

© S. Staab, 2003

AIFB

Practical Aspects of Metadata

Steffen Staab
Institut AIFB
Universität Karlsruhe (TH)

© S. Staab, 2003

AIFB

What is Metadata?

“Data about data”

“Metadata not only identifies and describes an information object; it also documents how that object behaves, its function and use, its relationship to other information objects, and how it should be managed” (Baca)

Slide 2

© S. Staab, 2003

AIFB

Metadata is commonly divided into three types:

- **DESCRIPTIVE** - that which describes the object, relating to what the object contains or is about
- **ADMINISTRATIVE** - that which indicates the who, what, where, how aspects associated with the object's creation and preservation
- **STRUCTURAL** - that which relates to the formal set of associations within or among individual information objects, i.e., how a system functions

Slide 3

© S. Staab, 2003

AIFB

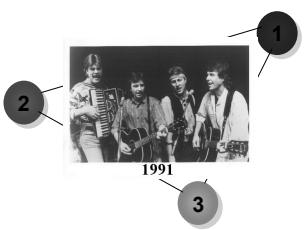
Concepts to know:

- Types of metadata

Descriptive
Title = Nitty Gritty Dirt Band

Structural
File type = jpg

Administrative
Rights holder = NGDB




Slide 4

© S. Staab, 2003

Concepts to know (cont.):

- Semantics
 - What's in a name?
- Syntax
 - We got grammar
- Interoperability
 - Sharing...

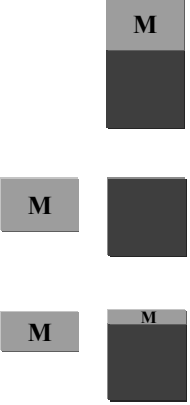


Slide 5

© S. Staab, 2003

Concepts to know (cont.):

- Metadata objects can be:
 - Embedded in the resource
 - Separate from the resource
 - Both embedded and separate



Slide 6

© S. Staab, 2003

Why is Metadata important?

- Increased accessibility
- Retention of context
- Expanding use
- Multi-versioning
- Legal issues
- Preservation
- System improvement and economics

Slide 7

© S. Staab, 2003

Types of Metadata Schema

- MARC (MACHINE-Readable Cataloging)
- Dublin Core
- MPEG-7
- EAD (Encoded Archival Description)
- RDF (Resource Description Framework)
- IEEE LOM (Learning Objects Metadata)
- ADL Scorm
- RSS (RDF Site Summary)

Slide 8

© S. Staab, 2003

MARC

AIFB

The MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form.

Slide 9

© S. Staab, 2003

MARC Tags

AIFB

- (008) Fixed fields
- 02x Reference number (ISBN, ISSN, manufacturer's no.)
- 041 Language code
- 1xx Main entry (author, composer, corporate name)
- 24x Title
- 260 Publication info.
- 300 Physical description
- 4xx Series
- 5xx Notes (general, summaries, contents, access, etc.)
- 6xx Subjects
- 7xx Added entries (additional authors, titles, relational works)
- 856 Electronic location and access

Slide 10

© S. Staab, 2003

Example of a MARC Bibliographic Record

AIFB

```

001 47809907
008 010820m1996uuuuouoac s 000 0 eng dnamla
040 UMK|cUMK
245 00 Club Kaycee[h[computer file] : jazz sites & sounds from the Sound Archives & music
collections of the Miller Nichols Library at the University of Missouri--Kansas City.
260 Kansas City, Mo. : University of Missouri--Kansas City, cc1996-
500 Title from title screen; as viewed on Aug. 20, 2001.
520 Information about the jazz music and musicians in Kansas City.
538 Mode of access: World Wide Web via the Internet. System requirements for audio:
RealAudioTM Player or Player Plus.
650 0 Jazz|zMissouri|zKansas City.
650 0 Jazz|zKansas|zKansas City.
700 1 Haddix, Chuck.|4aut
710 2 University of Missouri--Kansas City.|bMarr Sound Archives.
856 41 |zAccess University of Missouri--Kansas City Web site|uhttp://
www.umkc.edu/orgs/kcjazz/

