

Datenerhebung, -aufbereitung und -archivierung
aus **sprachwissenschaftlicher** Sicht



Standards der

Datenerhebung, -aufbereitung und -archivierung

aus **sprachwissenschaftlicher** Sicht

Thomas Schmidt

Projekt Z2

„Computergestützte Erfassungs- und Analysemethoden Multilingualer Daten“

SFB 538 „Mehrsprachigkeit“

<http://www.exmaralda.org>

thomas.schmidt@uni-hamburg.de



Vorstellung: Projekt Z2

- initiiert im Juli 2000 von Jochen Rehbein
- seit Juli 2005 offizielles Teilprojekt des SFB 538
,Mehrsprachigkeit' : **Computergestützte Erfassungs- und Analysemethoden multilingualer Daten**, bis Juni 2011
- Projektinhalte:
 - Texttechnologische Verfahren für die Repräsentation linguistischer Daten entwickeln
 - Werkzeuge für die Erfassung und Analyse von Korpora entwickeln (→ EXMARaLDA)
 - Voraussetzung für Archivierbarkeit von Daten schaffen (→ Standards)
 - Aufbereiten der SFB-Daten für „Nachnutzung“
 - Methodologische Reflexion



Gliederung

1. Allgemeines zu Standards

- Wofür?
- Qualitätsmerkmale / Entscheidungskriterien?
- Was?

2. Standards für Korpora gesprochener Sprache

- Aufnahmen
- Metadaten
- Transkriptionen

Wofür Standards?



- Daten austauschen
 - zwischen Menschen (Verständlichkeit)
 - zwischen Computern (Interoperabilität)
- Daten (wieder)verwenden
 - jetzt
 - später (Nachhaltigkeit, Archivierbarkeit)
- Konkreter:
 - Daten mit Betriebssystemen A, B und C bearbeiten
 - Daten mit Werkzeugen X, Y und Z bearbeiten
 - Daten im Archiv XY auffindbar, durchsuchbar machen



Standards und „Standards“

- „echte“ Standards: ISO, DIN / Unicode, DC, MPEG
- „proposed“ Standards: LAF, OLAC Metadata
- (W3C-)Recommendations: XML, HTML, CSS, HTTP
- de facto Standards: *.doc, *.pdf , *.wav / Praat, CHAT
- Richtlinien: TEI, IPA
- Konventionen: HIAT, GAT, ICOR, DT
- Etablierte („best“) Praktiken: ELAN, EXMARaLDA, IMDI

Beziehungen zwischen Standards

- Standards arbeiten zusammen / bauen aufeinander auf

<http://www.fiss-bmbf.uni-hamburg.de/zur-forschungsinitiative.html>

- **http** – W3C-Recommendation für Übermittlung von Hypertext
- **html** – W3C-Recommendation für Repräsentation von Hypertext
 - **html** verwendet **css** (W3C-Recommendation)
 - **html** ist eine Anwendung von **xml** (W3C-Recommendation)
 - **xml** verwendet **Unicode** (ISO-Standard)
 - **xml** ist eine Anwendung von **SGML** (ISO-Standard)

Beziehungen zwischen Standards

- Standards können (sollten) interoperabel sein

Recording : MyTheory.flv

Recording : MyTheory.mp3

Recording : MyTheory.wav

Transcription MyTheory

EXMARaLDA: [Transcription] [Segmented]

Visualisation: [Partiture] [RTF] [PDF][XML] [Utterances] [Words]

Export: [TEI] [AG] [EAF] [Praat] [Chat] [FOLKER]

Was standardisieren?



- „Primärdaten“
 - Audio-Aufnahmen / Video-Aufnahmen
 - Textvorlagen, Bildvorlagen, etc.
- Sekundärdaten
 - geschriebene Texte / Transkriptionen
 - Schriftzeichen / Text- bzw. Transkriptionsstruktur
 - Annotationen
- Metadaten
 - Katalogdaten / Sitzungsdaten
- Beziehungen zwischen Primär-, Sekundär- und Metadaten

Was standardisieren?



