

# ***TextGrid:***

**Modulare Plattform für verteilte und kooperative  
wissenschaftliche Textdatenverarbeitung. Erstellung eines  
Community-Grid für die Geisteswissenschaften**

**- Gekürzte Version des Antrags zum öffentlichen Gebrauch -**

## **Vorwort (Februar 2006)**

TextGrid ist bisher das einzige geisteswissenschaftliche Projekt im Rahmen der D-Grid Initiative. Obwohl das Projekt - verspätet zu den anderen D-Grid Projekten - im Februar 2006 gestartet ist, verweist es doch auf Vorarbeiten, die wesentlich länger zurückliegen. Der hier vorliegende Antrag wurde im Oktober 2004 für den 1. D-Grid Call geschrieben. Seitdem ist die Entwicklung weiter fortgeschritten. Gerade die in TextGrid einfließenden Grid- und Semantischen Technologien gehören zu den zurzeit aktivsten Forschungsbereichen. Auch im Bereich der Textwissenschaften ändern sich die Parameter fortlaufend, nicht zuletzt durch Kooperationen mit spannenden Initiativen. Diese laufenden Entwicklungen werden direkt in die Projektarbeit mit einbezogen.

Die hier vorliegende gekürzte Version des Antrages spiegelt den Stand von Oktober 2004 wider und wird nicht weiter aktualisiert.

## - INHALTSVERZEICHNIS -

1. Ausgangslage .....	4
2. Ziele.....	5
2.1. Gesamtziel des Vorhabens .....	5
2.2. Bezug des Vorhabens zu den förderpolitischen Zielen .....	6
2.3. Wissenschaftliche und/oder technische Arbeitsziele des Vorhabens .....	7
3. Stand der Technik, bisherige Arbeiten der Antragsteller .....	8
3.1. Einführung in die Problematik .....	8
3.2. Bisherige Arbeiten der Antragsteller.....	9
4. Beschreibung der Arbeitspakete.....	10
4.1. AP 1: Inhaltliche Studie mit Empfehlungen über die Nachnutzbarkeit internationaler Editionstools.....	11
4.2. AP 2: Entwicklung Community-spezifischer Werkzeuge (Annotations-, Analyse-Tools)...	13
4.3. AP 3: Anbindung der Community-Tools und Vorschläge für Entwicklungen an der Integrations-Plattform .....	18
4.4. AP 4: Entwicklung der Community Muster-Applikation.....	24
4.5. AP 5: Semantic Web und TextGrid = Semantic TextGrid .....	27
4.6. AP 6: Projektmanagement und Öffentlichkeitsarbeit.....	33
4.7. Ressourcenplan: Zusammenfassung.....	37
5. Verbundpartner (Konsortium).....	37
5.1. Beschreibung der Partner .....	38
5.1.1. Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB).....	38
5.1.2. DAASI International GmbH (DAASI).....	38
5.1.3. Saphor GmbH.....	39
5.1.4. Kompetenzzentrum für EDV-Philologie an der Universität Würzburg (U-Wür) .....	39
5.1.5. Institut für deutsche Sprache (IDS) .....	39
5.1.6. Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier (U-Trier).....	40
5.1.7. Institut für Sprach- und Literaturwissenschaft, TU Darmstadt (TU-Da) .....	40
5.1.8. Fachhochschule Worms (FH-Wo).....	41
5.2. Zusammenarbeit mit Dritten, assoziierte Partnerschaften .....	41

## 1. Ausgangslage

Die wissenschaftliche Beschäftigung mit Informationsobjekten, vor allem mit Textdokumenten, unter Einsatz moderner Informationstechnologie findet trotz erkennbarer Bemühungen um stärker kooperativ organisierte Lösungen derzeit vorwiegend in lokalen Installationen mit jeweils projektbezogenen Applikationen statt.

Die verschiedenen Textwissenschaften und die wissenschaftlichen Bibliotheken beschäftigen sich seit über zehn Jahren an der (Retro-)Digitalisierung und Auswertung von Texten, die Teil des kulturellen Gedächtnisses sind. So intensiv diese Bemühungen auch sind, so stehen sie – vor allem im internationalen Vergleich – doch erst am Anfang. Projektinitiativen wie beispielsweise 'Deutsch Diachron Digital (DDD). Ein historisches Referenzkorpus für das Deutsche' machen diesen Nachholbedarf in Deutschland deutlich. DDD als Community-Initiative zahlreicher Sprach- und Literaturwissenschaftler in Deutschland hat zum Ziel, für das Deutsche die digitale Quellenbasis und die Sichtbarkeit in der internationalen Wissenschaftslandschaft zu erlangen, die andere Sprachen und Kulturen längst haben.

Bislang lässt die unterschiedliche Wahl von Archiv- bzw. Präsentationsformaten und die dezentrale Bereitstellung ohne gemeinsame Schnittstellen keinen einheitlichen Zugriff auf eine sich bildende virtuelle Nationalbibliothek zu. Weder kann man ermitteln, welche Texte überhaupt schon digitalisiert vorliegen, noch in den umfangreichen Datenmengen integriert recherchieren. Zahlreiche Funktionalitäten werden deswegen immer wieder in proprietären Formaten und Anwendungen neu entwickelt, Insellösungen entstehen, Mehrfachdigitalisierungen werden vorgenommen usw. Insbesondere das Erstellen bzw. Bereitstellen von digitalen Texten geschieht ohne Bezug auf den schon heute ungeheuren Mehrfachnutzen, der durch die Integration in vorhandene Korpora oder Verknüpfung mit Erschließungsmaterialien wie z.B. Wörterbücher, Lexika, Nachschlagewerken und Nachweisinstrumenten erzeugt werden kann.

Die in ihren Zielvorstellungen stark an interdisziplinären Perspektiven interessierte Forschung stößt heute noch sehr rasch an ihre Grenzen, und zwar wegen des Einsatzes von proprietären Softwarelösungen, fehlender Standards bei Interfaces und Formaten sowie Beschränkungen der Speicherkapazitäten, um nur die wichtigsten zu nennen.

Unter den fünf Hauptarbeitsaspekten Publikation, Edition, Textdatenverarbeitung, Workflow und Rechtemanagement sind gegenwärtig keine Angebote zu sehen, die in der gewünschten Weise integrierte Instrumente mit definiertem Leistungsumfang enthalten, die den spezifischen Anforderungen der Textwissenschaften auch gerecht werden. So wird zum Beispiel der Begriff "Editieren/Edieren" im Zusammenhang philologischer bzw. geisteswissenschaftlicher (edv-gestützter) Bearbeitung wird nicht im Sinne eines gängigen Software-Editors mit einfachen Funktionen wie Text eintragen, löschen, korrigieren etc. verwendet, sondern nach geisteswissenschaftlicher Terminologie-Tradition, die ein wesentlich komplexeres Editieren/Edieren meint. Die entstehenden Texte sind nach diesem Verständnis Ergebnisse eines komplexen, erkenntnisgeleiteten Prozesses und damit Produkt und Basis der wissenschaftlichen Arbeit zugleich.

Ist dieser Sachverhalt schon innerhalb einzelner Disziplinen zu konstatieren, so ist er noch augenfälliger bei einem Blick über die Fachgrenzen hinaus, wenn spezifische Anforderungen ins Spiel kommen. In der Praxis wird ein großer Teil der Arbeitszeit gar nicht mit der Textarbeit im engeren

Sinne aufgewendet, sondern mit Recherche nach geeignetem Werkzeug, der Anpassung an individuelle Bedürfnisse, dem Erlernen der entsprechenden Technologien und ähnlichen Dingen belegt. Das ist um so bedauerlicher, als aus der rechnergestützten Textdatenbearbeitung, Edition und Publikation in allen geisteswissenschaftlichen und, sofern wissenschaftshistorisch orientiert, auch naturwissenschaftlichen Disziplinen Impulse für die Forschung kommen und noch stärker kommen werden.

Selbstverständlich gibt es auch heute schon Entwürfe für Normierung und Standardisierung von Formaten und Diensten, auf die für das geplante Vorhaben nach entsprechend gründlicher Prüfung gebaut werden kann.

## **2. Ziele**

### **2.1. Gesamtziel des Vorhabens**

Der Gesamtkomplex der wissenschaftlichen Textdatenverarbeitung wird mit der Einführung und Bereitstellung kollaborativer Methoden und durch Nutzung von Netzwerken mit verteilten Ressourcen und standardisierten Werkzeugen auf eine neue Basis gestellt. Diese Zielvorstellung trägt den Tendenzen einer zunehmend international vernetzten, interdisziplinären, mobilen und in virtuellen Arbeitsumgebungen operierenden Wissenschaft Rechnung.

Die Zielvorstellung kann in ihrer kürzesten Form mit dem Schlagwort „Virtuelle Forschungsbibliothek“ beschrieben werden, worunter – etwas präziser formuliert – eine grid-fähige Workbench für die philologische Bearbeitung, Analyse, Annotation, Edition und Publikation von Textdaten zu verstehen ist.

Eine grid-gestützte Architektur drängt sich für diese Zielstellung in zweierlei Hinsicht auf: Zum einen existieren schon heute als Ergebnis zahlreicher Digitalisierungs- und Texterfassungsvorhaben gewaltige Mengen an Daten, die verteilt vorgehalten werden, zusammen gesehen aber Datenquantitäten von mehreren Terabytes ausmachen. Durch die Verbindung der elektronischen Texte von wissenschaftlicher Referenzqualität mit ebenso für die wissenschaftliche Analyse qualifizierten digitalen Abbildungen der primären Textquellen kommen Volumina von Primärdaten (in philologischem Sinn kritischen Textdaten und Metadaten) zusammen, die sinnvoll nur noch über die Grid-Technologie mit transparent verteilter Rechenlast zu pflegen, zu verteilen und zu nutzen sind. Vor allem die Einbindung von Objekten in Bild- und Tonformaten erzeugt höchste Ansprüche an Speicher- und Netzkapazitäten. Komplexe Suchen in heterogen verteilten Textcorpora im Giga- bis Terabyte-Umfang profitieren erheblich von Grid-Technologien. Hier sei als besonderes Beispiel das einheitliche Durchführen von oftmals heterogen verteilten Prozeduren an verschiedenen, im Netz verteilten Texten genannt.

Die Verteilung der Ressourcen, die internationale Zersplitterung von Forschungs- und Arbeitsgruppen und lokal bereitgestellte, potentiell integrationsfähige Module für die wissenschaftliche Bearbeitung bilden den zweiten Grund für den Aufbau eines Community Grid. Nur auf diesem Weg kann eine Plattform entstehen, auf der zeit- und ortsunabhängig die weltweit gestreuten wissenschaftlichen Kompetenzen zur Bearbeitung von Texten gebündelt und mit einem Set von modularen Softwaretools ausgestattet werden, wobei bereits vorhandene partielle Lösungen adaptiert und integriert werden.

Für die Wissenschaftler ergeben sich höchst attraktive Konsequenzen, hauptsächlich im Hinblick auf die Schonung eigener Ressourcen, aber auch mit Blick auf die Weiterentwicklung von Methoden und die Verbesserung der Kommunikation. Die wissenschaftliche Diskussion wird in der Community bereits im Prozess der jeweiligen Vorhaben gefördert, in vielen Fällen, vor allem im interdisziplinären Kontext, überhaupt erst ermöglicht.

Je effizienter diese Ziele im Community Grid unterstützt werden sollen, desto wichtiger ist die konsequent nutzerorientierte, umfassende Bereitstellung nicht nur der Objekte, sondern vor allem auch der Dienste. Die noch so anspruchsvollen technischen Vorhaben können eine benutzerfreundliche Oberfläche haben. Um es mit dem Bild der „virtuellen Forschungsbibliothek“ auszudrücken: Angemessene, aber möglichst einfach gehaltene Orientierungs- und Ordnungssysteme sollen alle Nutzer in die Lage versetzen, sich schnell über die jeweils relevanten Angebote, Informationsobjekte wie Services zu informieren und diese auch zu nutzen.

Hinzuweisen ist zudem auf die Entwicklungsperspektiven mit der Zielvorstellung eines Semantic Grid. Die EDV-philologischen und textwissenschaftlichen Auswertungsmethoden, insbesondere hochgradig strukturierter Textsorten wie z.B. Wörterbücher, werden einen unverzichtbaren Grundstein für die zuverlässige Erfassung von Ontologien legen. Darauf wiederum können über Mapping-Verfahren Instrumente entwickelt werden, die ihrerseits wesentliche Dienste für eine Strukturierung der digitalen Welt nach semantischen Kriterien sind.