```

Slide 11

© S. Staab, 2003

Example of a MARC Bibliographic Record (OPAC)

AIFB

```

LANG: eng LOCATION: knb BIB LVL: m BCODE3: - CAT DA:08-21-01
MAT TYPE: a COUNTRY: mou
OCLC # 47809907
TITLE Club Kaycee [computer file] : jazz sites & sounds from the Sound Archives & music
collections of the Miller Nichols Library at the University of Missouri--Kansas City.
IMPRINT Kansas City, Mo. : University of Missouri--Kansas City, c1996-
NOTE Information about the jazz music and musicians in Kansas City.
NOTE Mode of access: World Wide Web via the Internet. System requirements for audio:
RealAudioTM Player or Player Plus.
NOTE Title from title screen; as viewed on Aug. 20, 2001.
SUBJECT Jazz -- Missouri -- Kansas City.
SUBJECT Jazz -- Kansas -- Kansas City.
ADD AUTHOR Haddix, Chuck, Author.
ADD AUTHOR University of Missouri--Kansas City. Marr Sound Archives.
MARC Access University of Missouri--Kansas City Web site http://www.umkc.edu/orgs/kcjazz/

```

Slide 12

MPEG-7

- Formally named “Multimedia Content Description Interface”
- Developed by MPEG (Moving Picture Experts Group)
 - A standard for describing the multimedia content data using eXtensible Markup Language (XML)
 - Offers a standardized set of descriptors, a set of description schemes, a language to specify description schemes (Description Definition Language), and one or more ways to encode descriptions

EAD

- The EAD Document Type Definition (DTD) is a standard for encoding archival finding aids using the Standard Generalized Markup Language (SGML).
- Basically consist of two segments:
 - 1) a segment that provides information about the finding aid itself (its title, compiler, compilation date, etc.)
 - 2) a segment that provides information about a body of archival materials (a collection, a record group, a fonds, or a series)

Dublin Core

- The Dublin Core is a metadata element set intended to facilitate discovery of electronic resources and was originally conceived for author-generated description of Web resources.
- The characteristics of the Dublin Core that distinguish it as a prominent candidate for description of electronic resources include **Simplicity**; Semantic Interoperability; International Consensus; Extensibility; and Metadata Modularity on the Web.

Dublin Core Metadata Elements

- | | |
|--------------|---------------|
| • Identifier | • Format |
| • Language | • Description |
| • Creator | • Contributor |
| • Title | • Subject |
| • Publisher | • Rights |
| • Date | • Relation |
| • Source | |

Example of a Dublin Core record

Identifier 47809907
Language en
Title Club Kaycee
Publisher University of Missouri—Kansas City
Date 1996
Source RCA LPV-511
Format Internet resource
Description Information about the jazz music and musicians in Kansas City.
Contributor Haddix, Chuck, author
Contributor Middleton, Scott, engineer
Relation IsPartOf University of Missouri—Kansas City University Libraries web site
Subject Jazz – Kansas City
Rights University of Missouri—Kansas City, copyright holder

Slide 17

Example of a Dublin Core record

Identifier 538
Language en
Creator King Pleasure
Title Parker's mood
Publisher Prestige Records
Date 1953-12-24
Source Prestige: 880
Format Digital audio file
Contributor Clarke, Kenny, 1914-, musician
Contributor Middleton, Scott, engineer
Relation IsPartOf Club Kaycee web site
Subject Jazz – Kansas City
Rights Prestige Records (Firm), copyright holder

Slide 18

Qualified DC

- DC semantics are defined very broadly.
- Possible to:
 - refine the meaning of elements using 'type':
 - Relation TYPE=IsPartOf
 - associate value with externally defined 'scheme':
 - Subject SCHEME=LCSH
 - Date SCHEME=ISO 8601
 - indicate 'language' of value
 - Title LANGUAGE=en

Slide 19

How is DC currently used?