- Oberflächenvariation (Form)
 - Geschlecht, Sex, sex : männlich/weiblich, m/f
 - Datum : 14/11/71, 14.11.1971, 14. November 1971
- Semantik
 - L1-Sprache = „Vor dem zweiten Lebensjahr im Familien-Umfeld erworbene Sprache“
- Beschreibungsmethode
 - Alter des Sprechers vs. Geburtsjahr des Sprechers

Was standardisieren?

- Oberflächenvariation

ich mache • eine Pause	(HIAT)
ich mache (.) eine Pause	(GAT)
ich mache * eine Pause	(DIDA)
ich mache .. eine Pause	(DT)
ich mache # eine Pause	(CHAT)
ich mache / eine Pause	(Czech system)



Was nicht standardisieren?

- Kategorien, Beschreibungen, die Gegenstand (nicht: empirische Basis) des Erkenntnisinteresses sind
- Beschreibungen / Kategorien, die sich einer systematischen Kategorisierung / Formalisierung (noch) entziehen

Beispiel



RV (Rudolf Völler)

Sex	male
Ausbildung: beruflich	Bürokaufmann, spielte seit dem 8. Lebensjahr Fuß
Ausbildung: schulisch	Realschule
Ausgeübte Berufe	bis 1996 Professi kickers Offenba München (1980- (1982-1987), AS Olympique Mar: 04 Leverkusen (:
Beruf	Teamchef der F seit 2000
Beruf der Mutter	Arbeit als Nähe
Beruf des Vaters	Gelernter Drehe Lagermeister, J Hanau 1860 (Ful
Familie	Drei Brüder. Sei einer Italienerin,
Funktion	Gast bei Walder

Rudi Völler: Wutausbruch (2 Speakers, 1 Transcription)

Aufnahmedatum	07.09.2003
folder	Rudi
Gesprächstyp	Fernsehinterview
project-name	EXMARaLDA DemoKorpus
Quelle	Sportschau (ARD) am 06.09.2003. Aufgenommen und bereitgestellt von Dr. Wilfried Schütte (Institut für Deutsche Sprache, Mannheim).
transcription-convention	HIAT (vereinfacht), für Nonverbales projekteigene: Mimik und Gestik der Sprecher werden nur ansatzweise angedeutet. Abkürzungen: LA= linker Arm, RA= rechter Arm, LH= linke Hand, RH= rechte Hand, KO= Kopf, OK= Oberkörper.
transcription-name	Rudi Völler Wutausbruch
Transkribent	Annette Schnieder
Transkription erstellt im	Januar 2004
Vorgeschichte	Das Fußball-EM-Qualifikationsspiels gegen Island

34▶	35▶ 36▶	37▶	38▶	39▶ 40▶	41▶
WH [v]		• • Rudi, darf ich zu dem Spiel heute noch mal zurückkommen?	Okay. • ((atmet ein)) Äh da		
WH [en]		• • Rudi, can I return to the game again?	Okay. • ((atmet ein)) So, there		
WH [nv]			lächelt kurz		
RV [sup]	leiser		leise		
RV [v]	Das sage ich euch ganz ehrlich.		Jà gerne.		
RV [en]	I honestly tell you that.		Yes, do so.		
RV [nv]	lässt die RH (Mikro) sinken				
WH [k]		betont ruhig und sachlich		betont ruhig und sachlich	
RV [k]			ruhig		

Kandidaten



- „Primärdaten“
 - WAV, MP3, MPEG, DIVX, MOV, ...
 - PDF, DOC, ...
 - JPG, PNG, BMP, ...
- Sekundärdaten
 - Unicode, UTF-8, UTF-16, ...
 - TXT, DOC, PDF, ...
 - TEI, CES, XCES, ...
 - EXMARaLDA, FOLKER, ELAN, PRAAT, CHAT, AG, ...
 - HIAT, GAT, CHAT, DIDA, IPA, ...
 - STTS, SUSANNE, ...
- Metadaten
 - Dublin Core, OLAC, IMDI, eTEI, ...
- Beziehungen zwischen Primär-, Sekundär- und Metadaten



Merkmale guter Standards

- Offen

- transparenter Entstehungs- und Veränderungsprozess
- dokumentiert
- veröffentlicht
- kostenlos
- Ansprechpartner, Feedbackmöglichkeit, Support



Merkmale guter Standards

- Praxiserprobt

- Unterstützung durch Tool(s)
- Dokumentierte Anwendungen / Beispiele
- Explizit eingeschränkter Anwendungsbereich, z.B.
 - Darstellung von Transkriptionen in Veröffentlichungen
 - Verwendung in Transkriptionstools
 - Auffinden von Transkriptionen in Archiven
- Stabil (nicht starr!)



