Project Deliverable Report

D6.3 – Social and informal learning support services v1.5

Work Package  WP6
Task  6.3
Date of delivery  Contractual: 01-09-2010  Actual: 15-10-2010
Code name  D6.3
Version: 1.0
Type of deliverable  Report
Sec.(distribution level)  Public
Contributors  Paola Monachesi (UU), Eline Westerhout (UU), Thomas Markus (UU), Vlad Posea (UPB-NCIT), Kiril Simov (IPP-BAS), Petya Osenova (IPP-BAS), Stefan Trausan-Matu (UPB-NCIT)
Authors (Partner)  Paola Monachesi (UU), Eline Westerhout (UU), Thomas Markus (UU), Vlad Posea (UPB-NCIT), Kiril Simov (IPP-BAS), Petya Osenova (IPP-BAS), Stefan Trausan-Matu (UPB-NCIT)
Contact Person  Paola Monachesi (UU)
WP/Task responsible  Paola Monachesi (UU)
EC Project Officer  Ms. M. Csap
Abstract (for dissemination)  The goal of this Workpackage is to develop services that facilitate learners and tutors in accessing formal and informal knowledge sources in the context of a learning task. To this end, a Common Semantic Framework (CSF) is being developed that allows the stakeholders to identify, retrieve and exchange the learning material. More specifically, it supports formal and informal learning by facilitating the re-use of course material, by supporting knowledge discovery through an ontology enhanced with the vocabulary of the community of practice (CoP), and by recommending material on the basis of the content, tags and users belonging to the CoP. Communication is facilitated through the use of social networks and new communities of learners can be established through the recommendations provided by the system. The CSF provides theoretically motivated technological solutions to educational problems that arise in a Lifelong learning context.

Keywords List  Social media applications, ontologies, tagging, social networks, conceptual annotation
## Table of Contents

EXECUTIVE SUMMARY............................................................................................................. 3

1  INTRODUCTION .................................................................................................................. 4

2  IMPLEMENTATION ................................................................................................................. 6
  2.1  SYSTEM OVERVIEW ................................................................................................. 7
  2.2  VALIDATION V1.0 .................................................................................................. 8
  2.2.1  The FLSS ........................................................................................................... 9
  2.2.2  The iFLSS: knowledge discovery component and social learning component .... 10
  2.3  SOFTWARE IMPROVEMENTS ............................................................................... 13
  2.3.1  The data layer .................................................................................................... 13
  2.3.2  The services layer .............................................................................................. 17
  2.3.3  The visualisation layer ...................................................................................... 32
  2.4  Widget integration ....................................................................................................... 39

3  VERIFICATION ...................................................................................................................... 41
  3.1  DOCUMENT ANNOTATION .................................................................................... 41
  3.2  ONTOLOGY ............................................................................................................. 42
  3.2.1  Ontology enrichment........................................................................................ 42
  3.2.2  Concept mapping ............................................................................................... 43
  3.3  QUALITY OF RESULTS ......................................................................................... 45
  3.3.1  Semantically annotated resources .................................................................. 45
  3.3.2  Informal resources ............................................................................................ 46
  3.3.3  Social resources ............................................................................................... 47
  3.4  SIZE OF THE SOCIAL NETWORK ........................................................................ 48
  3.5  PERFORMANCE ....................................................................................................... 49
  3.5.1  Scalability of the knowledge discovery component ....................................... 49
  3.5.2  Speed and accuracy of the crawler ................................................................. 51

4  NEXT STEPS .......................................................................................................................... 52
  4.1  Validation .................................................................................................................... 52
  4.1.1  Validation of the FLSS ....................................................................................... 52
  4.1.2  Validation of the iFLSS ...................................................................................... 53
  4.2  Threads ....................................................................................................................... 53
  4.2.1  Short thread ......................................................................................................... 53
  4.2.2  Long thread ......................................................................................................... 54
  4.3  TRANSFERRABILITY ............................................................................................... 56
  4.3.1  Technical ........................................................................................................... 56
  4.3.2  Pedagogical ....................................................................................................... 57
  4.3.3  Organizational ................................................................................................... 57
  4.4  PLANS FOR ROADMAP .......................................................................................... 58
  4.4.1  FLSS ................................................................................................................. 58
  4.4.2  iFLSS ................................................................................................................. 58

5  CONCLUSIONS ...................................................................................................................... 60

REFERENCES .......................................................................................................................... 62

APPENDIX A: WEB SERVICE DOCUMENTATION ................................................................ 63
Executive Summary

There is an increasing demand on individuals and organizations for knowledge and skills while the availability of information and content is growing exponentially. It is crucial to identify new ways to acquire, contribute and exploit knowledge and thereby facilitate learning. The objective of the Language Technologies for LifeLong Learning (LTfLL) project is to create next generation support services to enhance competence building and knowledge creation in educational and organizational settings. More specifically, one of the goals of the project is to support formal and informal learning by developing services that facilitate learners and tutors in accessing formal and informal knowledge sources in the context of a learning task.

The services are based on the idea that knowledge is a socially mediated product. Learners, therefore, build knowledge collaboratively and then internalize it in a personal knowledge building process. As a result, learners become skilled members of a Community of Practice (CoP), mastering the learning domain speech genre.

To this end, a Common Semantic Framework (CSF) has been developed, which includes the Formal Learning Support System (FLSS,) and Informal Learning Support System (iFLSS).

The services are motivated by the need to find a solution to educational problems that arise in a Lifelong Learning context. In D6.2, a detailed description of their technical characteristics has been given. We have also provided an overview of the pedagogical motivation behind their development and the expected impact on the stakeholders.

In the current reporting period, we have validated the FLSS with tutors and the iFLSS with learners. The stakeholders judged the services as being relevant for the learning process and provided us with useful suggestions for their further development. On the basis of the feedback received, an enhanced and extended version of the services has been developed (Version 1.5). The focus has been mainly on improving the usability, while we have also paid attention to a better scalability, interoperability and stability of the services. The most important contributions rely on the introduction of live data from social media, a more sophisticated visualization of the ontology, a disambiguation algorithm that improves the quality of the search results and a faster document annotation system.

The development of completely new functionalities is part of the road map for after the project. The main contribution with respect to the interoperability of the services concerns the technical integration of the different components within the ELGG widget framework.
1 Introduction

One of the objectives of the Language Technologies for Lifelong Learning (LTfLL) project is to develop services that facilitate learners and tutors in accessing formal and informal knowledge sources in the context of a learning task. To this end, a Common Semantic Framework (CSF) has been developed in WP6.

In the first phase of the project, reported in D6.1, we have identified the appropriate methodology and technology for the integration of the formal and informal learning components within the CSF. More specifically, we have:

- identified the theoretical foundations and the educational scenario underlying the CSF
- accomplished a study of the state of the art
- investigated and assessed existing tools and techniques, relevant for the CSF
- designed the architecture of the CSF
- identified services to be evaluated through showcases.

In the second phase of the project, described in D6.2, we have carried out an implementation of the relevant services underlying the CSF, which includes the Formal Learning Support System (FLSS, related to task 6.1) and Informal Learning Support System (iFLSS, related to task 6.2ab). We have worked towards an integration of these two systems within the CSF. A first implementation of (parts of) the services related to the CSF has been validated through the showcases reported in D7.2 and the feedback obtained has been employed to revise the services and the general architecture that resulted in Version 1 of the services. In particular, in D6.2, we presented the implementation of the:

- data aggregation services to crawl relevant data from social media sites;
- ontology enrichment services to enhance an existing ontology with the vocabulary of the Community of Practice;
- document annotation services based on linguistic information to improve retrieval and re-use of parts of learning material;
- search for peers and learning material:
  - ontology based search and browsing to support knowledge discovery;
  - tag and users based search to support social learning;
- visualisation which has been re-implemented on the basis of the showcases feedback.

In the current reporting period, we have validated the FLSS with tutors and the iFLSS with learners using the scenarios described in D3.2. The feedback provided by validation (reported in D7.3) has triggered revisions and improvements to the software and a new release (Version 1.5). The focus has been mainly on improving the usability, while we have also paid attention to a better scalability, interoperability and stability of the services. The main contribution with respect to the interoperability concerns the technical integration of the various components within the ELGG widget framework (described in D2.3).
Initially, we planned to explore new directions and techniques to improve the CSF in the third year of the project. However, since the focus of the project has shifted towards working on the threads and strengthening the previous version of the software, we were not able to do this. Some exploratory experiments have been carried out in the direction of personalizing ontologies and the definition of experts, but the implementation of new functionalities in which these topics are addressed has been put on the road map for after the project.

The deliverable is organized as follows. Chapter 2 presents the revisions to the implementation of the various components. It is mainly based on the feedback received during validation. In chapter 3, the verification of the V1.5 services is described. The plans for the next phase are described in chapter 4, in which attention is devoted to the validation of the V1.5 services (WP7), the integration with other services developed in the project in the long and short thread (WP4 and WP5), a description of the transferability of the software to other domains and institutions, and a roadmap with plans for the last phase of the project and the period after the project. In chapter 5, the deliverable ends with the conclusions.
2 Implementation

The main aim of WP6 is to provide services that support formal and informal learning. In formal learning, learners follow a specific curriculum. For each topic in the curriculum, there is a set of learning materials, provided by the tutor. In contrast to this, the curriculum in the informal learning process is not obligatory, and the role of the tutor is either absent or not obvious.

The learners thus exploit different types of content to succeed in their learning goals. From the formal learning perspective, they mainly rely on material which has been prepared by the tutor on a specific topic and is addressed by the curriculum. Learners are expected to access certain pieces of pre-stored information, which would lead them to the required knowledge. In the informal learning process, learners locate knowledge sources on the web, select the relevant ones, and finally study them. During this process, learners often need guidance and thus they might profit from finding appropriate people (peers) on the web and a more structured way to access content.

Needless to say, the two learning paradigms are complementary from the lifelong learner’s perspective. In WP6, we have developed services that support both the formal style of learning (Formal Learning Support System - FLSS) and the informal one (inFormal Learning Support System - iFLSS). The iFLSS consists of an ontology-based knowledge discovery component and a social network based component. In the previous deliverable (D6.2), we have described the FLSS and the iFLSS and our attempts to integrate them in a Common Semantic Framework (CSF).

In the reporting period, the FLSS and iFLSS have been validated separately. An improved version of the services (V1.5) has been developed on the basis of the validation. The changes involved scalability and usability improvements (e.g. improving the interface, including a help function), but we have also done research and experiments at a conceptual level (e.g. relation extraction and difficulty estimation). In addition to improving the individual components, progress has been made in the direction of further integrating the services at the widget level.
The picture shows the structure of the CSF. It contains three parts reflecting the different aspects that are addressed:

- Infrastructure: a collection of basic components for data management and other services not specific to a type of learning (formal or informal);
- Formal Learning Support System (FLSS): services and widgets on top of the infrastructure that support formal learning (i.e. multimedia document annotation);
- Informal Learning Support System (iFLSS): contains services and widgets for informal learning resources (i.e. ontology enrichment, search based on tags and users).

2.1 System overview

The CSF consists of a number of web services which have been structured into three layers to provide a clear overview of the functionalities of the framework and how they interact.

![Figure 2: Interaction between components and data in the CSF](image)
Data (section 2.3.1)
- Sources: lists the different sources of data that are either social networks which provide informal learning materials or manually maintained tutor-provided formal learning materials.
- Acquisition: data from social networks is acquired through either crawling a learner's personal social network and associated resource or by sending search queries to the services themselves. Finally, methods are supported to manually provide material to the system.
- Storage: shows that the data is stored using a single semantic repository which contains crawled informal learning resources, social network information, domain ontologies, reference ontologies, definitions and annotated documents.

Services (section 2.3.2): the services interact with the data repository to acquire information. They can be roughly decomposed into the following three categories:
- Ontology enrichment: combines data from the semantic repository and existing domain ontologies in order to enrich them.
- Document annotation: addresses the discourse-oriented annotation, such as anaphora and co-reference relations as well as the rhetorical structure of the texts. In addition, a service for the manual annotation of pictures is part of this module.
- Search: focuses on search for materials and peers based on ontologies and social networks to support social learning including the expert view on a domain.

Visualisation (section 2.3.3): aims to support the interpretability of the information by offering the learner a visualisation that fits his current learning task:
- Semantically annotated documents
- Knowledge discovery support
- Social learning support

The layers and its components (highlighted in bold) have been described in detail in Chapter 2 of D6.2 (Data: section 2.3.1; Services: section 2.3.2-2.3.4; Visualisation: section 2.3.5). We do not repeat the basic information in D6.3 but we refer to D6.2 for further details. The validation of the iFLSS and FLSS has led to the development of an improved version of the software, which is called V1.5. The enhancements are presented in the remainder of this chapter. Information on how to retrieve the software on Sourceforge can be found in Appendix A.

2.2 Validation V1.0

The CSF has been developed to propose solutions to real life problems identified in the context of (LifeLong) learning activities. To investigate whether our services can indeed be used for enhancing the learning process, we have assessed the validity of V1.0 of the FLSS and the iFLSS on the basis of two scenarios. The FLSS scenario
focused on the use of the service by tutors, while the iFLSS scenario addressed the learners.

The results show that the components of the (i)FLSS are appreciated by these different types of stakeholders. The validation of the FLSS and the iFLSS has been described extensively in D7.3 and in [Westerhout et al., 2010] for the iFLSS. We focus in this deliverable on the actions that have been undertaken to address the problems mentioned by the learners, which have led to the development of version 1.5 of the software.

In the following sections, we provide screenshots of the FLSS and iFLSS, as well as a summary of the feedback from validation and the resulting actions that had an effect on software development.