## **2.2. Bezug des Vorhabens zu den förderpolitischen Zielen**

Der Aufbau des hier vorgestellten Community Grid wird dazu beitragen, dass gerade in den Textwissenschaften die Kommunikationsnetze stärker als bisher als integraler Bestandteil wissenschaftlicher Arbeit verstanden werden. Dazu kann vor allem die Verbreitung der hier angebotenen neuen wissenschaftlichen Methoden erheblichen Mehrwert beitragen. Von optimierten Zugangsverfahren zu relevanten Dokumenten, deren wissenschaftlicher Erarbeitung und deren Präsentationen ist eine deutliche Verbesserung der Arbeitsökonomie zu erwarten. Mit der Entwicklung einer virtuellen Arbeitsumgebung durch Nutzung verteilter Datenressourcen und Einbindung von netzbasierten, synchronen und asynchronen Kommunikationsdiensten werden neue Formen der wissenschaftlichen Diskussion ermöglicht. Erleichterung und Beschleunigung des Meinungs austausches innerhalb der Fachgrenzen, aber vor allem darüber hinaus, können zur Belebung der Wissenschaftsdiskurse und in bisher noch kaum abzusehendem Maße zum wissenschaftlichen Fortschritt beitragen. Die geisteswissenschaftliche Community profitiert von den im Rahmen von e-Science und Grid entwickelten Technologien. Das Web-Services-Paradigma ("everything is a service") sorgt dafür, dass auch im geisteswissenschaftlichen Umfeld zukunftsfähige und international anschlussfähige Anwendungen zur Verfügung gestellt werden, die kompatibel zu Anwendungsentwicklungen im kommerziellen und naturwissenschaftlichen Bereich sind. Hervorzuheben sind hierbei die neuen Kollaborationstechnologien, die den gesamten wissenschaftlichen Diskurs nachhaltig beeinflussen werden. Die Geisteswissenschaften von der Entwicklung und dem Einsatz all dieser Möglichkeiten gerade auf einem ihrer Kerngebiete – Sprache und Texte – auszuschließen, würde national wie international die zukünftige Wettbewerbsfähigkeit eines wichtigen Wissenschaftszweig erheblich beeinträchtigen und dürfte auch nicht ohne Auswirkungen auf die anderen Wissenschaftszweige bleiben. Die Realisierung von TextGrid wird die

deutsche Wissenschaft in einem Bereich stärken, in dem sie traditionell eine international führende Rolle hat(te).

Von der Herstellung von Interoperabilität auf allen Ebenen innerhalb des Community Grid sind substantielle Beiträge zur Standardisierung und Normierung von Formaten und Verfahren zu erwarten. Ein zentraler Standard im Rahmen der Textwissenschaften ist mit der Text Encoding Initiative (TEI) etabliert. Eine aktive Mitarbeit im TEI-Consortium und anderen Standardisierungsgremien bildet daher eine wesentliche Voraussetzung für die internationale Anschlussfähigkeit der TextGrid-Gruppe. Ein weiteres Ziel stellt die Förderung von Trans- und Interdisziplinarität und Methodenvielfalt durch Integration von digitalen Medien verschiedenster Art bzw. netzgestützten Features zur Prüfung und Visualisierung von Arbeitsergebnissen zu nennen.

Und schließlich sei auf die vorauszusehende Akzeptanzverbesserung durch nutzerfreundliche, interaktive und weitgehend intuitiv verständliche Präsentationsformen verwiesen.

### **2.3. Wissenschaftliche und/oder technische Arbeitsziele des Vorhabens**

Insbesondere die Bereitstellung von Instrumenten für Textwissenschaftler bedeutet einen hohen technischen Aufwand, unabhängig davon, ob einzelne Module, deren Funktionalität im Community Grid angeboten werden soll, so gekapselt werden, dass eine Einbindung möglich ist, oder ob fehlende Elemente vollständig neu entwickelt werden müssen.

Bewähren muss sich das Gesamtangebot an integrierten Leistungen in der Praxis der textwissenschaftlichen Arbeit. Deshalb wird sich eines der Arbeitspakete mit der Herstellung von Musterapplikationen beschäftigen, an denen der Leistungsumfang der implementierten Instrumente kritisch geprüft und optimiert werden muss. Die zur Integration entwickelten offenen Schnittstellen zur Einbindung weiterer Services werden in einem europaweit akzeptierten Konsensverfahren in CEN/ISSS allen interessierten Kreisen in Europa vorgestellt.

Die im Community Grid bereitgestellten Features haben dann ihren Zweck erfüllt, wenn ihre Anwendung unmittelbar zur Generierung bislang noch nicht dokumentierter Zusammenhänge und damit zur Wissensvermehrung beiträgt. Mit Mehrwerten wie der Abbildung multivalenter Bezüge zwischen Informationsobjekten durch einen Link-Editor oder der Möglichkeit, Text- und Bildausschnitte einander zuordnen zu können, gewinnt die textbasierte Forschungsarbeit eine neue Qualität. Neue Sichtweisen werden durch Verbreiterung der Quellenbasis und rechnergestützte Anwendungen möglich.

Darüber hinausgehend wird angestrebt, TextGrid als Infrastrukturkomponente in den Textwissenschaften zu etablieren. Es soll über Projektgrenzen und nach Projektende ein zentrales Werkzeug für die Wissenschaften werden. Verbreitung und Nachhaltigkeit werden durch Kooperationen und Öffentlichkeitsarbeit aktiv verfolgt, und auch technische Sustainability-Konzepte wie offene Registry- und Repository-Lösungen werden Teil der TextGrid Architektur sein.

### 3. Stand der Technik, bisherige Arbeiten der Antragsteller

#### 3.1. Einführung in die Problematik

Textwissenschaftler verwenden den Computer schon seit über 50 Jahren, seit den 1970er Jahren sind Institutionen entstanden, die solche Anwendungen in Forschung und Lehre untersuchen und erstellen, z.B. die Abteilung 'Literarische und Dokumentarische Datenverarbeitung' am Rechenzentrum der Universität Tübingen. Heute hat sich ein breites internationales Forschungsfeld etabliert, das man umfassend als *Humanities Computing* [McCarthy] bezeichnet und innerhalb dessen sich Teilbereiche wie Computerphilologie [Jannidis], Korpuslinguistik [McEnery] u.a.m. ausdifferenziert haben.

Die textwissenschaftlichen Ansätze innerhalb der Geisteswissenschaften können mit rechnergestützten Methoden traditionelle Aufgaben neu lösen, z.B. Druckeditionen und Konkordanzen erstellen, oder mit ganz neuen Methoden arbeiten, z.B. digitale Editionen und Textkorpora erstellen und für die philologische, linguistische, sowie historische Recherche verwenden. In diesem Sinne sind Textwissenschaftler beteiligt an der gegenwärtig vollzogenen Digitalisierung kulturellen Wissens. Diese ist etwa im Bereich der Metadaten sehr weit fortgeschritten, im Bereich der Volltextdigitalisierung steckt sie hingegen noch in den Anfängen, zumindest wenn man das an der Menge des gedruckten Wissens misst.

Grundsätzlich lässt sich sagen, dass der Computer in den Geisteswissenschaften immer dann gebraucht wird, wenn große Datenmengen verarbeitet bzw. komplexe (statistische) Analysen erstellt werden müssen. Umfang der Datenmengen und Komplexität der Algorithmen erfordern immer besondere Speicher- und Rechenleistungen. Einzelprojekte können heute gut mit Standard-Arbeitsrechnern durchgeführt werden, müssen dann aber gerade auf den wesentlichen Vorteil verzichten, der sich aus dem Arbeiten im Netz ergibt: Erschließungsstrategien, die sich auf die gesamte verteilte Textmenge stützen und, was ein wichtiger Netzeffekt ist, durch die Verbesserungen und Ergänzungen der im Grid zugänglichen Korpora stets besser werden. Hinzu kommen die ganz neuen Auswertungsverfahren im Bereich der Philologie, die erst durch den Einsatz von Computern im Netz ermöglicht werden. Neue statistische Verfahren zum automatischen semantischen Clustering und zur Verschlagwortung von Texten, um nur ein Beispiel zu nennen, verbrauchen sehr viel Rechenzeit.

Die Situation der Softwareentwicklung in den Textwissenschaften ist noch weitgehend geprägt von lokalen und projektbezogenen Lösungen. Ganz ohne Zweifel liegt in philologischer Software wie dem seit den 1970er Jahren entwickelten Paket TUSTEP, das stets in enger Abstimmung mit den Anforderungen der Anwender weiterentwickelt wurde, eine unvergleichliche Sammlung von Wissen über Algorithmen, Lösungskonzepten und ganz unterschiedlichen Anforderungsprofilen vor. Ähnliches gilt für Programme wie WordCruncher, Tact usw. Der größte Teil der verfügbaren Software ist jedoch sehr spezifisch, unterstützt wichtige einschlägige Standards wie z.B. TEI und Unicode nicht oder nur unzureichend oder ist so schwer zu handhaben und einarbeitungsintensiv, dass sich viele interessierte Textwissenschaftler vor einer Anwendung dieser Tools scheuen. Nicht zuletzt hat sich das Abbrechen der Entwicklung von Programmen, die nicht als Open Source zugänglich waren, wiederholt als problematisch erwiesen. Vor allem aber ist kaum eines der vorliegenden Programme wirklich netzwerk- oder gar grid-fähig.

Das neue Paradigma der computergestützten Wissenschaften im Bereich eScience, also Wissenschaften unter Verwendung der neuen Gridtechnologien, hat sich zwar vorerst hauptsächlich in den Naturwissenschaften, wie Kernphysik, Meteorologie, Genetik etc. etabliert, es lassen sich

allerdings auch erste Ansätze für geisteswissenschaftliche eSciences ausmachen. So ist das 'Arts and Humanities Research Board' involviert in die UK-eScience Initiative (<http://www.nesc.ac.uk>), (vgl. [http://www.ahrict.rdg.ac.uk/informationresources/e\\_science.htm](http://www.ahrict.rdg.ac.uk/informationresources/e_science.htm)). Ebenfalls im Rahmen der eScience Initiative wurde ein Projekt gefördert, welches sich ausschließlich um Publikationsfragen kümmert: „A System for publishing Scientific Data“ ([http://www.nesc.ac.uk/action/projects/project\\_action.cfm?title=195](http://www.nesc.ac.uk/action/projects/project_action.cfm?title=195)). Selbstverständlich haben die vielen Arbeiten im Bereich der Computerlinguistik, insbesondere diejenigen, die unter den Begriff Semantic Web sowie unter automatisches Clustering von Textressourcen subsummiert werden können, eine weitreichende Konsequenz auf computergestützte Geisteswissenschaften. Überdies hat sich im Global Grid Forum eine Research Group namens „Humanities, Arts, and Social Science RG“ (<http://forge.gridforum.prg/projects/hass-rg>) gebildet, die sich allerdings in einem sehr frühen Status der Auslotung von Anwendungsmöglichkeiten der Grid-Technologie in den entsprechenden Communities befindet. Trotz dieser ersten Ansätze kann gesagt werden, dass das Grid-Computing noch nicht in den Geisteswissenschaften angekommen ist. Hier liegt eine Chance für die D-Grid-Initiative, neue Akzente im Grid-Computing zu setzen. Eine Anwendergemeinschaft mit einer langen EDV-Tradition steht jedenfalls bereit, die Errungenschaften des Grid-Computings auszunutzen. Einen Überblick über die deutschen Aktivitäten im Bereich der Philologie gibt die Bibliographie über Deutschsprachige Literatur zur Computerphilologie (<http://computerphilologie.uni-muenchen.de/jg99/bibliographie.html>).

### **3.2. Bisherige Arbeiten der Antragsteller**

Die Antragsteller arbeiten seit 2003 in der Würzburger Arbeitsgruppe <philtag> an der Definition und Entwicklung von philologischer und textwissenschaftlicher Software. Dort sind so verschiedene Kompetenzen wie die des Informatikers, des klassischen Philologen, des Mediävisten, des Linguisten, des Literaturwissenschaftlers und des Philosophen vereinigt. Ausgangspunkt für die Kooperation ist die Erkenntnis, dass TEI sich als zurzeit bestes Datenformat für die langfristige Archivierung von philologisch interessanten Texten erwiesen hat. Alle können auf Grund ihrer Beteiligung an Digitalisierungsprojekten auf intensive praktische Kenntnisse über die Verwendung von TEI zurückgreifen. Sie sind ebenfalls an der Weiterentwicklung der TEI-Richtlinien interessiert und verfolgen diese aktiv, etwa durch Mitarbeit im TEI-Council (Jannidis), in der Charters Encoding Initiative (CEI) (Rapp) oder in weiteren verschiedenen Arbeitsgruppen, die die TEI-Richtlinien überarbeiten und ergänzen. Dazu diente etwa das letzte Würzburger Treffen, zu dem mit Lou Burnard (European Editor der TEI, Oxford Computing Services) und Laurent Romary (TEI-Consortium und INRIA, Nancy) als Gastreferenten hochrangige Vertreter der TEI-Community anwesend waren. Die Orientierung an wesentlichen einschlägigen Standards bestimmt gleichzeitig die Mitarbeit an Gremien wie der Dublin Core Metadata Initiative (SUB im Advisory Board der DCMI) oder im Editorial Board des Metadata Encoding Transmission Standard (METS).

Die Antragsteller sind an der Digitalisierung und Erstellung von Editionen, von digitalen Wörterbüchern und von Korpora umfassend beteiligt (in Auswahl):

- Deutsches Wörterbuch von Jacob und Wilhelm Grimm: [www.dwb.uni-trier.de](http://www.dwb.uni-trier.de) (Trier)
- Mittelhochdeutsche Wörterbücher im Verbund: [www.mwv.uni-trier.de](http://www.mwv.uni-trier.de) (Trier)
- Hartmann von Aue-Portal: Auf der Basis umfangreicher Vorarbeiten, die Handschriftenabbildungen und –transkriptionen, kritische lemmatisierte und mit den mhd.

