Linking the tele-TASK video portal to the Semantic Web

Bert Baumann, Andreas Groß, Christoph Meinel, Harald Sack

Hasso Plattner Institute for Software Systems Engineering
Postfach 900460, D-14440 Potsdam, Germany
([bert.baumann|andreas.gross|christoph.meinel|harald.sack]@hpi.uni-potsdam.de)

Abstract: Audiovisual data have gained an enormous and ever-growing popularity in the world wide web. Also a growing number of educational content such as, e.g., lecture recordings or audiovisual learning material can be found recently. But, pinpoint and exhaustive retrieval of audiovisual e-learning content in the web is rather difficult as well as automated metadata interchange and integration. We demonstrate a use-case of metadata integration for audiovisual learning resources by complementing web pages of a video lecture portal with semantic RDFa annotations giving way to automated access and universal retrievability.

1 Introduction

Audiovisual data has become the predominant medium of the 20th century and the amount of video data available in the World Wide Web (WWW) is ever-growing day-to-day. Video portals and video search engines enable users to randomly access audiovisual data according to their demands and personal preferences. Among entertainment, news, or documentaries there is also a growing number of educational content available in the WWW. Numerous universities and institutions for higher education are publishing video recordings of lectures and seminars via streaming and podcasts, and some have become rather popular, such as, e.g., MIT open courseware or tele-TASK. The tele-TASK system offers an entire lifecycle infrastructure for lecture recording, archival, and retrieval.

But, for the ordinary user, retrieval and access to those lecture recordings, is not always trivial. First, one has to know, where to find educational content. If one is looking for a specific lecture from MIT, then MIT's web site certainly is a good starting point. Without preferring a specific educational institution, video portal's or video search engines are the next best choice. But, neither special interest video portals nor video search engines provide exhaustive information about the universe of available lecture recordings and potential interrelationships, because metadata exchange formats for web based audiovisual learning resources are not utilized consistently. Furthermore, to be available for search engines, metadata have to be included into the web pages directly.

Of course there exist metadata standards for documentation and e-learning content. Most prominent are Dublin Core (DC) metadata for bibliographical data and Learning Object Metadata (LOM) as well as the Sharable Content Object Reference Model (SCORM) for the description of learning resources. Even though there are several ways for integrating one of these XML-based metadata directly into web pages, individual practice often inhibits real data integration of heterogeneous audiovisual learning resources.

In this paper, we show how to use RDF-based semantic descriptions of DC and LOM, and how to integrate this metadata directly into (X)HTML web pages via RDFa. There exist simple XSLT transformations for extracting plain RDF metadata from RDFa-enriched web pages. We show, how to integrate e-learning related metadata schemata with microformats and other RDF-based metadata (e.g., FOAF, DBPedia, etc.) to enable data integration over heterogeneous schemata. Thus, giving way for the development of new mashup applications and linking own audiovisual content to the Semantic Web's Linked Open Data cloud.

The paper is structured as follows: Section 2 gives a short overview about metadata schemata for bibliographical and audiovisual resources in the e-learning context or semantic metadata. Section 3 introduces the tele-TASK video portal and lecture recording infrastructure, while Section 4 provides implementational details about RDFa integration of several metadata schemata into the tele-TASK web site and explains several examples on how to use RDFa-based semantic metadata for e-learning resources. Section 5 provides a short summary and outlook on future work.

2 Metadata Standards and Semantics

2.1 Metadata

Metadata are data about data, i.e. structured data describing the characteristics of information bearing entities. Metadata can be used for identification, retrieval, evaluation, and administration of the data they describe. In particular there is an emphasis today on automated metadata processing with the purpose of identification and retrieval, as e.g. being applied in web search engines [Dur85]. Usually, metadata are categorized by their degree of inherent structure [DSS93]. On the lower end, there are unstructured metadata such as, free text annotations or tags. Structured metadata follow a distinguished data schema. In addition, categories can be arranged by using taxonomic relationships for generalization and specialization. More complex metadata structures comprise relationships, dependencies, constraints, and rules that can be expressed with the help of ontologies. For efficient identification and retrieval, we have to follow a common, standardized metadata schema.

2.2 Metadata for Documentation, Bibliography, and e-learning

For the purpose of documentation and bibliography as well as for e-learning, several metadata schemata have been developed. Well-established for bibliography is the so called Dublin Core metadata standard, while for e-learning esp. Learning Object Metadata (LOM) has become popular.

