



Erwartungen der “Linguistik” Community an D-GRID

Peter Wittenburg, Hans Uszkoreit
Max-Planck-Institut für Psycholinguistik, Nijmegen - NL
DFKI, Saarbrücken – D

peter.wittenburg@mpi.nl
www.mpi.nl



Wer sind sie?



- zwei Institute stellvertretend für eine der Humanities Disziplinen, der Linguistik, mit einer breiten Verzahnung zur modernen Informationstechnologie (Human Language Technology) – insbesondere dem Semantic Web.
- das MPI ist seit Jahren mit führend auf dem Gebiet der Language Resources und des Resource Managements und eines der treibenden Kräfte zB zur Resource Integration
- das DFKI ist eines der führenden Institute auf dem Gebiet der HLT und der Erzeugung von integrierten Informationsangeboten
- beide Institute haben durch interdisziplinäre Projekte einen Einblick in die Komplexität der Aufgaben in DEN Humanities, können diesen Bereich jedoch nicht repräsentieren
- SW und Humanities sehr heterogen (Stand, Mind-Set, ...)



Sind sie GRID-fähig?



- beide Institute beteiligten sich an integrativen Initiativen, die sich heute unter dem Label GRID einordnen lassen
 - ISLE (International Standards for Language Engineering)
Metadata Infrastructure, world-wide Standards for L&LE
 - INTERA (INTEgrated European language Resource Area)
Tool-Resource Interchange, Open Metadata Domain
 - ISO TC37/SC4
Open Data Category Registries for Standardization of Semantics
Flexible Lexicon Infrastructure
 - DAM-LR (Distributed Access Management for LR - Europe)
DELAMAN (World-wide Network of LR Archives)
Data-GRID, URIDs, Shibboleth, ...
 - LT World (Language Technology Portal)
Integrated information



IMDI Metadata Netzwerk



- momentan arbeiten ~40 Inst. an einer integrierten MD Domäne
- insbesondere in Deutschland (Kiel, Berlin, Leipzig, Hamburg, Bonn, Erfurt, Bochum, ...)