Merkmale guter Standards

- Modular

- baut auf / nutzt / integriert andere Standards
- gestuft nach Grad der Konsensfähigkeit / Theoriespezifität, z.B.:
 - GAT Minimal-, Basis-, Feintranskript
 - HIAT 1, HIAT 2 (Intonation), HIAT 2 (Multimodalität)
 - MinCHAT, CHAT
 - Dublin Core, OLAC
 - TEI Lite, TEI



Merkmale guter Standards

- Formalisiert / formalisierbar
 - Konformität automatisch überprüfbar, z.B.
 - XML + Dokumentgrammatik (DTD, Schema)
 - EXMARaLDA + HIAT + Struktur-/Segmentierungs-Check
 - Abstraktion über Darstellung
 - Abstraktion über physikalische Speicherung (Datenformat → Datenmodell)



Merkmale guter Standards

- explizit, eindeutig, verbindlich, intersubjektiv

Als **Faustregel kann vielleicht** gelten: "so:" wird notiert, wenn eine Silbe gegenüber der **erwartbaren** "normalen" Länge **etwas** länger gedehnt wird. "so::", "so:::" **o.ä.** wird notiert, wenn eine **deutlich** größere Längung vorliegt [...]

Standards...



- für Tonaufnahmen

- WAV (PCM)

- + de facto Standard
 - + optimale Qualität
 - + gute Verarbeitbarkeit
 - Platzbedarf (ca. 10 MB / Minute bei 44kHz / Stereo)

- MP3 / OGG / ...

- + geringerer Platzbedarf (ca. 1 MB / Minute)
 - verlustbehaftete Komprimierung
 - eingeschränkte Verarbeitbarkeit

Standards...



- für Videoaufnahmen

- Containerformate („Verpackung“ für Video+Audio)

- MPEG, AVI (Audio Video Interleave), MOV (Apple Quicktime), WMV (Windows), RM (Real Media), ...

- Codecs (Komprimierung)

- ffmpeg, divx, xvid, H.264, ...

- Archivfähigkeit ↔ Verarbeitbarkeit

- Qualität ↔ Platzbedarf, Rechenleistung

Standards für Metadaten

- „Catalogue“ vs. „Session“ Metadaten
 - „Catalogue data“ – Einordnen/Auffinden einer Ressource in ein(em) Archivsystem, z.B.:
 - Name, Urheber
 - Rechteinhaber
 - Sprache
 - „Session data“ – Beschreibung einzelner Bestandteile der Ressource für Teilauswahl, Analyse, z.B.
 - Personenbezogene Daten (Alter, Geschlecht von Sprechern)
 - Kommunikationsbezogene Daten (Gesprächstyp, Datum)
 - Umstände der Erhebung (Aufnahmegerät, Transkribent)

Standards für Metadaten

- Struktur, Vokabular, Format

- Struktur

- Catalogue vs. Session,
- Personen vs. Kommunikationen

- Vokabular

- Attribute: „Alter“, „Age“, „age“
- Werte: „07;08,24“, „7 Jahre, 8 Monate, 24 Tage“

- Format

- XML
- Text
- RDB

Dublin Core



- Dublin Core Metadata Initiative (DCMI):
<http://dublincore.org/>
- Aus dem / für das Bibliothekswesen (auch digitale Archive)
- Katalogdaten
 - Technisch: format, type, ...
 - Inhalt: title, subject, coverage, description, ...
 - Personen & Rechte: creator, publisher, contributor, rights holder, ...
 - Vernetzung: source, relation, audience, ...
 - Lebenszyklus (Versionen etc.)