Dublin Core The Dublic Core (DC) metadata standard was developed for the description of text-based information objects. It consists of 15 core elements that are intended for the compilation of bibliographical data [Wei97]. In addition, DCMI metadata terms recommend additional fields (element refinements), which allow for a more detailed description or categorization according to the user's preferences. DC summarizes metadata for technical and content-based description of authors, related persons, intellectual property rights, as well as relationships among the described resources and life cycle information. Although intended to describe text-based resources, DC can be used to describe audiovisual objects such as, e.g., lecture recordings [HA99].

LOM Learning Object Metadata (LOM) is an open IEEE metadata standard for the description of learning objects [HD02]. The LOM metadata schema has been designed to support the reusability of learning objects, to aid discoverability, and to facilitate their interoperability, usually in the context of online learning management systems. It enables the description of entities related to the learning process such as, e.g., type of object, author, owner, terms of distribution, format, and pedagogical attributes, such as teaching or interaction style.

2.3 Semantic Metadata

The Semantic Web is an evolving extension of the World Wide Web (WWW) in which the meaning of information and services on the web is well defined [BLHL01]. Thereby, it will be possible for the web to understand and satisfy the requests of people and machines to use the web content. Key technology of the Semantic Web are semantic metadata (ontologies) representing commonly shared conceptualizations being specified with standardized, formalized languages [Gru93]. The World Wide Web Consortium (W3C)[1] has already standardized a set of knowledge representation languages of different semantic expressivity being arranged in a hierarchically layered model.

Resource Description Framework (RDF) und RDF Schema (RDFS) RDF and RDFS are simple knowledge representations for the definition of individual objects and their relationships as well as classes and their interrelationship among each other can be defined [LS99]. Individuals and concepts are identified via Uniform Resource Identifier (URI). RDF data consists out of simple triples `(a,b,c)`, where `a` represents some individual, `b` stands for a property of `a`, and `c` gives a distinct value to property `b`. Individuals are concrete realizations (instances) of concepts (classes). Concepts can be derived from more general concepts with the help of RDF Schema (RDFS) via generalization, specialization, or class extension. In that way, relationships among classes can be defined, also if they are not part of the given metadata schema [BG04].

Web Ontology Language (OWL) Semantic expressivity of RDF and RDFS is rather limited, as, e.g., there is no possibility to generalize statements for a group of individuals, or the definition of logical attributes and constraints. The Web Ontology Language OWL is the W3C standard for the specification of ontologies based on description logics [MvH04].

[1] http://w3c.org/

```
<div xmlns:dc="http://purl.org/dc/elements/1.1/"
    about="http://www.tele-task.de/view/3931">
  <span property="dc:title">Semantic Web</span>
  <span property="dc:date">2008-10-23</span>
</div>
```

```
@prefix dc:<http://purl.org/dc/elements/1.1/> .
<http://www.tele-task.de/view/3931> dc:title "Semantic Web" .
<http://www.tele-task.de/view/3931> dc:date "2008-10-23" .
```

Fig. 1: Example for RDFa and corresponding RDF extract (in RDF turtle syntax)

OWL comes in three different variants, OWL Lite, OWL DL, and OWL Full, according to its semantic expressivity, which is also related to its computational complexity. In addition to RDF(S) class and relationship definitions, OWL adds different class constructors, class and property constraints as well as (restricted) universal and existential quantification.

2.4 Online Integration and Interoperability

While the metadata schemata that have been described in the previous sections are structured data, web pages are semi-structured (X)HTML-encoded information resources. (X)HTML only provides information about document structure and not about the document's textual content. Also, (X)HTML cannot be extended to include other metadata. Nevertheless, there are different ways to incorporate additional metadata into (X)HTML-encoded documents. We will focus on microformats and RDFa:

Microformats Microformats define a specific markup format for semantic annotation of (X)HTML documents [Dub05]. Microformat annotations are encoded within (X)HTML tag attributes and can easily be extracted from web documents. Thus, applications are able to gather some information about the meaning of web page's content (such as contact information, geographic coordinates, calendar events, and the like) for subsequent processing. Microformat semantic is defined by common agreement and not by formal definition.