Wörterbüchern verknüpfte Texte enthalten, wird das Portal zu einem umfangreichen Informationssystem, zu einen zentralen Autor des deutschen Mittelalters ausgebaut (<http://www.fgcu.edu/rboggs/hartmann/Heinrich/AhMain/AhHome.htm>) (Trier).

- Der junge Goethe in seiner Zeit (<http://www.jgoethe.uni-muenchen.de>) (Jannidis)
- Das Projekts "Historisches Korpus": Sammlung historischer deutscher Texte der Zeit zwischen dem Ende des 18. und der Mitte des 20. Jahrhunderts (IDS)

Im Rahmen dieser und anderer Projekte haben sich die Beteiligten dabei mit den Anforderungen an philologische Software im allgemeinen vertraut gemacht und den Einschränkungen der vorliegenden Programme kennen gelernt. Dabei sind sie zu der Einsicht gelangt, dass lokale Lösungen entscheidende Vorteile der Digitalisierung verschonen und somit die Verbreitung des kulturellen Erbes, wie es Netzwerkeffekte bewirken könnten, verzögern.

Ein erster Schritt der Würzburger Arbeitsgruppe bestand daher darin, die Anforderungen an textwissenschaftliche Software und die Architektur einer modularen integrativen Lösung zu definieren, was in einer Reihe von Vorträgen und Treffen sowie der Diskussion in einem Online-Forum geschah.

Dies war der Ausgangspunkt für die Konzeption von Prototypen für einzelne Problemlösungen, die im Rahmen von Seminaren und Abschlussarbeiten entwickelt wurden.

DAASI International GmbH verfügt über einschlägige Technologieexpertise insbesondere in den Bereichen Middleware-Infrastruktur, Security und Informationsmanagement. Viele an Hochschulen durchgeführte Projekte können als Vorarbeiten angesehen werden, insbesondere weil hierdurch Open Source Module und Bibliotheken zu Fragen der Authentifizierung, Autorisierung, Public Key Infrastructure und Verzeichnisdienste erstellt worden sind. Die Entwicklung und der Pilotbetrieb einer Schema-Registrierung (<http://www.schemareg.org>) ist eine Plattform zur Veröffentlichung von Ergebnissen des Projekts. Schließlich sind die Arbeiten im Rahmen des GGFs, insbesondere zu Grid Information Services, Grid PKI und CIM-basierende Schema-Modellierung sowie die generelle Expertise zu Grid-Technologien, die durch langjährige Mitarbeit in den Arbeitsgruppen und Diskussionen mit Gridexperten entstanden ist, innerhalb des Projekts sinnvoll einsetzbar.

Die Firma Saphor GmbH verfügt über langjährige Erfahrung in den Bereichen Datenstrukturierung und -konvertierung, Publishing und DTD-Entwicklung. Für die Konvertierung und Aufbereitung von Textdaten existieren zahlreiche Programmbausteine, die für die Entwicklung der geplanten Module herangezogen werden können. Zur Publikation strukturierter Daten (SGML/XML) liegen Satzroutinen für TEI nach xsl:fo sowie für TEI nach LaTeX vor, die streng an den jeweiligen Standards orientiert sind. Web-Publishing mit Cocoon und Web Services sind weitere Bereiche, in denen Saphor über entsprechende Kompetenzen verfügt.

#### **4. Beschreibung der Arbeitspakete**

Das Ziel ist die Entwicklung eines Community-Grid für textbasiert arbeitende Disziplinen, welches die folgenden Basisfunktionalitäten unterstützt: Werkzeuge für wissenschaftliche Editionsarbeit, Definition von standardisierten Schnittstellen für Publikationssoftware, Module für wissenschaftliche Textdatenverarbeitung sowie Verwaltung definierter und kontrollierter Zugriffe auf Daten und Werkzeuge. Damit wird eine interdisziplinäre, internationale und vernetzte virtuelle Forschungsplattform konstituiert.

Gestützt auf Reports über die Nachnutzbarkeit vorhandener Software (AP1) werden einzelne Module für die Grid-Anwendung adaptiert oder neuentwickelt (AP2) und in die Integrationsplattform integriert (AP3). Ihre konkrete Leistungsfähigkeit wird durch Einsatz in Musterapplikationen demonstriert (AP4). Schwerpunkt des AP5 ist der Beitrag des TextGrid für die Bildung eines Semantic Grid.

Die für weitere Projekte offenen Schnittstellen garantieren Synergien in der wissenschaftlichen Textdatenverarbeitung sowie eine Rationalisierung des wissenschaftlichen Arbeitens u.a. durch optimierten Zugriff auf Primärquellen und Werkzeuge. Damit werden weitreichende technologische und inhaltliche Impulse für die textbasierte Forschung gegeben.

Die folgenden Kapitel beschreiben die einzelnen Arbeitspakete (AP).

#### 4.1. AP 1: Inhaltliche Studie mit Empfehlungen über die Nachnutzbarkeit internationaler Editionstools

<b>Nummer des Arbeitspaketes:</b>	<b>1</b>							
<b>Titel:</b>	<b>Inhaltliche Studie mit Empfehlungen über die Nachnutzbarkeit international relevanter Werkzeuge</b>							
<b>AP-Leiter</b>	FH Worms							
<b>Start-Monat:</b>	Februar 2006 (M1) bis Januar 2007 (M24)							
<b>Partner (Akronym):</b>	SUB	DAASI	Saphor	U-Wür	IDS	U-Tri	TU-Da	FH-W
<b>Personen-Monate pro Partner:</b>	5	2	1	2	5	2	5	2
<b>Gesamtanzahl PM</b>	<b>24</b>							

##### **Ziele (Kurzbeschreibung)**

Das Projekt soll international relevante Werkzeuge aus den Anwendungsbereichen Publishing, Text Processing, Text Retrieval und Linking sowie Workflow Tools sichten, auf ihre Nutzbarkeit für TextGrid überprüfen und relevante Tools auf ihre Verwendbarkeit in TextGrid testen.

##### **Beschreibung des Arbeitspaketes (Detail)**

Ziel des Arbeitspaketes ist es, arbeitsteilig die einschlägigen in internationalen Projekten eingesetzten bzw. in den diversen Open-Source-Gruppen entwickelten Werkzeuge im Hinblick auf das Anforderungsprofil von TextGrid zu sichten, ihr Grid-Funktionalität zu überprüfen, relevant erscheinende Produkte zu testen und eine Empfehlung über ihre Einsetzbarkeit in TextGrid abzugeben.

Im Vordergrund stehen dabei, analog zu AP2 und AP5, Arbeitsmodule zu den Diensten, Text Processing, Linking, Text Retrieval und Publishing sowie zu Ontologien. Management-Leistungen für die Bereiche Access, Workflow, Kommunikation und Benutzer werden jedoch ebenfalls in die Analyse einbezogen.

## **1. Text Processing**

Der Bereich 'Text Processing' benötigt sehr differenzierte Werkzeuge. Im Vordergrund steht zunächst die Überprüfung von XML-Editoren mit Validierung nach DTD und Relax NG, die sich auch für die Verwaltung von Metadaten unterschiedlicher Struktur eignen und den Zugriff auf Dateien unterschiedlicher Formate (XML, Tiff-Header) erlauben. Werkzeuge zur Lemmatisierung sollen eine ein- oder mehrstufige automatische Grundlemmatisierung ermöglichen, Werkzeuge zur Kollation sollen Texte miteinander vergleichen, die Unterschiede festhalten und ggf. deren Weiterverarbeitung zur Erstellung eines kritischen Apparats ermöglichen. Streaming-Editoren sollen für Datei-Transformationen eingesetzt werden, Tokenizer für die Wortsegmentierung je nach Wortbegriff und den jeweils sprachüblichen Regeln für Wortgrenzen. Für das Sortieren wird die Fähigkeit zu Anpassung an die kulturellen bzw. fachlichen Anforderungen erwartet. Die Vorgabe von Sortierkriterien durch den Nutzer sollte möglich sein.

In das Text Processing wird auch die Optische Zeichenerkennung (OCR) einbezogen, die mit Fortschreiten der Erkennungsalgorithmen und Verbesserung der Auflösung selbst für Frakturschrift wieder zu einer Option werden könnte. Hier soll auch ermittelt werden, ob ergonomische Nutzeroberflächen eine erste Plausibilitätskontrolle durch den Nutzer ermöglichen bzw. ob die Korrektur durch neuere Überblendtechniken ggf. zuverlässiger, schneller und weniger fehleranfällig erfolgen kann.

## **2. Linking**

Die Anforderungen an den Linking-Editor reichen von maskenbasierter Eingabe, über die Validierung von Links und die Anzeige möglicher Sprungziele bis zum Verlinken von Texten mit Bildausschnitten und der Unterstützung der komplexen Link-Syntax der TEI-Richtlinien.

Die Verwaltung und Eingabe bibliographischer Informationen soll im Hinblick auf international bzw. fachwissenschaftlich übliche Formate separat geprüft werden.

## **3. Text Retrieval**

Retrieval-Dienste sind einerseits besonders auf ihre Grid-Fähigkeit hin zu überprüfen, nämlich umfangreiche verteilte Datenbestände als ein virtuelles Dokument zu verarbeiten. Ferner muss auch ihre Fähigkeit, die komplexe generische XML-Information für differenzierte Abfragen nutzen zu können, evaluiert werden.

## **4. Publishing**

Ausgehend vom Anforderungsprofil von TextGrid (Verwaltung kritischer Apparate, Fähigkeit zur Verarbeitung mathematischer Formeln und zu parametergesteuertem Batch-Betrieb, Verarbeitung von Trennmarkierungen zur Layoutoptimierung, typographische Kontrollmechanismen für Qualitätssatz) sind Textsatzsysteme für konventionellen Druck zu überprüfen. Genauso sind Publisher für das Web auf ihre Nutzbarkeit hin zu bewerten.

## **5. Management von Workflow, Access, Kommunikation und Nutzer**

Lösungen für administrative Dienste wie Workflow, Access, Kommunikation und Nutzer werden von Fall zu Fall bewertet. Insbesondere die Ergebnisse einschlägiger internationaler Grid-Projekte werden bei dieser Evaluation zu berücksichtigen sein.

## 6. Ontologien

Die in vielen Bereichen derzeit vielfältig entwickelten Ontologien sowie bestehende Software zu deren Verwaltung sind im Hinblick auf ihre Verwendung für TextGrid zu bewerten. Ausgangspunkt sind dabei primär sprachbezogene Konzepte. Andere Ansätze sollen jedoch auch auf ihre Nutzbarkeit hin geprüft werden.

Recherche-Resultate und Empfehlungen werden in einer Datenbank gesammelt und regelmäßig aktualisiert.

### 4.2. AP 2: Entwicklung Community-spezifischer Werkzeuge (Annotations-, Analyse-Tools)

<b>Nummer des Arbeitspaketes:</b>	2							
<b>Titel:</b>	<b>Entwicklung Community-spezifischer Werkzeuge (Annotations- und Analyse-Tools)</b>							
<b>AP-Leiter</b>	TU Darmstadt (Saphor)							
<b>Start-Monat:</b>	Februar 2006 (M1) bis Januar 2008 (M36)							
<b>Partner (Akronym):</b>	SUB	DAASI	Saphor	U-Wür	IDS	U-Tri	TU-Da	FH-W
<b>Personen-Monate pro Partner:</b>	10	7	37	3	14	26	24	2
<b>Gesamtanzahl PM</b>	<b>123</b>							

#### Ziele (Kurzbeschreibung)

Mit dem Projekt soll eine grid-fähige Workbench für die Erstellung, Bearbeitung, Annotation und Analyse von XML-kodierten Textdateien aufgebaut werden. Damit entsteht ein wesentliches Instrument zur Erstellung und Verwendung einer virtuellen Nationalbibliothek. Die Workbench wird modular aufgebaut sein und ihre gesamte Funktionalität wird als erweiterbares Set von Modulen enthalten sein. Die Workbench wird alle Textwissenschaften zu einer Community zusammenschließen, die ihre Ressourcen gemeinsam erarbeitet und einander zugänglich macht.

#### Beschreibung des Arbeitspaketes

Ziel dieses Arbeitspaketes ist eine grid-fähige Workbench für die Erstellung, Bearbeitung, Annotation und Analyse von XML-kodierten Textdateien. Aufgabe dieses Rahmens ist es a) die sehr großen Datenmengen, die durch die Bildung einer virtuellen Nationalbibliothek anfallen, den Textwissenschaftlern unter einer einheitlichen Arbeitsoberfläche zugänglich zu machen und b) die sehr rechenzeitintensiven Anwendungen zur Annotation und Analyse von Texten zu verteilen. Die internationale Community von Textwissenschaftlern hat in den letzten Jahrzehnten eine Reihe von Programmen entwickelt, die für kleinere bis mittlere lokal verfügbare Korpora verwendet wurden und die einige ausgewählte der im folgenden

beschriebenen Funktionen zur Verfügung stellen.