- Embedded into HTML Web pages
 - <META> tag
 - limited functionality
 - no structure
 - version 4.0 support for SCHEMES
 - syntax for qualified DC in <META> tags not well established

Slide 20

© S. Staab, 2003

DC in HTML

- <HTML><HEAD>
- <TITLE>UKOLN Home Page</TITLE>
- <META NAME="DC.Title" CONTENT="UKOLN: UK Office for Library and Information Networking">
- <META NAME="DC.Subject" CONTENT="national centre, network information support, library community, awareness, research, information services, public library networking, bibliographic management, distributed library systems, metadata, resource discovery, conferences, lectures, workshops">
- <META NAME="DC.Description" CONTENT="UKOLN is a national centre for support in network information management in the library and information communities. It provides awareness, research and information services">
- <META NAME="DC.Creator" CONTENT="UKOLN Information Services Group">
- </HEAD>
- ...

Slide 21

© S. Staab, 2003

DC in RDF

```

<RDF:RDF>
  <RDF:Description
    RDF:HREF="http://www.ukoln.ac.uk/metadata/">
    <DC:Title>The UKOLN Metadata Home Page</DC:Title>
  </RDF:Description>
</RDF:RDF>

```

Slide 22

© S. Staab, 2003

DC in RDF

```

<?xml:namespace
ns="http://purl.org/dublin_core/schema/" prefix="DC"?>

<RDF:RDF>
  <RDF:Description
    RDF:HREF="http://www.ukoln.ac.uk/metadata/">
    <DC:Title>The UKOLN Metadata Home Page</DC:Title>
  </RDF:Description>
</RDF:RDF>

```

Slide 23

© S. Staab, 2003

Central Characteristics of the Dublin Core Metadata Element Set

- Descriptive metadata for resource discovery (15 elements)
- Extensible (a starting place for richer description)
- Interdisciplinary (semantic interoperability)
- International (20 languages and growing)

Slide 24

© S. Staab, 2003

Extensibility (refined semantics)

- Ukrainian Doll model
 - improve description precision with sub-structure (sub-elements and schemes)
 - should degrade gracefully to preserve interoperability

Creator

Given Name	Affiliation
Surname	Contact Info

Slide 25

© S. Staab, 2003

Extensibility The Lego Metaphor

- Modular extensibility
 - additional elements to support local or discipline-specific requirements
 - complementary packages of metadata:

Slide 26

© S. Staab, 2003

DCMI Matrix of Semantics and Communities

	Title	Creator	Publisher	Contributor	Date	Relation	Source	Identifier	Language	Subject	Description	Coverage	Format	Type	Rights	Audience	Jurisdiction	MARC-Relater Codes	LCNAF
DCMI	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
DC-Education	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
DC-Government	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
DC-Library	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■

Slide 27

© S. Staab, 2003

Alternative Representations of DC Metadata

Japanese

English

Portuguese

Danish

Slide 28

© S. Staab, 2003

The Complete Matrix: A registry of metadata Semantics

Slide 29

© S. Staab, 2003

Target Point: Ontology based Metadata

Ontology

Annotation

```

<swrc:PhDStudent; rdf:ID="person_sha">
  <swrc:name>Siegfried Handschuh</swrc:name>
  <swrc:cooperateWith rdf:resource =
    "http://www.aifb.uni-karlsruhe.de/WBS/est#person_est"/>
</swrc:PhDStudent>
  
```

Web Page

URL

Slide 30

© S. Staab, 2003

Bad Experiences / Challenges

- **Consistency:** adhere to given ontology e.g. check of type constraints when creating instance relationships
- **Proper Reference:** Unique Ids e.g. at KA² three different IDs for our colleague Dieter Fensel
- **Avoid Redundancy:** Identify and reuse existing annotation
- **Relational Metadata:** Simple annotation tools only template like annotation
- **Maintenance:** Maintenance of knowledge markup
- **Ease of use:** Navigation of semantic structure, HCI related
- **Efficiency:** Automation of annotation process

