.java
$
 */

package gate;

import java.io.*;
import java.net.URL;
```

- GATEArtefact
- Thing
- EntitySource
- LexicalResource
- GATEPlugin
- Entity
  - Happening
    - Group
      - Organization

Organization

http://proton.semanticweb.org/2005/04/protonu#University-of-Sheff
instance
http://proton.semanticweb.org/2005/04/protonu#Organization
http://proton.semanticweb.org/2005/04/protonu#Organization

Document: DocumentFormat.java___1200307305338___3157

(b) Adding a new annotation

**Fig. 4.** CA interface

***Ontology population***

Some of the Content Augmentation Tools can also identify a set of potential instances for the classes in the domain ontology from the legacy content. User will decide whether or not to accept these assertions.

During above processes, whenever the domain ontology is revised, users need to ensure the ontology is still correct.

Finally the legacy contents and the result of related semantic augmentation is stored in the heterogeneous knowledge store. This is important if users are working with large datasets. It is also the safest way to ensure that the annotations can be reloaded as same as before. The annotation can be saved either separately with the legacy content or embedded within the legacy files.

In the SOA lifecycle, the next phases are service deployment and service descriptions evaluation and refinement.

Service Deployment refers to the process of deploying services within a service execution environment and service evaluation refers to the ongoing monitoring of an SOA system to determine whether it meets its design goals. During the course, if users encounter any problems, the domain ontology and SA-WSDL definitions are revised. As these phases are outside of the scope of the TAO project (the scope of TAO is just generating the semantic descriptions), we will not give more details about them here. [3] presents some general guidelines for these tasks.

## 4 Methodology evaluation

The methodology presented in this paper has been validated in several high-profile case studies. For example, the GATE system, as presented in this paper, is a comprehensive open source platform (with thousands of users) and an aircraft maintenance application form Dassault Aviation is a data-intensive business process application (managing a multi-million business). More information about these case studies can be found in [1, 4].

The evaluation of TAO methodology and tools is conducted following the criteria proposed in [14]. We carry out the evaluation from several aspects:

**What is the performance of the methodology and tools: ontology extraction and annotation performance.** During experimenting with Ontology Learning tools, due to the nature of this process, we have not measured the time spent for ontology acquisition, therefore we report only empirical results here. During the experiments with LATINO, the trickiest part was setting the weights. This requires a lot of experimenting and testing, and there is no common rule to be followed in order to derive the most accurate hierarchy of concepts that can serve as a core of an ontology. The better the weights settings, the more precise concepts are derived and less time is left for the expert to check automatically derived ontology and manually perform changes. However, OntoSight software which is currently being developed by Josef Stefan Institute might speed up this process for future usage. This software will, given the golden standard ontology, provide means for automatically identifying which set of weights gives best results, which will enable users to reduce time for experimenting with setting the weights and changing them until the best results are achieved following the intuition.

Created GATE domain ontology was further used and evaluated with Content Augmentation Tools, specifically KCIT. Given the time limitations, we could not report on the precision and recall regarding the produced annotations over the GATE software artefacts. However, these tools are available online and could be tested with any other ontology[9]. To show the quantitative side regarding the KCIT performance, we measured

[9] `http://www.tao-project.eu/`

the time to initialise it and run it over a corpora with 32 Java Classes randomly chosen from the GATE source code, and also with the GATE User Manual[10]. Table 1 shows the results in seconds.

**Table 1.** Initialization time for KCIT with GATE knowledge base(http://gate.ac.uk/ns/gate-kb) and execution times for producing ontology-aware annotations automatically. Shown times are in seconds.