### 2.2.1 The FLSS

![Figure 3: V1.0 of the FLSS component](image)

The issues that have been proposed in the validation and have been addressed in version 1.5 are summarized in Table 1.
## 2.2.2 The iFLSS: knowledge discovery component and social learning component

### Table 1: FLSS’ validation results

<table>
<thead>
<tr>
<th>Version 1.0</th>
<th>Version 1.5</th>
<th>Component</th>
<th>Aspect</th>
<th>Related section/chapter</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept Annotation Language Pipe</td>
<td>New implementation of three pipes</td>
<td>Concept Annotation</td>
<td>Faster online annotation</td>
<td>Language pipes</td>
<td>Implemented</td>
</tr>
<tr>
<td>Learning Object Names</td>
<td>Snippets in the search results</td>
<td>Search, Visualisation</td>
<td>New functionality</td>
<td>Semantically annotated documents</td>
<td>Implemented</td>
</tr>
<tr>
<td>Full Document Indexing</td>
<td>Annotation based indexing</td>
<td>Search, Visualisation</td>
<td>Faster document retrieval</td>
<td>The data layer</td>
<td>Implemented</td>
</tr>
<tr>
<td>XML document view</td>
<td>HTML document view</td>
<td>Visualisation</td>
<td>Usability</td>
<td>Semantically annotated documents</td>
<td>Implemented</td>
</tr>
<tr>
<td>Document Concept Statistics</td>
<td>Course Concept Statistics</td>
<td>Visualisation</td>
<td>Usability</td>
<td>Semantically annotated documents</td>
<td>Implemented</td>
</tr>
</tbody>
</table>

### Figure 4: V1.0 of the knowledge discovery component
The issues that have been tackled in version 1.5 are summarized in Table 2 and Table 3. The remainder of this chapter describes the changes that have been made to the software. The enhanced version (V1.5) of the services will be validated again using a scenario-based approach (section 4.1).

<table>
<thead>
<tr>
<th>Version 1.0</th>
<th>Version 1.5</th>
<th>Component</th>
<th>Aspect</th>
<th>Related section/chapter</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Own ranking method for Delicious resources</td>
<td>Delicious ranking</td>
<td>Data, Search</td>
<td>New approach</td>
<td>Dynamic social data retrieval for knowledge discovery, Retrieval of informal learning materials</td>
<td>Implemented</td>
</tr>
<tr>
<td>Delicious resources</td>
<td>Delicious, SlideShare, YouTube resources</td>
<td>Data, Search</td>
<td>New functionality</td>
<td>Dynamic social data retrieval for knowledge discovery, Retrieval of informal learning materials</td>
<td>Implemented</td>
</tr>
<tr>
<td>no scientific documents</td>
<td>Bibsonomy documents</td>
<td>Data, Search</td>
<td>New functionality</td>
<td>Dynamic social data retrieval for knowledge discovery, Retrieval of informal learning materials</td>
<td>Implemented</td>
</tr>
</tbody>
</table>

Figure 5: V1.0 of the social learning component
<table>
<thead>
<tr>
<th>Version 1.0</th>
<th>Version 1.5</th>
<th>Component</th>
<th>Aspect</th>
<th>Related section/chapter</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>no disambiguation in ontology enrichment</td>
<td>disambiguation integrated in ontology enrichment</td>
<td>Ontology enrichment</td>
<td>New approach</td>
<td>Including a disambiguation component</td>
<td>Implemented</td>
</tr>
<tr>
<td>not enough relations in ontology</td>
<td>relation extraction investigated</td>
<td>Ontology enrichment</td>
<td>New approach</td>
<td>Relation extraction</td>
<td>Research</td>
</tr>
<tr>
<td>no disambiguation of queries</td>
<td>disambiguation of keyword based searches</td>
<td>Search</td>
<td>New approach</td>
<td>Retrieval of informal learning materials</td>
<td>Implemented</td>
</tr>
<tr>
<td>definitions directly retrieved from dbpedia using the term</td>
<td>mapping term to lt4el/dbpedia concept more carefully</td>
<td>Search</td>
<td>New approach</td>
<td>Retrieval of definitions</td>
<td>Implemented</td>
</tr>
<tr>
<td>only one definition for multiple terms</td>
<td>one definition for each term</td>
<td>Search</td>
<td>Usability</td>
<td>Retrieval of definitions</td>
<td>Implemented</td>
</tr>
<tr>
<td>static visualisation</td>
<td>dynamic visualisation</td>
<td>Visualisation</td>
<td>Usability</td>
<td>Ontology visualisation</td>
<td>Implemented</td>
</tr>
<tr>
<td>graph creation slow</td>
<td>graph creation faster</td>
<td>Visualisation</td>
<td>Optimization, Scalability</td>
<td>Ontology visualisation</td>
<td>Implemented</td>
</tr>
<tr>
<td>graph fragment for multiple terms problematic</td>
<td>show shortest path from concept1 to concept2</td>
<td>Visualisation</td>
<td>Usability</td>
<td>Ontology visualisation</td>
<td>Implemented</td>
</tr>
<tr>
<td>long definitions</td>
<td>short definitions</td>
<td>Search</td>
<td>Usability</td>
<td>Definitions</td>
<td>Implemented</td>
</tr>
<tr>
<td>separate components</td>
<td>technical integration</td>
<td>All</td>
<td>Interoperability</td>
<td>Widget integration</td>
<td>Implemented</td>
</tr>
<tr>
<td>bug fixing</td>
<td>Visualisation, Search</td>
<td>Usability</td>
<td></td>
<td></td>
<td>Implemented</td>
</tr>
<tr>
<td>software scalability improvements</td>
<td>Visualisation, Search, Data</td>
<td>Scalability</td>
<td>Verification chapter</td>
<td></td>
<td>Implemented</td>
</tr>
<tr>
<td>no testing with large group of users</td>
<td>measured system performance for 100 simulated users</td>
<td>All</td>
<td>Scalability</td>
<td>Verification chapter</td>
<td>Tested</td>
</tr>
</tbody>
</table>

Table 2: Summary of improvements in the knowledge discovery component of the iFLSS
### Table 3: Summary of improvements in the social learning component of the iFLSS

<table>
<thead>
<tr>
<th>Version 1.0</th>
<th>Version 1.5</th>
<th>Component</th>
<th>Aspect</th>
<th>Related section/chapter</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>crawlers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delicious, SlideShare, YouTube, Flickr crawlers</td>
<td>Delicious, SlideShare, YouTube, Flickr crawlers - updated as some API's changed</td>
<td>Data acquisition</td>
<td>Update</td>
<td>Crawling personalised social data</td>
<td>Implemented</td>
</tr>
<tr>
<td>users appeared multiple times in the search results</td>
<td>detection of multiple accounts for same user</td>
<td>Search</td>
<td>Usability</td>
<td>User search</td>
<td>Implemented</td>
</tr>
<tr>
<td>no path to user shown in the widget</td>
<td>show connection between searcher and user returned</td>
<td>Visualisation</td>
<td>Usability</td>
<td>Social search results</td>
<td>Implemented</td>
</tr>
<tr>
<td>no document titles</td>
<td>document titles</td>
<td>Visualisation</td>
<td>Usability</td>
<td>Social search results</td>
<td>Implemented</td>
</tr>
<tr>
<td>recommender algorithm not scalable</td>
<td>recommender algorithm scalable</td>
<td>Search/Recommendation</td>
<td>Scalability</td>
<td></td>
<td>research</td>
</tr>
<tr>
<td>no caching systems</td>
<td>caching for search and recommendation</td>
<td>Search/Recommendation</td>
<td>Scalability</td>
<td></td>
<td>Implemented</td>
</tr>
<tr>
<td>various bugs</td>
<td>bugfixing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.3 Software improvements

As already mentioned, in this reporting phase, effort has been mainly dedicated to enhance the usability, scalability and stability of the software. Three types of improvements have been made in V1.5:

1. improvements driven by validation outcomes (described in Table 1-Table 3);
2. new functionalities driven by conceptual changes;
3. improvements related to the usability, scalability and stability on the basis of internally testing the software.

This section presents for each of the modules a description of the improvements. A summary of the changes per module is presented at the beginning of each section (Table 4-Table 9).

2.3.1 The data layer

2.3.1.1 Pedagogical impact

**Data:** For informal learning: (a) data from the learner’s peers is crawled and transformed into semantic data; (b) resources from social media sites are dynamically
retrieved. For formal learning: the semantic repository includes learning objects approved by the tutors.

**Impact on learners:** in the case of informal learning, the learners have access to their documents and the public documents of their peers in a single place and indexed with semantic data. These aggregated data represent their community knowledge. In the case of formal learning, the learners can have access to course material and material provided by the tutors that are also located in the semantic repository.

**Stakeholders:** learners and tutors. The tutors can have available (and indexed) the public documents produced or uploaded by their students.

**Formal/Informal Learning:** both

**Relevance with respect to theoretical model:** this service aggregates the knowledge of the communities the learner or tutor is part of. More specifically, we can identify the communities’ artifacts, the development of the personal knowledge, the evolution of the learner's ideas towards the shared understanding of the community and also the learner's attention profile which is another element of the Stahl cycle [Stahl 2006].

**2.3.1.2 Feedback from validation**

The validation has yielded some feedback that concerned the data component of the CSF. One issue involved the resources that could be retrieved using the knowledge discovery component. In the previous version of the software, we had opted to create a dataset of crawled data from Delicious, because of the freedom this offered with respect to the transformation and statistics that one can generate. However such a dataset also appeared to have important drawbacks that, in part, negate the benefits of social media:

1. The data collection is static and doesn't update itself
2. A crawled dataset is only a subset of some larger data collection. This entails a reduced coverage when compared to the original data.

Especially the second disadvantage was confirmed by the stakeholders (learners and teaching manager), who indicated in the validation that the quality and the ranking of the resources from the static dataset were not satisfactory. Regarding the types of resources that could be retrieved, the learners indicated that they preferred the documents of the social learning component over the results obtained with the knowledge discovery component. One of the reasons for this was that there was a larger diversity of data types (slides and videos in addition to Delicious bookmarks). The learners also suggested to include more scientific results in addition to the generally practical, tutorial-like documents from Delicious.

The validation of the iFLSS also revealed that among the students that participated in the validation Facebook and Twitter are the most commonly used social network sites. YouTube is often used for watching videos (not for learning), whereas SlideShare and Delicious were new to most of the students.
The data format for the document annotation developed for V1.0 was appropriate with respect to the representation of all the information related to the annotated documents in RDF. However, during the validation the users noticed that the resulting RDF graphs are very large and thus slow to transfer over the web.

In Table 4, we summarize the improvements in the ‘Data’ layer, as result of validation and internal testing of the software.

<table>
<thead>
<tr>
<th>Version 1.0</th>
<th>Version 1.5</th>
<th>Aspect</th>
<th>FLSS / iFLSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Document Indexing</td>
<td>Annotation based indexing</td>
<td>Scalability</td>
<td>FLSS</td>
</tr>
<tr>
<td>Serial crawler</td>
<td>Distributed crawler</td>
<td>Scalability</td>
<td>iFLSS</td>
</tr>
<tr>
<td>Delicious, SlideShare, YouTube, Flickr crawlers</td>
<td>Delicious, SlideShare, YouTube, Flickr, Twitter and Facebook crawlers</td>
<td>Usability</td>
<td>iFLSS</td>
</tr>
<tr>
<td>Delicious, SlideShare, YouTube, Flickr crawlers</td>
<td>Updated Delicious, SlideShare, YouTube, Flickr crawlers - some API's changed</td>
<td>Update</td>
<td>iFLSS</td>
</tr>
<tr>
<td>Data stored in static dataset</td>
<td>Data retrieved dynamically</td>
<td>Usability, Quality, Stability</td>
<td>iFLSS</td>
</tr>
</tbody>
</table>

**Table 4: Improvements in the 'Data' layer**

### 2.3.1.3 Data acquisition

#### 2.3.1.3.1 Dynamic social data retrieval for knowledge discovery

In version 1.0 of our software, a static dataset was used as informal learning material. However, as already mentioned, the validation revealed that this has an important drawback: it is limited. It is for this reason that we now acquire data directly from various social media and integrate them with the existing data model on the fly. This ensures that the results are up-to-date and extensive. In addition, we take advantage of the ranking algorithms of the social services to extract the most popular resources and in this way solve another issue mentioned during the validation.

In the new version, we have web services to include live data from SlideShare, YouTube, Delicious and Bibsonomy\(^1\). The corresponding widgets automatically update and display the latest resources from those websites when interacting with the ontology or entering a keyword based search.

Learners are thus able to get accustomed with a new domain and find and store relevant learning materials. In the long term, learners will be connected with peers relevant to the domain in their respective Community of Practice. This results in a switch to the social recommendation system for more personalized content as opposed to a more general overview of the domain.

The web services for retrieving data perform the necessary throttling in order to prevent (temporarily) overloading of the external services. To make it easy to see

---

\(^1\) We have investigated the possibility to implement Google Scholar or Cite-u-like web services for retrieving scientific data. However Google Scholar and Cite-u-like don't provide an API for external applications to use which in turn means that we are unable to integrate them within our system.
what the origin of the learning resources is, the logo of the original social media application is shown on the top of each list.

Figure 6: Search results from various social media as displayed in the application

2.3.1.3.2 Crawling personalised social data

Social search uses a dynamic crawling approach that crawls personalized data and combines them together for personalised search and recommendation. The social data repository only contains tags and resources that are in the direct vicinity of the learner.

This makes the results both relevant and trusted, but imposes limitations on the scope and coverage of the data. The crawling process results in a local data repository that allows us to use our own ranking, recommendation algorithms and data integration.

The scalability of the crawler has been improved to run in a distributed environment in the cloud. This approach uses the Hadoop technology\(^2\) and the Map-Reduce approach to distribute the crawling tasks. In this way, we can both reduce the time it

takes to acquire the information and minimize the chances of our crawler being blocked by the social web sites. The interval after which the crawler revisits a user's profile has also been diminished significantly.

The previous version supported Delicious, SlideShare and YouTube accounts to obtain the personalised data. However, the validation revealed that Twitter and Facebook are the most commonly used social media applications within the learners own networks. The data aggregation module has therefore been extended and now crawls the Twitter and Facebook accounts from the learners. The validation also showed that the Flickr resources were not valuable enough from the learning point of view. It was thus important to discover alternate platforms where learners share learning objects.