### **Allgemeine Anforderungen**

Die Digitalisierung von Texten, deren Qualität wissenschaftlichen, bibliothekarischen und archivarischen Ansprüchen genügt, ist aufwendig und kostenintensiv. Entsprechend wichtig sind Standards, die eine hardware- und softwareunabhängige Speicherung von Texten ermöglichen. Für jedes der folgenden Tools wird daher vorausgesetzt, dass sie Daten gemäß den Standards XML und Unicode verarbeiten, die dieses Ziel unterstützen. Das Tagset der Text Encoding Initiative, das heute in zahlreichen elektronischen Texten Verwendung findet, soll das grundlegende Datenformat für die Tools in Arbeitspaket 2 darstellen, soweit diese nicht generische XML-Tools sind.

Aus dem Anspruch des Community-Grid lässt sich ableiten, dass die Tools untereinander und mit dem Textgrid per Webservice kommunizieren müssen.

Die Workbench wird modular aufgebaut sein, d.h. ihre gesamte Funktionalität wird in einem Set von Modulen enthalten sein, das leicht zu ergänzen ist. Die Workbench soll es der Community der Textwissenschaften ermöglichen, an einem virtuellen Nationalkorpus zu arbeiten und ihre disziplinspezifischen Abfragen und Auswertungen zu formulieren, vorhandene Daten mit disziplinspezifischem Markup zu erstellen und zu ergänzen sowie neue Daten zu erzeugen.

### **Workbench:**

Die im Folgenden näher beschriebenen Funktionen der Workbench operieren transparent und einheitlich auf lokalen und Netz-Daten. Die Workbench ist Autorenwerkzeug, aber auch Recherche- bzw. Arbeitswerkzeug für Endbenutzer. Sie integriert in modularer Form alle notwendigen Funktionalitäten und ist leicht erweiterbar. Die Workbench erlaubt auch die graphische Definition von Arbeitsabläufen, die dann als Folge von Batchprozessen abgearbeitet werden.

*Vorarbeiten:* Die Würzburger Arbeitsgruppe <philtag> (bestehend aus: Hans-Werner Bartz, Dr. Thomas Burch, Prof. Fotis Jannidis, Dr. Klaus Prätor, Dr. Andrea Rapp, Prof. Dietmar Seipel, Prof. Werner Wegstein) hat sich mehrfach getroffen und Architektur, Arbeitsziele und einzelne Funktionen der Workbench festgelegt. Weitere Vorarbeiten gibt es bei den Arbeitsmodulen.

### **Arbeitsmodule**

Die Arbeitsmodule lassen sich fünf Diensten (siehe Abbildung 1 "Architektur der TextGrid-Middleware") zuordnen: 1. Publishing, 2. Text Processing, 3. Text Retrieval, 4. Linking und 5. Administrative Module

#### *1. Publishing*

##### **Text Publisher für den Druck:**

In diesem Arbeitsmodul werden TEI-Daten für den Druck vorbereitet. Aufbauend auf der Analyse in API werden geeignete Textsatzsysteme (LaTeX, ggf. XSL-FO) in einem oder zwei Grid Services verkapselt. Das gewählte Werkzeug muss als Mindestanforderung mit kritischen Apparaten und mathematischen Formeln umgehen können. Es muss sich im Batch-Modus aufrufen lassen. Aufbau einer interaktiven grid-basierten Anwendung zur typographischen Umsetzung von Tags und zur Verfeinerung der Typographie (etwa

Hinweis auf übervolle Zeilen und Möglichkeit, Trennkennstellen in den Text einzubringen).

*Vorarbeiten:* Saphor GmbH: Stylesheets TEI nach XSL-FO und TEI nach TeX für einfache TEI-kodierte Texte (aktuell nicht für Editionen geeignet, aber im Gegensatz zu den Stylesheets des TEI-Consortiums standardkonform).

### **Text Publisher für das Web:**

Ziel ist die Aufbereitung von TEI-Daten für die wissenschaftlich adäquate Publikation im WWW.

Anwendung Cocoon-basiert (XML-Daten per Grid-Services übergeben, idealerweise konfigurierbare XSLT-Stylesheets als Grundlage für die Präsentationsform).

## *2. Text Processing*

### **XML-Editor:**

XML-Editor mit Syntax-Coloring und Validierung nach DTD, XML Schema und Relax NG.

*Vorarbeiten:* Es existieren zahlreiche XML-Parser und mit Eclipse auch Tools, die sich ohne großen Aufwand einrichten lassen.

### **Metadaten-Annotation:**

Eintragen von stark strukturierten Daten insbesondere den Metadaten wie Autor, Titel usw. Struktur der Datenfelder und Validierungsschema sollen frei definierbar sein. Realisierung via Implementierung des Xforms-Standards. Rückspeicherung via offene Plug-ins; Basisfunktionalität: XML-Daten nach TEI, Tiff-Header.

### **Lemmatisierung:**

Implementierung eines Werkzeugs zur Lemmatisierung mit automatischer Grundlemmatisierung und manueller Korrektur / Überarbeitung. Erarbeitung umfassender Lemmalisten für das Deutsche und seine historischen Phasen.

*Vorarbeiten:* Es existieren in der linguistischen Community eine Reihe von Lemmatisierungsprogrammen für die modernen Sprachstufen des Deutschen; für die historischen Sprachstufen s. auch 'Lemmatisierungstool' unter Eigenleistungen U-Trier.

### **Kollationierung:**

Zwei bis beliebig viele TEI-kodierte Dokumente werden verglichen und ihre Unterschiede werden nach TEI kodiert und notiert.

*Vorarbeiten:* Prototyp wird als Abschlussarbeit entwickelt.

### **Streaming-Editor:**

Transformationen von Dateien aufgrund von Regeln, z.B. automatisierte Anreicherung potentiell unstrukturierter Texte mit XML-Strukturen. Die Eingabe muss nicht XML sein, sondern kann ein beliebiger

Datenstream sein (etwa OCR-Rohdaten, reiner Text), Ausgabe wird üblicherweise, muss aber nicht XML sein. Das System wird allerdings auf die Ausgabe von XML-Daten optimiert sein. Module werden Streams aus ausgewählten Anwendungsformaten unterstützen.

*Vorarbeiten:* Im Rahmen einer Seminararbeit betreut von Prof. Jannidis wurde ein Prototyp erstellt, der aufgrund von Regelbeschreibungen in einer XML-notierten Konfigurationsdatei Tags einträgt. Saphor hat zahlreiche Konvertierungswerkzeuge für die Explizierung der Datenstrukturen ihrer Kunden erarbeitet.

#### **Tokenizer:**

Zerlegt Texte in Wörter nach den jeweils sprachüblichen Regeln und dem Verständnis von Wort und Wortgrenze. Internationalisierung.

#### **Sortieren:**

Anordnung von gegebenen Zeichenketten gemäß kulturellen und fachlichen Erwartungen. Transformation ungeordneter Zeichenketten in geordnete. Konventionen vom Nutzer vorgebar. Bekannte nationale und europäische Standards, etwa DIN 5007, NFZ44-001 (AFNOR), TK 34.1 (SIS) und ENV 13710 (CEN) werden direkt unterstützt.

*Vorarbeiten:* Probleme des Sortierens gut bekannt und normiert (etwa ISO/IEC 14651 / Unicode Collation Algorithm); schnelle Lösungen in Java-Bibliotheken.

#### **OCR-Daten mit Qualitätsanreicherung**, evtl. mit Abgleich gegen Wörterbücher:

Bereinigung von OCR-typischen Erfassungsfehlern und Anreicherung der Daten mit ersten Strukturinformationen. OCR-erfasste Daten gerade älterer Drucke (oft in Fraktur) weisen typische Erfassungsprobleme auf, etwa bestimmte Verlesungen, Verwechslungen von Schaft-s und f usw. Dieser Service wird solche Erfassungsprobleme semi-automatisch korrigieren. Die Daten können mit zeitspezifischen Wörterbucheinträgen abgeglichen werden, wie sie vom IDS für die Lemmatisierung erstellt wurden und im AP 5 generalisiert werden. Gleichzeitig bringt der Service elementare Strukturinformationen wie Absätze und Überschriften TEI-konform ein.

#### **OCR:**

Prototypische Einbindung eines OCR-Moduls. Die Universität Würzburg setzt eine teure Lizenz der Finereader Engine ein, die u. a. auch Fraktur unterstützt. Der Service macht die Dienstleistung "automatisierte OCR" testweise für eine verteilte Nutzerschaft verfügbar, deren Institute nicht über eine der kostspieligen Lizenzen verfügen. Auf diese Weise werden unnötige Doppellizensierungen an deutschen Hochschulen vermieden. Dieser Dienst wird prototypisch für vergleichbare textwissenschaftliche Spezialdienstleistungen als Grid-Service getestet. Erwerb der FineReader-Lizenz durch die Universität Würzburg und Evaluierung.

*Vorarbeiten:* Evaluierung von Scan-Software wie z.B. Macrolog Optopus, Xerox Scanworx, Textbridge, durch U-Wür; Erwerb der AGORA-Konverter Lizenz durch die SUB. Evaluierung der docWorks-Software durch SUB. Prüfroutinen für im double-keying-Verfahren erfasste Volltexte für verschiedene Sprachen und Sprachstufen auf der Basis von TUSTEP werden an der U-Trier eingesetzt.

### 3. Text Retrieval

#### **Query-Interface:**

Dieser Service verkapselt den von der Integrationsplattform bereitgestellten Zugriff auf heterogene Datenquellen unter einer einheitlichen, XPath-basierten Abfrage-Syntax.

#### **Text Retrieval (Volltext):**

Dieses Service realisiert auf der Basis des „Query Interface“ die verteilte Suche im Gesamt-TextGrid oder definierten Untermengen davon unter Berücksichtigung bekannter Textstrukturen.

*Annahmen:* Verteilte Corpora von TEI-Dokumenten können, auf Grid-Technologien aufbauend, transparent als ein virtuelles Dokument verstanden werden, dessen genaue Zusammensetzung vom Benutzer vorgegeben werden kann. Es ist möglich, dieses virtuelle Dokument, das je nach Forschungsinteressen des Textwissenschaftlers, etwa die gesammelten Werke Goethes, den Werkbestand einer mittelalterlichen Manuskriptsammlung oder aber gleich die Gesamtheit aller erfassten Texte einer Nationalliteratur umfassen kann, nach benutzerdefinierten Kriterien zu durchsuchen. Die vorhandene TEI-Struktur kann dabei für semantisch differenzierte Abfragen ausgenutzt werden.

*Vorarbeiten:* Es existieren als frequente bis hochfrequente Open Source Projekte XML-Datenbanken mit Volltextsuchfunktionen.

### 4. Linking

#### **Link-Editor:**

Plug-in für Editor. Maskenbasierte Eingabe von Sprungzielen und Links. Überprüft Validität von Links. Zeigt bei internen Links Liste aller möglichen Ziele an. Ermöglicht das Verlinken von Text und Bildausschnitten. Unabhängigkeit von Bildskalierung. XML-basierte Kodierung der Kodierung. Abspeichern: a) als Teil des Texts, b) als Standoff Markup. Implementierung via SVG. Unterstützung der komplexen Linksyntax von TEI. Außerdem soll es möglich sein, die Linkstruktur zu prüfen und automatisiert zu erweitern, in dem z.B. die Transitivität (Befund A -> B, B -> C erzeugt: A -> C) ausgenutzt wird.

*Vorarbeiten:* Für das Verlinken von Text und Bild ist im Rahmen einer Abschlussarbeit, betreut von Prof. Wegstein, ein Prototyp entwickelt worden. Im Rahmen einer Diplomarbeit, betreut von Dr. Burch, wird das WB-Link-Tool weiterentwickelt (s. Eigenleistungen U-Trier).

#### **Bibliographie:**

Aufbauend auf den Services „integrierter XML-Editor“ und unter Nutzung der XForms-Technologie (vergl. „Metadaten-Annotation) als selbst-validierender Eingabemechanismus für hochstrukturierte Daten bauen wir ein Modul für die Verwaltung bibliographischer Daten, wie sie nicht nur in textwissenschaftlichen Projekten auftauchen. Es wird hochverteilte und kollaborative Bibliographien unterstützen.

Die intern im TEI-Bibliographie-Format gehaltenen Daten werden über eine XPath-Schnittstelle abfragbar sein. Sie wird es außerdem erlauben, automatisch über xpointer oder xptr u. a. eine mit TEI-Dokumenten verknüpfte Bibliographie zu generieren.

Daten können in verschiedenen Formaten (TEI, DocBook, TeX) und Formatierungskonventionen ausgeliefert werden. Dieses Modul kann somit mittelfristig auch zu einem möglichen Ersatz für das in den Naturwissenschaften populäre BibTeX werden.

#### 5. Administrative Dienste

##### **Editor für den technischen Workflow (EtW):**

Mit dem EtW kann der Nutzer den Workflow der Grid-Services für seine konkreten Bedürfnisse orchestrieren. Der EtW ermöglicht es, bestimmte typische Workflows zu speichern und als Module zu verwalten, sowie in komplexere Gesamt-Workflows zu integrieren. Der EtW fungiert als GUI für den in AP3 zu entwickelnden Workflow Editor.