RDFa Similar to microformats, RDFa (RDF in (X)HTML attributes) utilizes unused (X)HTML attributes to include RDF metadata into simple web pages [adi08]. RDFa uses attributes from (X)HTML's `meta` and `link` elements, and generalizes them so that they can be used for all (X)HTML syntax elements (cf. Fig. 1).

While microformats are always fixed to a special topic (calendar, geographic data, address data, etc.), RDFa annotations can make use of any RDF ontology and thus, it is much more flexible and allows to annotate (X)HTML markup with semantics. A simple mapping is defined with GRDDL (Gleaning Resource Descriptions from Dialects of Languages) [W3C07] and XSLT (Extensible Stylesheet Language Transformations) [Kay07] so that plain RDF may be extracted.

3 Video lecturing with tele-TASK

This chapter introduces the tele-TASK system and its components, i.e. the tele-TASK recording system and the tele-TASK web portal.

3.1 The tele-TASK Recording and Distribution System

The tele-TASK[2] recording system [SM02] is a sophisticated technology for the creation and transmission of advanced video presentations via the internet. This state-of-the-art solution is outstanding for its simplicity and dependability. In addition to high-quality video and audio of the lecturer the system delivers a synchronous video feed of the lecturer's computer screen without installing any additional software on the lecturer's computer. This ability is singular and separates tele-TASK from any competitive lecture recording devices. With the help of the tele-TASK technology users worldwide can access to teaching courses and presentations using live streams or archived recordings. The presentations are available via internet and can be downloaded on portable devices [WLM07] such as, e.g., PDA, mobile video players, 3G mobile phones, or lean-back consumer electronics also.

The tele-TASK content is published via several distribution channels. The main distribution platform is the tele-TASK web portal (Fig. 2) for the publication of lecture and event recordings at the „Hasso Plattner Institute for Systems Engineering" (HPI). In addition, podcasts of tele-TASK recordings can also be accessed via iTunes U[3]. Recorded tele-TASK lectures are available in various formats such as, e.g., RealMedia, Flash Video, and MP4. The portal offers different post-processing steps for cutting, synchronization, and media conversion. Currently the tele-TASK database comprises more than 2.200 lecture recordings and 2.600 podcasts from 500+ different speakers. All tele-TASK content can be accessed and downloaded for free.

Since 2009 tele-TASK lectures are part of the popular Apple iTunes U repository, which is part of the iTunes Store. But, as being part of the educational section of iTunes, tele-TASK content is freely available. The HPI's tele-TASK pool on iTunes U is one out of four selected german elite education centers distributing their learning materials on iTunes U[4]. The iTunes store can only be used via the proprietary Apple iTunes client software. The main drawback of iTunes U lies in its strictly proprietary nature, preventing worldwide searchability and data integration.

3.2 Restrictions of the tele-TASK Portal

Audiovisual lecture recordings in terms of streaming media are receiving an ever-growing popularity among learners. One of the reasons for it's popularity is that the learner might

[2] (**TeleT**eaching **A**nywhere **S**olution **K**it)

[3] http://itunes.hpi.uni-potsdam.de

[4] HPI iTunes U portal page: http://deimos3.apple.com/WebObjects/Core.woa/Browse/hpi-de-public

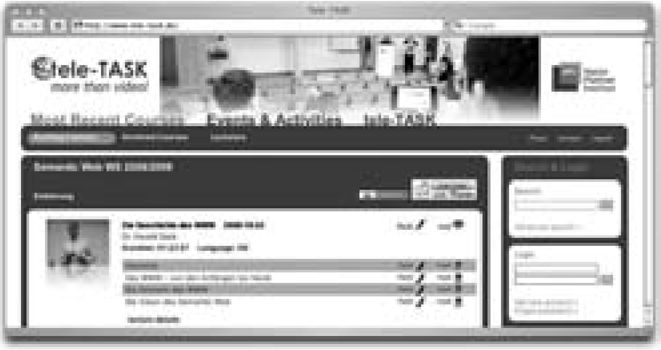

Fig. 2: Reference of a video lecture within the tele-TASK web portal

access and learn the video lecture everywhere at anytime, independent from the live event. Therefore, the retrievability of lecture recordings has become most decisive. Even though tele-TASK provides search formulas with automated online completion, it is difficult for the user to specify the accurate search terms. In particular, the integration of additional external data resources such as, e.g., calendar, address books, or others, to increase the retrievability is not possible. One way to achieve better search results and simplify the retrieval process is to provide semantic metadata to enable a flexible and dynamic data integration. Moreover the integration of semantic metadata enables the automatic connection e-learning resources worldwide.