The Twitter extension analyzes the tweets posted by the peers and the peers of the peers of the learner and stores the links in the semantic repository as RDF. The twitter crawler also stores tags added by the twitter users (the so called hash-tags). This characteristic makes the twitter data very similar to the rest of the social web data that our system is using.

The Facebook extension allows the system to index and store the pages "liked" by the learner and his peers. Due to certain characteristics of the Facebook system the crawler isn't able to index data that is further away than the learner's peers. We have also implemented a system which allows the learner to set permission for our crawler on the Facebook website itself.

### 2.3.1.4 Data storage

The data models that have been developed for V1.0 were appropriate to accommodate the improvements carried out in this phase. Some minor adjustments have been made in the web service definitions to improve usability.

We have changed the data format to include both the XML and RDF representations of the annotated documents. The XML format is used to represent the structure of the document, the text content and the annotation of the document whereas in RDF, we only represent the concept annotation and the location of this concept annotation within the annotated document. The semantic repository is used to perform the semantic search. The results from semantic search (locations of matching elements in the XML structure of the document) are then used to process the XML representation and deliver the final results as snippets.

### 2.3.2 The services layer

#### 2.3.2.1 Ontology enrichment

#### 2.3.2.1.1 Pedagogical impact

**Data:** computing ontology, lexicon files (English and Dutch), reference ontologies (DBpedia, Yago, openCyc), crawled social media data (Delicious)
Impact on learners: socially driven ontology enrichment integrates vocabulary from the communities that the learner is part of with the expert view of the domain. New, socially relevant concepts are included in an existing domain ontology. This makes the enriched ontology more accessible because of the use of the vocabulary of the CoP that the user is part of.

Stakeholders: learners and tutors. There is no distinction made between the two in the ontology enrichment process. Tutors' influence is implicit in the ontology enrichment process because of their social relation to the learners expressed within the social networking site. The automatically enriched ontology contains new concepts which have gained prominence since the original development of the ontology. The enriched ontology is used for annotating learning objects which cover new topics or contain new concepts.

Formal/ Informal Learning: both

Relevance with respect to theoretical model: the ontology enrichment service improves the shared understanding of the domain concepts and their relations. Improved internalization of new knowledge is supported by selecting the group perspective of the domain closest to the learner. The group perspective is determined by considering cultural artifacts from the web available within the system. The collaborative knowledge expressed through tags from social media is integrated with the expert view from the reference ontology, thus enhancing personal comprehension.

2.3.2.1.2 Feedback from validation

The validation revealed that a number of concepts within the FLSS were not correctly disambiguated. The learners especially noticed this when definitions were shown, which were often incorrect. For example, the definition of the island Java was shown while the learners intended the programming language. To solve this issue, the current version of the software includes a disambiguation component.

The learners were very positive on the use of the ontology fragment for knowledge discovery. The ontology assisted them in identifying important concepts within a domain and helped to find out which concepts are related to each other. For some of the concepts, it is indicated exactly which type of relation there exists between them. This was also highly appreciated by the users, who indicated that they would like to see more types of relations. The majority of the relations in the original LT4eL domain ontology were of the type 'subclassOf' (shown as 'isa' relations in the visualisation). During the enrichment process, focus has been on discovering new related concepts, while not much effort has been undertaken to detect the type of relations between the existing concepts and the newly added concepts. As a result, most of the relations were simply labeled as 'related'. Currently the fragment only states that, for example, HTML and CSS are related, but it is not indicated in which way. In the validation, the learners thus clearly indicated that they would like to see how concepts are related to each other in order to make the knowledge discovery component even more useful to them.
In Table 5, we summarize the improvements to the ‘Ontology enrichment’ services, as a result of validation and internal testing of the software.

<table>
<thead>
<tr>
<th>Version 1.0</th>
<th>Version 1.5</th>
<th>Aspect</th>
<th>FLSS / iFLSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>no disambiguation in ontology enrichment</td>
<td>disambiguation integrated in ontology enrichment</td>
<td>Quality, Usability</td>
<td>iFLSS</td>
</tr>
<tr>
<td><em>New functionality</em></td>
<td>ontology based personalization</td>
<td>Usability</td>
<td>iFLSS</td>
</tr>
<tr>
<td>(research phase)</td>
<td>relation extraction</td>
<td>Usability</td>
<td>iFLSS</td>
</tr>
</tbody>
</table>

Table 5: Improvements to 'Ontology enrichment' services

2.3.2.1.3 Disambiguation component

The ontology enrichment methodology was previously described in D6.2, but we will now summarize it for convenience of the reader. Seed concept lexicalisations from the domain ontology trigger a similarity measure (e.g. resource co-occurrence) which generates socially relevant terms using social data. These relevant terms are then processed by the disambiguation algorithm. The disambiguation algorithm needs a context in order to disambiguate the socially relevant term. This context is provided by the domain ontology concepts that are directly associated with the seed concept that triggered the similarity measure. The context provided by the domain ontology will ensure that the word sense (reference ontology concept) that is most related to the existing concepts in the domain ontology will be selected. This allows it to select the proper word sense (reference ontology concept) for concept lexicalisations that are ambiguous. After this step taxonomic information, relations and alternate lexicalisations are retrieved by analysing data sets from reference ontologies such as DBpedia. The increased quality of the ontology enrichment process due to word sense disambiguation will improve learner's ability to benefit from the enriched ontology.

Figure 1 highlights the important parts of the ontology enrichment process that have been updated in V1.5:
This allows us to enrich the ontology with ambiguous terms such as 'python' and 'java' amongst others. For example; the similarity measure generates the tag 'python' in response to the seed concept lexicalisation 'java' as a socially relevant term. The concepts which are close to java in the ontology (DAO, JSF, Service-Oriented Architecture, Object oriented programming language) will provide the proper context to select the correct reference ontology concept for the socially relevant term 'python'. The correct concept (word sense) in this example for the term 'python', in the context of the java programming language, would be [http://dbpedia.org/resource/Python_(Programming_Language)](http://dbpedia.org/resource/Python_(Programming_Language)).

This process, although more precise, does mean that the ontology enrichment process will take more time due to its computational cost. However, this is not an important issue, because ontology enrichment is done in a batch before hand and not on-the-fly during user interaction.

### 2.3.2.1.4 Ontology based personalisation

Personalisation support was added to the ontology enrichment methodology, in connection with the disambiguation components. They now support the automatic generation of a MOAT\(^3\) RDF fragment given semantic modeling of data from social media. The MOAT vocabulary allows us to store the meaning of a tag as used by the learner in a specific context in connection with a resource. For example the tag 'java'
used by Marie in the context of a programming manual will be linked to a different concept than when John uses the same tag with a resource about islands. MOAT allows us to store these differences in meaning and context in a semantic format. More specifically; each tag is linked to its appropriate word sense if possible. We can thus automatically determine the meaning of a tag in the original context that a particular learner used it in. This information is then exported as a MOAT RDF fragment which can be stored in the semantic repository and therefore be used by other modules or external applications.

These data can then be used by other components to facilitate accessing social media content through ontologies. It can also be used for semantic analysis of social media content on the conceptual level instead of on the lexical (tag) level. This would constitute an explicit bridge from the formal modelling through ontologies to social tags. The MOAT-modelling vocabulary is described in more detail in D6.2.

### 2.3.2.1.5 Relation extraction

The learners indicated during validation that they wanted to see more relations in the ontology fragment. We have therefore carried out some experiments in the direction of relation extraction. The approach that we have adopted is a three level procedure:

1. Check whether the terms are related in DBPedia or other reference ontologies (e.g. XML is extended to XHTML)
2. Check whether the DBPedia definitions of the terms (i.e. the definitions shown in the visualisation) contain both terms. E.g. for the terms XLink and XML: "The XML Linking Language, or XLink, is an XML markup language used for creating hyperlinks in XML documents."
3. Search the Internet for a sentence containing both terms (e.g. "The purpose of a DTD is to define the legal building blocks of an XML document")

So far, we have developed prototypes for each of the three steps separately. We have implemented a web service for finding DBpedia relations between two DBpedia concepts. To make this web service more useful, the concept disambiguation component discussed in the previous section will be employed in order to link the concepts in our ontology to DBpedia concepts. For the second and third step, we use an approach based on [Schutz and Buitelaar, 2005], which investigates the relations between two concepts within the same sentence. To detect sentences containing the two concepts, we use either the DBpedia definition or consult the Internet. We have developed a webservice that searches the first $n$ pages returned by the Yahoo search engine for sentences containing the two terms that are related in the ontology. Some of these sentences are useless, since they don't express a relation between the concepts. Manual inspection of the remainder of the sentences revealed that a large diversity of verbal phrases can be employed to express relations between concepts. In order to identify which syntactic relations exist between the two terms, the Stanford NLP parser has been implemented to generate deep syntactic parses of the sentences containing the two terms of a concept pair. For further processing, the plain text output parses are converted to XML dependency structures.
2.3.2.1.6 Customizing ontologies

Ontologies provide a clear conceptual structure of a domain, which facilitates learning. However, an obvious shortcoming is that they represent a generic domain conceptualization that does not take a learner's existing background into account. For modelling a learner's personal knowledge in its original context, MOAT can be employed to attach to each term (tag) a conceptual interpretation. However, using the MOAT vocabulary we cannot express the explicit relations among concepts that we have in an ontology. To fill the gap between a well structured, but generic, expert view of a domain (represented in an ontology) and a learner's personal knowledge representation (using the MOAT vocabulary), an approach needs to be developed that customizes one or more domain ontologies for a specific learner.

The work that has been carried out in WP4 (see D4.3) provides a relevant starting point for our work in this direction. It highlights the differences and overlap that exist between the topic composition of a corpus of learning materials approved by an expert and the text provided by the learner. Both the corpus and the ontology provide a generic expert view of a domain, whereas text provided by the learner and knowledge formalized through the MOAT vocabulary express a learner's personal knowledge.

We aggregate a large number of learner provided non-textual feedbacks instead of using learner provided text in order to minimize the disruption of the normal workflow of the learner. The topic composition of the resources for which we received feedback allows us to, indirectly, identify the differences between the topics already familiar to the learner and compare them to the concepts, as present in a domain ontology. The domain ontology can then assist the learner in the acquisition of new domain concepts that were either absent or indicated by the learner as unfamiliar. At present this is done by colour-coding the domain concepts based on a learner's model. In addition, the topic composition of resources combined with the personal conceptual model of the learner allows us to rank resources based on how familiar a learner is with the topics. In both cases, we use a personalized model that has been created using a machine learning classifier. The model learns to predict the relation between the topic composition of a document and the learner provided feedback. This information is extracted from all the available feedback that a learner provided about whole documents. This work, although promising, is not yet mature enough to be validated and has been included in the roadmap, for future development.

2.3.2.2 Document annotation

2.3.2.2.1 Pedagogical impact

Data: (a) (non)-annotated learning objects, which can contain text and images; (b) domain ontology; (c) lexicons; (d) language NLP pipe.

Impact on learners: depending on his specific aim, the learner can choose on which level to derive relevant information – only conceptual or contextualized conceptual.
Stakeholders: additionally to observing the annotated objects that encode semantic information, tutors and learners can perform annotations on their own. This might be done manually (image and discourse annotations) or automatically (over newly uploaded materials).

Formal/Informal Learning: both.

Relevance with respect to theoretical model: the annotated data support the creation of contextualized knowledge. The annotations presuppose the tacit pre-understanding activity in the Stahl cycle [Stahl 2006]. They explicate inter-concept and rhetoric relations, in which the relevant concepts to be learned are considered as 'focus of attention'. Consequently, these concepts are brought into a Community of Practice for further elaboration, addition, illustration, attitude expression, etc.

2.3.2.2.2 Feedback from validation

The validation has been carried out on English and Bulgarian data. The results on English showed that although the ontology-based search returns relevant learning materials, more markers of relevance were required by the tutors (statistics, frequency, etc.). The tutors felt that they spend too much time in assessing the retrieved documents.

Another result was the sparse coverage of the data due to only domain annotation. The co-reference annotation on top of the semantic annotation layer improved the concept density with 30%. This problem will be further addressed in the future (and included as point in the roadmap) by means of an extended domain ontology and lexicon, as well as a common-sense annotation.

We also realized that the NLP pipe in its full capacity happened to be too slow in a real-time setting. For example, a text of 350 tokens has been processed in 3 minutes. Thus, for the next round the pipe has been divided into 3 sub-pipes. The first two which have to do with tokenization, POS tagging and concept annotation are fast enough (10-15 sec), while the third (which includes co-references and WordNet information) is envisaged as an offline automatic service.

In Table 6, we summarize the improvements to the ‘Document annotation’ services, as result of validation and internal testing of the software.

<table>
<thead>
<tr>
<th>Version 1.0</th>
<th>Version 1.5</th>
<th>Aspect</th>
<th>FLSS / iFLSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept Annotation Language Pipe</td>
<td>New implementation of three pipes</td>
<td>Scalability</td>
<td>FLSS</td>
</tr>
</tbody>
</table>

Table 6: Improvements to 'Document annotation' services

2.3.2.2.3 Language pipes

The language pipe for English underwent significant improvements. The capabilities of the CLaRK NLP system with the existing software package OpenNLP were
combined. Since the main criticism of the stakeholders as well as the EU reviewers was the poor performance in real time, the pipe was divided into three versions with respect to richness of the linguistic information and processing time required:

- Pipe 0.1, uses CLaRK-components. It includes tokenization and concept annotation, and is very fast. Thus, ontology-based search is facilitated on newly added learning objects.
- Pipe 0.2, uses CLaRK as well as OpenNLP-components. It includes the above information from the previous step and in addition POS and syntactic information. It is also reasonably fast.
- Pipe 0.3, uses CLaRK and OpenNLP-components. It includes all the modules, i.e. in addition to the above: Annotation with synsets from WordNet, co-reference annotation and distribution of conceptual information on the basis of co-reference annotation. However, it is slow due to the use of the underlying algorithm in OpenNLP over the co-reference, and Wordnet information.