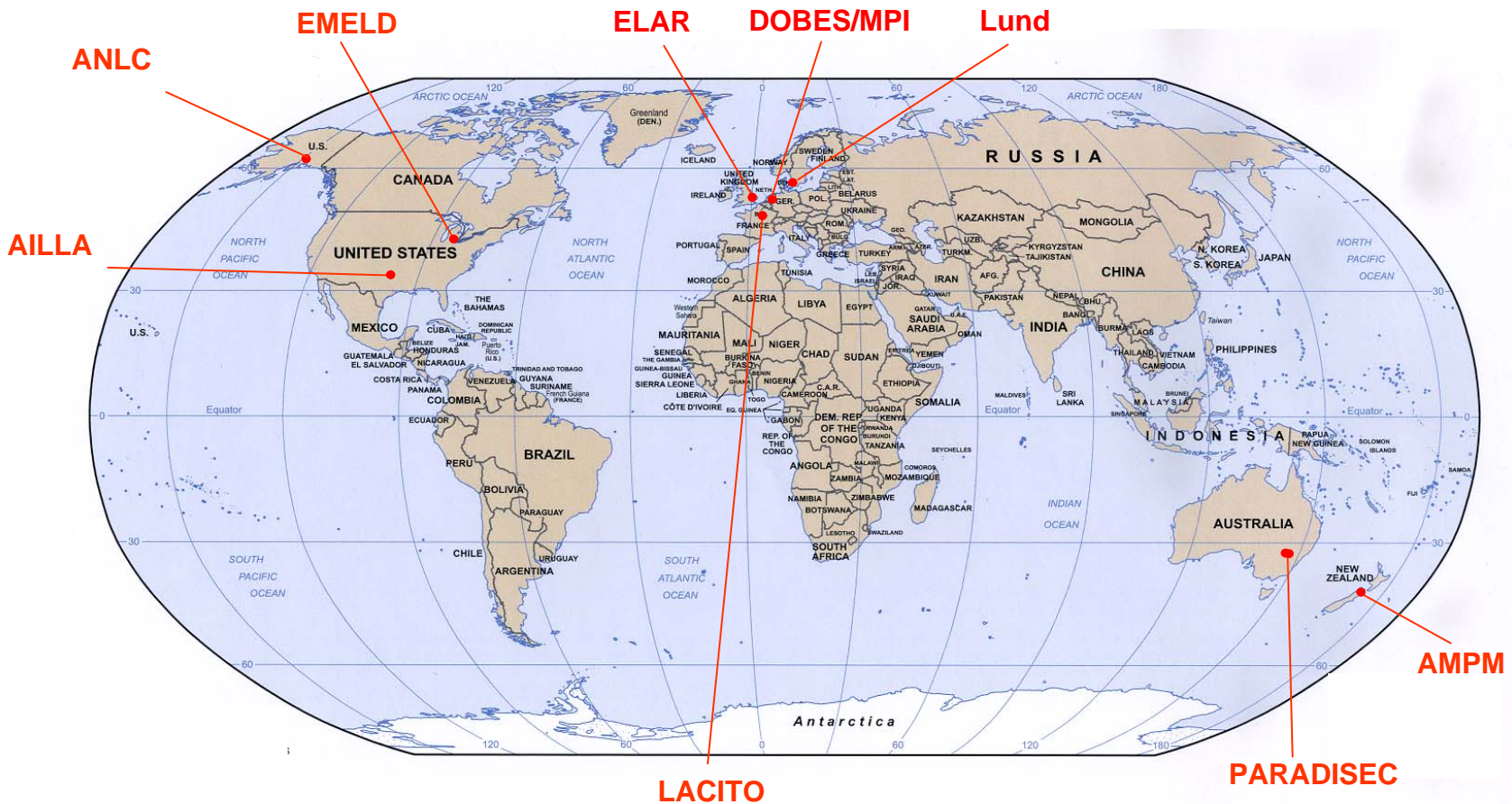




DELAMAN Network



- Digital Endangered Languages and Music Archive Network
- data distribution for LT preservation purposes
- common user and access management domain
- November Internationaler Workshop in Nijmegen