| **Initialization time** | 6.265 |
|---|---|
| **Corpora** | **Execution time** |
| *GATE User Manual* | 0.688 |
| *32 Java classes from GATE source code* | 16.422 |
| **Average execution time** | 0.518 |

**Is the extracted ontology a good basis for ontology building: expert evaluation?** To measure the quality of the ontology that has been created, we have collected 36 questions from GATE user mailing list, where the GATE users enquire about various GATE modules, plugins, processing resources, and problems they encounter while using these modules. We manually examined these questions, and identified that out of these 36 questions, 61.1% (22 questions) were *answerable*: the GATE domain ontology that is developed following the methodology described in this paper, contained the answers for these questions. For the rest of the 14 questions (38.9%), the answer was not in the knowledge base. This is due to the questions that were usually enquiring about too specific details such as: *Can someone send me the codes for all characters that are classified as DEFAULT_TOKEN?* or *What is the the correct location of the lexicon files for Hepple Tagger?*. Another set of questions for which answer was not in the knowledge base enquired about specific features that were usually not included inside user manuals and documentation, but were related to the knowledge of experienced GATE users, e.g. the familiarity with GATE GUI, or types of parsers for language processing available in GATE. For example, *Which Gate plugin parsers can find object and subject of a sentence?* or *Is it possible to see the POS tagging from the GATE GUI?*. Another set of queries were enquiring about Processing Resources, that have not been yet included in the GATE knowledge base due to various reasons (i.e. they have been included in the GATE distribution recently). This is when the Ontology Refinement phase of our methodology comes to place: we need to refine the ontology if needed and enrich the knowledge base so that we reach 100% of *answerable* questions.

**Does the extracted ontology and semantic annotated resources support a certain task, such as more effective query and answer? – appropriateness for a task.** After finalising GATE domain ontology and populating the knowledge base with relevant concepts, we need the means for accessing this knowledge in a user-friendly manner. Although the ontology and semantically annotated software artefacts have potential to ease the process of finding relevant information for domain experts i.e. GATE developers or GATE users in this case, we cannot expect from them to learn formal languages for querying ontologies such as SeRQL or SPARQL. We also cannot assume

[10] http://gate.ac.uk/sale/tao/index.html

that they understand Semantic Web technologies, ontologies and ontology languages. That is why, in the context of the TAO project, we have been working on developing a Question-based Interface to Ontologies (QuestIO). QuestIO accepts text-based Natural Language queries as an input, transforms them into SeRQL queries, executes them and render results in the user-friendly manner.

For the purpose of evaluating QuestIO as a part of the GATE case study, we have ran it with previously collected 22 *answerable* questions and examined results. The overall performance of QuestIO system was 68% of correctly answered questions, which, in comparison with 58% accomplished by AquaLog system [10] in similar evaluation, we consider quite a good performance. More details about this evaluation is outside the scope of this paper, and we refer reader to [2] and [18].

## 5 Related work

A number of ontology-design methodologies that have been proposed to date to guide the process of ontology development from scratch have been listed in a comprehensive survey in [6, 11]. While, Mariano et. al. have identified seven of the most commonly used methodologies for designing ontologies from scratch[11]. [5, 17] have outlined a set of principles and design criteria that have been proved useful in developing domain ontologies. During the last decade several ontology-learning systems have been developed such as ASIUM [9], OntoLearn, Text2Onto, OntoGen, and others. Most of these systems depend on linguistic analysis and machine learning algorithms to find potentially interesting concepts and relations between them.

Whilst several methodologies exist to develop domain ontologies either from scratch or from text, there is no widely accepted methodology for transitioning existing applications to SOA based on domain ontologies. In our methodology, the domain ontology plays a key role in the transition process as it contains all the semantics required for annotating the services of the new SOA. Our methodology and tools are focused on legacy application transitioning. We use various kinds of function related resources to derive the domain ontology. Since most existing applications tend to have documentation describing their functionality and APIs, it is possible to use automatic processing tools to abstract domain concepts from terms used in such documentation and build the domain ontology. Furthermore, our methodology is fully supported by an integrated tool studio.

## 6 Conclusion and future works

A key requirement of transitioning applications to Semantic Web Services has promoted the urgent need of systematic methodologies and tools to assist the migration process. In this paper we have taken an ontological view of Semantic Web Services. Both the lifecycle of SOA and building ontologies were examined, in order to understand the requirements for transitioning legacy systems to SOAs. This has been used for developing an initial methodology for the transitioning process based on domain ontologies learned

[11] `http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/OntoWeb_Del_1-4.pdf`

from such applications. To support this methodology, a set of tools and a detailed cookbook style guide are presented as well. This transitioning process has been validated by a few large case studies. Next, these tools will be integrated as a single suite which can provide a one-stop service to assist users to migrate their legacy applications to semantic enable services. More intensive evaluation is currently under performing.

**Acknowledgements.**

This work is partially supported by the EU-funded TAO project (IST-2004-026460).