Because of the different language pipes, the tutor can choose the most appropriate version of automatic linguistic analysis for his learning material. The tutor might either perform a quick semantic annotation for orientation, or a more extensive analysis for getting richly annotated texts. The main trade-off of the newly added pipes (0.1 and 0.2) is the improved speed of the annotation in real-time tasks. On the other hand, the most complex pipe (0.3) provides better precision. Thus, the tutor can decide which pipe to use in various tasks. For example, just getting an impression of the semantic annotation would take a short amount of time. If the results are acceptable, the tutor might proceed by storing the annotated document or by modifying the annotation where necessary. If, however, the annotation, provided by the faster pipelines, requires substantial manual work from the tutor, a richer analysis might be performed by the more extensive annotation pipeline.

The work within the document annotation is distributed among the two systems - CLaRK and OpenNLP, respectively - in the following way:

- The concept annotation grammars have been extended on the basis of a lexicon, which is aligned with the ontology. The grammar relies on information from the tokenization step. The information from POS tagging and syntactic tagging was used for disambiguation purposes (for example, "help" annotated as a verb can not be connected to the concept "help" within a software system).
- The information from WordNet and co-reference annotation establishes chains of equivalent phrases within the text. These chains are used in the distribution of conceptual information from the concept annotation. The typical cases of co-reference which we encountered were: (1) co-reference between a term and a pronoun in which case the concept assigned to the term is distributed to the pronoun; and (2) two terms are connected, in which case, the more general term received, as annotation, the concept assigned to the more specific term.

The Bulgarian pipe has only been implemented in the CLaRK system and currently does not support annotation using WordNet information and co-references.
2.3.2.3 Search and recommendation

2.3.2.3.1 Ontology-based search

Pedagogical impact

**Data:** (a) formal and informal learning resources. For the formal learning part, the ontology-based search can return semantically annotated formal content on a sub-document level (paragraphs and sentences), that matches the learner's query. Informal resources come from social media applications. The resources recommended by the social search service have been created and tagged by the learner's peers. (b) domain ontology and lexicons.

**Impact on learners:** learners are supported in finding relevant learning content from formal and informal learning resources. The ontology based search can return content on sub-document level: paragraph and sentences, which matches the learner's query when the resources are properly annotated. This gives the learner a highly focused result on their specific topic of interest. Ontology browsing constitute the core of the knowledge discovery process since it gives the learner structure and a good overview when exploring and searching either formal or informal learning objects.

**Stakeholders:** tutors can structure their courses more efficiently through ordering and annotating their learning objects, because they would have the possibility to browse the domain ontology to give structure as well as to get related material via ontology based search. The learners can exploit the ontology for getting a better understanding of the topic in the given domain and are facilitated in finding relevant informal content.

**Formal/ Informal Learning:** both

**Relevance with respect to theoretical model:** the ontology based search supports the ‘building personal knowing’ process in the Stahl cycle being the mediation between the cognitive and cultural artifacts, on the one hand, and the learner on the other [Stahl 2006]. The informal search helps the learner validate his ideas against the ideas of the community. The learner gains access to what Stahl calls "other utterances" - the ideas of the people around the learner in the community that help him develop a shared understanding with others.

Feedback from validation

The validation has provided some useful feedback to improve the ontology-based search. Part of the comments involved minor issues, and were often closely related to the visualisation module (section 2.3.3). However, there were also some more relevant issues mentioned by the learners and the tutors in the validation.

The evaluation of the FLSS revealed that offering only the document titles did not provide the tutor with enough information to decide on the usefulness of the learning
material. The tutors asked for more specific information about the content of the documents.

Regarding the knowledge discovery component, a first comment concerned the types of learning materials returned. The learners classified the majority of the learning materials as being superficial and non-scientific, while in learning often scientific, in-depth resources are needed. More specifically, it was only possible to retrieve Delicious resources whereas the stakeholders indicated that they would prefer to get access to more types of resources, such as Google Scholar results. The ranking of the learning materials retrieved in the knowledge discovery component was also not satisfactory.

The definition retrieval has become an important tool for successfully navigating an ontology in situations where the learner is unfamiliar with most or all concepts. The validation showed that all learners very much appreciated this functionality. The definition gives them a quick explanation as to what the concept is about. It is for this reason that the definition retrieval needs to be reliable. In the previous version there was an unsophisticated method for retrieving definitions from DBpedia based on the search query or the lexicalisation(s) of the currently selected concepts in the ontology. This had some shortcomings as identified during testing and validation such as:

1. Only one definition was returned for a multi-concept search request
2. Possibly incorrect definition selection for concepts from the domain ontology without a definition. In these situations the wide-scope knowledge resource (DBpedia) either returned an incorrect word sense, and thereby an unsuitable definition, or no definition at all.

In Table 7, we summarize the improvements to the ‘Ontology-based search’ services, as result of validation and internal testing of the software.

<table>
<thead>
<tr>
<th>Version 1.0</th>
<th>Version 1.5</th>
<th>Aspect</th>
<th>FLSS / iFLSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning Object Names</td>
<td>Snippets in the search results</td>
<td>Usability</td>
<td>FLSS</td>
</tr>
<tr>
<td>Own ranking method for Delicious resources</td>
<td>Using Delicious ranking</td>
<td>Usability, Quality</td>
<td>iFLSS</td>
</tr>
<tr>
<td><em>New functionality</em></td>
<td>Disambiguation of results by linking queries to concepts</td>
<td>Usability, Quality</td>
<td>iFLSS</td>
</tr>
<tr>
<td><em>New functionality</em></td>
<td>Delicious, SlideShare, YouTube resources</td>
<td>Usability</td>
<td>iFLSS</td>
</tr>
<tr>
<td><em>New functionality</em></td>
<td>Bibsonomy documents (scientific)</td>
<td>Usability, Quality</td>
<td>iFLSS</td>
</tr>
<tr>
<td>definitions directly retrieved from DBpedia using the term</td>
<td>mapping term to lt4el / DBpedia concept more carefully</td>
<td>Usability, Quality</td>
<td>iFLSS</td>
</tr>
<tr>
<td>no caching systems</td>
<td>caching for search and recommendation</td>
<td>Scalability</td>
<td>iFLSS</td>
</tr>
<tr>
<td>various bugs</td>
<td>bug fixing</td>
<td>Usability</td>
<td>iFLSS</td>
</tr>
<tr>
<td>problems with large numbers of users</td>
<td>software scalability improvements</td>
<td>Scalability improvement</td>
<td>iFLSS</td>
</tr>
</tbody>
</table>

Table 7: Improvements to 'Ontology-based search' services
Retrieval of formal learning materials

Since the learners considered the document titles of the formal learning materials to be not sufficient in order to decide on their usefulness, a different search approach has been implemented which enables the selection of relevant document snippets. The user starts with selecting a concept from the ontology. The system then performs query expansion which adds the subconcepts of the selected concept from the ontology. The extended search query is evaluated over the RDF repository. For each extracted document appropriate snippets are generated and the user is able to examine them in order to select the appropriate documents for further use.

Retrieval of informal learning materials

As already mentioned, there was room for improvement with respect to the retrieval of informal learning materials. The learners clearly indicated which aspects could be improved:

1. the ranking of the learning materials was not satisfactory;
2. the majority of the learning materials was non-scientific;
3. it was only possible to retrieve Delicious resources whereas the stakeholders indicated that they would prefer to get access to more types of resources.

To solve the first problem, we decided to implement a web service that includes live Delicious resources for a given concept or lexicalisation. The other problems have been addressed by including more types of social media resources in addition to the Delicious bookmarks. More specifically, we have added web services communicating with YouTube (videos), SlideShare (slides), and Bibsonomy (scientific papers) as described in section 2.3.1.3.1. Another advantage of accessing the social services directly is that it will always be possible to provide search results through keyword based search. The quality of the search results is then further improved by applying concept filtering using the domain ontology (section 2.3.2.1.3).
Figure 8 graphically illustrates the changes to the integration of live results from social media as opposed to a static repository. A search query is triggered by the learner through either the ontology or by entering a keyword based search. The query is sent as-is to the social media service to acquire resources. The received results from the social media service are then associated with their proper concepts by employing the disambiguation algorithm (section 2.3.2.1.3) after which a filtering step is performed.

The search query is linked to concepts from the domain ontology through a lexical lookup using the domain ontology's lexical entries. These domain ontology concepts are then mapped to matching concepts from a reference ontology. The domain ontology concepts need to be mapped to the reference ontology, because the search results should be assigned an unbiased word sense. This is because there might be search results which, though having a matching lexicalisation, actually belong to different domain.

In parallel, reference ontology concepts are linked to each search result with the disambiguation algorithm which considers all the tags associated with each specific search result. The co-occurrence of the tags allows the disambiguation algorithm to assign the correct word sense, because the other tags provide the proper context. These word senses come only from a broad reference ontology, as opposed to a domain ontology, in order to assign an unbiased word sense.
The filtering process compares the concepts from the disambiguated search results with the original query which has been linked to the domain ontology. The domain ontology provides a conceptual interpretation of the search query and only retains search results that match the concepts linked to the search query. A query refinement step using the domain ontology is then performed if not enough results remain after filtering. For example: if an initial query for 'java' only results in information about the Indonesian island instead of the programming language a query refinement step is performed to automatically extend the query with additional lexicalisations from the ontology in order to make it more specific.

As an example, we show the filtered and unfiltered results for the query 'ajax' as sent to YouTube. Results that have been crossed out have not been accepted by the concept filtering algorithm. Both struck and unstriked search results will be included in a regular keyword based search, but only unstriked search results will be present after ontology based filtering yielding more relevant resources.

- Ajax amsterdam
- Whiteboard//Session: What is AJAX?
- Ajax
- AJAX Tutorial 1 - Introduction
- Cristiano Ronaldo Vs Ajax Home By CrixRonnie
- Suarez, socio van Ajax
- Zlatan Ibrahimovic Goal for Ajax
- Ajax Fair Play
- Ajax Amsterdam Vs Chelsea FC (3-1) All Goals & Highlights 23.07.10
- Ajax Tutorial
- Ajax Amsterdam AFCA hooligans
- De Graafschap 0-5 Ajax Full Highlights 29th August 2010
- Ajax: Introduction
- Ajax in kolkende ArenA naar Champions League
- Google I/O 2009 - Advanced Techniques, AJAX API Playground
- ajax-zingt-shaffy
- 1973 European Cup Final Ajax v Juventus
- Flash Gordon Hawkmen vs Ajax
- Introduction to Debugging AJAX Application with Firebug
- Ajax Fair Play

The false negatives that appear during concept filtering are in many cases the result of poor resource tagging which doesn't allow the disambiguation to associate the right word sense. More details on the verification of these web services can be found in section 2.3.1.3.1.

Retrieval of definitions

The definition retrieval proved to be problematic during the validation, especially for ambiguous concepts. For example, for the concept 'Java programming language' a definition was retrieved from Wikipedia that described the island named 'Java' that is part of Indonesia instead of the computer programming language. In order to solve
this shortcoming, a method for lightweight concept mapping was introduced. It works by considering all the lexicalisations of concepts associated with the current concept from the domain ontology. The resulting set of lexicalisations is then fed to the word sense disambiguation algorithm (section 2.3.2.1.3) which selects the most likely word sense (DBpedia resource) for each of the lexicalisations. The definitions are then retrieved by using the assigned word senses (DBpedia resources) from the disambiguation results. In this setup, the domain ontology's structure is used to provide a context for disambiguating the seed term and to improve definition retrieval.

The retrieval process results in a list of definitions for each of the terms entered. Terms are chunked as much as possible during this process. I.e. 'web application' retrieves the definition for 'web application' and not for 'web' and 'application' separately.

In the previous version, there was a bug for queries with two or more terms. In such situations, only the definition of the first term was shown whereas one would like to see the definitions of all terms. Version 1.5 has solved this issue and the learner is now provided with a definition for each of the concepts that exists in the ontology.

Verification results on the reliability of this type of concepts mapping can be found in section 3.2.2.

2.3.2.3.2 Social search and recommendation

Pedagogical impact

Data: personalised social data that has been crawled using a learner’s neighborhood. The data refer to the documents created and tagged by the learner's peers and has been transformed into semantic data.

Impact on learners: we help the learners discover relevant materials that have been created or bookmarked by their peers and that could help them in their learning process. When indicating what resources could be useful we also indicate the peer who created or bookmarked the resource and how the learner could contact him for further information.

Stakeholders: learners and tutors. Learners are facilitated in finding relevant and trusted content and peers. We not only provide a list of resources to the learner, but also indicate who created or bookmarked the resource and how the learner could contact this person for further information. In this way, learners are facilitated in finding relevant and trusted content and peers. As a consequence, tutors will have to spend less time in providing recommendations and feedback to learners. They would just spend some time bookmarking interesting content and the learners will automatically receive this content as recommendation or as answer to queries.

Formal/ Informal Learning: informal (but could be extended to formal as well).
Relevance with respect to theoretical model: the search helps the learner validate his ideas against the ideas of the community. The learner gains access to what Stahl calls "other utterances" [Stahl 2006], that is, the ideas of the people around him in his own network that help him develop a shared understanding with his peers.

Feedback from validation

The validation gave no rise to make any changes to the method for retrieving learning materials. The learners were satisfied with both the quality and the ranking of the documents. Regarding the search for relevant people, a problem identified during validation is that users can have accounts on multiple web sites and therefore often appear more than once in the list with recommended people. We have developed a web service that will allow administrators to identify similar accounts on different web sites and to state that these accounts belong to the same person.

In Table 8, we summarize the improvements to the ‘Social search’ services, as result of validation and internal testing of the software.