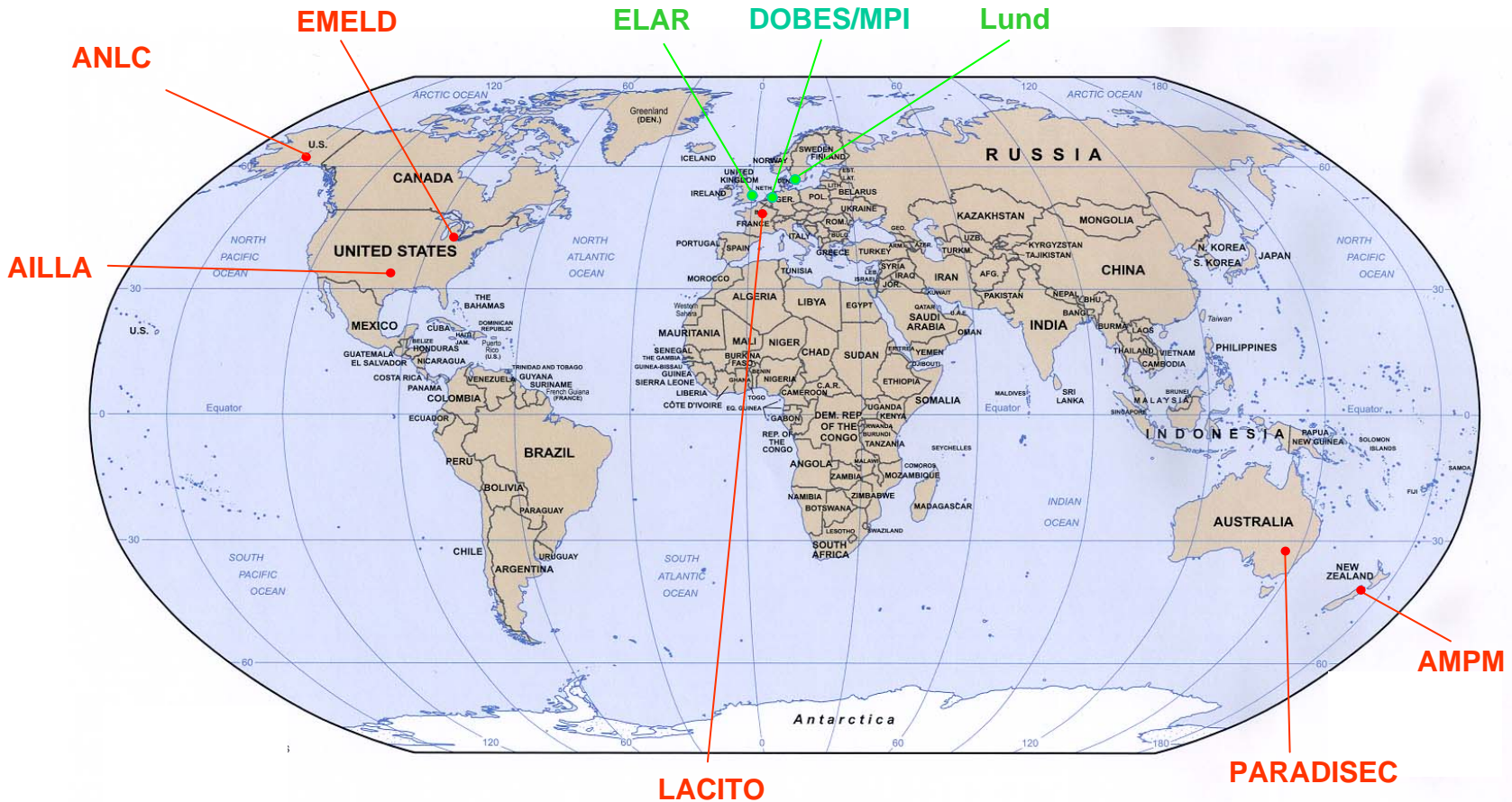




DAM-LR Partners



- Distributed Access Management for Language Resources
- very highly rated EU proposal – just got the evaluation results
- first step towards DELAMAN goals

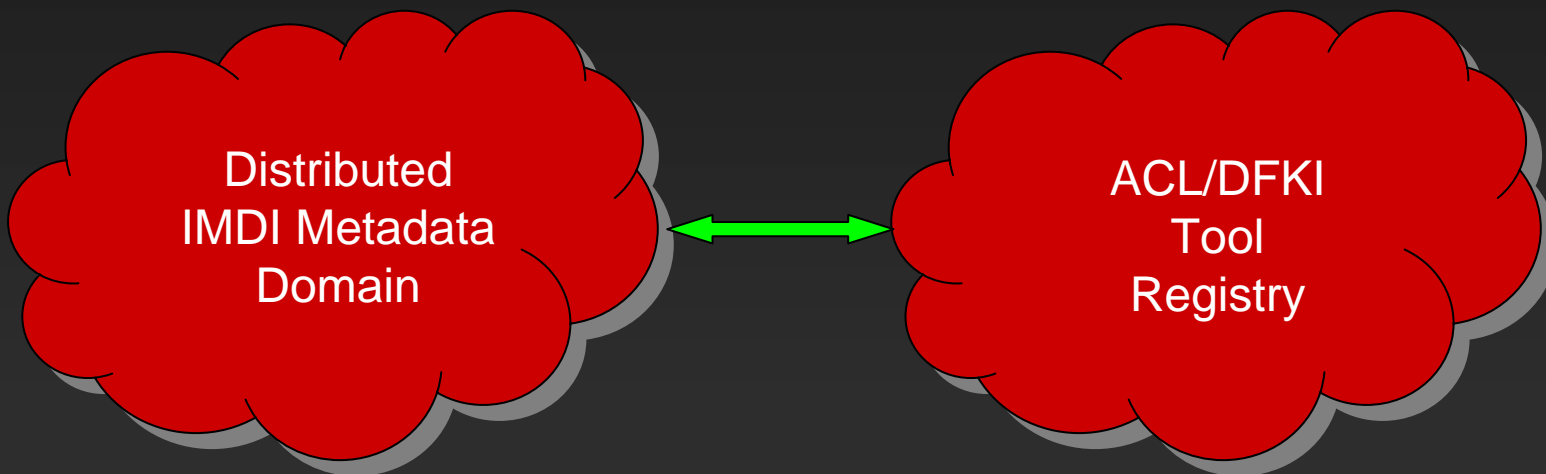




INTERA Integration



- users discover a useful resource and want to know about useful tools
- profile exchange and matching – LREP protocol



