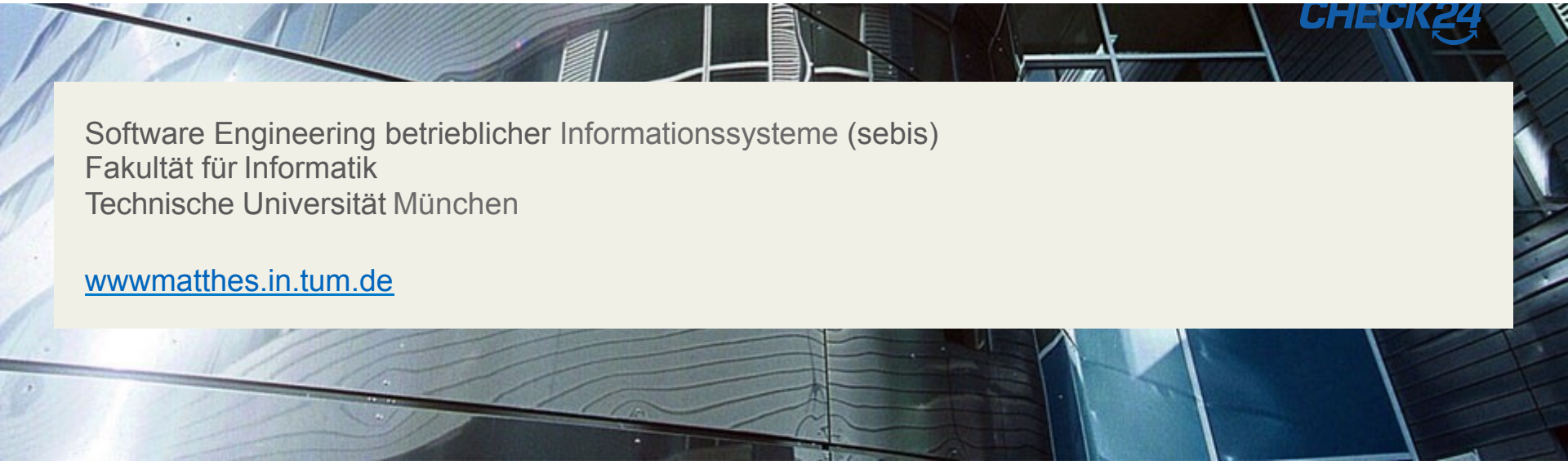
Using Web Annotations to Represent Relations between Structured and Unstructured Information in Semantic Wikis

Master thesis - Kickoff presentation

Shivguru Rao

Advisor: Daniel Braun

Munich, 28.01.2018



Software Engineering betrieblicher Informationssysteme (sebis)
Fakultät für Informatik
Technische Universität München

wwwmatthes.in.tum.de

Motivation

What is structured data?

Attributes of this Student Project

Title (de)	Nutzung von Web Annotations zur Repräsentation von Relationen zwischen strukturierten und unstrukturierten Informationen in Semantic Wikis
Title (en)	Using Web Annotations to Represent Relations between Structured and Unstructured Information in Semantic Wikis
Project	
Type	Master's Thesis
Status	started
Student	
Advisor	Daniel Braun
Supervisor	Prof. Dr. Florian Matthes

Motivation

What is un-structured data?

Master's Thesis Bhimasenarao Shivguru Rao

Last modified Dec 21, 2017

No tags assigned

Using Web Annotations to Represent Relations between Structured and Unstructured Information in Semantic Wikis

Hybri Wikis combine unstructured information in the form of Wiki texts with structured information. However, there is no link between the structured and the unstructured information. Therefore, if a piece of information is represented as both, structured and unstructured information, it has to be update manually in both formats in order to keep the content consistent.

Aim of this thesis is to use the metamodel of SocioCortex and Machine Learning to automatically link structured and unstructured data in Hybrid Wikis and also extract structured information from unstructured content. A web application should be built in order to allow users to manually annotate unstructured data from SocioCortex in order to create training data for the Machine Learning algorithm and visualize the links between the data format.

Motivation contd..

Linking structured and unstructured data?

Master's Thesis Bhimasenarao Shivguru Rao

Last modified Dec 21, 2017

No tags assigned

Using Web Annotations to Represent Relations between Structured and Unstructured Information in Semantic Wikis

Hybri Wikis combine unstructured information in the form of Wiki texts with structured information. However, there is no link between the structured and the unstructured information. Therefore, if a piece of information is represented as both, structured and unstructured information, it has to be update manually in both formats in order to keep the content consistent.