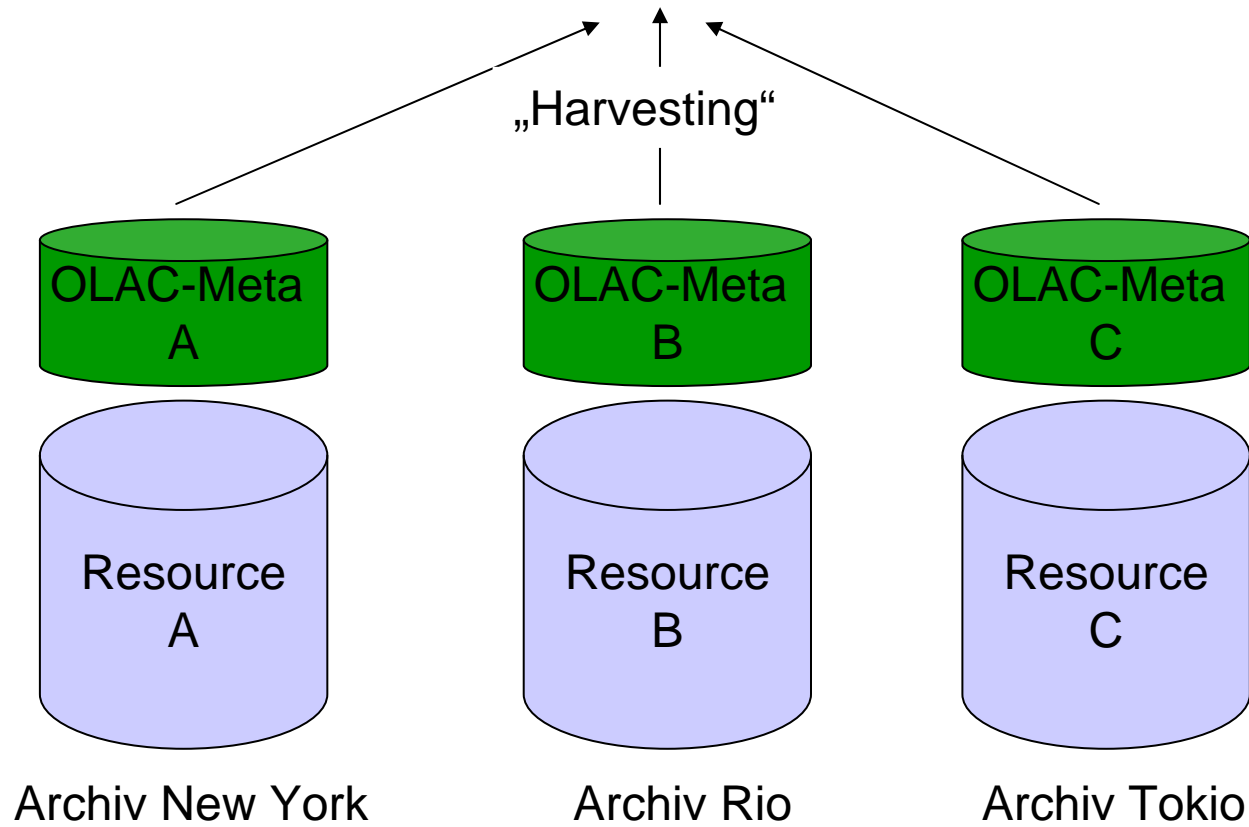


OLAC

- Open Language Archiving Community (<http://www.language-archives.org/>)
- Katalogdaten
- Dublin Core + zusätzliche Elemente für **linguistische** Ressourcen
- + Harvesting-Protokoll
- + Konzept für eine Infrastruktur („Verbundkatalog“)