Slide 31

© S. Staab, 2003

Design of CREAM

- Annotation in our context is a set of instantiations of:
 1. Classes
 2. Attributes: Properties from class instance to datatype instance
 3. Relationship: Properties from class instance to class instance
- Relational Metadata
- relationship instances, e.g. cooperatesWith(Siegfried,Steffen)

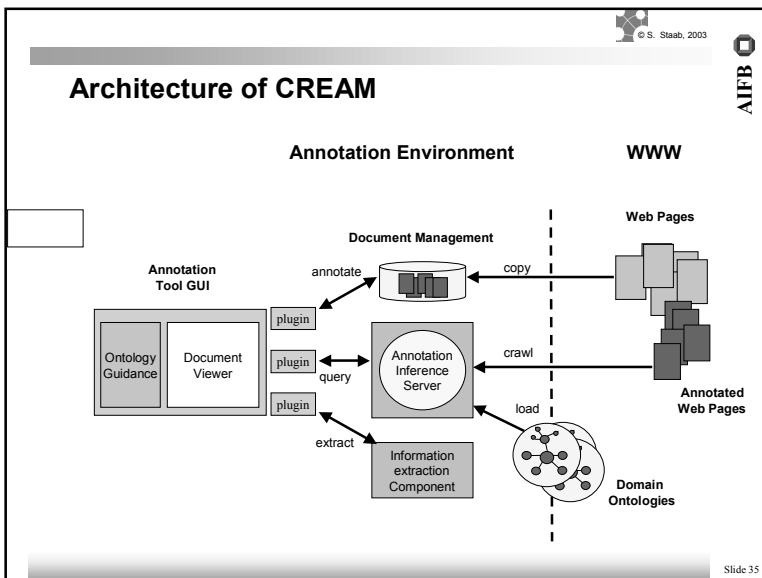
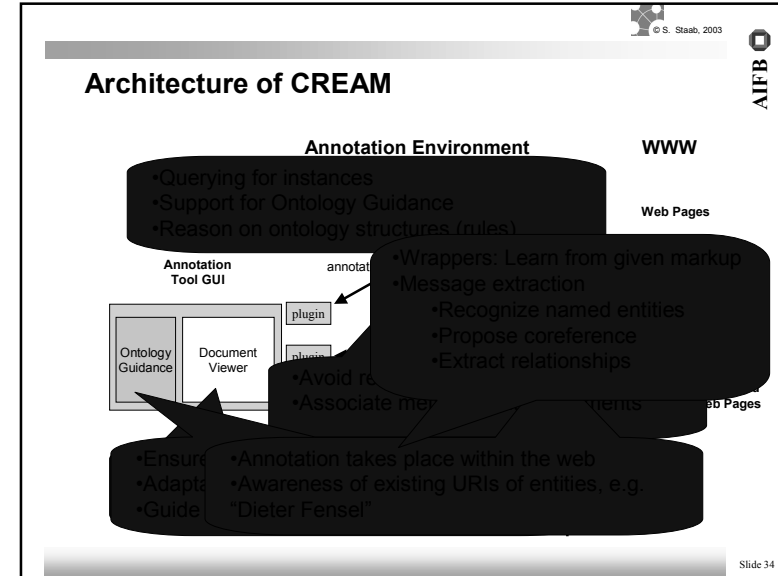
Slide 32

© S. Staab, 2003

Design Rationale - Linking Challenges with Required Modules

Requirement	Document Viewer	Ontology Guidance	Replication		Storage		Information Extraction
			Crawler	Annotation Inference Server	Document Management		
Consistency		X		X			
Proper Reference			X	X			
Avoid Redundancy			X	X	X		
Relational Metadata		X	X	X			
Maintenance					X	X	
Ease of use	X	X					X
Efficiency	X	X	(X)	(X)	(X)		X