<table>
<thead>
<tr>
<th>Version 1.0</th>
<th>Version 1.5</th>
<th>Aspect</th>
<th>FLSS / iFLSS</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>New functionality</em></td>
<td>web service that identifies similar accounts on different web sites belonging to the same person</td>
<td>Usability</td>
<td>iFLSS</td>
</tr>
<tr>
<td>no caching systems</td>
<td>caching for search and recommendation</td>
<td>Scalability</td>
<td>iFLSS</td>
</tr>
<tr>
<td>various bugs</td>
<td>bug fixing</td>
<td>Usability</td>
<td>iFLSS</td>
</tr>
</tbody>
</table>

Table 8: Improvements to ‘Social search’ services

Detecting similar accounts

In version 1.0, accounts for the same user on different web sites are treated separately. However, a problem mentioned during validation is that some users appeared multiple times in the search results because these users have relevant accounts on multiple web sites. Since the various accounts pertain only to one individual, they actually should be bundled. We have therefore developed a web service together with a simple and easy to use interface that will allow administrators to detect which accounts are similar and to state that these accounts actually belong to the same person.

The detection of similar users is based on a similarity measure $S$ defined as:

$$S(AccountA, AccountB) = p \times SimName(AccountA, AccountB) + (1-p) \times SimContent(AccountA, AccountB)$$

- $SimName(AccountA, AccountB)$ is computed as the Jaro-Winkler similarity between the usernames of the users holding the two accounts.
- $SimContent(AccountA, AccountB)$ is computed as the cosine distance between the normalized APML profiles of the two users, computed as described in D6.2
- $p$ is a ratio whose value needs to be chosen arbitrarily. As a rule of thumb $p$ should have values close to 0.8 for small networks and to 0.4-0.5 for very
large networks. The variation in $p$ should take into account the probability for two very similar names to belong to different users. This probability increases as the network grows.

The reason we have decided to implement this service was because we believed that by joining accounts of the same person, the importance of this person in the network will increase and therefore the search results from this person will be better ranked, thus improving search. However this assumption has not been validated yet and it was added to the roadmap.

Figure 9: Screenshot of the tool that allows administrators to search for similar users and to link accounts

2.3.3 The visualisation layer

2.3.3.1 Pedagogical impact

Data: visualisation of (a) semantically annotated learning objects for formal learning; (b) a domain ontology for both formal and informal learning; (c) definitions; (d) relevant peers; (e) formal and informal search results.

Impact on learners: the learner has the possibility to observe the cognitive artifacts, stored as learning objects and ontology in the way, which is more familiar and comfortable to him, or most appropriate for his tasks: for the visualisation of the formal learning objects, the visualisation includes HTML-like view, and a concept map-like view. For the informal learning resource, the user is directed to the original resource. The ontology is either visualized using a tree-like view (FLSS) or a graph-oriented view (iFLSS).

Stakeholders: tutors and learners
Formal/ Informal Learning: both

Relevance with respect to theoretical model: the visualisation is the direct way of connecting the personal learning process with the knowledge that is made explicit within the Community of Practice.

2.3.3.2 Feedback from validation

The validation showed that the visualisation of a system plays a crucial role for its adoption and acceptance. Much of the feedback from the stakeholders was related to the visualisation. Relevant questions that needed to be answered regarding the visualisation are: "Is the system attractive to the learners?" "How are the cognitive artifacts visualized?", "Is it easy to understand?", "What is the added value of the system over existing systems?". The visualisation constitutes the direct connection between the personal learning process and the knowledge which is made explicit within the ontology and the CoP. The CSF contains several types of data, which all need to be visualized in a proper way. The stakeholders have provided feedback on the visualisation of different parts of the system, which has led to the implementation of several improvements in this module.

For the FLSS, the tutors wanted the visualisation of the search results to be improved. In the previous version, the search service only returned the names of the documents that satisfied the search query, while the tutors indicated to prefer getting more information on the document. More specifically, they indicated that they wanted to get insight in the context in which a concept is used and an overview of the concepts present in a document. Apart from the result list, the visualisation of the semantically annotated formal learning objects could be improved as well. These documents were visualized in the previous version using a synchronized view of the learning object together with its annotation. The documents were hard to read in this way, since the learners could see only one sentence at a time and could not easily access the text of the complete document.

The major point for improvement in the case of the knowledge discovery component, concerned the graph fragment. For the majority of the learners, the domain ontology fragment improved their insights in the domain. The learners considered this fragment especially useful when one is a beginner in a domain. An obvious point of improvement of the validated version of the software, was that it was only possible to generate a static fragment for a search query, which then didn't change until the user entered a new query. This meant that, when entering a search request, a graph was generated that only included the nodes that were up to N degrees away from the concepts matching the search query. Browsing was restricted to this fragment. More manual effort was asked from a user that wanted to explore a concept and its relations that was at the end of the static graph. In that case a new keyword search request had to be formulated with that specific concept in order to obtain a new static graph that contained the concept of interest. The learners indicated in the validation that they would prefer a dynamic graph.
A minor issue mentioned in the validation concerned the length of the definitions that were shown in the knowledge discovery component. Some definitions consisted of more than one sentence (generally 2 or 3), which was considered to be too long.

The social learning component provided a list of learning materials and recommended potentially relevant peers to the learner on the basis of a search query. However, it was up to the learner to decide whether these recommendations are actually useful. The validation revealed that some learners had problems deciding whether the peers and documents suggested were to be trusted or not, since the service did not provide feedback on this. Some learners proposed to show the APML profiles of the peers as mouse-over (e.g. as cloud tags) to give the learner a quick and compact impression of their area(s) of expertise.

In Table 9, we summarize the improvements in the ‘Visualisation’ layer, as result of validation and internal testing of the software.

<table>
<thead>
<tr>
<th>Version 1.0</th>
<th>Version 1.5</th>
<th>Aspect</th>
<th>FLSS / iFLSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>document titles in search results</td>
<td>document titles and snippets in</td>
<td>Usability</td>
<td>FLSS</td>
</tr>
<tr>
<td></td>
<td>search results</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XML document view</td>
<td>HTML document view</td>
<td>Usability</td>
<td>FLSS</td>
</tr>
<tr>
<td><em>New functionality</em></td>
<td>Concept list for a document or a</td>
<td>Usability</td>
<td>FLSS</td>
</tr>
<tr>
<td></td>
<td>set of documents</td>
<td></td>
<td></td>
</tr>
<tr>
<td>static visualisation</td>
<td>dynamic visualisation</td>
<td>Usability</td>
<td>iFLSS</td>
</tr>
<tr>
<td>graph creation slow</td>
<td>graph creation much faster</td>
<td>Optimization,</td>
<td>iFLSS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Scalability</td>
<td></td>
</tr>
<tr>
<td>problems with large number of users</td>
<td>software scalability improvements</td>
<td>Scalability</td>
<td>iFLSS</td>
</tr>
<tr>
<td>long definitions</td>
<td>short definitions (max 1 sentence)</td>
<td>Usability</td>
<td>iFLSS</td>
</tr>
<tr>
<td>only one definition for multiple</td>
<td>one definition for each term (e.g.</td>
<td>Usability</td>
<td>iFLSS</td>
</tr>
<tr>
<td>terms (e.g. HTML or JavaScript)</td>
<td>HTML and JavaScript)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>graph fragment for multiple terms</td>
<td>show shortest path from concept1</td>
<td>Usability</td>
<td>iFLSS</td>
</tr>
<tr>
<td></td>
<td>to concept2</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>New functionality</em></td>
<td>widget shows how the searcher is</td>
<td>Usability</td>
<td>iFLSS</td>
</tr>
<tr>
<td></td>
<td>connected to the user returned</td>
<td></td>
<td></td>
</tr>
<tr>
<td>social learning results represented</td>
<td>social learning results</td>
<td>Usability</td>
<td>iFLSS</td>
</tr>
<tr>
<td>by tags and document type</td>
<td>represented by tags and</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>document title</td>
<td></td>
<td></td>
</tr>
<tr>
<td>various bugs</td>
<td>bugfixing</td>
<td>Usability</td>
<td>iFLSS</td>
</tr>
</tbody>
</table>

Table 9: Improvements in the ‘Visualisation’ layer

---

**2.3.3.3 Knowledge discovery support**

The learners indicated that they would prefer a dynamic graph fragment instead of a static one. We have therefore improved the visualisation by dynamically extending the graph during a user's interaction with it. Currently, when a user clicks on any
concept in the graph, the graph is morphed\(^4\) to a new graph that only contains the concepts that are N degrees away from the concept that the user clicked on. This means that older concepts will disappear as the user navigates away from them. An animated graphical transformation informs the user of the new nodes and the nodes that are too far away from the current focus that have been removed.

Figure 10 illustrates the difference between the old version and the new one. With the old version the user would need to formulate a new query for each new graph fragment. In the new version, a simple mouse click is enough to morph the graph into a new one. This greatly improves the potential for exploring the ontology by the user while trying to reduce the complexity at the same time.

In addition the visual complexity of the graph was reduced by only visualizing the path between two or more concepts. These concepts were generated either from a multi-keyword query from the user or a single term that is actually an ambiguous lexicalisation. In the old version the directly associated concepts of the endpoints of the path(s) were displayed along with the path(s) that connected the concepts to each other. Currently only the path between the concepts is shown and nothing else. A user can display the additional concepts that surround the main concept by clicking on it if needed. Clicking a concept will automatically extend the graph with the additional concepts.

Bottlenecks have been identified during graph generation and caching of intermediate results is performed to guarantee satisfactory performance with an increased amount of users. Numerous other small visual improvements have been carried out. One example is the way longer lexicalisations are presented in the graph such as "object-oriented programming language" (as shown in the example). Instead of displaying them on a single line, the words are displayed on multiple lines. This reduces the number of concept lexicalisations that overlap with others.

A minor issue mentioned in the validation regarding the visualisation concerned the length of the definitions that were shown. Some definitions consisted of more than one sentence (generally 2 or 3). However, since the definitions are only intended to support the learner by giving him an idea of the meaning of a concept, they don't need to be very long. We have therefore decided to restrict them in Version 1.5 to a maximum length of one sentence. If the learner wants more information, he can click on 'Read more' to visit the Wikipedia page on the concept.

\(^4\) a technical term, which means: change or cause to change smoothly from one image to another by small gradual steps using computer animation techniques
Figure 10: Visualized domain ontology interaction in V1.0 (old version) and V1.5 (new version)
2.3.3.4 Semantically annotated documents

In version V1.0, the search of formal documents only returned the names of the documents that satisfied the search query. To support the users, three changes have been made regarding the visualisation of the search results and the documents.

First, the search results are now also displayed as a list of snippets, where each of them represents the context of the concepts that have been found. In this way, the users can better evaluate which document is appropriate for their task. Figure 11 demonstrates this new feature. How the snippets are obtained has been described in section 2.3.2.3.1.

A second change concerns a new type of document view, based on suggestions from the tutors in the validation: they wanted to be able to get an overview of the concept annotations in the documents. Therefore, a service has been developed that enables the tutor to see a list of concepts and their frequencies in the document. The number of occurrences can be shown either for a particular document or for a set of documents. For the tutor, these statistics provide evidence for the conceptual coverage of a particular learning object or course. A course is represented as a set of learning objects. Figure 12 shows a table in which the first column contains the concept identifiers from the ontology. The second column is used for the corresponding lexicalization of the concept and the third column shows the number of occurrences of the concept within the learning object.
The third change has been implemented because the users experienced problems reading the documents. The semantically annotated formal learning objects in the previous version were visualized only using a synchronized view of the learning object together with its annotation. In version 1.5, a new type of visualisation of the documents has been created. Compared to the previous version, the XML elements are deleted and the concept lexicalization (or its coreferent) is highlighted in bold. The next screenshot shows such a visualisation of an annotated document.

2.3.3.5 Social learning support

The validation revealed that some learners had problems deciding whether the peers and documents suggested were to be trusted or not, since the service did not provide feedback on this. We did not follow the suggestion from the learners here -- i.e. show the APML profiles of the peers as mouse-over (e.g. as cloud tags) -- since several
experts consider cloud tags not very good from the usability point of view. Jakob Nielsen, for example, tells about cloud tags: "Just because something looks cool, and just because it's a current fad in Web design, don't mean that it's necessarily best for usability". Also in [Hearst & Rosner, 2008] the authors conclude that tag clouds are used as they are appealing and indicators of social activities and that they are inferior to other ways of organizing information for just understanding the information.

In order to solve the problem that the user actually noticed - the lack of trust in results provided by unknown users - we decided to implement a service and a small html+javascript script that shows how the learner connects to the person from which the result came. This way we keep the interface simple and we offer the information needed in order for the user to keep trusting the process.

A minor issue concerned the presentation of the list of learning materials. In version V1.0, the learner had to decide whether a link could be relevant on the basis of a number of tags and an indication of the origin of the resource (YouTube, Delicious, SlideShare) were presented. The learners indicated that they would prefer to see the document title as well. In the current version, this change has been implemented.

### 2.4 Widget integration

WP6 shares a common data model and basic data management services through the Common Semantic Framework (CSF). Services in the CSF are used by both the iFLSS and FLSS for a variety of tasks. User interaction however does not happen at the level of the individual web services, but by a combination of different web services from which the results are combined in one or widgets. The learner interacts with one or more widgets in order to accomplish his or her learning task. This type of user interaction calls for another type of integration on the user interface level that has a different set of requirements. The widgets should communicate with as little delay as possible and should support different compositions of widgets without a learner having to worry about the compatibility of the various web services that are in the background.

We have implemented widgets according to the emerging W3C standards which can communicate using a widget communication framework, as implemented within WP2 for the Wookie widget container. The inter-widget communication allows interactions which take place in one widget to affect the state of other widgets in the same window almost immediately. This means for example that a keyword search in one widget can update the ontology visualisation, document search and definition retrieval widgets simultaneously. This greatly enhances the user experience, because only a single action is required instead of entering the same search query repetitively in multiple widgets in order to update their state. Another advantage of using the widgets is that it allows the user to construct a Personalised Learning Environment (PLE) as thoroughly covered in D2.3. This allows the learner to remove widgets that display results not relevant to his or her current learning task.