- user can select one of the tools and execute it
- what is exchanged: tool or resource?
- work in progress – meant to better understand the problems
- based on web-services



LT World at DFKI



Language Technology World

Search Projects

[Help / Search globally](#)

Name/Acronym

Person

Organization

All Fields

go

[Enter your projects into our Database](#)

Information & Knowledge

[Technologies](#)

[Abbreviations](#)

Players & Teams

[People](#)

[Projects](#)

[Organizations](#)

Systems & Resources

[R & D - Systems](#)

[Products](#)

[Links](#)

Communication & Events

[News](#)

[Conferences](#)

About LT World

Provided by



sponsored by



bmb+f

Bundesministerium für
Bildung und Forschung

General LT

[CLASSiKs Collaboration in Language and Speech Science and technology](#), [CoLLaTE](#), [COLLATE Computational](#)

[Linguistics and Language Technology for Real Life Applications](#), [EUROMAP Language Technologies](#), [KCL](#)

[Kompetenzzentrum Computerlinguistik](#), [NIPUK National Inventory Project](#), [SLATE Swedish Language Technology Information Centre](#)

Authoring Tools

[BibEdit](#), [Construction industry specification analysis and understanding](#), [FLAG Flexible Language and Grammar Checking](#), [Granska Grammar checking and proof-reading](#), [IDAS](#), [NSCOPE](#), [SAK](#), [SEATS specialised english author training system](#), [SMART Source Media Authoring Resources and Tools](#), [TETRIS Technologie-Transfer intelligenter Sprachtechnologie](#), [The Editor's Assistant](#), [Whiteboard](#), [Workbench WWW-based hypertexts for on-line scientific discussion and publication](#), [YPPS Yellow-Pages Pagination System](#)

Automatic Hyperlinking

[NSCOPE](#), [WWW-Persona](#)

Automatic Indexing

[OLIVE](#)

Automatic Language Identification

[MEMPHIS Multilingual Content for Flexible Format Internet Services](#)

Categorial Grammar

[CCG Parsing](#)

Categorization

[ASTRA Automatic creation of structured archives](#), [BINDEX Bilingual Automatic Parallel Indexing and Classification](#), [DESIRE](#), [ICC Innovation at the Call-Center](#), [PEKING People and Knowledge Cross-Linqual Information Gathering](#), [READ Recognition and Document Analysis](#)

Classification Clustering

[Search Support for Unfamiliar Metadata Vocabularies](#)

Clustering

[CLASS Collaboration in Language and Speech Science and Technology](#), [CLASSiKs Collaboration in Language and Speech Science and technology](#), [DESIRE](#), [GRACE Graphical Communication in HCI](#), [LILOG](#)

Communication

[MAP Multimedia Arbeitsplatz der Zukunft](#), [UMMI Understanding multiparty multimedia interactions](#)