Aim of this thesis is to use the metamodel of SocioCortex and Machine Learning to automatically link structured and unstructured data in Hybrid Wikis and also extract structured information from unstructured content. A web application should be built in order to allow users to manually annotate unstructured data from SocioCortex in order to create training data for the Machine Learning algorithm and visualize the links between the data format.

Attributes of this Student Project

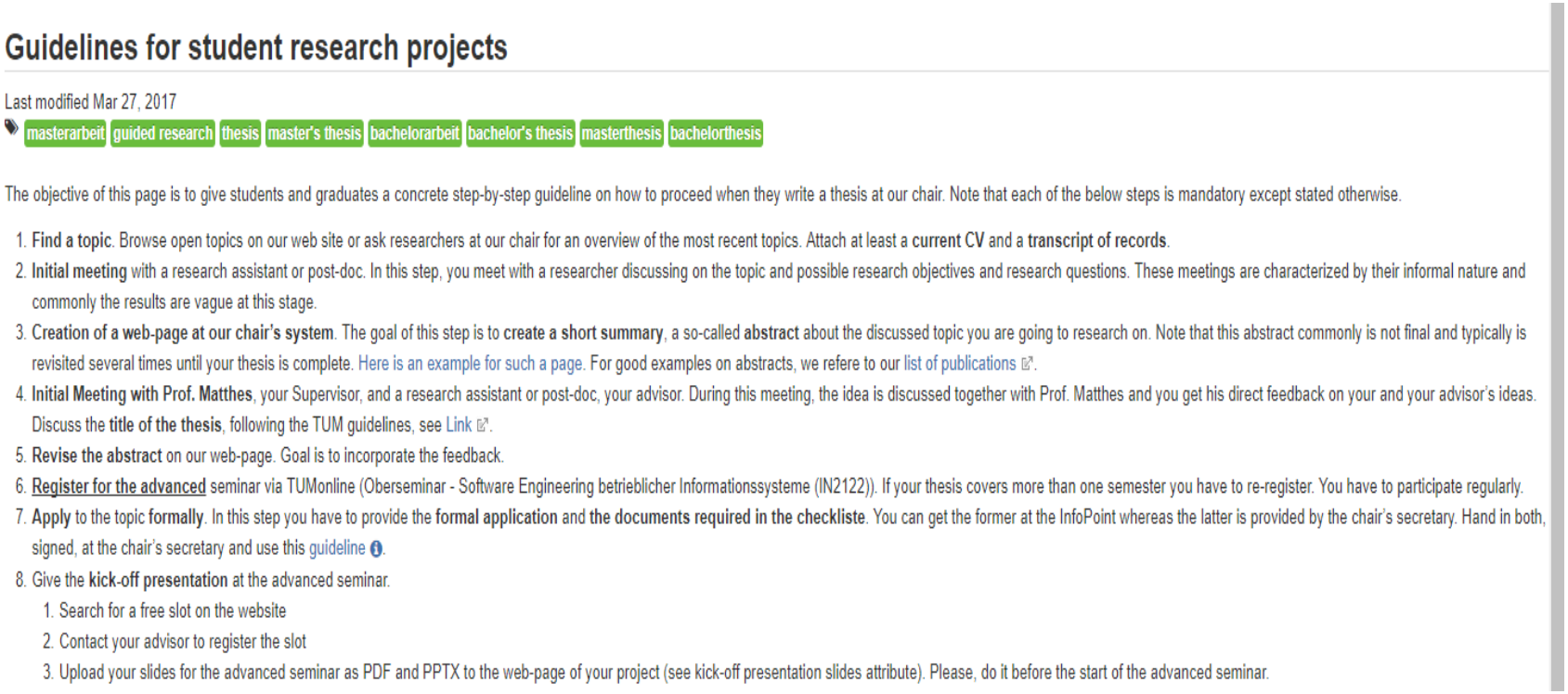
Title (de)	Nutzung von Web Annotations zur Repräsentation von Relationen zwischen strukturierten und unstrukturierten Informationen in Semantic Wikis
Title (en)	Using Web Annotations to Represent Relations between Structured and Unstructured Information in Semantic Wikis
Project	
Type	Master's Thesis
Status	started
Student	
Advisor	Daniel Braun

Motivation contd..

- Why do we need to link structured and unstructured data?
  - Same data, different representation
  - Handle changes to data
  - Ensure data consistency

Proposed solution

- Web Annotations



Guidelines for student research projects

Last modified Mar 27, 2017

masterarbeit guided research thesis master's thesis bachelorarbeit bachelor's thesis masterthesis bachelorthesis

The objective of this page is to give students and graduates a concrete step-by-step guideline on how to proceed when they write a thesis at our chair. Note that each of the below steps is mandatory except stated otherwise.