##### **Editor für den administrativen Workflow (EaW):**

Neben der technischen Orchestrierung von Web Services ist es auch in textwissenschaftlichen Projekten oft notwendig, einen administrativen Workflow zu definieren (z. B. Erarbeitungs- und Korrekturphase von Daten, Freigaben für spezielle Nutzergruppen, generelle Freigabe zur Publikation, jeweils zu autorisieren durch die dafür verantwortlichen Stellen). Der EaW fungiert als GUI für den in AP3 zu entwickelnden Workflow Editor.

#### **Abhängigkeiten zu anderen APs**

Dieses Arbeitspaket bildet die Grundlage für die Entwicklung der Community-Musterapplikationen (AP 4) und definiert die Spezifikationen für die Integrations-Plattform (AP 3).

Auch zu AP 5 bestehen Abhängigkeitsverhältnisse (Ontologien).

### **4.3. AP 3: Anbindung der Community-Tools und Vorschläge für Entwicklungen an der Integrations-Plattform**

<b>Nummer des Arbeitspaketes:</b>	<b>3</b>							
<b>Titel:</b>	<b>Anbindung der Community-Tools und Vorschläge für Entwicklungen an der Integrationsplattform</b>							
<b>AP-Leiter</b>	SUB Göttingen (DAASI)							
<b>Start-Monat:</b>	Februar 2006 (M1) bis Januar 2008 (M36)							
<b>Partner (Akronym):</b>	SUB	DAASI	Saphor	U-Wür	IDS	U-Tri	TU-Da	FH-W
<b>Personen-Monate pro Partner:</b>	17	32	6	-	3	3	3	4
<b>Gesamtanzahl PM</b>	<b>68</b>							

### **Ziele (Kurzbeschreibung)**

Die in den anderen Arbeitspaketen erstellten Softwarebausteine werden über eine TextGrid-spezifische Middleware-Infrastruktur an die Integrationsplattform (IP) angebunden. Hierzu muss eine Schnittstelle zu den Tools spezifiziert und implementiert werden sowie eine Schnittstelle zu den Diensten der IP. Diese Vermittlerfunktion erfordert einen hohen Kommunikationsaufwand sowohl zu den Projektpartnern als auch zu den Betreibern der IP, insbesondere bei der Vermittlung der TextGrid-Community Anforderungen an die IP.

### **Beschreibung des Arbeitspaketes (Detail)**

Grid-Technologien befinden sich nach wie vor in einem Entwicklungsprozess, im Global Grid Forum (GGF, <http://www.ggf.org>) werden weiterhin neue Integrationsstandards entwickelt. Es zeichnet sich eine eindeutige Richtung dieser Entwicklungen ab, nämlich eine Angleichung an das in der Industrie zunehmend an Bedeutung gewinnende Web-Service-Paradigma, welches auf den Standards „Simple Object Access Protocol (SOAP)“, „Web Service Description Language (WSDL)“ und „Universal Description, Discovery, and Integration (UDDI)“ beruht. Das Paradigma heißt Grid Service, also eine logische Abstraktion von Client-Server-Architekturen. *"Everything is a service"*. Hierdurch war es möglich, eine ganz neue Flexibilität einzuführen, indem man beliebige Grid-Dienste den Anforderungen der Grid-Community entsprechend zur Verfügung stellte. Das Konzept Grid Service war hierbei sehr dem in der Industrie zunehmende Bedeutung erlangten Konzept des Web-Service entlehnt, wobei jedoch ein wesentlicher Unterschied bestand: Ein Web-Service ist entsprechend dem zugrunde liegenden Transportprotokoll HTTP statuslos, ein Grid Service wurde aber als ein statusbehafteter sessionbasierter Dienst aufgefasst.

Die internationalen Entwicklungen im Grid-Computing lassen sich sehr gut in den Standardisierungsarbeiten des Global Grid Forums bzw. in den damit zusammenhängenden Entwicklungsarbeiten des Globus-Projekts ([www.globus.org](http://www.globus.org)) nachvollziehen. Letztes Ergebnis dieser Arbeiten ist das Globus Toolkit 4, das ein neues Konzept, das Web Service Resource Framework (WSRF), implementiert. Hierbei können Web-Services, wie sie in der Wirtschaft gängig sind, verwendet werden, d.h. am WSDL waren keine Erweiterungen mehr notwendig. Die spezifischen Anforderungen der Grid-Dienste werden durch ein Framework befriedigt, welches außerhalb der Web-Services-Beschreibung Zustandsinformationen über die von den Web Services bearbeiteten Ressourcen beschreiben kann. WSRF ist eine Gruppe von fünf Standards zur Beschreibung von Ressourcen, persistenter Referenzen, Gruppierung und Fehlerbehandlung von Web-Services, die auf Standards zur Adressierung, Notification und Brokering aufbauen. Es gibt bereits eine erste auf NET-Technologie basierende WSRF-Implementierung an der Virginia Tech University (<http://www.cs.virginia.edu/~gsw2c/wsrif.net.html>). Eine starke Beteiligung der Industrie bei der Spezifizierung des WSRF hat dazu geführt, dass es relativ einfach mit Standard-Web-Service-Technologie realisierbar ist.

Globus Toolkit 4 ist mittlerweile als produktionsreif einzustufen und kann im Rahmen eines Projekts, dessen vornehmliches Ziel es ist, einer Community für die alltägliche Arbeit verwendbare Tools zur Verfügung zu stellen, verwendet werden.

Das in Globus Toolkit 4 implementierte WSRF wird als die Zukunft im Grid-Computing angesehen.

Um TextGrid aber auch prinzipiell unter anderen Grid-Infrastrukturen lauffähig zu machen, wird die

Abstraktionsschicht Grid Application Toolkit (GAT) implementiert, die im EU-Projekt GridLab – A Grid Application Toolkit and Testbed ([www.gridlab.org](http://www.gridlab.org)) entwickelt wurde. GAT kapselt Zugriffe auf das Grid und macht somit Anwendungsentwicklungen unabhängig von der zugrunde liegenden Grid-Infrastruktur. Durch eine den Entwicklungen des Grid-Computings angepasste Außenschnittstelle bleiben somit alle mit GAT entwickelten Dienste in Zukunft kompatibel und sorgen somit für ein Höchstmaß an Interoperabilität. Das GAT leistet aber auch eine weitere Abstraktion, indem es Komplexität in der Anbindung kapselt und mittels einer API durch einfache Funktionsaufrufe komplexe Grid-Schnittstellen zur Verfügung stellt.

Bei der Implementierung des GAT wird TextGrid sich auch in einem hohen Masse mit der D-Grid Integrationsplattform (IP) koordinieren, um Synergien optimal auszunutzen und mit anderen Community-Projekten auf eine gemeinsame Infrastruktur bauen zu können.

Generell, für TextGrid gesamt und für das AP 3 im Speziellen, ist es von zentraler Bedeutung, eine möglichst gute Kooperation mit dem IP-Projekt zu suchen, um einerseits Anforderungen aus der im Grid-Bereich eher exotischen TextGrid-Community in die IP-Arbeiten hineinzubringen und um andererseits als Tester der von der IP bereitgestellten Dienste zu fungieren sowie um auch eigene Lösungsvorschläge einzubringen. TextGrid soll softwaretechnisch eng an die Integrationsplattform gekoppelt werden.

### **Die Architektur der TextGrid-Middleware**

Der im Folgenden beschriebene Vorschlag einer Architektur muss als eine erste Annäherung verstanden werden, da ein gesichertes Modell erst nach der Evaluation der Pläne der Integrationsplattform aufgestellt werden kann. Im Projekt wird auf existierende Grid-Software Lösungen, die sich im Rahmen von TextGrid als sinnvoll und nützlich erweisen, aufgebaut. Wichtige Evaluationskandidaten sind z.B. die im Rahmen von GridLab entwickelten Softwarepakete. Wenn diese im Folgenden zitiert werden, so sind sie als eine Möglichkeit unter meist mehreren aufzufassen. Da davon ausgegangen werden kann, dass die IP, deren Aufgabe es ist, heterogene Community-Grids zentral mit Basisdiensten zu versorgen, ähnliche Kapselungen bzw. eine Weiterentwicklung von GAT zur Verfügung stellen wird, wird die Integration von TextGrid über diese Technologie hergestellt. Hierdurch können die TextGrid-spezifischen Dienste über eine einheitliche Schnittstelle als Web Services zur Verfügung gestellt werden und mittels der TextGrid-Middleware an die Grunddienste der IP angebunden werden.

Die TextGrid-Middleware setzt sich also aus verschiedenen gleichartigen Interfaces zusammen, die in Abbildung 1 dargestellt werden. Die in den anderen Arbeitspaketen erstellten Einzeldienste für Textwissenschaften sind dunkelgelb gekennzeichnet, hellgelb die entsprechenden Interfaces der Middleware. Diese stellen einerseits ein Benutzerinterface dar, andererseits versorgen sie die Einzeldienste mit grundlegenden Grid-Diensten, wie zum Beispiel Datei-Dienste oder Rechenleistung. Auf diese Dienste sollte bevorzugt über GAT zugegriffen werden. Dort, wo es sinnvoll ist, könnte TextGrid aber auch direkt die Grid-Protokolle ansprechen, wie z.B. GridFTP. Rosa sind die generischen, also nicht Textwissenschaftspezifischen Dienste der Middleware gefärbt, die höchstwahrscheinlich von den entsprechenden IP-Diensten unterschieden werden müssen, da Dienste wie Zugriffskontrolle und Kommunikationsdienste für jede Community separat über eine eigene Benutzerverwaltung verwaltet werden sollten. In diese generische Middleware-Schicht gehört auch eine weitere Abstraktionsschicht, durch die die TextGrid-spezifischen Dienste und die generischen, über die IP zur Verfügung gestellten Dienste, gleichwertig in die Workflows eingebaut werden können. Hierdurch werden die generischen Grid-Dienste zu TextGrid-Diensten. Die Schnittstellen zur Integrationsplattform erscheinen hellgrün. Letztere ist wiederum hellblau gekennzeichnet. Es wird sicherlich mehr als ein Grid geben, welches ähnliche Dienste zur

Verfügung stellen kann. Durch die Verwendung der GAT-Architektur wird es einfacher sein, das TextGrid an andere Grids anzubinden.

Das Konzept von TextGrid setzt konsequent auf die XML-basierte Web-Services-Technologie, die auch im geisteswissenschaftlichen Bereich als Zukunftstechnologie z.B. bei Virtual Libraries gesehen wird. Alles wird als ein logischer Dienst aufgefasst, der im Netz auf Rechnern instanziiert werden kann. Hierbei werden die einzelnen in AP2 und AP5 erstellten Dienste soweit wie sinnvoll modularisiert. Zusammenfassende Dienste ("Services") stellen Schnittstellen zu verschiedenen auf Grundfunktionalitäten aufgeteilte Module her. Hierdurch können einerseits bereits vorhandene Module mit einem Web-Service-Wrapper in die Workbench mit aufgenommen werden, andererseits, und das ist der viel wesentlichere Vorteil, alle Module durch ein Workflow-Management miteinander in einzelnen Arbeitsabläufen verbunden werden. Dies ermöglicht es dem Textwissenschaftler, mit einigen Mausklicks neue Bausteinanordnungen aufzubauen – man könnte sie auch als Experimentanordnungen bezeichnen, wenn z.B. mittels verschiedener Text-Processing-Services und Auswertungsmodulen Hypothesen zu einem Text bestätigt oder widerlegt werden können. Die Kommunikation zwischen TextGrid-Middleware und den einzelnen TextGrid-Diensten findet über WSDL-Beschreibungen statt sowie über eine im Projekt zu spezifizierende TextGrid Processing Markup Language, die TextGrid-spezifische XML-Tags definiert, soweit sie nicht in der Text Encoding Initiative (TEI), oder in der Mark-Up-Language von OpenOffice (OpenOfficeXML) spezifiziert sind. Kompatibilität mit den beiden letztgenannten Standards wird jedoch angestrebt.