---

[5](http://www.useit.com/alertbox/tag-clouds.html)
The following widgets were developed and integrated with Wookie & ELGG making use of the developed Inter-Widget-Communication framework within WP2:

- Keyword search (either concepts, people, resources or tags)
- Widgets for displaying live keyword based search results from YouTube, SlideShare, Delicious, Bibsonomy including ontology based filtering.
- Formal search results
- Social search results
- Social people search results
- Social personalised recommendations
- Definition retrieval
- Ontology visualisation
- Learner profile
- Annotations

The individual results have been completely decomposed and can be recombined by learners in new ways that best fit the learning task.
3 Verification

The revisions to the implementation of the CSF have been discussed in the previous chapter. In order to test how well the different services perform, a number of verification experiments have been run, which are described in this chapter. More specifically, we have (a) verified the accuracy of the document annotation (section 3.1), ontology enrichment and concept mapping services (section 3.2), (b) assessed the quality and suitability of the (formal and informal) resources and peers suggested by the CSF (section 3.3), (c) investigated the size of the social network (section 3.4) and (d) conducted experiments to test the scalability and performance of the software with large numbers of users and verified the speed of the social network crawler (section 3.5).

3.1 Document annotation

The verification of the FLSS semantic annotation service aimed at evaluating the output of the three versions of the language pipe (see section 2.3.2.2.3). The verification was carried out in the period 1-5 October 2010. The motivation for this comparison is that the user needs to know how many errors he can expect when using a specific language pipe version. Only Pipe 0.3 was available in version V1.0 of the software. Some users reported during the validation of this pipe that the annotation process was too slow. Therefore, the semantic annotation service has been split into different language pipes. The verification was performed by comparing the output of the three pipes with each other and with the gold standard. We selected three learning objects in the sub domain of HTML and annotated each of them using all three pipes.

The results are as follows: Pipe 0.1 annotated 316 concepts; Pipe 0.2 annotated 282 concepts; and Pipe 0.3 annotated 299 concepts. The set of annotations produced by Pipe 0.3 contains the set produced by Pipe 0.2. Pipe 0.3 assigned 17 concepts to pronouns by means of co-referential chains, and 11 terms were annotated with more specific concepts. For the 11 more specific concepts, we considered the old annotations when we compared Pipe 0.2 and Pipe 0.3. This comparison shows that usage of Pipe 0.2 reduces the number of the annotated phrases in the text, but is comparable with Pipe 0.3 concerning the concept coverage. Pipe 0.1 wrongly annotated 34 concepts compared to the output of Pipe 0.2. From these 34 concept annotations, 6 concepts are unique. The error rate is 10.7%.

Thus, the conclusion from the verification is:

- Pipe 0.3 produces as much accurate annotations as Pipe 0.2, but also adds new annotations on the basis of co-referential chains.
- Pipe 0.2 produces annotations without the ones due to the coreferential chains;
- Pipe 0.1 produces annotations which include all the correct annotations from Pipe 0.2. It also produces wrong annotations. According to this experiment, the errors are less than 15%.
The user has the possibility to select the most appropriate pipe for the task at hand. If he wants to get a quick overview of the conceptual content of the learning object, then the Pipe 0.1 is most appropriate, while Pipe 0.3 might be used in cases where the accuracy is more important. Pipe 0.3 can be used offline in situations where the text to be annotated is large. Pipe 0.2 can be seen as an intermediate solution: it is faster than Pipe 0.3, and produces more reliable results than Pipe 0.1.

3.2 Ontology

3.2.1 Ontology enrichment

The goal of the verification of the ontology enrichment process (discussed in section 2.3.2.1 and in D6.2) is to test whether the automatically added concepts and relations are correct and whether they overlap with manual enrichment. The verification has been performed in April 2009 using the 1.0 version of the ontology enrichment pipeline [Monachesi & Markus, 2010].

The ontology that has been enriched is the LT4eL domain ontology on computing [Monachesi et al., 2008], which was last updated in 2008. This is an excellent candidate for ontology enrichment due to the fast-evolving computing domain. The ontology enrichment methodology uses a similarity measure (resource cooccurrence) to acquire related terms based on a social data set crawled from Delicious. This data set contains 598379 resources, 154476 users and 221796 tags on a wide range of subjects, but with an emphasis on computer related terminology (see D6.2 for more details).

In order to evaluate our ontology enrichment methodology, we have compared three different ontologies:

1. the original LT4eL computing ontology with the related English lexicon (1200 classes);
2. a manually enriched ontology which takes the LT4eL one as basis (1336 classes and 1672 lexical entries). This is our gold standard.
3. The automatically enriched ontology, which takes the original LT4eL ontology as a start (2016 classes and 2325 lexical entries).

A first analysis of the lexical differences between (1) and (2) shows a difference of 80 lexicalisations. The aim of our evaluation was to assess whether the automatic enrichment process would add lexicalisations (and related concepts) that overlap with manually added lexicalizations given a similar sub-domain. The automatically enriched ontology has been generated by considering each co-occurring tag in our Delicious data set as eligible for enrichment. Even though we considered every co-occurring tag as eligible for use in ontology enrichment, the lexical overlap between the manually enriched ontology and the automatic one is minimal. More specifically, 69 terms which have been added manually to the LT4eL ontology are multi-word units and are not attested in Delicious. They are representative of the expert view of the domain given their level of specificity and include terms such as: "NMTOKEN
attribute", "XML element type declaration", "XML attribute list declaration". The remaining 21 terms are attested in Delicious but only 13 of them are generated by the similarity measures and are attested in DBPedia.

Regardless of the minimal lexical overlap between the manually and the automatic enriched ontology, it is not the case that the terms added automatically are inappropriate misplaced in the ontology, as the following evaluation that filters upper ontology concepts reveals:

- Total number of unique statements: 1265
- Accurate enough for ontology enrichment: 1010
- Too inaccurate: 255

This brings the amount of usable additions to about 80%. We have analyzed the added relations further and will divide the results into two separate groups:

- 598 triples which use the ltfl:related relation. This relation only indicates that two terms are related but it doesn’t say in which way.
  - Correct: 497 (83%)
  - Incorrect: 101

- 667 Clear ontological relations (rdfs:subclassof or either DBpedia specific ones)
  - Correct: 513 (77%)
  - Incorrect: 154

The comparison between the manually and automatically enriched ontology showed that there is minimal overlap between them and that the enrichment results are of appropriate quality. The automatically enriched ontology includes the vocabulary of the community of users, while manually enriched ontology includes very specialized terms provided by a domain expert. It is exactly this complementarity that we wanted to achieve by embedding tags into an existing ontology and that we want to exploit in eLearning applications.

### 3.2.2 Concept mapping

In order to assess the efficiency of the concept mapping (discussed in section 2.3.2.1.3) the following verification experiment was performed on 31 September 2010. The goal of the concept mapping is to associate the most likely concepts (word senses) from a reference ontology (such as DBpedia) to the concepts from a domain ontology – the LT4eL ontology in this experiment. Mapping domain ontology concepts to those in a reference ontology discloses wider access to other information, such as unambiguous retrieval of definitions (section 2.3.2.3.1). It is, however, unavoidable that there are parts of the domain ontology which are just too specific when compared to a reference ontology. As a consequence, only part of the domain ontology will have overlapping concepts with a reference ontology due to differences in domain specificity. We expect this to show up in the verification experiments.
The experiment has been performed by considering 200 randomly selected ontology concepts from the LT4eL ontology (before its enrichment) on computing. We manually determined for each concept mapping whether it was:

- Incorrect - An unrelated concept has been selected
- Acceptable, a strongly similar concept was selected
- Correct - An equivalent concept was selected
- Unmapped - A concept is said to be unmapped when the algorithm didn't associate a concept with it. For example when there is no lexical overlap between domain and reference ontology lexicalisations.

For each concept the directly associated concepts in the domain ontology connected through any relation up to 2 steps away from the seed concept have been considered as the disambiguation context.

Figure 14: Concept mapping results: (a) mapping of 200 random samples; (b) reasons for unmapped concepts

The results show that the majority of concept mappings is correct and that this provides useful feedback, even though not perfect. Further evaluation of the unmapped concepts confirmed our hypothesis about the domain specificity of some of the terms. Correctly mapping a concept to another concept automatically entails that the definition that we retrieve from DBpedia is also correct. This is due to the fact that definitions are currently retrieved using concepts as opposed to lexicalisations, which improves the quality of definition retrieval for synonyms. Improvements to the concept mapping will therefore lead to improved performance during definition retrieval. These verification results are therefore interchangeable with experiments about the performance of definition retrieval from reference ontologies given a domain ontology concept. The results are quite satisfactory for definition retrieval and ontology based conceptual filtering of search results.
3.3 Quality of results

3.3.1 Semantically annotated resources

The ontology-based search verification aims at assessing the relevance of the returned (formal) learning materials for a query. The verification was organized in the period 27 September 2010 to 1 October 2010 as a workshop with five tutors from IPP-BAS using version 1.5 of the FLSS services. The tutors have been divided into two groups of respectively two and three tutors.

The first group was shown a list of learning materials related to HTML available within the FLSS. They were asked to choose three topics within the HTML domain and to associate these with all the relevant learning objects on this topic in the FLSS. The topics selected were "tables in HTML", "fonts in HTML" and "images in HTML", obtaining thus a gold standard consisting of the relevant learning materials related to these three topics.

The second group was asked to formulate queries with respect to one of the topics specified by the first group and to perform an ontology-based search for relevant material. The retrieved materials were then automatically compared to the gold standard for each topic and the precision and recall were calculated. The recall is the number of retrieved gold standard documents divided by the total number of gold standard documents for the relevant topic. The precision is the number of retrieved gold standard documents divided by the total number of retrieved documents.

The gold standard documents are relevant to the general topics that were defined by the two tutors. The more specific the query, the lower the precision could be, since a smaller amount of documents from the gold standard has an impact on the query, in case of specific queries. The recall can increase for the same reason. We have therefore performed additional checks to see whether the precision decreases with more specific queries. We did this check for queries for which the precision was higher than 50%, because we assume that a precision of 50% means that the results are satisfactory.

Our hypothesis was: If a query expresses a general concept the precision will be low and recall high. With more specific queries which denote more specific concepts, the precision improves, but recall decreases. Better results can be achieved in cases where specific queries are formulated. General queries express more generic concepts, which usually lead to a larger set of results. Specific queries are defined by more specific concepts or denote more than one concept. For example, the query "lt4el:Table" is more general than the query "lt4el:TableTag" because the "lt4el:Table" denotes many kinds of tables - in HTML, MS Word, LaTeX, etc. The "lt4el:TableTag" denotes the HTML tag for a HTML-table and is thus related to a subset of the results for "lt4el:Table". The example illustrates that more specific queries are not necessarily defined by more than one concept. Table 10 shows the average results for all queries.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>lt4el:Table</td>
<td>16.5%</td>
<td>93.8%</td>
</tr>
</tbody>
</table>
As mentioned above, in the case of precision over 50%, we checked the relevance of the corresponding documents manually. It turned out that 8 of the 12 queries resulted in a precision higher than 50%. Following the assumption that more than 50% precision is a good measure for the semantic search the conclusion of this verification is that the FLSS semantic search could support the user in the retrieval of relevant learning materials. The users have to be instructed to start with general queries when they are not sure about the specific topic of interest. They then need to gradually make their queries more specific on the basis of the annotation within the returned documents and by using the ontology.

### 3.3.2 Informal resources

Assessing the quality of the informal learning resources as provided by social media retrieved through knowledge discovery is crucial in order to determine whether such a system has educational value. We need to determine whether the quality of resources, although noisy, is high enough to add value to the learning process. We have therefore determined the precision (relevancy of the search results) and recall (how many of the relevant search results are retained after filtering) in order to get a sense of the utility of the social services and the live integration of search results combined with concept filtering (section 2.3.2.1.3) in practice. The experiments have been performed between 21 and 22 September 2010.

Empirical testing has shown that YouTube has the smallest amount of suitable learning material in the first 20 search results which are relevant in the context of the LT4eL domain ontology on computing. I.e., a query for ‘ajax’ will include a lot of soccer-related resources, whereas SlideShare, Delicious and Bibsonomy all return resources about the web technology word sense. We have therefore run our verification using the YouTube service to determine the relevance of learning resources as returned by this service. YouTube is thus the worst case performance that our system is likely to encounter. On the other hand YouTube is also the social service which could benefit most from concept filtering using a domain ontology.

To test the concept filtering algorithm, the first 20 search results of five single keyword queries and five multi-word queries have been independently determined to be either relevant or non-relevant by two domain experts. Since the filtering algorithm
aims at distinguishing between correct and incorrect senses of ambiguous terms in a certain context, three ambiguous terms have been selected as queries: "java", "soap", and "python". In addition, two non-ambiguous terms have been tested as well: "html" and "website". Five combinations of two term queries have been analysed to verify the results of the algorithm on multi-word queries. The queries combine two ambiguous terms ("python java") an ambiguous and a non-ambiguous term ("soap xml", "java programming") or two non-ambiguous terms ("html website", "html xml"). Table 11 and Table 12 show the average precision and recall scores obtained.

<table>
<thead>
<tr>
<th></th>
<th>Expert 1</th>
<th>Expert 2</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision - unfiltered</td>
<td>0.63</td>
<td>0.56</td>
<td>0.60</td>
</tr>
<tr>
<td>Precision - filtered</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>Recall - filtered</td>
<td>0.80</td>
<td>0.85</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Table 11: Concept filtering using one term queries

<table>
<thead>
<tr>
<th></th>
<th>Expert 1</th>
<th>Expert 2</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision - unfiltered</td>
<td>0.81</td>
<td>0.84</td>
<td>0.83</td>
</tr>
<tr>
<td>Precision - filtered</td>
<td>0.83</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>Recall - filtered</td>
<td>0.93</td>
<td>0.91</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Table 12: Concept filtering using multiple term queries

As expected, the suitability of the resources for more specific queries (more than one keyword) increases. We also see that concept based filtering increases the quality of the results, because of the ambiguities of the simple queries. These ambiguities disappear with more specific queries and as expected the additional improvement attained with concept filtering is minimal.