Slide 33



- © S. Staab, 2003
- ## Implementation: Ont-O-Mat
- Instantiation of the CREAM framework
 - Ont-O-Mat Annotizer (Ontology Automat, Magic Annotation Tool)
 - Download Beta Version (0.1): <http://km.aifb.uni-karlsruhe.de/annotation/index.html>
 - Generic plug-in and service mechanism
 - Example: **Full Inference Server Plugin** or **Extended RDF API Plugin**
- Slide 36

© S. Staab, 2003

Components

- Ontology Guidance and Document Viewer: **Ontology Browser, HTML Browser**
- **Document Management**: straight forward, file system based
- **Annotation Inference Server**: Ontology-API plugin based on RDF-API or SiIRI (Ontobroker/F-Logic-based) plugin
- **RDF Crawler**: <http://ontobroker.semanticweb.org/RDFCrawler>
- **Information Extraction**: about to be integrated

Slide 37

© S. Staab, 2003

Ont-O-Mat

Ready: 1616.0x free

© S. Staab, 2003

Functions of Ont-O-Mat

1. Instance identification
2. Instance-Class relationships
3. Instance-Attribute relationships
4. Instance-Instance relationships

Slide 39

© S. Staab, 2003

Modes of Interaction

1. Annotation by Typing Statements
 - Exclusively with Ontology & Fact Browser
2. Annotation by Markup
 - Reuse Data from the Document Editor in the Fact Browser
3. Annotation by Authoring Web Pages
 - Reuse Data from the Fact Browser in the Document Editor

Slide 40

Annotation by Typing Statements

© S. Staab, 2003

The screenshot shows the OntoMat-Annotizer interface. On the left, an ontology browser displays a hierarchy of classes including Employee, Faculty_Member, Associate_Professor, Full_Professor, Lecturer, Researcher, Student, and PhD_Student. Below this is a table for 'Attributes' and 'Values' with columns for address, email, fax, first_name, and homepage. An input dialog is open, showing a text field with 'Steffen Staab' and buttons for 'Enter' and 'Cancel'. Below the dialog, a list of classes is shown, with 'Employee' and 'Faculty_Member' selected. Callouts with arrows point to the 'Generate Class Instance', 'Attribute Instance', and 'Relationship Instance' actions.

Slide 41

Annotation by Typing Statements

© S. Staab, 2003

This screenshot is similar to Slide 41, showing the same document and ontology browser. The input dialog is more prominent, and the class list below it is expanded to show 'Agent_Systems', 'Agents', 'Artificial_Intelligence', 'Business_Engineering', and 'Data_Mining'. The 'Generate Class Instance' callout now points to the 'Employee' class in the list.

Slide 42

Annotation by Typing Statements

© S. Staab, 2003

This screenshot is similar to Slide 42, showing the same document and ontology browser. The class list is further expanded to show 'Agent_Systems', 'Agents', 'Artificial_Intelligence', 'Business_Engineering', and 'Data_Mining'. The 'Generate Class Instance' callout now points to the 'Agent_Systems' class in the list.

Slide 43

Annotation by Markup

© S. Staab, 2003

This screenshot is similar to Slide 43, showing the same document and ontology browser. The class list is further expanded to show 'Agent_Systems', 'Agents', 'Artificial_Intelligence', 'Business_Engineering', and 'Data_Mining'. The input dialog is more prominent, and callouts with arrows point to 'Generate Class Instance', 'Attribute Instance', and 'Relationship Instance' actions.

Slide 44

Markup Class Instances

The screenshot shows the OntoMat Annotator interface. On the left, a tree view shows the ontology structure with 'Towards Semantic Web Mining' selected. The main window displays the HTML source of the document, with several elements highlighted in yellow to indicate class instances: 'year', 'title', 'year', 'Abstract', and 'Research Paper at International Semantic Web Conference (ISWC) 2002, June 9-12th, 2002, Sardinia, Italia'. A table at the bottom left shows the mapping between these elements and their corresponding class instances.