1. **Find a topic**. Browse open topics on our web site or ask researchers at our chair for an overview of the most recent topics. Attach at least a **current CV** and a **transcript of records**.
2. **Initial meeting** with a research assistant or post-doc. In this step, you meet with a researcher discussing on the topic and possible research objectives and research questions. These meetings are characterized by their informal nature and commonly the results are vague at this stage.
3. **Creation of a web-page at our chair's system**. The goal of this step is to **create a short summary**, a so-called **abstract** about the discussed topic you are going to research on. Note that this abstract commonly is not final and typically is revisited several times until your thesis is complete. Here is an example for such a page. For good examples on abstracts, we refere to our list of publications.
4. **Initial Meeting with Prof. Matthes**, your Supervisor, and a research assistant or post-doc, your advisor. During this meeting, the idea is discussed together with Prof. Matthes and you get his direct feedback on your and your advisor's ideas. Discuss the **title of the thesis**, following the TUM guidelines, see Link.
5. **Revise the abstract** on our web-page. Goal is to incorporate the feedback.
6. **Register for the advanced** seminar via TUMonline (Oberseminar - Software Engineering betrieblicher Informationssysteme (IN2122)). If your thesis covers more than one semester you have to re-register. You have to participate regularly.
7. **Apply** to the topic **formally**. In this step you have to provide the **formal application** and **the documents required in the checkliste**. You can get the former at the InfoPoint whereas the latter is provided by the chair's secretary. Hand in both, signed, at the chair's secretary and use this guideline.
8. Give the **kick-off presentation** at the advanced seminar.
    1. Search for a free slot on the website
    2. Contact your advisor to register the slot
    3. Upload your slides for the advanced seminar as PDF and PPTX to the web-page of your project (see kick-off presentation slides attribute). Please, do it before the start of the advanced seminar.

Proposed solution contd.

- JSON-LD based embedded Web Annotations

```
<script type='application/ld+json;profile="http://www.w3.org/ns/anno.jsonld"'>
{
    "@context": "http://www.w3.org/ns/anno.jsonld",
    "id": "http://example.org/annotations/anno-01",
    "type": "Annotation",
    "motivation": "describing",

    "target": {
        "source": "http://example.org/Emblem004.html",
        "selector": {
            "type": "CssSelector",
            "value": "div#mottoTranscription"
        }
    },

    "body": {
        "type": "TextualBody",
        "value": "Fiunt, quae posse negabas.",
        "format": "text/plain",
        "language": "la"
    },

    "created": "2017-01-26T17:30:04.639Z",
    "creator": {
        "type": "Person",
        "email": "mailto:mara@example.org",
        "name": "Mara"
    }
}
</script>
```

Proposed solution contd.

- Saving Annotation Information as a JSON document
  - Uses XPath and relative positioning of data to annotate texts

```
{
  "id": "39fc339cf058bd22176771b3e3187329",  # unique id (added by backend)
  "annotator_schema_version": "v1.0",        # schema version: default v1.0
  "created": "2011-05-24T18:52:08.036814",   # created datetime in iso8601 format (added by backend)
  "updated": "2011-05-26T12:17:05.012544",   # updated datetime in iso8601 format (added by backend)
  "text": "A note I wrote",                  # content of annotation
  "quote": "the text that was annotated",    # the annotated text (added by frontend)
  "uri": "http://example.com",               # URI of annotated document (added by frontend)
  "ranges": [                                # list of ranges covered by annotation (usually only one e
    {
      "start": "/p[69]/span/span",           # (relative) XPath to start element
      "end": "/p[70]/span/span",             # (relative) XPath to end element
      "startOffset": 0,                      # character offset within start element
      "endOffset": 120                       # character offset within end element
    }
  ],
  "user": "alice",                           # user id of annotation owner (can also be an object with
  "consumer": "annotateit",                  # consumer key of backend
  "tags": [ "review", "error" ],             # list of tags (from Tags plugin)
  "permissions": {                           # annotation permissions (from Permissions/AnnotateItPermi
    "read": ["group:__world__"],
    "admin": [],
    "update": [],
    "delete": []
  }
}
```

Manual Annotation

- Web based Tool
  - Type Script and Angular 2
- Annotate structured and unstructured data
- Link them together for future use
- Store the annotations in SocioCortex