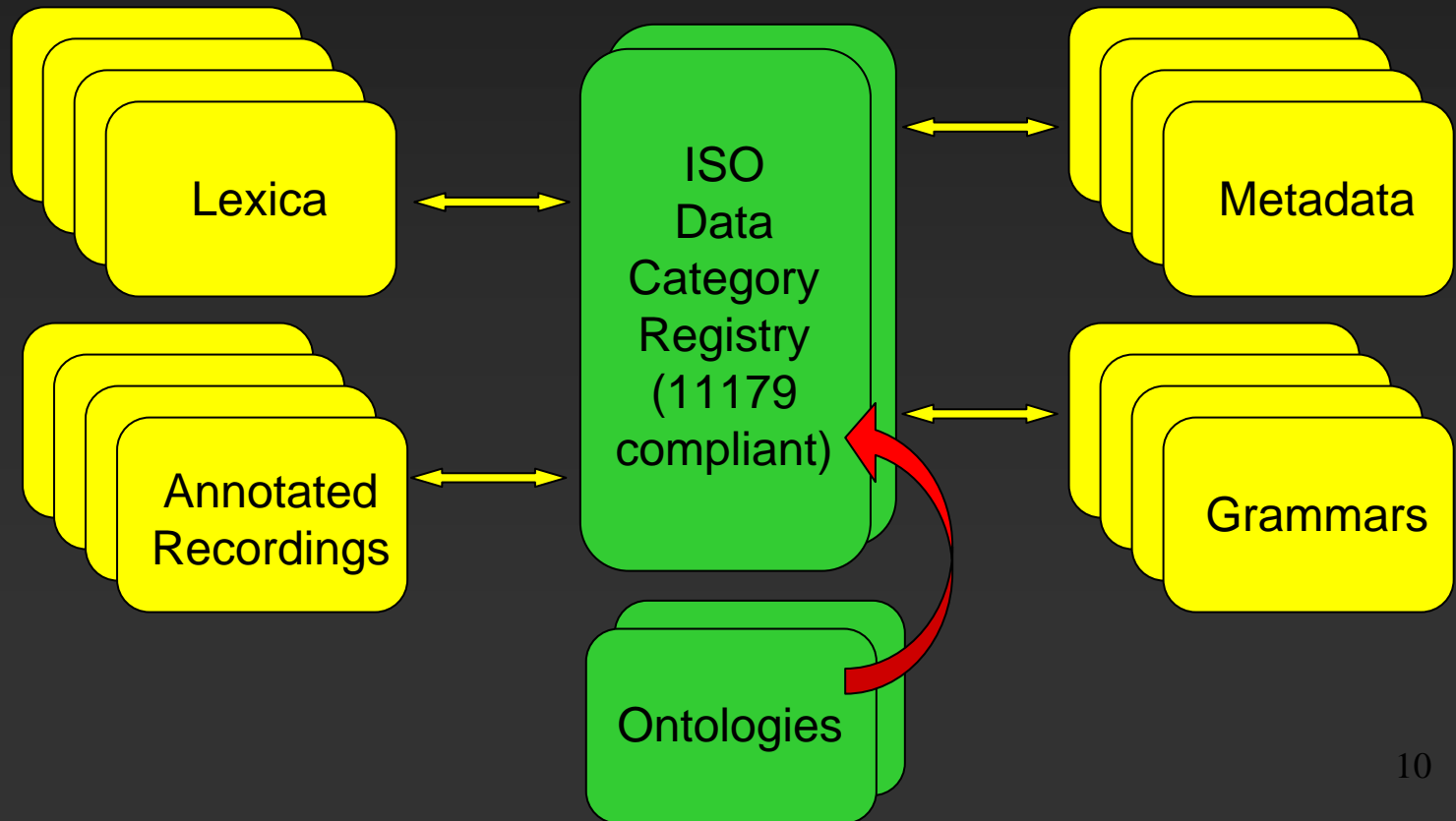
- typical information portal
- integrates information about persons, projects, patents, organizations, technologies, news
- built upon terminologies, thesauri, multidimensional classifications, ontologies
- integrating various domains



Current ISO TC37/SC4 Work



- filling slowly the Semantic Gap
- standardization beyond XML, MPEG, RDF, OWL, ...
- all focusing around the creation of Data Category Registries as the backbone for linguistic terminology
- in addition some standard formats for lexicons, annotations, ...





Verändert sich L<?