Attributes	Values
year	2002

Slide 45

Markup Attribute Instances

The screenshot shows the OntoMat Annotator interface. The main window displays the HTML source of the document, with several elements highlighted in yellow to indicate attribute instances: 'title', 'year', and 'Abstract'. A table at the bottom left shows the mapping between these elements and their corresponding attribute instances.

Attributes	Values
title	Towards Semantic ...
year	2002

Research Paper at International (ISWC) 2002, June 9-12th, 200

Abstract

Slide 46

Markup Relationship Instances

The screenshot shows the OntoMat Annotator interface. The main window displays the HTML source of the document, with several elements highlighted in yellow to indicate relationship instances: 'Semantic Web Mining aims at to improve, on the one hand, th', 'Web Mining, on the other h', and 'meet today, and sketches way:'. A table at the bottom left shows the mapping between these elements and their corresponding relationship instances.

Attributes	Values
year	Towards Semantic ...
year	2002

Slide 47

Annotation by Authoring Web Pages

The screenshot shows the OntoMat Annotator interface. The main window displays the HTML source of the document, with several elements highlighted in yellow to indicate annotation by authoring web pages: 'Towards Semantic Web Mining', 'written in 2002', and 'Towards Semantic Web Mining has author Gerd Szurawski'. A table at the bottom left shows the mapping between these elements and their corresponding annotation by authoring web pages instances.

Attributes	Values
year	Towards Semantic ...
year	2002

Create Text and if possible Links out of a Class Instance

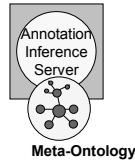
Attribute Instance

Relationship Instance generates simple text

Slide 48

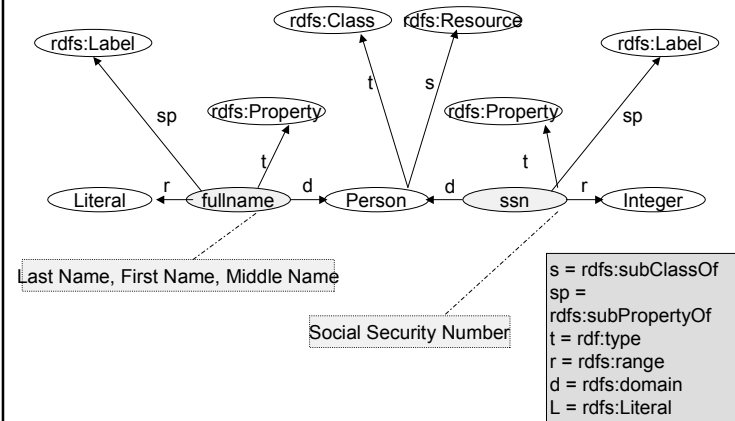
Meta Ontology

- **Modularization** of ontology development and use
- Describes how classes, attributes and relationships from the ontology should be used by the CREAM environment.
- The Meta Ontology supports the modes of interaction
- Meta Ontology characterizations:
 - Label
 - Default Pointing
 - Property Mode



Slide 53

Meta Ontology - Label



Slide 54

Meta Ontology - Label

- Labels are used at (at least) two points of interaction
 - Instance Generation
 - **RDF-API for a new URN as ID**
 - **Assign piece of Text with Attribute recorded as `rdfs:label`, e.g. `fullname` or `ssn`**
 - Content Generation
 - **Text is produces by `rdfs:label`, e.g. `fullname` or `ssn`**

- The connection is not objective.
- Their linkage depends on usage in a particular scenario.

Slide 55

Meta Ontology – Default Pointing

- Specify the default pointing behavior for class instances
- **Exploiting** the XPointer candidate recommendation: `CREAM:UniqueDPointer`, `CREAM:AutoDPointer` and `CREAM:AutoUniqueDPointer`
- **Instance-Generation** (Annotation by Markup):
 - `CREAM:AutoDPointer` or `CREAM:AutoUniqueDPointer`
- **Content-Generation** (Annotation by Authoring):
 - `CREAM:UniqueDPointer` or `CREAM:AutoUniqueDPointer`
- **Example:**
 - Person with properties `hasHomepage` (`CREAM:AutoUniqueDPointer`) and `fullname` (`Label`)