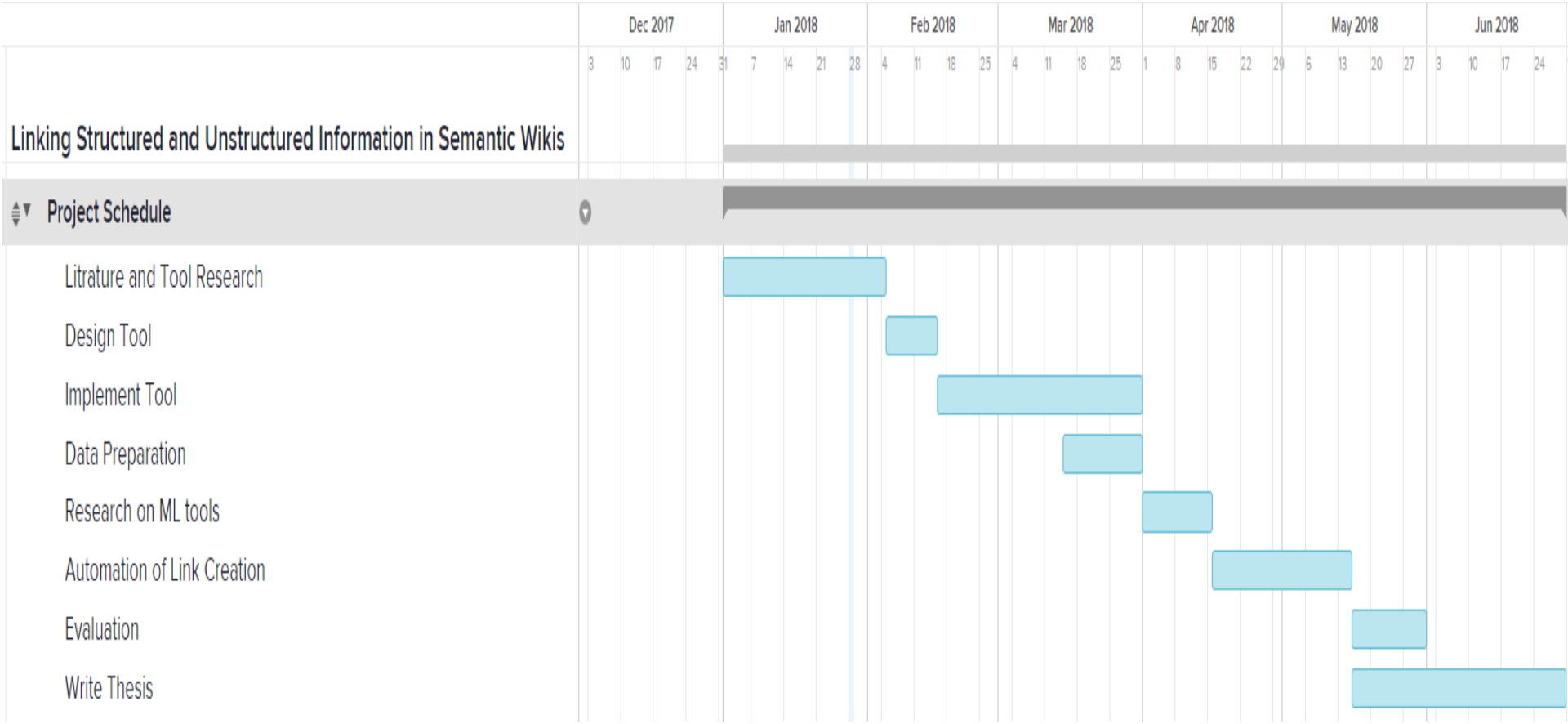
Automatic Generation of Links

- Machine learning based Approach
- The tool will act as a method to input Training Dataset
- Which data set is used for linking?
  - Manually adding links to Training dataset
  - Automatically generate links with help of training data

Research questions

- How can we link structured and unstructured?
- How can we store these Links?
- Can we use annotation standards to link data?
- How can we adapt to the data changes?
- How effective it is to generate these links automatically?

Time Line


Dec 2017
Jan 2018
Feb 2018
Mar 2018
Apr 2018
May 2018
Jun 2018
3 10 17 24 31 7 14 21 28 4 11 18 25 4 11 18 25 1 8 15 22 29 6 13 20 27 3 10 17 24
Linking Structured and Unstructured Information in Semantic Wikis
Project Schedule
Litrature and Tool Research
Design Tool
Implement Tool
Data Preparation
Research on ML tools
Automation of Link Creation
Evaluation
Write Thesis

Thank You

References and Sources

- https://www.w3.org/TR/annotation-html/
- https://en.wikipedia.org/wiki/Web_annotation
- https://wwwmatthes.in.tum.de/pages/xx3rsj4r3pcg/Guidelines-for-student-research-projects
- https://wwwmatthes.in.tum.de/pages/ykew47mumc1b/Master-s-Thesis-Bhimasenarao-Shivguru-Rao