4 Integrating tele-Task into the Semantic Web

This chapter addresses the implementation of semantic metadata for the tele-TASK web portal via RDFa annotations and linking to the Semantic Web.

4.1 Standardized Metadata for tele-TASK Data and RDFa Integration

The core data element of the tele-TASK database is the video recording of a single lecture event. Lectures can be combined into groups of lectures or lecture series. A lecture series comprises all single lectures of a distinct topic within the time frame of a semester. Likewise, a single lecture may be sectioned into several chapters, which are utilized for the production of podcast contributions.

Metadata for lectures, such as, e.g., title, abstract, language, date, duration, lecturer name(s), etc., are complemented by lecture series metadata such as, e.g., keywords, series type, place, institution, etc. All tele-TASK metadata can be mapped to standardized XML-based metadata schemata such as DC and LOM (cf. Section 2). To be publicly available

```
<div id="canvasleft" about="/lectures/view/3933/" typeof=
 "lom:LearningObject">
 <h1 property="dc:title dcterms:title" xml:lang="de">
     Universelle Vokabularien mit XML (german)</h1>
 <span property="lom:keyword" content="computer,informatics,semantic web"/>
 <span property="dcterms:extent" datatype="lom:Duration">01:32:35</span>
 <a href="/people/15/" property="dc:creator">Dr. Harald  Sack</a>
 <table id="lectureabstracttable" rel="rdf:Bag" resource="#podcast-list">
  <tr about="... SEW_2008_11_06_part_0_podcast.mp4">
  <td property="dcterms:title" xml:lang="de">
     Die Vision des Semantic Web</td>
  <td>
   <span rel="dcterms:hasPart"
     resource="... SEW_2008_11_06_part_0_podcast.mp4"/>
   <span rel="dcterms:isPartOf" resource="/lectures/view/3933/"/>
   <span property="dc:format dcterms:format" datatype="dcterms:IMT"
     content="video/mp4"/> ...
  </td> ...
 </tr>
 </table> </div>
```

Fig. 3: Dublin Core and LOM as Part of the (X)HTML Web Page

on the WWW these metadata schemata must be integrated into (X)HTML encoded web pages. One way to achieve this is to store metadata as a separate XML-file being linked by the original web page. But, this approach prohibits proper assignment of single metadata to their corresponding representatives in the (X)HTML file. Therefore, we decided to include metadata directly into the (X)HTML document via tag attributes and already available text information of the (X)HTML document. DC and LOM metadata schemata are already available as RDF encoded metadata [NPB03] and can be directly included into (X)HTML documents via RDFa. For further processing, RDF syntax can be extracted automatically via XSLT or by using applications such as, e.g., W3C RDF extractor[5].

By including DC and LOM RDF Schema definitions via namespaces into the (X)HTML document, these metadata schemata can be utilized with RDFa annotations. One way to supply data values can be achieved by integrating the already displayed textual content of the (X)HTML web page into the RDFa annotation. RDF tripels represent assertions about resources (identified via URI) in the same way as simple natural language statements consisting of subject (resource `a`), predicate (relationship or property `b`) and an object value (literal or resource `c`). Via RDFa annotation subject `a` and property `b` can be embedded as attributes within an (X)HTML-tag that envelopes a text value, the object value `c`, i.e. `<div about="a" property="b">c</div>.` If a RDF triple is about to contain data values that are not displayed in the (X)HTML web page, the `content`-attribute can be used to hold the data, i.e. `<div about="a" property="b" content="c"/>.` Fig. 3 shows a typical example of a tele-TASK (X)HTML web page with embedded RDFa annotation displaying metadata about video lectures.