Slide 56

© S. Staab, 2003

Meta Ontology – Property Mode

- Distinguishes between different roles
- Reference:**
e.g. refer to the current U.S. president at <http://www.whitehouse.gov>
- Quotation:**
e.g. "Bill Clinton" as president of U.S. in 1999 at <http://www.whitehouse.gov>
- Unlinked Fact:**
e.g. a fact-attributes may be filled with "Spanish Civil War" for the reference pointing to the picture "Guernica", <http://www.guernica.swinternet.co.uk/guerni/ca.jpg>.

Slide 57

© S. Staab, 2003

Meta-Ontology & Modes of Interaction

- Annotation by Typing Statements
 - rdfs:label
 - default pointer: AutoDPointer
 - property mode: unlinked fact
- Annotation by Markup
 - rdfs:label
 - default pointer: AutoDPointer
 - property mode: reference, quotation
- Annotation by Authoring Web Pages
 - rdfs:label
 - default pointer: UniqueDPointer
 - property mode: reference, quotation, unlinked fact

Slide 58

© S. Staab, 2003

Related Work

knowledge acquisition frameworks

knowledge markup in the semantic web

annotation frameworks

SHOE (Knowledge Annotator)

Ontobroker (OntoPad)

WebKB

OntoEdit

Protégé 2000

CritLink

Amaya/Annotea

Slide 59

© S. Staab, 2003

Stages of Evaluation

- Instance identification

Prof. Dr. Rudi Studer

Institute for Applied Informatics and Formal Description Methods - AIFB
University of Karlsruhe
D-76128 Karlsruhe

Knowledge Management
- Instance-Class relationships
- Instance-Attribute relationships

Attributes	Values
address	D-76128 Karlsruhe
age	49
email	rstuder@aifb.uni-
fax	+49 721 608-658
firstName	Rudi
homepage	studer@aifb.uni- 115716
- Instance-Instance relationships

Slide 60

© S. Staab, 2003

Evaluation Setting

SWRC Ontology

Property	Data Type
age	INTEGER
address	STRING
email	STRING
fax	STRING
firstName	STRING
lastName	STRING
mailAddress	STRING
name	STRING
photo	STRING
position	Organization

Annotator 1: SAKB₁

instance-of (AIFB, INSTITUTE).
instance-of (Karlsruhe Univ, UNIVERSITY).
instance-of (KM, RESEARCHGROUP).
HASPARTS(AIFB, KM).

instance-of (SteffenStaab, EMPLOYEE).
NAME (SteffenStaab, EMPLOYEE).
AFFILIATION (SteffenStaab, AIFB).

instance-of (Alexander Maedche, PERSON).
instance-of (PAKM2000, CONFERENCE).
LOCATION(PAKM2000, Basel).

Example Webpage

Annotator 2: SAKB₂

instance-of (AIFB, INSTITUTE).
instance-of (SteffenStaab, ASSISTANTPROFESSOR).
NAME (SteffenStaab, ASSISTANTPROFESSOR).
instance-of (NLP, RESEARCHTOPIC).
HASTOPIC (SteffenStaab, NLP).

instance-of (AlexanderMaedche, PHDSTUDENT).
NAME (AlexanderMaedche).

instance-of (PAKM2000, EVENT).
LOCATION(PAKM2000, Basel).

Slide 62

© S. Staab, 2003

Preliminary Cross-annotator Comparison

- Semantic Annotation Knowledge Base I
- Semantic Annotation Knowledge Base II
- Set A1 of object identifiers
- Set A2 of object identifiers
- Set B1 of object-class relshps
- Set B2 of object-class relshps
- Set C1 of object-attribute relshps
- Set C2 of object-attribute relshps
- Set D1 of object-object relshps
- Set D2 of object-object relshps

