TIB AV-Portal: Semantic content mining with semi-automatic metadata editing

Felix Saurbier (TIB) / Jörg Waitelonis (yovisto GmbH)
Patras, July 7th, 2017
LIBER 2017
Agenda

1. TIB and TIB AV-Portal
2. Semantic Video Analysis
3. Named Entity Linking
4. Data Quality
5. Semi-automatic Metadata Editing
6. Versioning of RDF-Data
German National Library of Science and Technology (TIB)

- Leibniz Information Centre for Science and Technology and University Library at Hannover
- German National Library for science and technology
- The world’s largest science and technology library
- An infrastructure provider for the scientific work process
- TIB-Strategy: Move beyond text
- Competence Centre for Non-Textual Materials
TIB AV-Portal (av.tib.eu)

- Platform for **quality-tested scientific videos**
- Online since April 2014
- Developed by TIB and Hasso Plattner Institute

- **Automatic metadata enrichment**, DOI/MFID, long-term preservation, semantic search

- More than 10,000 Videos (June 2017)
- Conference recordings, lectures, experiments, video abstracts, simulations, animations
- Videos predominantly under open access licences
Agenda

1. TIB and TIB AV-Portal
2. Semantic Video Analysis
3. Named Entity Linking
4. Data Quality
5. Semi-automatic Metadata Editing
6. Versioning of RDF-Data
Scene Recognition (SBD) 
Speech Recognition (ASR) 
Text Recognition (OCR) 
Image Recognition (VCD) 
Named Entity Linking (NEL)
Semantic Video Analysis – Results

Video Sequences

Audio Transcript

Named Entities

02:46

hitting science well about what people images that in software to crimes this data right you need these tools there is no signs without software today and minus catching maybe butterflies that at some point you would use this to a generator a database of the butterflies again and researchers need maybe to classify all this so software is essential for science today the it can also

02:44

Resultant  Point (geometry)  Database  Service (economics)

Information  Whiteboard  Computer animation  Software

Software  Right angle  Sign (mathematics)  Game controller

Medical imaging  Function (mathematics)
Agenda

1. TIB and TIB AV-Portal
2. Semantic Video Analysis
3. Named Entity Linking
4. Data Quality
5. Semi-automatic Metadata Editing
6. Versioning of RDF-Data
Named Entity Linking – Overview

- Based on Audio/OCR-Transcripts
- Semantic analysis and annotation
- Knowledge base:
  - 63,356 GND subject headings
- Cross-lingual mapping (DE – ENG)
  - e.g. DBpedia, LCSH
all 3 of the 6 principals that doesn't actually satisfy the persistence part of it because they get a repository that may not be persistent it may actually change we may go away from get up at some point to move to I don't know and I different version control system that has an understandable API potentially on or something else on but so we recommend that the identifier's resolved to persistent landing pages that contain metadata and link to the software rather than directly to the source code.
Named Entity Linking – Annotation Data Model

- **Video**
  - media-ID: 31031

- **Media Fragment**

- **Annotation**
  - annotated by: ASR-1.0.0
  - transcript: “[…] directly to the source code itself … ”

- **Body**
  - annotated by: NEL-1.0.0
  - text-anchor: "source code"

- **NEL**
  - entity: http://d-nb.info/gnd/4488209-9
Named Entity Linking – Data Model

Resource Description Framework (RDF)

```
tib:asr/31031_26463533_23007#char=483,494 itsrdf:taAnnotatorsRef tib:annotator/NEL-1.0.0 ;
itsrdf:taIdentRef <http://d-nb.info/gnd/4488209-9> ;
nif:anchorOf "source code".
```

Vocabularies

- Bibframe Vocabulary
- DCMI Metada Terms
- DCMI Type Vocabulary
- Friend of a Friend Vocabulary
- Open Annotation Data Model
- NLP Interchange Format
- Internationalization Tag Set (ITS) Ontology

https://av.tib.eu/opendata
Agenda

1. TIB and TIB AV-Portal
2. Semantic Video Analysis
3. Named Entity Linking
4. Data Quality
5. Semi-automatic Metadata Editing
6. Versioning of RDF-Data
Data Quality – Typical Annotation Errors

**Missing Entity**

*different version control system that has an understandable* [API] [API]

**Compound Split**

*directly to the* [source code] [source code] [source code] [source code] [source code] [source code] *itself*

**General/Specific**

*particularly because of* [landing pages] [landing pages] [landing pages] [landing pages] [landing pages] [landing pages] [landing pages]

**Wrong Entity**

*because they get a* [repository] [repository] [repository] [repository] [repository] [repository] [repository]
Data Quality – Solutions

- Improving OCR and ASR transcripts
- Improving automatic NEL
- Blacklisting of common/frequent annotation errors
- Manual editing and correction of semantic annotations
- Semi-automatic editing of semantic annotations
Agenda