## References

1. Damljanovic D., Bontcheva K., Tablan V., Roberts I., Agatonovic M., Andrey S., and Sun J. Gate case study: Domain ontology and semantic augmentation of legacy content. Technical Report D6.2, TAO project, 2008.
2. Danica Damljanovic, Valentin Tablan, and Kalina Bontcheva. A text-based query interface to owl ontologies. In *6th Language Resources and Evaluation Conference (LREC)*, Marrakech, Morocco, May 2008. ELRA.
3. Amardeilh F., Vatant B., Gibbins N., Payne T., Saleh A., and Wang H. H. Sws bootstrapping methodology. Technical Report D1.2, TAO project, 2008.
4. Cerbah F. Case study 2: Domain ontology building and semantic augmentation of legacy content. Technical Report D7.2, TAO project, 2008.
5. Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, 1993.
6. D. Jones, T. Bench-Capon, and P. Visser. Methodologies for ontology development.
7. Bontcheva K., Damljanovic D., Aswani N., Agatonovic M., , and Sun J. Key concept identication and clustering of similar content. Technical Report D3.1, TAO project, 2007.
8. Bontcheva K., Roberts I., Agatonovic M., Nioche J., and Sun J. Gate case study: Requirement analysis and application of tao methodology in data intensive applications. Technical Report D6.1, TAO project, 2006.
9. Holger Lausen and Joel Farrell. Semantic annotations for WSDL and XML schema. W3C recommendation, W3C, August 2007. http://www.w3.org/TR/2007/REC-sawsdl-20070828/.
10. Vanessa Lopez, Victoria Uren, Enrico Motta, and Michele Pasin. Aqualog: An ontology-driven question answering system for organizational semantic intranets. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):72–105, June 2007.
11. M. Lpez. Overview of the methodologies for building ontologies, 1999.
12. Grcar M. Ontology learning services library (v1). Technical Report D2.1-v1, TAO project, 2007.
13. Grcar M. Ontology learning services library (v2). Technical Report D2.1, TAO project, 2008.
14. Sabou M. Learning web service ontologies: an automatic extraction method and its evaluation. *Ontology Learning from Text: Methods, Evaluation and Application*, 2005.
15. S. McIlraith, T. Son, and H. Zeng. Semantic web services, 2001.
16. Dumitru Roman, Uwe Keller, Holger Lausen, Jos de Bruijn, Rubn Lara, Michael Stollberg, Axel Polleres, Cristina Feier, Christoph Bussler, and Dieter Fensel. Web services modeling ontology. *Journal of Applied Ontology*, 39(1):77–106, 2005.
17. Bill Swartout, Ramesh Patil, Kevin Knight, and Tom Russ. Toward distributed use of large-scale ontologies. In *10th Workshop on Knowledge Acquisition*, Canada, June 1996.
18. Valentin Tablan, Danica Damljanovic, and Kalina Bontcheva. A natural language query interface to structured information. In *Proceedings of the 5h European Semantic Web Conference (ESWC 2008)*, Tenerife, Spain, June 2008.