3.3.3 Social resources

We have performed a qualitative analysis of the results returned for two users of the platform - both social networking users with multiple accounts (Delicious, YouTube, SlideShare). This analysis represents a verification for the social search module and was performed at the end of September 2010. The results returned depended strongly on the connections of the users in the social network and on the content posted by these ones. The first user - Vlad is interested in semantic web while the second - Eline is interested in machine learning. Eline added Vlad as a friend in at least one social network while Vlad didn't add Eline as a friend. The results that Eline gets when searching for "semantic web" are related to the subject but not as good as the results returned when Vlad searches for the same topic. The fact that Eline gets results related to the subject is due to his connection to Vlad. The situation changes when they search for "machine learning". Eline gets better results because she is better connected with persons interested in this topic while Vlad who is not connected to Eline or to other user interested in machine learning gets no results at all.

This experiment shows that the software is able to return very good results (all the results are relevant for the domain) but only if the network of the user contains the resources needed. Also according to the network some relevant results are also returned in the native language of the user (Dutch for Eline and Romanian for Vlad).
This experiment will be extended in the validation phase where we will focus on the quality of results for a larger group of learners, mostly undergraduates.

<table>
<thead>
<tr>
<th>keywords/users</th>
<th>Vladspea</th>
<th>elinevenerhoudt</th>
</tr>
</thead>
<tbody>
<tr>
<td>machinelearning</td>
<td>no results as no relevant users are in the user's network</td>
<td><a href="http://www.youtube.com/watch?v=GB8mVppsc4z">http://www.youtube.com/watch?v=GB8mVppsc4z</a></td>
</tr>
<tr>
<td></td>
<td><a href="http://staff.science.uva.nl/%7Ecroftz/#twartb">http://staff.science.uva.nl/%7Ecroftz/#twartb</a> ools</td>
<td>page of a professor in a Dutch university with many interesting articles (in Dutch)</td>
</tr>
<tr>
<td></td>
<td><a href="http://rekenhule-basisschool-pabo.nl/">http://rekenhule-basisschool-pabo.nl/</a></td>
<td>page with resources on teaching computing (in Dutch)</td>
</tr>
<tr>
<td></td>
<td><a href="http://www.talenteekenen.nl/">http://www.talenteekenen.nl/</a></td>
<td>page on math and learning education (in Dutch)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>semanticweb</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.ted.com/talks/ho">http://www.ted.com/talks/ho</a> m-bearers-lee-on-the-next web.html</td>
<td>talk from the inventor of the web and of the semantic web about the semantic web</td>
</tr>
<tr>
<td><a href="http://pipes.deri.org/">http://pipes.deri.org/</a></td>
<td>semantic web app</td>
</tr>
<tr>
<td><a href="http://dheer.net/">http://dheer.net/</a></td>
<td>semantic web recommender system</td>
</tr>
<tr>
<td><a href="http://www.slideshare.net/taddod/dataincubator">http://www.slideshare.net/taddod/dataincubator</a></td>
<td>semantic web tool presentation</td>
</tr>
<tr>
<td><a href="http://www.slideshare.net/taddod/howing-date-through-the-talls-platform-presentation">http://www.slideshare.net/taddod/howing-date-through-the-talls-platform-presentation</a></td>
<td>semantic web tool presentation</td>
</tr>
<tr>
<td><a href="http://www.slideshare.net/taddod/dating-date-through-the-talls-platform-presentation">http://www.slideshare.net/taddod/dating-date-through-the-talls-platform-presentation</a></td>
<td>semantic web tool presentation</td>
</tr>
</tbody>
</table>

Table 13: Social search results

3.4 Size of the social network

A very relevant question asked about our software was - do people actually use social networking sites for learning purposes enough to take advantage of the content created in our learning application? Does the social network offer sufficient relevant items? Does this service help the learner or the tutor actually gain time? This verification aims to address some of this issues, mostly formulated during the review meeting. The verification was performed in July 2010.

We tried to show that the social networking sites are used for learning by showing how actual tutors from different universities have accounts on social networks and we examined their activities using our crawler.

The experiment is further described in [Stoica et al., 2010] and the results can be summarized as following. We crawled the accounts of 11 professors or PhD students in European universities in the field of Computer Science and Technology Enhanced Learning. The crawling gathered data about the tutors, the peers and the peers of the peers. After the initial crawl of these tutors’ networks we have monitored the networks of the tutors using our crawler over a period of 15 days. From the initial...
crawl and the 15 days monitoring, we have obtained a number of 100,000 RDF triples, 456 users accounts, 5886 resources, 4633 being bookmarks on delicious, 497 presentations on slideshare.net, 200 images on flickr, 58 tweets and 205 videos. The main common characteristic of these resources is that most of them are tagged. The tagging statistics are the following: There were 19765 distinct taggings with 6933 distinct tags, most of these being on Delicious (75%), then on Youtube and Slideshare (11% each). The low number of resources (5886) collected is partly because all the APIs return only a small number of the last resources posted by the user (50-100). We analyzed also the evolution in time and we discovered that in the network of each tutor we had 4 new items posted every day (47 items/day and 11 tutors). The table that shows the evolution of the network is given below.

<table>
<thead>
<tr>
<th></th>
<th>Initial number</th>
<th>Final number (after 15 days)</th>
<th>Avg. number of items/day</th>
</tr>
</thead>
<tbody>
<tr>
<td>User relations</td>
<td>2973</td>
<td>4867</td>
<td>126.26</td>
</tr>
<tr>
<td>Resources</td>
<td>5180</td>
<td>5886</td>
<td>47.06</td>
</tr>
<tr>
<td>Tags</td>
<td>6205</td>
<td>6933</td>
<td>48.53</td>
</tr>
</tbody>
</table>

Table 14: Evolution of the network

The growth in the number of users seem to be very high when comparing with the growth of resources. This might be due to the fact that during the time interval when the experiment was carried out, some user with a very large network, with very few active users, was added by the tutor. This explains the huge increase in the average number of users and the rather slower increase in the number of resources. The preliminary numbers found at this step indicate that tutors do not use content sharing sites very much. They have an average of less than 10 connected peers per account and their networks have less than 100 resources per account. However, the content that they generate is enough to represent an important source of learning objects for the average learner.

This experiment shows that a small and periodic investment of time (posting his presentation online, linking to other tutors, bookmarking interesting content) from the part of a tutor can provide valuable resources to his network of peers, resources that can be retrieved and that can be put to good use by our tools.

Most of the educational software on the market requires effort to set up at least from the tutors’ part. Our approach takes advantage of the power of the social networking sites to ease the workload on the tutor and still offer many valuable results for the learner. The experiment alone can not prove that the results returned improve learning, this will be verified in the validation, it just aims to show that the tools can help the tutor gain time.

### 3.5 Performance

#### 3.5.1 Scalability of the knowledge discovery component

Performance testing of the knowledge discovery component has been carried out using the Apache Jmeter application. This tool is able to simulate web service
requests performed by a configurable number of users. This allows us to determine whether the application's response times are acceptable when many users employ the knowledge discovery component at the same time. The experiment was carried out in the period 2-3 September 2010.

The tool was configured with the following parameters:

- Maximum amount of simulated concurrent users: 100
- Ramp up time: 500 seconds (time until full system load is achieved)
- User interactions were simulated through gaussian random interactions with an average of 5.0 seconds and a 1 second standard deviation
- Five different types of user interactions were sampled. Each user interaction consists of three separate web service calls:
  1. Generation of an ontology fragment which gets visualized as a graph (graph)
  2. Retrieving a definition for the currently selected seed concept or search term (definition)
  3. Resolving the search term or lexicalisation of the concept to a concept from the domain ontology (concepts)

The following table illustrates the response times of the various web services and their parameters.

<table>
<thead>
<tr>
<th>Label</th>
<th># Samples</th>
<th>Average</th>
<th>Min</th>
<th>Max</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>graph: excel word</td>
<td>190</td>
<td>110</td>
<td>65</td>
<td>4082</td>
<td>290.41</td>
</tr>
<tr>
<td>graph: xml</td>
<td>200</td>
<td>68</td>
<td>57</td>
<td>879</td>
<td>59.85</td>
</tr>
<tr>
<td>graph: java</td>
<td>199</td>
<td>115</td>
<td>70</td>
<td>1168</td>
<td>80.87</td>
</tr>
<tr>
<td>graph: web html</td>
<td>208</td>
<td>67</td>
<td>58</td>
<td>1166</td>
<td>77.23</td>
</tr>
<tr>
<td>graph: haml html python</td>
<td>206</td>
<td>97</td>
<td>64</td>
<td>1577</td>
<td>119.13</td>
</tr>
<tr>
<td>concepts: excel word</td>
<td>168</td>
<td>69</td>
<td>56</td>
<td>1466</td>
<td>108.52</td>
</tr>
<tr>
<td>concepts: java</td>
<td>174</td>
<td>62</td>
<td>55</td>
<td>193</td>
<td>18.91</td>
</tr>
<tr>
<td>concepts: haml html python</td>
<td>185</td>
<td>64</td>
<td>56</td>
<td>901</td>
<td>62.10</td>
</tr>
<tr>
<td>concepts: web html</td>
<td>176</td>
<td>67</td>
<td>56</td>
<td>486</td>
<td>41.70</td>
</tr>
<tr>
<td>definition: java</td>
<td>153</td>
<td>120</td>
<td>58</td>
<td>8774</td>
<td>702.02</td>
</tr>
<tr>
<td>definition: haml html python</td>
<td>157</td>
<td>109</td>
<td>58</td>
<td>16478</td>
<td>1305.87</td>
</tr>
<tr>
<td>concepts: xml</td>
<td>175</td>
<td>60</td>
<td>56</td>
<td>130</td>
<td>8.76</td>
</tr>
<tr>
<td>definition: web html</td>
<td>150</td>
<td>150</td>
<td>58</td>
<td>13604</td>
<td>1082.92</td>
</tr>
<tr>
<td>definition: xml</td>
<td>150</td>
<td>65</td>
<td>57</td>
<td>231</td>
<td>22.12</td>
</tr>
<tr>
<td>definition: excel word</td>
<td>144</td>
<td>295</td>
<td>57</td>
<td>33603</td>
<td>2785.34</td>
</tr>
<tr>
<td>TOTAL</td>
<td>2647</td>
<td>102</td>
<td>55</td>
<td>33603</td>
<td>2797.09</td>
</tr>
</tbody>
</table>

Table 15: Response times of the various web services and their parameters

Each row corresponds to one of the web service calls (graph, definition, concepts) for each of the five different user interactions. The label column shows the type of service that was called (graph, definition, concepts) followed by the arguments. The number of samples shows the total number of requests made of that specific request to the respective web service. The numbers in the table are the number of milliseconds it took in best case (see column ‘Min’ in Table 15), on average (column ‘Average’) and worst case (column ‘Max’) for a web service to respond to the request. These results have been produced from a cold start, meaning that no information was cached in advance. The maximum response request times were all within the first calls to the web service due to the unavailability of cached results. The results show that there are
initial slowdowns due to the cold start, but that these are quickly resolved for subsequent access through proper caching. Further access to the system is quite swift with average response times well under a second.

### 3.5.2 Speed and accuracy of the crawler

The crawler that extracts data from the social networking sites has been improved to distribute crawling tasks between multiple machines. The challenge for this crawler is to be able to get as much data from the social networks as possible without disturbing the servers that offer the APIs for social networking. Also the crawler should be able to fetch data about the users in a relatively short time - in order not to miss relevant actions of active users. The way to solve these problems is to be able to run the crawler in a distributed environment in order to distribute the crawling tasks between more computers. This will improve the number of users fetched/day and will decrease the risk of being blocked by the social networking sites. The system gathers all the relations with other peers and all the resources posted by the users and his peers on 2 levels (peers and peers of peers). In this context, by relation we understand the existence in the social network of a link from one user to another with the significance that the first user follows the actions of the second on the site. By level we understand the number of relations between users.

The experiment showed that the time needed for exploring only one level of the network is higher using the distributed crawler. This is the case because the overhead introduced by the distributed crawler doesn't pay off with small numbers of users. However for larger numbers of peers (for example for two levels of peers instead of one), the time improvement is significant (over 50%)

The following table shows the time difference for getting all the data for a single user using a distributed system and a serial system. The users on the first level are linked directly to our central user while the ones on the level 2 have an intermediate user between them and the central user. The results posted below represent the time needed for a distributed crawler and for a serial crawler to extract data around an average user. This experiment was performed to verify the crawling module in September 2010.

<table>
<thead>
<tr>
<th></th>
<th>Serial time (s)</th>
<th>Distributed time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 level crawling</td>
<td>20.3</td>
<td>31.2</td>
</tr>
<tr>
<td>2 levels crawling</td>
<td>90.1</td>
<td>70.7</td>
</tr>
<tr>
<td>3 levels crawling</td>
<td>237.4</td>
<td>165.0</td>
</tr>
</tbody>
</table>

**Table 16: Performance of serial and distributed crawler**

The experiment shows that the crawler can run on a single machine and in a distributed environment as well as the benefits of it running in the distributed environment.
4 Next steps

The previous chapters have presented the implementation of version 1.5 of the CSF and the verification of the software. To investigate how the improved services contribute to the learning and teaching process, the enhanced FLSS and iFLSS will be validated using a scenario-based approach (section 4.1). Compared to the previous validation, two changes will be implemented in the design of the experiments. First, in contrast to the previous validation round, which was a short term validation, the learners and tutors will be asked to use the software over a longer period of time. Furthermore, to strengthen the results and to make it possible to draw stronger conclusions, the scenario will be run with a larger group of participants.