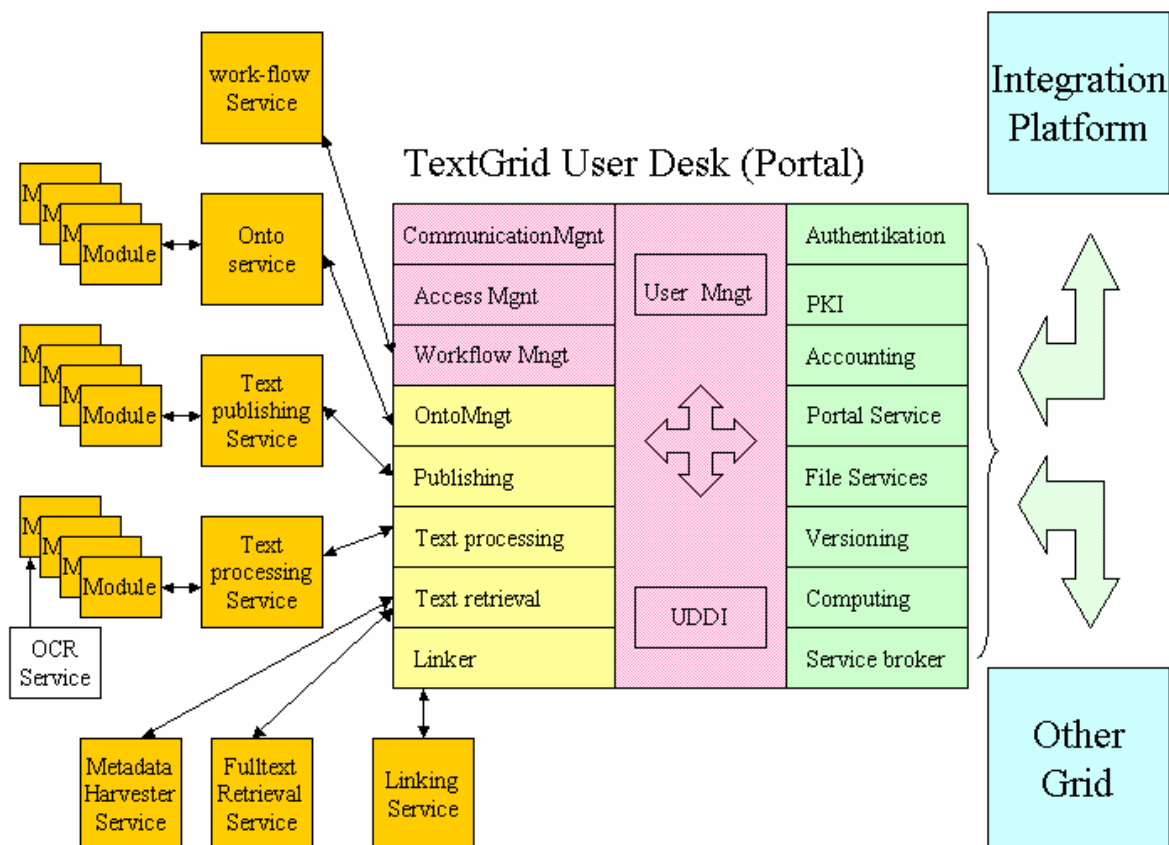


Abbildung 1: Architektur der TextGrid-Middleware

## **Die Bausteine der TextGrid-Middleware**

Das *TextGrid-Portal* ist die zentrale Schnittstelle zu den Diensten der TextGrid-Middleware. Neben einer zentralen Benutzerverwaltung bietet es ein Community-Portal mit Features wie Personalisierung und Logging, über welches der Zugriff auf die im Folgenden spezifizierten Dienste erfolgt. Für die Implementierung wird auf vorhandene Software aufgebaut, wie sie z.B. Gridsphere bietet.

Der *Workflow Manager* ermöglicht es, Service/Modul-Anordnungen über eine graphische Benutzeroberfläche zu definieren und zu testen, sowie bereits spezifizierte Anordnungen zusammen mit Erklärungen und Metadaten zu speichern bzw. gespeicherte Anordnungen zu suchen und auszuwählen. Solche Workflows können nicht nur aus einer linearen Serie von Web-Services bestehen, es kann auch komplexere Strukturen Verzweigungen, Schleifen etc. geben. Der Workflow-Service führt den eigentlichen Workflow durch. Als Technologien stehen eine ganze Reihe von Workflow-Beschreibungssprachen und deren Implementierungen zur Verfügung, wie z.B. die Industrie-nahe Business Process Execution Language (BPEL) oder die Web Services Choreography Description Language (WSCDL). Ein Teil der Projektarbeiten wird die verschiedenen Beschreibungssprachen und deren Implementierungen evaluieren, um den am besten geeigneten Kandidaten in die TextGrid-Middleware zu integrieren. Die fertige Implementierung Triana wird ebenfalls zu evaluieren sein. Der Workflow Manager wird mit den von der IP gelieferten Service-Broker-Funktionalität verknüpft, um dynamische Suchen nach geeigneten Web Services zu ermöglichen. Des Weiteren wird eine Monitoring-Funktion integriert, mittels derer man den Ablauf eines Workflows überwachen kann.

Der *Access Manager* sorgt dafür, dass nur berechtigte Benutzer die jeweiligen Dienste verwenden dürfen. Einerseits können hier entsprechende Rechte vergeben werden, andererseits kann z.B. der Workflow Manager schon bei der Erstellung eines Workflows überprüfen, ob der Benutzer über ausreichende Rechte verfügt. Die Rechte werden nicht einzelnen Benutzern vergeben, sondern über Rollen definiert, welche Benutzern zugeordnet werden können. Grid-Computing erfordert ganz neue Konzepte des Identity Managements, da im Gegensatz zu herkömmlichen domainspezifischen Diensten die Benutzer von Ressourcen aus anderen Domains stammen als die Anbieter dieser Ressourcen. Da die Zugriffsrechte von den Betreibern der Einzeldienste spezifiziert werden, die beliebig im Netz verteilt sein können, müssen hier Konzepte der Federated Identities in Betracht gezogen werden. Die Dynamik von Virtual Organisations muss abbildbar sein. Technologien, wie die Security Assertion Markup Language (SAML), die Grundlage für domainübergreifende Authentifizierungs- und Autorisierungssysteme wie Liberty Alliance und Shibboleth ist, sowie Privilege Management Infrastructure (PMI) in Verbindung mit Public Key Infrastructure (PKI) müssen bei der Implementierung in Betracht gezogen werden. Hierbei zeichnet sich im Grid-Umfeld eine besondere Bedeutung von Shibboleth ab, das im Rahmen des Projekts GridShib in Grid-Infrastrukturen integriert wird. Eine weitere neuere Entwicklung, die in diesem Bereich zu evaluieren sein wird, ist der Virtual Organization Membership Service (VOMS) zur Verwaltung von VOs. Es ist davon auszugehen, dass die IP solche Dienste bereitstellen wird, so dass die Text-Grid-Middleware nur für die Verwaltung der Zugriffsregeln und Benutzer zuständig ist. Wichtige weitere Technologien sind in diesem Zusammenhang Lightweight Directory Access Protocol (LDAP), Service Provisioning Markup Language (SPML), Directory Services Markup Language (DSML) und eXtensible Access Control Markup Language (XACML).

Der *Communications Manager* kombiniert eine Reihe von Diensten, die zum Teil ganz allgemeine Kommunikationsdienste sind wie Diskussionsplattformen, Videokonferenzen, gemeinsames gleichzeitiges Arbeiten an einem Dokument über das Netz etc. Bei diesen ist davon auszugehen, dass sie von der IP zur Verfügung gestellt werden. Darüber hinaus wird es aber auch mehr textspezifische Kommunikationsdienste

geben, die sich weniger durch die Technologie, sondern vielmehr durch den organisatorischen Ansatz auszeichnen. Kollaborative Edition von Texten, Metadaten und Ontologien, Entscheidungsstrukturen für Editorial Boards von elektronischen Zeitschriften, Verknüpfung und Annotationen von Texten etc. werden durch spezielle TextGrid-Kommunikationsdienste koordiniert.

Der *Ontology Manager* wird das Interface zu den Arbeiten des AP 5 sein. Er besteht aus drei Teilen: einem Metadata Manager, einem Wörterbuch Manager und dem eigentlichen Ontology Manager. Verschiedene Anwendungsmöglichkeiten von Ontologien werden hierbei durch die TextGrid-Middleware zur Verfügung gestellt. Ontologien können z.B. im Bereich Metadaten und Retrieval verwendet werden. Werden Metadaten zu Textressourcen durch Verweise auf Ontologie-Einträge angereichert, können Suchen intelligenter durchgeführt werden, vorausgesetzt, die Ontologien werden in die Suchinterfaces integriert. Dann weiß die Retrieval-Engine zum Beispiel, ob sie beim Stichwort "Bank" nach einem Finanzinstitut oder nach einem Sitzmöbel suchen soll. Ontologien können aber auch verwendet werden, um ganze Texte zu analysieren. Gerade in Verbindung mit Wörterbüchern lassen sich Ontologien auch für spezifische Arbeitsschritte der Textdatenverarbeitung nutzen, zum Beispiel um automatisch Namen zu extrahieren, vor allem historische Sprachstufen zu lemmatisieren und zu taggen oder Satzstrukturen zu analysieren etc. Insbesondere die Einbindung von Germanet (<http://www.sfs.uni-tuebingen.de/lzd/Intro.html>), einer allgemeinen Ontologie der deutschen Sprache, wird in diesem Zusammenhang sehr hilfreich sein. Durch den Ontology Manager werden also verschiedene Ontologien, die im Projekt (eventuell auch über die IP) zur Verfügung stehen, verwaltet. Komplexere Themen, wie bzw. die Verknüpfung von verschiedenen Ontologien, werden im TextGrid-Projekt vorerst jedoch nicht angegangen. Wichtige zu berücksichtigende Technologien werden hierbei die Web Ontology Language (OWL) und das Resource Description Framework (RDF) sein.

Die textspezifischen Interfaces zu *Text Processing*, *Text Retrieval* und *Text Publishing* stellen dem Benutzer die entsprechenden in AP 2 entwickelten Dienste unter Einbeziehung des Workflow Managers (zum Teil auch des Ontology Managers) zur Verfügung. Neben einer intuitiven Benutzeroberfläche zum Steuern bzw. Konfigurieren dieser Dienste, ist es v.a. die Aufgabe der TextGrid-Middleware, die Grid-Grunddienste, wie Computing, Fileservices (Verteilung der Daten, Erstellung einer logischen Datei aus beliebig vielen im Netz verteilten Einzeldateien, Replikamanagement) und Versioning, den TextGrid-Diensten zur Verfügung zu stellen, damit die bis zu Terabytes großen Corpora mit beliebig komplexen Workflows verarbeitet werden können. Ein in diesen Schnittstellen zu lösendes Problem ist die Generierung logischer Texte aus Hunderten oder Tausenden von Einzeldokumenten, um sie gemeinsam verarbeitbar zu machen. Es kann nicht vorausgesetzt werden, dass alle Dokumente nach einheitlichen TEI-Tags codiert sind. Es müssen deshalb Routinen entwickelt werden, die voll- und halbautomatische Transformationen durchführen, um auf Basis einer vereinfachten Struktur einen möglichst einheitlichen und logischen Text zu erzeugen. Über den Workflow Manager können auch TextGrid-fremde Dienste, wie z.B. ein OCR-Service eingebunden werden, um z.B. über Text-Processing-Module, die wiederum mit Wörterbüchern verknüpft werden, eine automatische Fehlerkorrektur des gescannten Textes durchzuführen. Die Steuerung dieser Dienste wird über die im Projekt zu definierende TextGrid Processing Markup Language realisiert.

Der ebenfalls in AP 2 entwickelte *Linking Service* wird von der TextGrid-Middleware für verschiedenste Anwendungen zur Verfügung gestellt. Hierbei werden nicht nur eindeutige Referenzen auf logische Texte verwaltet, sondern auch Relationen zwischen Ontologien und Wörterbüchern, Wörterbüchern und Texten etc. Wichtige zu evaluierende Technologien stellen hierbei Xpath und XLink dar.

Eine weitere Aufgabe der TextGrid-Middleware ist es, die vom Projekt erstellten Dienste zu veröffentlichen. Ein Registry-Service für TextGrid-Dienste stellt gleichzeitig einen wichtigen Schritt zur Verbreitung und

Nachhaltigkeit von TextGrid dar. Um im Web Services Paradigma zu bleiben, wird hierfür ein UDDI-Server implementiert, dessen Basis z.B. ein LDAP-Server sein kann. Zwar hat sich die UDDI-Technologie noch nicht umfassend durchgesetzt, das verteilte Prinzip ist jedoch eine gute Möglichkeit, die Dienste verschiedener Grids zu entdecken und zu verwenden. Die Alternativ-Technologie Web Service Inspection Language (WSIL) ist jedoch ebenfalls zu berücksichtigen.

In die zentrale Komponente werden weitere Informationsdienste implementiert, wie sie z.B. im GridLabMDS umgesetzt wurden.

Das Problem der Infrastrukturfinanzierung wird im Rahmen eines generischen Businessplans zum Betrieb eines Community-Grids dargestellt und ein Lösungsvorschlag erstellt, der von einer Open Source Lizenz der in diesem AP erstellten Software ausgeht.

Die Entwicklungsarbeiten werden soweit wie möglich mit OpenSource-Tools auf Linux-Plattform realisiert.

#### **Abhängigkeiten zu anderen APs**

Die Abhängigkeiten zu den anderen APs sind vielfältig. Zum einen müssen die Anforderungen aus AP 2 und AP 5 eruiert, die Schnittstellen zu AP 2 und AP 5 gemeinsam entwickelt und das Testszenario mit AP 4 abgestimmt werden, sowie die Standardisierungsaktivität mit AP 2 koordiniert werden. Zusammen mit AP 6 und AP 4 sollen TextGrid Schulungen zur Einbindung der Community durchgeführt werden. Schließlich bestehen auch zur Integrationsplattform starke Abhängigkeiten.