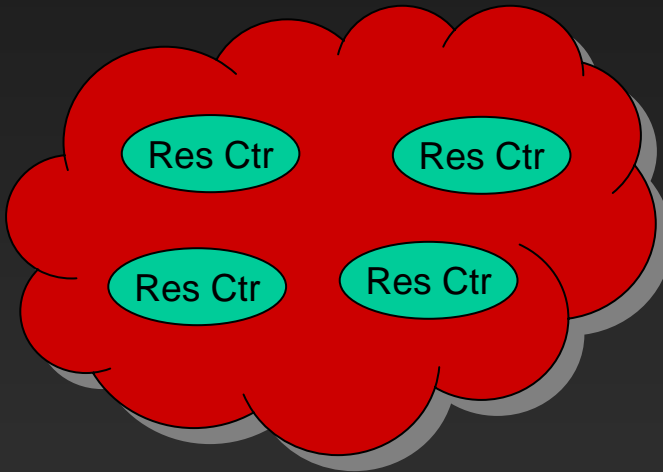
- starke Veränderungen in linguistischer Methodik und in der Sprachtechnologie (LT)
- hin zu empirischen Methoden gestützt durch größer werdende Sprachressourcen (LR)
- Beispiele typischer großer Kollektionen von LR
 - historical corpus (HU)
 - dialogue modeling (Campbell)
 - multimodal interaction (BAS, MPI, ...)
 - developmental corpora (CMU, MPI, ...)
 - documentation of Endangered Languages (AILLA, SOAS, MPI, ...)
 - Sign Language Studies (U Hamburg, U Stockholm, ...)
 -
- hin zu kollektiver Forschung und virtuellen Kollaborationen
eScience Paradigma



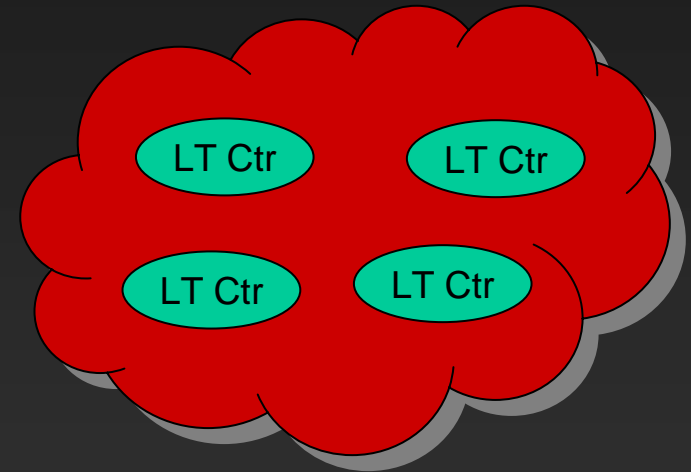
Wesentliche Herausforderung



- haben Zentren mit immer größer werdenden Korpora
- haben Zentren mit komplexer werdenden HLT Tools



Language Resource
Centers



Language Technology
Centers

- wir müssen **innerhalb** jeder Domäne integrieren
- wir müssen **zwischen** den Domänen integrieren



- Integration auf mehreren Ebenen
 - organisatorisch (OAI MHP, ...)
 - syntaktisch/Formate (XML Schema, MPEG, ...)
 - Semantisch (Semantic Web, ISO, Semantic GRID, ...)
- eine virtuelle instituts-übergreifende Metadaten Domäne
 - vergleichbar mit distribuiertem File System
 - aber disziplin-spezifische reichhaltige Beschreibung
- Open Access so weit wie möglich
 - aber auch IPR und Ethik (Video-Aufnahmen, ...)
- eine Benutzer-Identität zum flexiblen Kombinieren
- flexibler Einsatz und Kombination von HLT-Funktionen
- weitgehendst auf bestehendem aufbauen !!!
 - dh. sehr viele Mönchsarbeit (Adaptionen, Scripts, ...)
 - nicht nur ein Middleware Framework
- Distribuierung von Objekten wegen LZA



- towards resource services
 - easy metadata integration services
 - easy ingest and long-term preservation services (exchange, migration)
 - seamless discovery (browsing, searching, ...)
 - easy access to content (WSDL)
 - mining services on content
 - commentary services
 - ...
- associated aspects
 - user and access management services
 - user authentication services
 - common legal and ethical ground
 - URID association and resolving services
 - accounting services



- towards functional services
 - services registry and discovery (UDDI)
 - common API language (WSDL)
 - function request protocol and profile matching (brokerage)
 - all clear for singular requests
- many open questions for complex scenarios (no deep thoughts yet – or?)
 - how to evaluate offers (automatically)
 - how to define complex workflows (example Info Extraction)
 - how to find suitable conversion services (automatically)
 - how to find suitable ontologies (automatically)
- associated aspects
 - accounting
 - monitoring
 - load sharing / balancing
- collaborative frameworks (eScience) are based on p2p (?)