The services that have been developed in the LTfLL project (described in D5.3 and D4.3) all aim to enhance different aspects of the learning process in a lifelong learning context. In the validation, we have shown already how the CSF can support the formal and informal learning process. In the last phase of the project, we will be working on the integration of LTfLL services from different WPs to illustrate how they can be used together to enhance the learning process. To this end, two possible use cases of the services are currently being worked out in the so-called short (integration with WP4.1) and long (integration with WP4.2, WP5.1 and WP5.2) threads (section 4.2).

Not only the integration with other services is relevant, but also the transferability of the software to other domains and institutions is an important aspect for future adoption and extensibility of the tools and methodologies developed. It is thus important that the barriers to adoption in either new institutions or novel domains are timely identified. This can guide the development of both software and training material to address future requirements. In section 4.3, a number of technical, pedagogical and organisational points are described which may affect the transferability of the services.

Although we managed to implement a lot of new features and improved on existing ones, a number of suggestions from the validation feedback had to be moved to the roadmap (section 4.4). Some of these issues will be addressed during the last phase of the project, while others will be dealt with after the project has ended (e.g. during PhD projects).

4.1 Validation

4.1.1 Validation of the FLSS

In contrast to the previous validation round, where only the separate steps have been tested of the scenario, in the next one the tutors will validate the FLSS by developing a course unit on a chosen topic. This time two groups of tutors will be involved - 5 people from IPP-BAS and 5 people from the Sofia University. First, the tutors will be introduced to the system and their task. The tutors will be shown a list of topics in the IT domain within FLSS (for example: Formatting in HTML or Elements and
Attributes in XML), and will be asked to select a topic for construing a course. After familiarization, the tutors will be left for two weeks to design the course. At this time, all the necessary assistance will be provided. Instruments will be questionnaires and focus groups/interviews. Additionally, the services will be popularised on an open dissemination seminar at IPP-BAS.

4.1.2 Validation of the iFLSS

In the previous validation scenario, the learners used the iFLSS to acquire knowledge on website building and to identify content and peers that could help them to create such a website. The scenario had been designed especially for the validation. The current version of the software has improved considerably compared to the previously validated version and seems to be ready for inclusion in a real learning environment. The iFLSS tools will be integrated in the Moodle platform of the Politehnica University of Bucharest and used for a course with 125 students. The tools will be used to help the students visualize and understand the concepts and they will also be used to allow the students to find resources from their tutors' social networks. Also they will be able to personalize the social search if they use social networks themselves. The students will be free to use our tools and we will evaluate their usage and also the effect the usage has on them. We will also measure the adoption rate for our tools and we will compare the results of the students that used our tools with the results of the students who did not use our tools. Afterwards the students who adopted the tools will be asked to answer a questionnaire and participate in a focus group. This experiment will be carried out in a similar way at Utrecht University.

4.2 Threads

4.2.1 Short thread

The short thread integrates the concept annotation and search services provided by FLSS with the positioning services in task 4.1. The positioning services are based on the notion of questionnaires that require the learner to have a certain degree of knowledge in order to answer the questions.

4.2.1.1 Integration with 4.1

The 4.1 services provide support for the tutor to prepare a questionnaire as well as to check the answers from the learners. Additionally, they support the learner in receiving live feedback on the relevance and exhaustiveness of their answers. It is necessary to access learning materials and conceptual information for each of these tasks. The development of the integration of the 4.1 services and FLSS will focus on providing access to this information.

The tutor has the possibility to search for learning materials during the preparation of a questionnaire. The tutor can also add new learning materials, suggested by the services in task 4.1, to the FLSS repository (which is part of the CSF). Finally the learner can access the relevant learning materials based on the live feedback.
There are two main interactions between the services:

1) Search for learning material

- Input for FLSS services:
  - query for the semantic search

- Output generated by FLSS services
  - list of relevant documents

- Changes required in the implementation of FLSS services:
  - none

2) Addition of new learning material

- Input for FLSS services:
  - learning object

- Output generated by FLSS services
  - annotated document

- Changes required in the implementation of FLSS services:
  - when the tutor adds manually new annotations, the additions are stored for later inspection in order to be added as new lexicalizations to the lexicon

4.2.2 Long thread

The iFLSS offers data and support to all the other tasks in the long thread. The purpose of the iFLSS is to supply the other modules with links and explanations gathered from the social networking sites, whenever the modules request it. In this section, we describe the necessary integration between iFLSS and the other tasks (5.1, 5.2, 4.2) and address how the integration will be performed.

4.2.2.1 Integration with 5.1

One of the challenges of the 5.1 Polycafe tool is to offer information about issues that the learner did not address in his chat. In order to do that, the system identify key terms that have not been mentioned or that have not been mentioned enough by the learner. These terms are displayed in one of the 5.1 widgets. This widget will also call the iFLSS services using these terms and will provide learning resources regarding the missing terms. The integration will be very simple, the 5.1 widgets only needing to send the request containing the query terms to the ontology-based search service and the query terms and the username to the social search service. In order to take advantage of the social search services the user should already use social networking tools and has to be registered as a user of the social search service. This integration doesn't require any more development from our workpackage as we just provide results for the requests coming from a widget.

- Input for iFLSS services:
4.2.2.2 Integration with 5.2

The integration of the social learning tools will be made by improving the 5.2 Pensum tool with document search functionality. Within Pensum the learner will perform a search query which will be executed by iFLSS search services. This search request will return only text documents (neither Youtube videos, nor Slideshare presentations can be processed by Pensum). The learner will select one of the results returned by the search move to Pensum to process the resource.

- Input for iFLSS services:
  - list of terms
  - optional - user id (OpenID)
- Output generated by iFLSS services
  - list of links containing only documents from given social networks (delicious, bibsonomy)
- Changes required in the implementation of some iFLSS services:
  - filter added to the search service to return documents only from a given social networking site

4.2.2.3 Integration with 4.2

The integration with Conspect happens similarly to the integration with 5.1’s Polycafe. Conspect generates lists of terms that are either points of overlap with others or, more importantly, 'conceptual' differences. The learner is then able to click on a term for which he/she would like to receive more information about. Depending on the 4.2 implementation, the learner can receive a definition of the term, the visualisation of the ontology regarding the term or a list of resources retrieved through the social and semantic search. The results are obtained by calling the iFLSS services from the Conspect widget.

- Input for iFLSS services:
  - term
  - optional - user id (OpenID)
- Output generated by iFLSS services
  - list of links containing only documents from given social networks (delicious, bibsonomy)
  - definition of term
A generic change that affects all work packages is the need for a centralized system for both identification using OpenID and authorisation as provided by the LMS. This affects interaction between iFLSS and all the other work packages.

### 4.3 Transferability

Transferability is an important aspect for future adoption and extensibility of the tools and methodologies developed. Timely identifying the barriers to adoption in either new institutions or novel domains can guide the development of both software and training material to address future requirements. We will now shortly address a number of identified technical, pedagogical and organisational points which affect transferability of the services.

#### 4.3.1 Technical

- The ontology-based search and document annotation depend on the availability of an appropriate domain ontology, lexicons aligned to the domain ontologies in the required language(s), and suitable learning materials. We assume that there are enough NLP resources for most languages and domains already. The effort for transferring such tools into the format of semantic annotation grammars and linguistic pipes will depend on the original representation of the resources and their availability.

- In order to assess the transferability to a completely new domain, a pedagogically suitable ontology (note that this is quite different from a formally correct ontology for a given domain) needs to be created and loaded into the system. The ontology doesn’t necessarily need to be big given that the ontology enrichment service could improve on its size if social data in the given domain are available.

- The concepts from the domain ontology need to be actively used in social media in order for the social ontology enrichment and social search to work appropriately.

- The development of the crawler needs to be carried out on multiple servers, especially when using the distributed crawler in order to achieve acceptable performance when crawling social networks. We recommend using cloud computing services for the development of the crawler due to their scalability.
• At present, the graph-based ontology JavaScript visualisation only works in browsers with proper HTML-canvas support. In practice, this means that the visualisation only works in Firefox, Chrome and Safari. Microsoft Internet version 9 (currently in beta version) should bring increased support for HTML canvas and will probably be able to display the visualisation. We consider this technical transferability issue as temporary.

4.3.2 Pedagogical

• Learners need to be familiar with the use of social media websites and their inherent rules, advantages and drawbacks in order to appropriately use iFLSS.
• Tutors have to integrate the social content into their courses and to encourage research on the internet for new and relevant learning materials as opposed to a curriculum exclusively based on pre-determined formal resources.
• The courses have to encourage collaboration and networking between learners, tutors and members of the relevant Community of Practice.
• The iFLSS is best suited for exploratory or self-directed learning settings. It is less suited for curricula where a strict chronological order of the learning materials is enforced.
• The social search component of the iFLSS is dependent on a culture of trust and open sharing of content. The added value of the system will decrease with a reduction in either availability of public profiles (due to stricter privacy settings) and or a lack of accessible social structure in social networks. The debate with respect to the openness of social networks and the privacy challenges that this involves is currently still a matter of hot debate.

4.3.3 Organizational

• Stakeholders should differentiate between their dedicated LMS and the (i)FLSS. They should view the architecture of the (i)FLSS as a useful complementary service to the LMS.
• Management would need to provide long term commitment to the maintenance and extension of domain ontologies which cover the mean areas of interest. While the ontology enrichment service provides a means to enrich an existing domain model with socially relevant concepts, tutors should be in control of what domain knowledge is relevant in a specific course. This functionality can be realized through already existing ontology editing tools such as Protegé or others.
• Management needs to have a policy about the integration of social media into pedagogical practice and the proper application of it during teaching. Such a policy would need to address the use of social media for personal and institutional use through, for example, separate certified accounts.
• Learners have to already use content sharing websites for education use as well as tutors.
4.4 Plans for roadmap

4.4.1 FLSS

The Roadmap activities are concerned with visualisation, connection between the ontology and the learning objects, addition of more statistics over the retrieved material and taking into account the tutors' feedback.

- Visualisation: the same object (ontology or learning material) will be enriched with more visualisation options. For example, the ontology can be presented as a tree-like view or as a graph as in iFLSS. Thus, the stakeholder might choose the most relevant one for a given task.

- Visualisation: more relations among the resources will be visualized. For example, the tutor will be able to follow all the occurrences of the same concept in the learning material. Also, if she/he clicks on a concept, its place in the relevant subpart of the ontology will be displayed.

- Semantic search: A toolkit of statistical measures over the retrieved learning material will be added. At the moment only the standard frequency measures are reported.

- Semantic annotation: A service will be developed for tracking the tutors' feedback. It will record the logs of tutors' activities - queries for searches, searches, annotations, deletions, etc. It will reflect the idea of keeping a 'teaching history', which will improve the implementation as well as the pedagogical work.

4.4.2 iFLSS

Social learning component:

- Include a rating possibility
  - to rate documents: makes the selection of appropriate resources easier
  - to rate the quality of peers: feedback can be based on several aspects, e.g. level of expertise, availability, willingness to answer questions, quality of answers, quality of documents

- Give more information on recommended people: together with the ratings, this would help the learner to decide who is the best person to contact. Information that would be useful: educational background (studies, level or academic degree), frequency of use, questions answered. In contrast to the rating figures, these are objective data.

- Verify if joining accounts of different accounts belonging to the same person improves or not the search results.

Knowledge discovery component
- Include relation extraction in ontology enrichment (see Section 2.3.2.1.5)
- Make ontology fragment user specific (see Section 2.3.2.1.6)
- Indicate strength of relationships between concepts in the ontology fragment
5 Conclusions

The goal of WP6 is to develop services that facilitate learners and tutors in accessing formal and informal knowledge sources in the context of a learning task. A Common Semantic Framework has been developed that includes a Formal Learning Support System (FLSS) and an inFormal Learning Support System (iFLSS).

In the current reporting period, we have validated the FLSS with tutors and the iFLSS with learners. The stakeholders judged the services as being relevant for the learning process and provided us with useful suggestions for their further development. On the basis of the feedback received, an enhanced and extended version of the services has been developed (Version 1.5). The focus has been mainly on improving the usability, while we have also paid attention to a better scalability, interoperability and stability of the services.

We have made a major effort in improving the visualization component, this is especially the case for the ontology. The learners indicated that they would prefer a dynamic graph fragment instead of a static one. We have therefore improved the visualization by dynamically extending the graph during a user's interaction with it. The visualization of the search results of the semantically annotated learning objects has also been modified to include snippets that represent the context of the concepts that have been found, as well as a list of concepts and their frequencies in the document.

The need of learners for more types of content has also received attention and has led to additional crawlers to extract data from Twitter, Flickr and Facebook. The possibility for live data acquisition from YouTube, Bibsonomy, Delicious and SlideShare has been implemented. The integration of live data as opposed to crawled data fills the gap of a very recent and global perspective on information as opposed to strongly personalised content. The combination of the two (live and crawled data) allows the learner to benefit from both methods of finding content.

We have also worked on a disambiguation algorithm that facilitates a more robust link between concepts and their lexicalisations. The quality of the live integration of data resources can be increased by applying it in the context of a domain ontology. This greatly improves the quality of the search results, as shown by our verification. In addition, the disambiguation algorithm allows for a first step in the direction of personalized ontologies thanks also to the adoption of the MOAT vocabulary.

Document annotation has improved in the current version with the introduction of different language pipes which support tutors in making a trade-off between time and completeness. The slower, but more extensive annotation can be used when precision and coverage are essential whereas the faster language pipe can be used when there is only little time to annotate learning materials.

In this reporting period, we have carried out several verification experiments of the revised software with generally encouraging outcomes. A detailed study on the
evolution of social networks has been carried out. Performance and reliability has been tested and found to support the long term usage for the next validation round. The services will be used for a longer period of time and by a larger group of participants than in the previous validation.

Although many revisions and enhancements have been made to the existing services some research directions have not been fully explored in this project. This includes research on customizing individual domain ontologies according to learner profiles, extended user profiles as well as relation extraction. They are part of our road map.
References


Appendix A: Web service documentation