#### **4.4. AP 4: Entwicklung der Community Muster-Applikation**

<b>Nummer des Arbeitspaketes:</b>	<b>4</b>							
<b>Titel:</b>	<b>Entwicklung der Community Muster-Applikation</b>							
<b>AP-Leiter</b>	Universität Würzburg							
<b>Start-Monat:</b>	November 2006 (M10) bis Januar 2008 (M36)							
<b>Partner (Akronym):</b>	SUB	DAASI	Saphor	U-Wür	IDS	U-Tri	TU-Da	FH-W
<b>Personen-Monate pro Partner:</b>	11	5	9	21	5	12	-	4
<b>Gesamtanzahl PM</b>	<b>67</b>							

#### **Ziele (Kurzbeschreibung)**

Das Projekt soll die Testmaterialien bereitstellen, an denen die Nutzbarkeit der grid-fähigen Werkzeuge für die Arbeitsbereiche Publishing, Textverarbeitung, Text Retrieval und Linking in der erforderlichen Komplexität erprobt und in Tests mit Anwendern unterschiedlicher Profile evaluiert werden kann, so dass die Ergebnisse iterativ in den Entwicklungsprozess der Software zurückfließen.

### **Beschreibung des Arbeitspaketes (Detail)**

Das Arbeitspaket soll Testmaterialien zur Erprobung und Evaluierung der grid-fähigen Workbench für Publishing, Textverarbeitung, Text Retrieval und Linking bereitstellen, welche die erforderliche Komplexität für den iterativen Softwareentwicklungsprozess aufweisen. Drei Typen von Materialien sind dafür vorgesehen:

#### **1. Stark strukturierte Textdaten**

Idealtypisch für diesen Typ ist die Textsorte 'Wörterbuch'. Es beschreibt zu jedem Stichwort nicht nur metasprachlich die Bedeutung durch Paraphrasen oder Synonyme, sondern veranschaulicht die Wortbedeutung jeweils auch objektsprachlich durch Kontextbeispiele, meist in Gestalt von Quellenzitaten (mit genauer Referenz), ergänzt durch stilistische Bewertungen nach Sprachschichten und Stilfärbungen, zeitliche, räumliche und fachsprachliche Bewertungen von Gebrauchsbedingungen, erweitert durch Informationen zur grammatischen Form, ggf. auch der Aussprache (mit besonderen Anforderungen an den Zeichensatz), der Etymologie, der Wortgeschichte usw.

Aus dem Blickwinkel von TextGrid sind Wörterbuchdaten, und da besonders historische Wörterbücher mit nicht völlig konsistenter Struktur, ein ideales Testmaterial für Textverarbeitung, Text Retrieval und Linking, dies auch im Hinblick auf das zugrundeliegende Datenformat (XML) und die Markup-Konzeption (valide TEI-Dateien).

#### **2. Textdaten mit besonderen Anforderungen an die Präsentationsform**

Idealtypisch hierfür ist die 'Historisch-kritische Edition', die den Text eines Autors nicht nur auf der Basis aller erhaltenen Überlieferungszeugen erstellt, sondern – je nach Ausrichtung – überdies die Besonderheiten des einzelnen Textzeugen durch einen textkritischen Apparat, die Variation des Textes in der Überlieferung durch einen zusätzlichen überlieferungskritischen Apparat, ggf. auch entsprechende Druckanordnung (z.B. Paralleldruck mit Apparat) nachweist und durch einen Kommentarteil erschließt. Die Werke unserer bedeutenden Dichter werden durch solche historisch-kritische Editionen erschlossen, von denen nicht wenige bis heute noch weiter bearbeitet und erweitert werden. Aus der Sicht von TextGrid eignen sich die textphilologisch außerordentlich hochwertigen historisch-kritischen Editionen besonders als Testmaterial für Publishing, Textverarbeitung und Linking, da sie einerseits in ihrer elektronischen Form die Verbindung zu Abbildungen der Textzeugen erfordern, wodurch hohe Datenvolumina zusammenkommen und zusätzlich das Markup im Editionstext, bei der erforderlichen feinen Granularität wie sie die TEI-Richtlinien vorsieht und ermöglicht, eine beträchtliche Ausweitung des Datenumfangs mit sich bringt. Als Resultat entstehen damit digitale Texte von absoluter Referenzqualität für die weitere Arbeit in der wissenschaftlichen Community.

#### **3. Bilddaten mit hohen Anforderungen an Qualität und Metadatenverwaltung**

Die Verbindung des Editionstexts mit Abbildungen der benutzten Textüberlieferung in einer Qualität, die für die meisten Fragestellungen die Konsultation des Originals entbehrlich macht, gehört bei wissenschaftlichen

elektronischen Editionsprojekten heute zum internationalen Standard. Verwaltung, Archivierung, Erschließung stellen hohe Anforderungen an die EDV-Struktur und machen damit die Kombination von Textdaten und Bilddaten zu einer Grid-Anwendung, eben TextGrid.

### *Textauswahl*

Die ausgewählten Texte machen dies deutlich. Als Modellfall aus dem Bereich 'Wörterbuch' wird aus mehreren Gründen das 'Wörterbuch der deutschen Sprache' von Joachim Heinrich Campe (erschienen 1807 – 1811) eingesetzt. Es ist mit 6000 Seiten ausreichend umfangreich, bisher noch nicht digitalisiert, ergänzt aber die bereits digitalisierten Wörterbücher des 19. Jahrhunderts in idealer Weise. Die Bilddaten sind bereits in hochauflösenden Farbscans in Würzburg erstellt worden. Die hohe Auflösung ist wegen der Buchstabengröße erforderlich: die kleinsten Buchstaben, z.B. übergeschriebenes e, sind kleiner als 1 mm, Farbbilder sind nötig, weil das Papier so von unterschiedlich farbigen Hadern durchzogen ist, dass bei Graustufen- oder Schwarzweiß-Aufnahmen einzelne Stellen unleserlich werden. Jede Aufnahme benötigt daher 100 MB Speicherplatz, das Gesamtvolumen der Bilder macht damit 600 GB aus.

Als Modellfall der historisch-kritischen Edition für das Testbed kann die Jean-Paul-Ausgabe verwendet werden, die seit Jahren von der DFG-geförderten Jean-Paul-Arbeitsstelle am Institut für deutsche Philologie in Würzburg bearbeitet wird. Für die Edition der in Würzburg zu erstellenden Bände sind ca. 4000 Seiten handschriftliches Material zu digitalisieren, hinzu kommt die Digitalisierung sämtlicher Drucke, die zu Lebzeiten Jean Pauls erschienen sind.

Zusammen mit den Campe-Bilddaten dürfte damit ein Datenvolumen im Terabyte-Bereich erreicht werden.

### *Arbeitsmodule*

1. Erstellen der Textgrundlage zur Erprobung der Workbench an stark strukturierten Daten:

Texterfassung von Campes Wörterbuch durch eine in der Erfassung von Texten in Frakturschrift erfahrene Firma.

Einsatz der TextGrid-Module zur Erprobung von Textverarbeitung, Text Retrieval und Linking und der verwendeten Kodierungskonzepte auch der Metadaten im Rahmen des Projekts.

### *Vorarbeiten:*

Graphischer Arbeitsplatz (Java-Programme unter Eclipse-Oberfläche) zur Qualitätskontrolle von elektronischen Texten und zur Verwaltung von Bilddateien (Diplomarbeit Thienel am Institut für Informatik, betreut von Prof. Wegstein).

6000 Bildseiten des Campe-Wörterbuchs in hoher Auflösung (600 dpi).

2. Erstellen von geeigneten Evaluationsszenarien für die historisch-kritische Jean-Paul-Edition: Erprobung und Evaluation der Module für Publishing, Textverarbeitung und Linking mit den Mitarbeitern der Jean-Paul-Edition sowohl für konventionellen Druck wie Web Publishing einschließlich der Kodierungskonzepte

für Texte und Metadaten.

Texterfassung und Digitalisierung der Drucke des Romans 'Hesperus', Band 1 der Ausgabe sind von der Jean-Paul-Arbeitsstelle geleistet, Satz und Textvergleich erfolgen derzeit mit TUSTEP-Prozeduren und eignen sich damit sehr gut für eine Systemevaluierung. Graphischer Arbeitsplatz wie unter 1.

3. Dokumentation der Tools und Analyse und Beschreibung des Workflows.

4. Entwicklung und Auswertung von Tests der TextGrid-Benutzeroberfläche von Studierenden mit unterschiedlichem EDV-Profil (Philologie-Studierende, Informatik-Studierende mit Nebenfach 'Linguistik', Studierende der philologischen Aufbaustudiengänge 'EDV-Philologie' bzw. 'Linguistische Informations- und Textverarbeitung' an der Universität Würzburg)

5. Aufbereiten der Daten des Campe-Wörterbuchs zur gemeinsamen Nutzung für das Semantic Grid

Die übrigen Mitglieder des Konsortiums werden die entwickelten Module, soweit dies sinnvoll erscheint, in ihren eigenen Vorhaben einsetzen und evaluieren. Die Evaluationsergebnisse werden ebenfalls in den iterativen Prozess in die Softwareentwicklung einfließen.

#### **Abhängigkeiten zu anderen APs**

Der Erfolg des AP 4 hängt naturgemäß maßgeblich von den Resultaten aus den APs 1, 2, 3 und 4 ab und führt die Ergebnisse zusammen. Gemeinsam mit AP 6 und AP 3 sollen TextGrid Schulungen zur Einbindung der Community durchgeführt werden.

#### **4.5. AP 5: Semantic Web und TextGrid = Semantic TextGrid**

<b>Nummer des Arbeitspaketes:</b>	<b>5</b>							
<b>Titel:</b>	<b>Semantic Web und TextGrid = Semantic TextGrid</b>							
<b>AP-Leiter</b>	Universität Trier							
<b>Start-Monat:</b>	Februar 2007 (M13) bis Januar 2008 (M36)							
<b>Partner (Akronym):</b>	SUB	DAASI	Saphor	U-Wür	IDS	U-Tri	TU-Da	FH-W
<b>Personen-Monate pro Partner:</b>	8	6	-	6	5	10	-	-
<b>Gesamtanzahl PM</b>	<b>35</b>							

### **Ziele (Kurzbeschreibung)**

Im Rahmen von TextGrid sollen hochstrukturierte Textmaterialien aus verschiedenen digitalen Wörterbüchern und genuin philologische Methoden und Technologien eingesetzt werden, um eine "Meta-Lemmaliste" zu generieren, die den übergreifenden gemeinsamen Einstieg in die verschiedenen Wörterbuchinformationen ermöglicht. Sie dient als Basis-Ontologie für die Erschließung geisteswissenschaftlicher Primärquellen.

### **Beschreibung des Arbeitspaketes (Detail)**

Eines der aktuellsten und interessantesten Forschungsfelder im Bereich neuer Web-Technologien lässt sich mit dem Schlagwort 'Semantic Web' umreißen. Dieses Arbeitspaket soll die für die geisteswissenschaftliche Community bedeutende Brücke schlagen zwischen Grid-Technologien, Semantic Web und den damit verbundenen Technologien (Stichwort hier auch im Antrag OWL, RDF/XML, Topic Maps etc.).

In den Informationswissenschaften werden eine Fülle von Forschungsprojekten initiiert und durchgeführt und zahlreich Technologien und Methodiken erprobt und weiterentwickelt. Bislang ist diese Szene geprägt von Forschungen und Anwendungen aus den Bereichen der Wirtschaft (Bsp. Automobilindustrie) und der Naturwissenschaften (Bsp. Biotechnologie etc.). Im Bereich der geisteswissenschaftlichen Traditionswissenschaften werden die Möglichkeiten dieser neuen Technologien bislang noch kaum genutzt. Allerdings stehen Vorarbeiten hierzu aus dem Bereich Computerlinguistik zur Verfügung. Insbesondere Wortnetze, generelle Sprachontologien, können sinnvoll mit elektronischen Ausgaben von Wörterbüchern kombiniert werden. Das Konzept Wortnetz geht auf die Arbeiten der University of Princeton zurück, die ein Wortnetz der englischen Sprache (WordNet, <http://www.cogsci.princeton.edu/~wn/>) aufgebaut hat. Ein Wortnetz der deutschen Sprache liegt mit GermaNet (<http://www.sfs.nphil.uni-tuebingen.de/lsd/>) vor. Solche Wortnetze bilden die häufigsten Wörter einer Sprache und ihre bedeutungstragenden Beziehungen zu anderen Wörtern der Sprache ab. Sie werden z.B. zu Lesartendisambiguierung für Anwendungen im Bereich der Informationerschließung und maschinellen Übersetzung und zur semantischen Annotierung von Sprachdaten genutzt.