1. TIB and TIB AV-Portal
2. Semantic Video Analysis
3. Named Entity Linking
4. Data Quality
5. Semi-automatic Metadata Editing
6. Versioning of RDF-Data
Semi-automatic Metadata Editing

Advantages

- Combines human and machine assets
  - **Machines**: localizing / recognizing entities
  - **Humans**: disambiguation / linking the right entity

Solution

- Based on refer Wordpress plugin (refer.cx)
- Developed by yovisto GmbH
- Web-based WYSIWYG editor and NEL tool
- **Auto-suggest** of candidate entities for semantic annotations
all 3 of the 6 principals that doesn't actually satisfy the persistence part of it because they get a repository may not be persistent it may actually change we may go away from get up at some point to move to I don't know and I different version control system that has an understandable API potentially on or something else on but so we recommend that the identifier resolved to persistent landing pages that contain metadata and link to the software rather than directly to the source code itself and particularly because of landing pages are persistent of this is intended to ensure the longevity of the longevity of softer metadata even beyond the offers lifespan the may still on the in the in the paper there's a lot of other
Semi-automatic Metadata Editing – refer

Modify annotations: Add Entity

all 3 of the 6 principals that doesn't actually satisfy the persistence part of it because they get a repository may not be persistent it may actually change we may go from get up at some point to move to I don't know and I different version control system that has an understandable API potentially on or something else on but so we recommend that the identifier resolved to persistent landing pages that contain metadata and link to the software rather than directly to the source code itself and particularly because of landing pages are persistent of this is intended to ensure the longevity of the longevity of softer metadata even beyond the offers lifespan the may still on the in the in the paper there's a lot of other
Semi-automatic Metadata Editing – refer

Modify annotations: Add Entity

all 3 of the 6 principals that doesn't actually satisfy the ... may not be persistent it may actually change we may go from ... and I different version control system that has an understandab ... so we recommend that the identifier resolved to persistent the software rather than directly to the source code itself and persistent of this is intended to ensure the longevity of the longevity lifespan the may still on the in the in the paper there's a lot of other
Modify annotations: Add Entity

all 3 of the 6 principals that doesn't actually satisfy the persistence part of it because they get a repository may not be persistent it may actually change we may go away from get up at some point to move to I don't know and I different version control system that has an understandable API potentially on or something else on but so we recommend that the identifier resolved to persistent landing pages that contain metadata and link to the software rather than directly to the source code itself and particularly because of landing pages are persistent of this is intended to ensure the longevity of the longevity of softer metadata even beyond the offers lifespan the may still on the in the in the paper there's a lot of other
all 3 of the 6 principals that doesn't actually satisfy the persistence part of it because they get a repository may not be persistent it may actually change we may go away from get up at some point to move to I don't know and I different version control system that has an understandable API potentially on or something else on but so we recommend that the identifier resolved to persistent landing pages that contain metadata and link to the software rather than directly to the source code itself and particularly because of landing pages are persistent of this is intended to ensure the longevity of the longevity of softer metadata even beyond the offers lifespan the may still on the in the in the paper there's a lot of other
Semi-automatic Metadata Editing – refer

Modify annotations: Edit Entity
Agenda

1. TIB and TIB AV-Portal
2. Semantic Video Analysis
3. Named Entity Linking
4. Data Quality
5. Semi-automatic Metadata Editing
6. Versioning of RDF-Data
Versioning of RDF Data – Reasons

- Availability of prior versions and data consistency
  - Reversibility
  - Linking/querying specific versions
  - Investigation of changes
  - Assessment of data quality
  - Representation of dynamic processes
Versioning of RDF Data – TailR

- Revisioning System for Linked Open Data
- Developed by Hasso Plattner Institute
- Storage
  - Snapshots and delta storage
- Push-API
  - Push and store RDF versions
- Memento-API
  - Link and access prior RDF versions (mementoes)
TailR – Snapshot

RDF-Store

SNAPSHOT

SNAPSHOT 1

TailR System

http://../video/43127
TailR – Revision

TailR System

SNAPSHOT 1

REVISION 1
http://av.tib.eu/.../video/43127
Felix Saurbier
2017-06-22T14:00:24

DIFF

DELETE
INSERT

RDF-Store
Metadata Editor
TailR – Access & Rollback

Access Revision 1

- \(?\text{datetime}=2017-06-22T14:00:24\)

Revert to Revision 2

- Replay revisions from last snapshot

TailR System

SNAPSHOT 1

REVISON 1

http://av.tib.eu/.../video/43127
Felix Saurbier
2017-06-22T14:00:24

REVISON 2

http://av.tib.eu/.../video/43127
Jörg Waitelonis
2017-07-21T12:10:14

REVISON 3

http://av.tib.eu/.../video/43127
Jane Doe
2017-07-24T11:20:51
More Infos
KNM@tib.eu
av.tib.eu

Contact
Felix Saurbier
T +49 511 762-14645, felix.saurbier@tib.eu