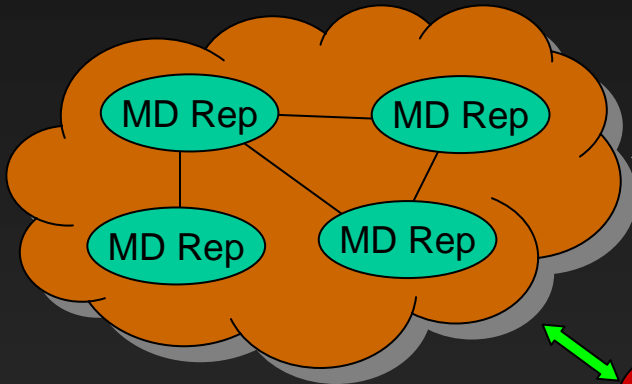
Was zu tun in D?



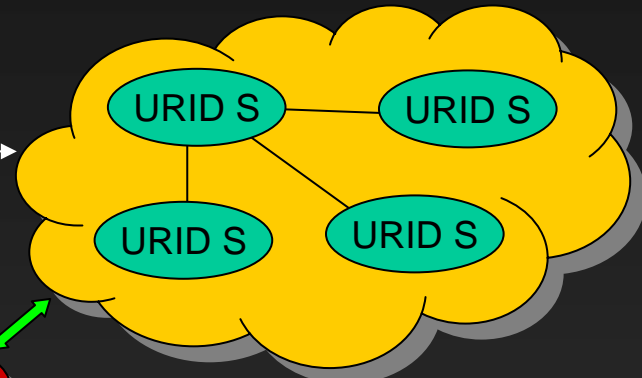
- es passiert ja schon einiges oder?
- was fehlt denn noch?
 - Bewußtsein in der SW und Hum Community!!!
 - Methodenschulung (Kategorisierung, Standards, ...)
 - IRISS Institut in NL, eScience Programm in UK, ...
 - ECHO Beispiel als Situationsbeschreibung
 - noch viele Insel-Projekte mit projekt-spezifischen Lösungen
 - zu stark auf sich selbst bezogen mit Web-Seite als Ziel
 - D sollte wieder mit an der Pole Position sein bez. Integration
 - Mittel für Institute für verschiedene Ebenen
 - Mitarbeit an Middleware – nicht alles der US überlassen
 - viele Adaptionen vor Ort



Distributed Metadata Domain



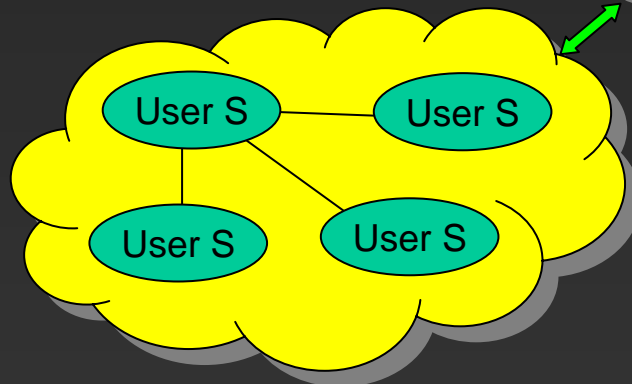
URID Resolving Service



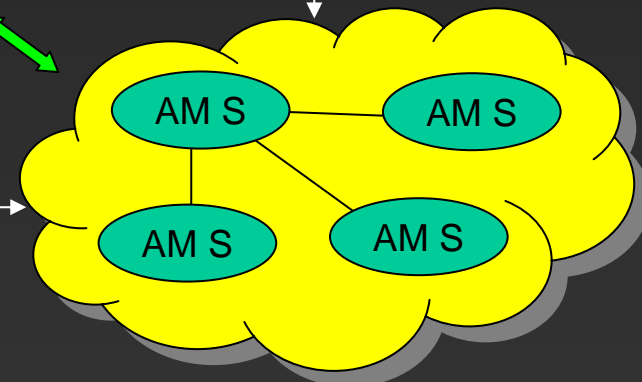
makes use of



is associated with



has access rights



User & group Management

Access Management

This is about integration middleware.