OLAC

Archiv-Service



OLAC

Corpus Data Basket Settings

Skandinavische Semikommunikation

Unique Speaker Distinction //speaker/abbreviation

Key	Value
dc:contributor	Kurt Braunmüller, Institut für Germanistik I, Skandinavistik, Von Melle Park 6, D-20146 Hamburg, braunmueller@uni-hamburg.de Ludger Zeevaert,
dc:creator	Thomas Schmidt, thomas.schmidt@uni-hamburg.de
dc:format	XML
dc:publisher	Kurt Braunmüller, Institut für Germanistik I, Skandinavistik, Von Melle Park 6, D-20146 Hamburg, braunmueller@uni-hamburg.de
dc:rights	to be elaborated
dc:rightsHolder	Kurt Braunmüller, Institut für Germanistik I, Skandinavistik, Von Melle Park 6, D-20146 Hamburg, braunmueller@uni-hamburg.de
dc:subject	linguistic corpus of spoken language
dc:title	Korpus 'Skandinavische Semikommunikation'
dc:type	Collection
olac:compiler	Kurt Braunmüller, Institut für Germanistik I, Skandinavistik, Von Melle Park 6, D-20146 Hamburg, braunmueller@uni-hamburg.de
olac:data-inputter	Hanna Hedeland
olac:developer	Thomas Schmidt, thomas.schmidt@uni-hamburg.de, Kai Wörner, kai.woerner@uni-hamburg.de
olac:researcher	Kurt Braunmüller, Ludger Zeevaert, Per Warter, Franziska Watzke, Gerard Doetjes, Bernadette Golinski
olac:sponsor	Deutsche Forschungsgemeinschaft (DFG)

IMDI



- ISLE (International Standard for Language Engineering) Metadata Initiative (<http://www.mpi.nl/IMDI/>)
- „a proposed metadata standard to describe multi-media and multi-modal language resources“
- Eingesetzt am MPI für Psycholinguistik, Nijmegen
- Catalogue Data + Session Data

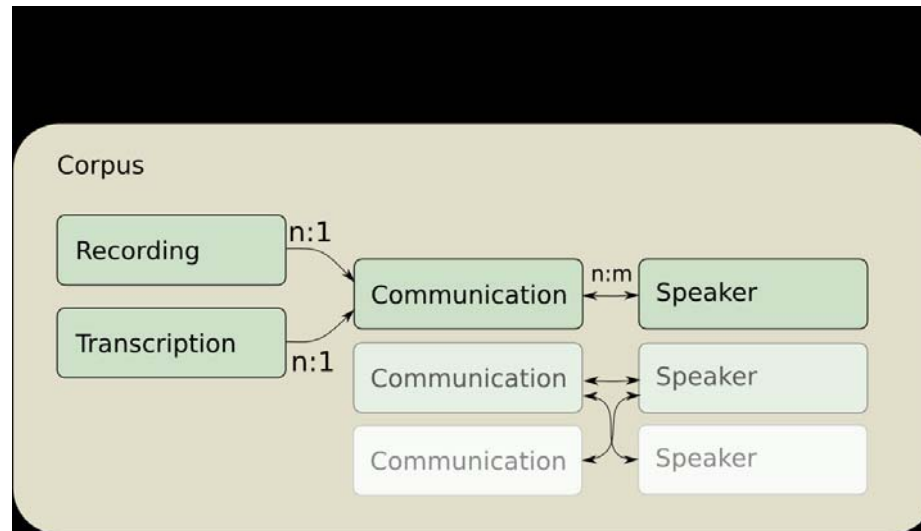
IMDI



- Sehr umfangreiches Vokabular
- Unterstützung durch IMDI-Editor, „Browsable Corpus“
- Praxiserprobt...
- ... mit mäßigem Ergebnis
 - großer Anteil aller Datensätze unvollständig
 - projektspezifische Spezifizierungen / Erweiterungen

EXMARaLDA Coma-Metadaten

- Catalogue Data mit beliebigem Vokabular (z.B. DC + OLAC)



- Session Data:

- feste Struktur, kein festes Vokabular
- einige wenige feste Datentypen (Location, Language, Media, etc.)

EXMARaLDA Coma-Metadaten



ENDFAS/SKOBI Gold Standard



[DC/OLAC]

dc:contributor Jochen Rehbein, Institut für Germanistik I, Von Melle Park 6, D-20146 Hamburg, rehbein@uni-hamburg.de Annette Herkenrath annette.herkenrath@tu-dortmund.de, Birsal Karakoc, Erkan Özdil

dc:creator Thomas Schmidt, thomas.schmidt@uni-hamburg.de

dc:format XML

dc:publisher Jochen Rehbein, Institut für Germanistik I, Von Melle Park 6, D-20146 Hamburg, rehbein@uni-hamburg.de

dc:rights free to use for research purposes provided permission has been granted by rights holder

dc:rightsHolder Jochen Rehbein, Institut für Germanistik I, Von Melle Park 6, D-20146 Hamburg, rehbein@uni-hamburg.de