Slide 62

© S. Staab, 2003

Definition

- Agreement-precision

$$AP(Q_i, Q_j) := |Q_i \cap Q_j| / |Q_i|$$

Slide 63

© S. Staab, 2003

Problem: Sliding Agreement

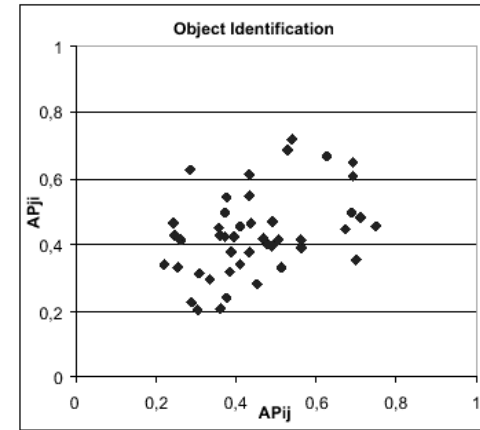
- Intuition:
- If I say „Siggi is a Person“ and Alex says „Siggi is a Man“, we differ only slightly
- If I say „Siegfried-Handschuh“ and Alex says „Siegfried_Handschuh“, we differ only slightly
- Sliding measures „IMA“, „OMA“, „RIAA“

Slide 64

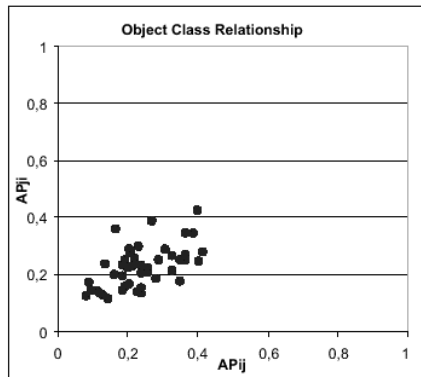
Core figures

Subject	$ I_i $ and $ c_i $	$ a_i $	$ r_i $
$SAKB_0$ (Gold standard)	124	237	183
$SAKB_1$ (anso)	111	206	97
$SAKB_2$ (eryi)	118	162	102
$SAKB_3$ (hela)	82	159	17
$SAKB_4$ (makr)	72	121	29
$SAKB_5$ (mama)	157	293	165
$SAKB_6$ (mari)	97	150	57
$SAKB_7$ (midu)	80	137	46
$SAKB_8$ (stse)	126	226	86
$SAKB_9$ (taso)	104	173	114
mean	107	186	90
standard deviation	26	53	55

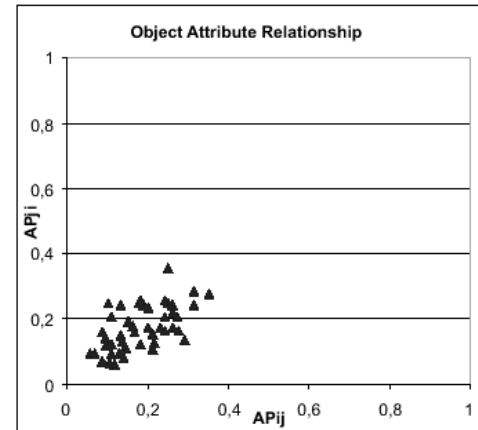
Comparing object identifiers



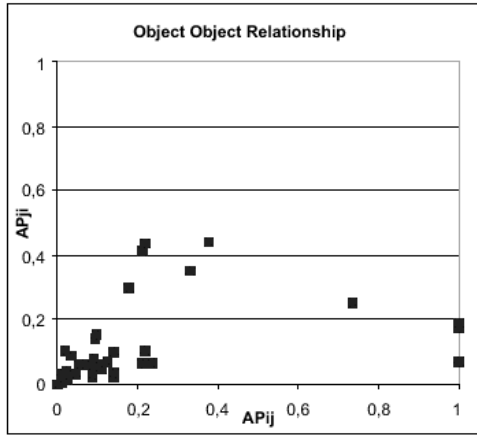
Comparing object-class relationships



Comparing object-attribute relationships



Comparing object-object relationships



Comparing comparisons