[5] http://www.w3.org/2007/08/pyRdfa/extract

```
PREFIX foaf:<http://xmlns.com/foaf/0.1/>
SELECT ?title ?lecture
WHERE {
    ?lecture dc:title ?title FILTER regex(str(?title), "rdf", "i").
}
```

Fig. 4: SPARQL example query to search all lectures that contain the string »rdf« in the title

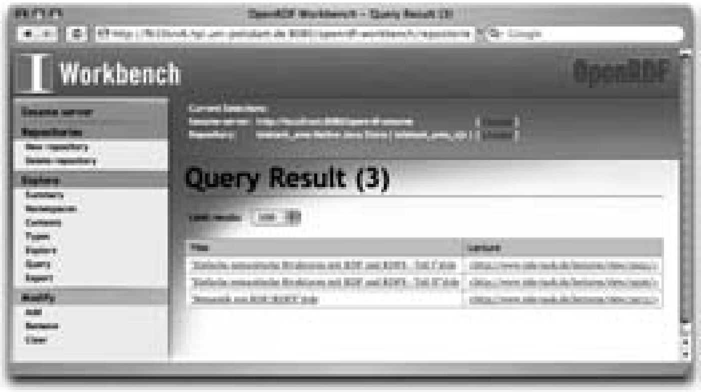

Fig. 5: Sesame server web portal displaying the result of the query in Fig. 4

4.2 Using RDFa Annotation for Information Integration in the Semantic Web

By providing semantic metadata via RDF/RDFa annotation the tele-TASK data can be combined and complemented with a large variety external data sources. On the one hand, autonomous software agents, search engines, or applications can link their own resources with tele-TASK data by using a SPARQL endpoint, while on the other hand, tele-TASK data also can be connected and augmented with external semantic data being matched with tele-TASK metadata.

n addition to the RDFa annotation being included within the tele-TASK (X)HTML web pages, we decided to store RDF data also in the RDF triple store database Sesame[6]. Sesame is an open source RDF database with support for RDF Schema inferencing and querying providing also a SPARQL endpoint. A SPARQL endpoint enables users (human as well as applications) to query a RDF knowledge base via the SPARQL language. A SPARQL endpoint typically returns query results in various machine-processable formats. Fig. 4 shows a simple example query in SPARQL and Fig. 5 show the result being displayed in the Sesame web user interface.

By providing a SPARQL endpoint and by deploying RDF-based Dublin Core and LOM metadata, tele-TASK video resources are linked to the semantic web, i.e. they provide a meaningful interface that can be accessed by human users as well as by software applica-

[6] http://www.openrdf.org/

```
<html xmlns="http://www.w3.org/1999/xhtml" ...
 xmlns:vcard="http://www.w3.org/2006/vcard/ns#"
 xmlns:foaf="http://xmlns.com/foaf/0.1/" ...>
<span datatype="" resource="#hpi-vcard-adr">
 <span property="vcard:street-address" content="Prof.-Dr.-Helmert-Str. 2-3"/>
 <span property="vcard:postal-code" content="14482"/>
 <span property="vcard:locality" content="Potsdam"/>
</span>
<span typeof="vcard:VCard foaf:Person" resource="#author-entity">
 <span property="vcard:fn foaf:name" content="Harald  Sack"/>
 <span rel="vcard:url foaf:homepage" href="/meinel/sack/"/>
 <span rel="foaf:weblog" href="http://moresemantic.blogspot.com/"/>
 <span rel="foaf:depiction" href="/uploads/pics/harald_min.jpg"/>
 <span rel="rdfs:seeAlso" href="/meinel/sack/foaf.rdf"/>
</span>
```

Fig. 6: Personal information with vCard and FOAF

tions. Applications can automatically search for authors, titles, or media formats, and can evaluate additional metadata provided for each video lecture or video lecture series, such as, e.g., keywords, duration, date, etc. Authors, as being persons, can also be described with the help of alternative metadata schemata for personal and address information, such as vCard[7] or FOAF [8] . FOAF defines a set of terms for letting users describe persons, their activities and their relations to other people and objects [BM07]. Anyone can use FOAF to describe himself or herself. In difference to other social networking services, FOAF allows groups of people to describe social networks without the need for a centralized database. FOAF is one of the largest projects on the Semantic Web which has an estimated 2–5 million users. vCard is an electronic format for the consistently exchange of business information. vCard elements can be freely reused, as e.g., within different LOM-attributes as being shown in fig. fig:vCard, which gives an example of vCard and FOAF data integration via RDFa annotation.