19. Marinova Z. Heterogeneous knowledge store. Technical Report D4.2, TAO project, 2008.

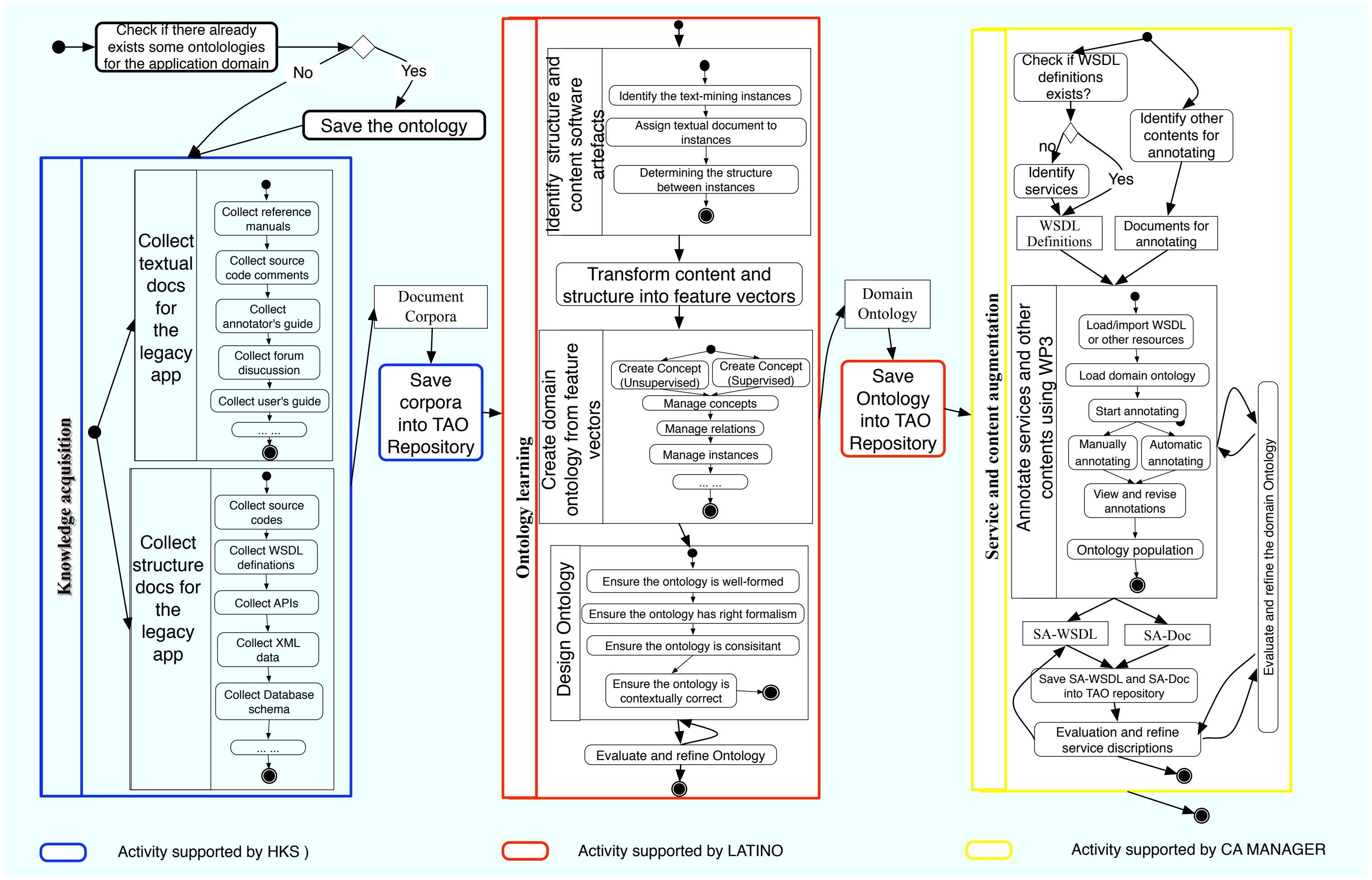

Check if there already exists some ontolologies for the application domain
No
Yes
Save the ontology
Knowledge acquisition
Collect textual docs for the legacy app
Collect reference manuals
Collect source code comments
Collect annotator's guide
Collect forum disucussion
Collect user's guide
... ...
Collect structure docs for the legacy app
Collect source codes
Collect WSDL definations
Collect APIs
Collect XML data
Collect Database schema
... ...
Document Corpora
Save corpora into TAO Repository
Ontology learning
Identify structure and content software artefacts
Identify the text-mining instances
Assign textual document to instances
Determining the structure between instances
Transform content and structure into feature vectors
Create domain ontology from feature vectors
Create Concept (Unsupervised)
Create Concept (Supervised)
Manage concepts
Manage relations
Manage instances
... ...
Design Ontology
Ensure the ontology is well-formed
Ensure the ontology has right formalism
Ensure the ontology is consisitant
Ensure the ontology is contextually correct
Evaluate and refine Ontology
Domain Ontology
Save Ontology into TAO Repository
Service and content augmentation
Check if WSDL definitions exists?
no
Yes
Identify services
Identify other contents for annotating
WSDL Definitions
Documents for annotating
Annotate services and other contents using WP3
Load/import WSDL or other resources
Load domain ontology
Start annotating
Manually annotating
Automatic annotating
View and revise annotations
Ontology population
SA-WSDL
SA-Doc
Save SA-WSDL and SA-Doc into TAO repository
Evaluation and refine service discriptions
Evaluate and refine the domain Ontology
Activity supported by HKS )
Activity supported by LATINO
Activity supported by CA MANAGER

**Figure 5.** Cookbook methodology overview.