dc:subject linguistic corpus of spoken language

dc:title Korpus 'Rehbein-ENDFAS / Rehbein-SKOBI'

dc:type Collection

olac:compiler Jochen Rehbein, Institut für Germanistik I, Von Melle Park 6, D-20146 Hamburg, rehbein@uni-hamburg.de

olac:data-inputter Secil Yusun, Nurkan Daricali, Hatice Yildirim, Tuba Özcan, Eylem Sentürk, and more

olac:developer Thomas Schmidt, thomas.schmidt@uni-hamburg.de, Kai Wörner, kai.woerner@uni-hamburg.de

olac:researcher Jochen Rehbein, Annette Herkenrath, Birsal Karakoc, Erkan Özdil

olac:sponsor Deutsche Forschungsgemeinschaft (DFG)

3 Communications

0604 : EFE01 tk - Schneewittchen (3 Speakers, 1 Transcription)

Bearbeitungsstand	f - Fertig
Diskursmodus	EFE01 tk - Schneewittchen
DOrt räuml.	Wohnung
Familie (Pseudo)	Akin
FileMakerID	1035
KassettenNr	0604
Kommentar	Weil sich Simge nicht an das ganze Märchen erinnern kann, fällt ihr das Nacherzählen schwer und sie wirkt z.T. gelangweilt.
Konvertierung	fertig konvertiert
Projekt	SKOBI
Thema Inhalt	Simge wurde das Märchen "Schneewittchen" vorgelesen. Sie soll es Kenans Mutter nacherzählen. Sie kann sich nicht an das ganze Märchen erinnern, sodass ihr Kenan z.T. beim Nacherzählen hilft. Die Kinderstimmen im Hintergrund stammen von Simges Geschwisterchen und werden nicht mittranskribiert.

Speakers: Ken; Sim; Oma6;

Location: Ankara, TR
Start: 2000-02-09T00:00:00
Duration:

Recording (13.581 minutes): EFE01tk_Akin_0604_f_SKO_090200.mp3

Recording (13.579 minutes): EFE01tk_Akin_0604_f_SKO_090200.wav

Transcription 06040

EXMARaLDA: [Transcription] [Segmented]
Visualisation: [Partiture] [RTF] [PDF][XML] [Utterances] [Words]
Export: [CHAT] [TEI] [AG] [EAF] [Praat] [FOLKER]

0736 : EFE04 tk - Maulwurf (3 Speakers, 1 Transcription)

0760 : EFE07 dt - Kennenlerngespräch (2 Speakers, 1 Transcription)

Sim (Simge: Kind Türkei)

Sex	male
Eigenschaft	ProbandIn/ Kind monolingual
Familie (Pseudo)	Akin
Nachname	--- anonym ---
Projekt	SKOBI
SprecherNr.	305
Versuchsgruppe	Kind Türkei
Vorname	--- anonym ---

Language: TRK TR

Sprachstatus MSprache

Location: Ankara, TR

Typ Geburt

Start:1992-03-06T00:00:00

Duration:

Location: Ankara, TR

Bemerkung1 Wohnort

Start:

Duration:

In Communications: 0604;

Ken (Kenan: ErwachseneR Deutschland)

Sex	male
Eigenschaft	InterviewerIn
Familie (Pseudo)	Güleç
Nachname	--- anonym ---
Projekt	SKOBI
SprecherNr.	293
Versuchsgruppe	ErwachseneR Deutschland
Vorname	--- anonym ---

Language: TRK TR

Sprachstatus MSprache

Language: GER DE

Bemerkung1 spricht Deutsch mit türkischem Akzent

Bemerkung2 spricht Türkisch besser als Deutsch

Sprachstatus FrSprache

Location: Konya, TR

Typ Geburt

Start:1963-01-01T00:00:00

Duration:

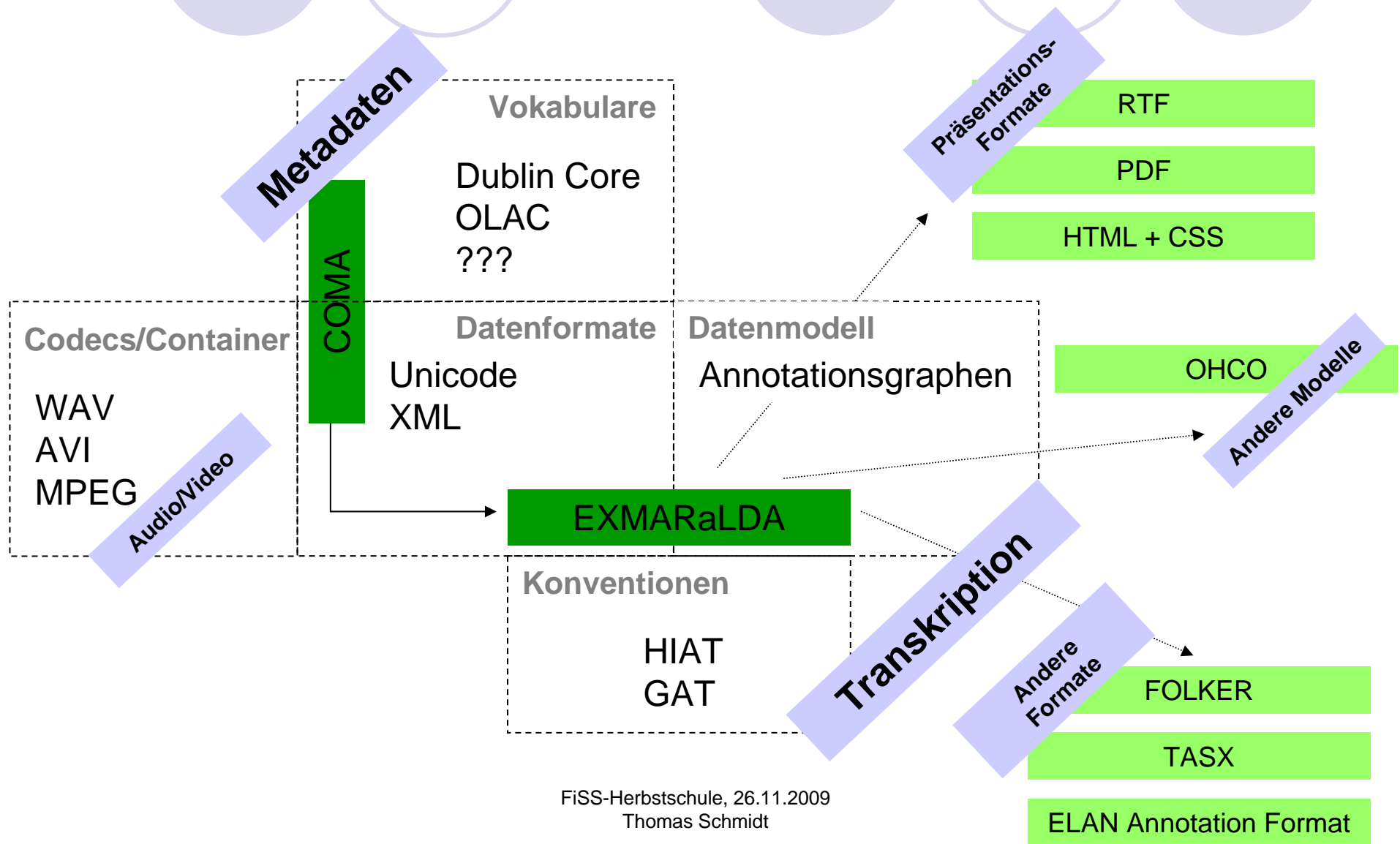
In Communications: 0604;