5 Summary and Outlook

We have shown a use-case of semantic data Integration via RDFa by complementing the tele-TASK web portal with semantic metadata that is already available within the ordinary tele-TASK database by deploying Dublin Core and LOM metadata schemata for RDFa data integration. As a next step, the SPARQL endpoint being described in the previous section will also be opened up for the public. This enables tele-TASK to link with the worldwide semantic web. Up to now, only the video data being provided by the tele-TASK web portal are described via standard metadata schemata. But, to be able to access the meaning of its content, domain ontologies must be deployed to annotate also the video content. A first step will be the mapping of already existing keywords that describe the video data's content to concepts, classes, and instances of the global wikipedia encyclo-

[7] http://www.w3.org/TR/vcard-rdf

[8] Friend of a Friend (FOAF) project homepage, `http://www.foaf-project.org/`

pedia, in particular to it's semantic counterpart, the DBPedia[9]. This would be the first step for tele-TASK to participate in the Linked Open Data (LOD) community[10].

Furthermore, if tele-Task entities can be mapped to LOD entities such as, e.g. DBPedia entities, numerous additional information stemming from the popular online encyclopedia 'Wikipedia'[11] can be used to complement and to enrich tele-Task data on the tele-Task portal site as well as for iTunes U. E.G., information about famous speakers and presenters, who are also present at wikipedia, can be extracted in an automated way and presented on the tele-Task website for that speaker.

Acknowledgment

The authors would like to thank students of the seminar „Semantic Web enabled Software“ of winter semester 2008/09, Andreas Motl, Paul Schröder and Dennis Brutski.

References

[adi08] RDFa in XHTML: Syntax and Processing, W3C Recommendation, 2008.

[ADL04] Advanced Distributed Learning ADL. *Sharable Content Object Reference Model (SCORM) 2nd Edition Conformance Requirements Version 1.1.* 2004.

[BG04] Dan Brickley und R.V. Guha. Resource Description Framework (RDF) Schema Specification, February 2004.

[BLHL01] Tim Berners-Lee, James Hendler und Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, 2001.

[BM07] Dan Brickley und Libby Miller. The Friend Of A Friend (FOAF) Vocabulary Specification, November 2007.

[DSS93] Randall Davis, Howard Shrobe und Peter Szolovits. What is a Knowledge Representation. *AI Magazine*, 14(1):17–33, 1993.

[Dub05] Micah Dubinko. What are Microformats, March 2005.

[Dur85] William R. Durrell. *Data Administration: A Practical Guide to Data Administration*. McGraw-Hill, 1985.

[Gru93] Thomas R. Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisitation*, 5(2):199–220, 1993.

[HA99] Jane Hunter und Liz Armstrong. A comparison of schemas for video metadata representation. *Comput. Netw.*, 31(11-16):1431–1451, 1999.

[9] http://dbpedia.org/

[10] http://linkeddata.org

[11] http://www.wikipedia.org/

[HD02] Wayne Hodgins und Erik Duval. Draft standard for learning technology - Learning Object Metadata - ISO/IEC 11404. Bericht, 2002.

[Kay07] Michael Kay. XSL Transformations (XSLT) Version 2.0. W3C recommendation, W3C, January 2007. http://www.w3.org/TR/2007/REC-xslt20-20070123/.

[LS99] O. Lassila und R. R. Swick. Resource Description Framework (RDF) Model and Syntax Specification . W3C Recommendation, World Wide Web Consortium, 1999.

[MvH04] Deborah L. McGuinness und Frank van Harmelen. OWL Web Ontology Language: Overview, W3C Recommendation, 10 February 2004.

[NPB03] Mikael Nilsson, Matthias Palmer und Jan Brase. The LOM RDF binding - principles and implementation. In *Proc. of 3rd Annual Ariadne Conference*, 2003.

[SM02] V. Schillings und Ch. Meinel. Tele-TASK – tele-teaching anywhere solution kit. In *Proceedings of ACM SIGUCCS*, Providence, USA, 2002.

[W3C07] W3C. Gleaning Resource Descriptions from Dialects of Languages (GRDDL). W3c recommendation, W3C, September 2007.

[Wei97] Stuart Weibel. The Dublin Core: A Simple Content Description Model for Electronic Resources, 1997.

[WLM07] Katrin Wolf, Serge Linckels und Christoph Meinel. Teleteaching anywhere solution kit (Tele-TASK) goes mobile. In *SIGUCCS '07: Proceedings of the 35th annual ACM SIGUCCS conference on User services*, Seiten 366–371, New York, NY, USA, 2007. ACM.