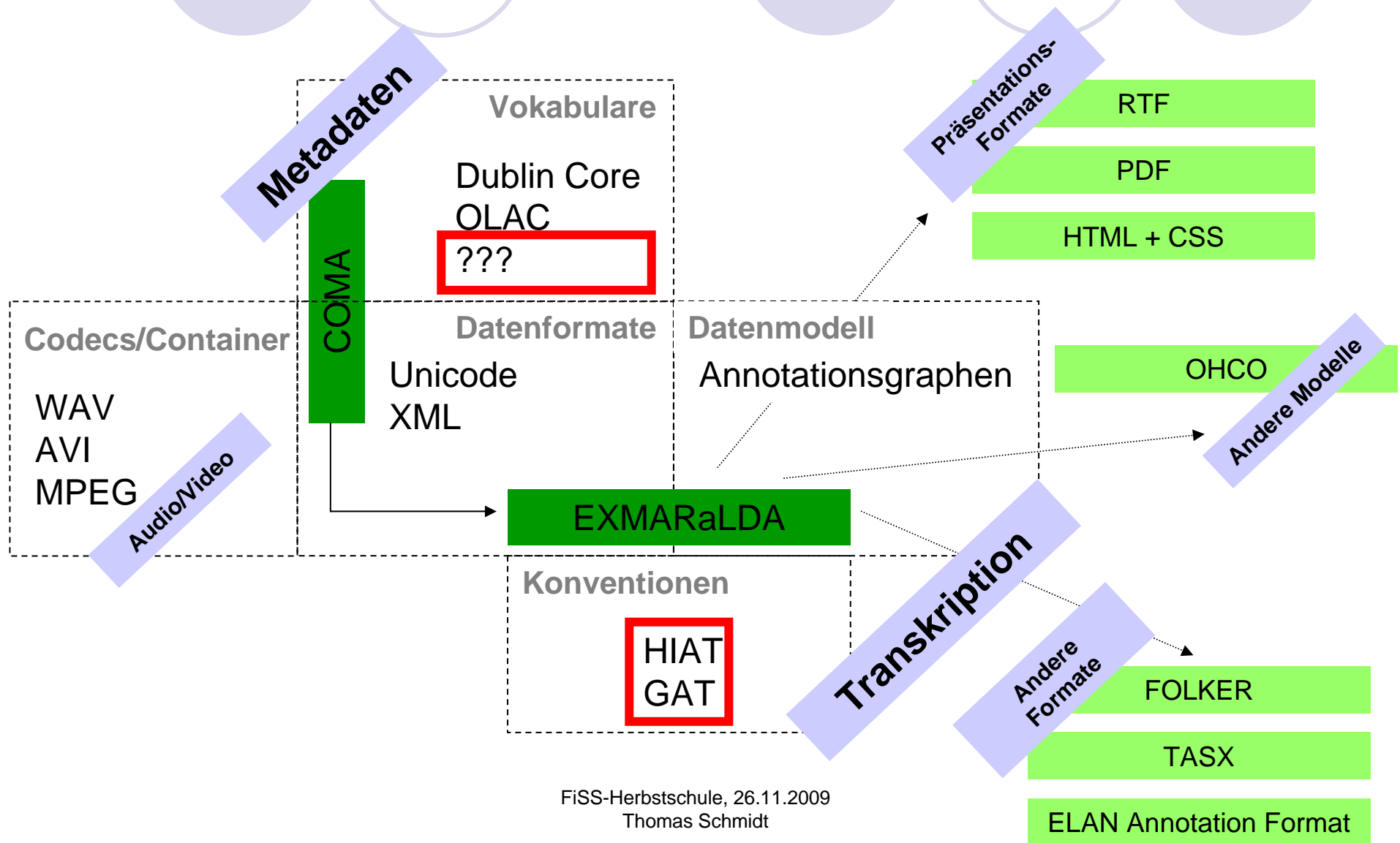
Standards für Transkriptionen

- Schriftzeichen (Unicode)
- Transkriptionsstruktur
 - Datenmodelle (Annotationsgraphen, OHCO)
 - Datenformate (XML)
- Transkriptionseinheiten und deren Beschreibung
 - Äußerungen/Intonationsphrasen, Wörter, Pausen, Nonverbales
 - Konventionen: HIAT, GAT, ...

Fundamental Interconnectedness Of All Things



Fundamental Interconnectedness Of All Things



Literatur



- **Bird, S. & Simons, G. (2002)** Seven Dimensions of Portability for Language Documentation and Description. In: Language (79) 557-582.
- **Lehmberg, T. & Wörner, K. (2008)** Annotation Standards. In: Lüdeling, A. & Kytö, M. (ed.): Corpus Linguistics - An international handbook, 484-501. Walter de Gruyter.
- **Schmidt, T. (2005)** Datenarchive für die Gesprächsforschung. Perspektiven, Probleme und Lösungsansätze. In: Gesprächsforschung (6) 103-126.