Im Bereich der Ontologie-Beschreibungssprachen haben sich die unter dem Stichwort Semantic Web zusammengefassten Arbeiten im Rahmen des W3Cs als Standard durchgesetzt. Das Resource Description Framework (RDF, <http://www.w3.org/RDF/>) wurde entwickelt, um Ressourcen über Metadaten computerlesbar zu beschreiben, wobei eine Grundstruktur aus Subjekt-Prädikat-Objekt verwendet wird. Die Web Ontology Language (OWL, <http://www.w3.org/2004/OWL/>) baut auf RDF auf und ermöglicht es, Ontologien zur semantischen Anreicherung von Web-Informationen zu beschreiben. Die Semantic Grid Research Group des Global Grid Forums hat sich zur Aufgabe gemacht, diese Semantic-Web-Technologien zur Beschreibung von Grid-relevanten Ressourcen zu verwenden, wobei die damit kompatible und in der Industrie weit verbreitete Technologie Common Information Model (CIM, <http://www.dmtf.org/standards/cim>) berücksichtigt wird. Erst wenn diese Arbeiten über den Bereich des Computational Grid hinaus gehen, also in die Bereiche Information Grid und Knowledge Grid kommen, werden sie für die Arbeiten dieses AP von Bedeutung.

Im Rahmen der hier geplanten Plattform sollen Materialien und genuin philologische Methoden und Technologien in Verbindung mit GermaNet und den Semantic-Web-Technologien eingesetzt werden, um

ebenfalls in diesem Forschungsfeld aktiv mitzuwirken bzw. diese Aspekte bei der Entwicklung neuer Werkzeuge mit zu berücksichtigen.

Ein Ergebnis dieses Arbeitspakets wird eine offene (weil verschiedene Standards unterstützende) und unter Einbezug von entsprechend implementierten Administrationstools wartbare, durch Datenverteilung im Netz höchst skalierbare und durch Redundanz ausfallsichere, die verschiedenen TextGrid-Anwendungen eine semantische Anbindung (z.B. für Analyse- und Suchvorgänge, für Metadatenproduktion) ermöglicht.

Traditionelle Wörterbuchmaterial, zum Beispiel, das bereits in tief strukturierter digitaler Form und für verschiedene Wörterbuchtypen vorliegt, soll in die ontologischen Strukturen der neuen Semantic-Web-Technologien zu integrieren. Hierbei werden automatische Konvertierungsroutinen und die Aufnahme in entsprechend modellierten Datenbanken eine wesentliche Rolle spielen, aber auch Benutzerinterfaces, die manuelle semantische Anreicherungen erleichtern. Insbesondere die flexible Abbildung von Objekten und einer großen Anzahl verschiedener Relationsarten zwischen den Objekten in einer Datenbank wird hierbei eine der Herausforderungen sein.

Im Bereich der Lemmatisierung (einschließlich Bedeutung) und Vernetzung der produzierten Materialien mit besonderer Berücksichtigung von sprachlichen Varietäten (diatopisch, diachronisch, diastratisch) wird eine neue Infrastruktur geschaffen. Lemmatisierungsprobleme der "Standardsprache" des 20. und 21.

Jahrhunderts betrachten Computerphilologen als "gelöst", daher liegt genau in dieser neuen "varietätenlinguistischen" Dimension des Semantic TextGrid ein entscheidender Mehrwert.

## **Ausgangslage innerhalb der Arbeitsgruppe TextGrid**

### **Wörterbücher und Nachschlagewerke**

Vor allem an den Universitäten Trier und Würzburg wurde über viele Jahre hinweg eine herausragende Expertise in Bezug auf die Produktion und die intelligente Retrodigitalisierung von verschiedenen Typen von Wörterbüchern und Nachschlagewerken aufgebaut. Dank intensiver philologischer Auseinandersetzung mit den Strukturen und Inhalten der Wörterbücher sowie des konsequenten Einsatzes und der Entwicklung modernster EDV-philologischer und informationstechnologischer Methoden und Werkzeuge ist der Anschluss an internationale, beispielsweise anglo-amerikanische und französische Standards gelungen.

## **Vorhandene digitale Wörterbücher im XML/TEI-Format, bzw. aktuelle Wörterbuchprojekte**

### **1. Sprachstadienwörterbücher**

#### **1.1 Joachim Heinrich Campe: Wörterbuch der deutschen Sprache (6 Bde.), 1807 – 1811**

Projekt von Prof. Wegstein, Universität Würzburg, hochauflösende Farbimages vorhanden, s. AP 4. Eine TEI-kodierte Version existiert aktuell noch nicht.

#### **1.2 Jacob und Wilhelm Grimm: Deutsches Wörterbuch (1837-1961)**

Das von den Brüdern Grimm begründete und an der Universität Trier digitalisierte „DWB“ gilt mit seinen ca. 325.000 Stichwörtern als das umfangreichste deutschsprachige Wörterbuch, das systematisch den deutschen Sprachschatz von der Mitte des 15. Jhs. bis zum Bearbeitungszeitpunkt der einzelnen Bände dokumentiert.

#### **1.3 Mittelhochdeutscher Wörterbuchverbund**

Der ebenfalls an der Universität in Trier erarbeitete „Mittelhochdeutsche Wörterbuchverbund“, enthält mit dem „Mittelhochdeutschen Wörterbuch“ von Benecke/Müller/Zarncke (BMZ), dem „Mittelhochdeutsche

Handwörterbuch“ von Matthias Lexer (Lexer) und dem „Findebuch zum mittelhochdeutschen Wortschatz“ (Findebuch), die z.T. seit dem Ende des 19. Jh. bis heute in Forschung und Lehre grundlegenden Nachschlagewerke zur mittelhochdeutschen Sprache für den Zeitraum von ca. 1050 bis 1350. Diese Wörterbücher sind über eine mittelhochdeutsche Meta-Lemmaliste multidirektional untereinander vernetzt. An der Universität Würzburg wurde ein Neuhochdeutsch-Mittelhochdeutsches Umkehrwörterbuch zu Lexers Wörterbuch erstellt; diese Daten stehen ebenfalls zur Verfügung.

#### **1.4 Neues mittelhochdeutsches Wörterbuch mit digitalem Quellenarchiv**

Auf der Basis eines digitalen Quellenarchivs von mittelhochdeutschen Texten vom 11. bis zum 14. Jh. ist in Göttingen und Trier ein neues mittelhochdeutsches Wörterbuch im Entstehen, das erstmals einen die ganze Periode zeitlich und räumlich gleichmäßig berücksichtigenden Überblick über die Verwendungsbedingungen und die Bedeutungsentwicklung des mittelhochdeutschen Wortbestandes gewährt. Die Lemmata sind direkt mit allen Vorkommen im digitalen Belegarchiv verknüpft.

### **2. Spezialwörterbücher**

#### **2.1 Dialektwörterbücher im Verbund**

Derzeit befindet sich ebenfalls in Trier ein Dialektwörterbuchverbund im Aufbau, der aufgrund der digitalen Erfassung und Verknüpfung des „Rheinischen“ und „Pfälzischen Wörterbuchs“, des „Wörterbuchs der deutsch-lothringischen Mundarten“ und des „Wörterbuchs der elsässischen Mundarten“ eine sprachgeographisch umfassende Dokumentation des (süd-)westdeutschen Dialektwortschatzes bietet und durch Verknüpfung mit Sprachkarten (Sprachatlas) eine einzigartige räumliche Differenzierung des deutschen Wortschatzes erlaubt.

### **3. Autorenwörterbücher**

#### **3.1 Goethe-Wörterbuch**

Mit fast 90.000 Stichwörtern, basierend auf allen erhaltenen Werken Goethes, ermöglicht das in Trier zu digitalisierende Goethe-Wörterbuch einen systematischen Einblick sowohl in den Individual- Wortschatz als auch in die Sach-, Begriffs- und Vorstellungswelt des Dichtersfürsten und seiner Zeit.

#### **3.2 Jean-Paul Wörterbuch**

Der Wortschatz von Jean-Paul wird in Würzburg in Verbindung mit den Editionen der Jean-Paul-Arbeitsstelle erschlossen. Erster Schritt dazu ist die Erstellung einer Lemmaliste und deren Abgleich mit den bereits zugänglichen Materialien.

### **4. Enzyklopädien**

#### **4.1 Oekonomische Encyklopädie Krünitz**

Die von Johann Georg Krünitz begründete, zwischen 1773 und 1858 in 242 Bänden erschienene und in Trier digital erfasste „Oekonomische Encyklopädie oder allgemeines System der Staats- Stadt- Haus- und Landwirtschaft“ dokumentiert in einzigartiger Weise den wissenschaftlichen Wortschatz des Deutschen in der Übergangszeit von der Agrar- zur Industriegesellschaft.

Diese Wörterbücher werden in einem integrierten Verfahren mit philologischen und informatorischen Methoden und Techniken miteinander vernetzt, sodass ein übergreifendes Informationssystem zur deutschen

Sprache – ihrer Struktur, Entwicklung und Semantik – entsteht, das weit über die Leistung und Möglichkeiten der bisherigen Wörterbücher hinausgeht.

Dies wird möglich durch die intensive Strukturierung und Auszeichnung der Wörterbücher bzw. durch die Explizierung der entsprechenden Informationspositionen. Markiert werden z.B. die Informationspositionen Lemma, grammatische Angabe, Bedeutung, Belegzitate, Belegstelle etc. Mit den Belegen werden darüber hinaus in vielen Fällen Symptomwerte wie Zeit, Raum, Gattung, Schicht etc. verbunden. Die Explizierung der im Druckmedium nur immanent gegebenen Informationsstruktur der Wörterbücher erlaubt nun neue Formen der Informationsgliederung und damit des Informationszugriffs. Diese Explizierung der Informationsstruktur kann nicht nur für die Entwicklung intelligenter Suchmöglichkeiten genutzt werden, sondern auch zu völlig neuen Fragestellungen an das Material führen. Jenseits von direkten (also bereits expliziten) Verweisen untereinander werden die Wörterbücher über Meta-Lemmalisten vernetzt. Diese Erstellung der Meta-Lemmalisten wird durch automatische Zuordnungsverfahren unterstützt und mit philologischer Kompetenz geprüft. Die Vernetzung der Wörterbücher untereinander ermöglicht dann einen gemeinsamen (Such-)Einstieg in alle Wörterbücher.

Die 'Meta-Lemmaliste' aus den Wörterbüchern mit den zahlreichen bereits vorhandenen Vernetzungen in andere Wörterbücher und Primärquellen bildet eine reiche Kern-Ontologie für die Textwissenschaften des Deutschen (bzw. die Geisteswissenschaften allgemein).

An einem konkreten Beispiel lässt sich die Sprachvarianz zeigen und damit das Potenzial, das in der Auswertung der Wörterbücher steckt. So können z.B. zum Begriff „Brombeere“ der übergeordneten Stichwortliste (Metalemmaliste) u.a. folgende Einträge gefunden werden [Die Einträge aus den Wörterbüchern sind im Folgenden stark gekürzt]:

#### **1.1 Campe: Wörterbuch der deutschen Sprache, Bd. 1, Sp. 624:**

**Die Brombeere**, Mz. Die --n, die Frucht der Brombeerstaude [...] heißt im gemeinen Leben auch **Kratzbeere, Bicksbeere, Bocksbeere, Fuchsbeere, Traubenbeere** etc. In andern und gemeinen Sprecharten wird diese Staude mit ihrer Frucht auch **Rambeere, Brummelbeere, Brommen, Braembesie, Breme und Bremenbeere** genannt.

#### **1.2 Jacob und Wilhem Grimm: Deutsches Wörterbuch, Bd.2, Sp. 396, Z. 79 ff.**

BROMBEERE, *f. rubum*, brambeere, *schwankt in* branbeere, branbire, bramber, braunbeere, brobeere, braubeer, brommer. DASYP. 309<sup>b</sup> brombeer *morum rubi*: [...]

#### **1.3 Mittelhochdeutscher Wörterbuchverbund**

**a) Lexer, Bd. 1 Sp.340, Z.6: brâm-ber** *stn.* (I. 104<sup>a</sup>) *brombeere* MONE *schausp.* EILH. 1717. swarz geverwet als ein zîtic brâmbere TROJ. 32743; bromber HPT. 5,14;

**b) BMZ, Bd. 1 Sp. 104a, Z. 36: brâmbere** *stn. brombeere. sumerl.* 40,70. 56,77. 57,53. *Schmeller* 1, 258. daz hulfe in niht ein brâmbere *Mone altd. schausp.* 3,446.

**c) Findebuch, S. 54: brâmbere** *stn.* ENIK. HVNST.

**2.1 Dialektwörterbuchverbund** (große Varianz, zahlreiche Belege für räumlich begrenzte Phraseologien und Semantiken)

**a) RheinWb, Bd. 1, Sp. 901-904:**

**Brame** III f.: [...] **2.** Brombeerranke, -strauch, -gestrüpp Malm, Monsch, Eup, Aach, Geilk, Selfk; aber auch der Dorntrieb der wilden Rosen Eup, Aach- Aach-Walh, Heinsb-Karken; Brennessel Heinsb-Kirchhv. RA.: *Do ös all lang de Br. drüöver gewaæse. De lett sech an Br. dörch jen Fott trecke för en Pennek* der Geizhals. -- Mosfrk hier u. da auch für Brombeere: [...] **Maren, Moærdeln, Schmelsbærn, Perdsbiren.** [...]

**b) PfWb, Bd.1, Sp. 1245f.**

**Brombeere** [...] Syn.: Brummels-, Dorn-, Heckenbeere, Heckenpraume, Himbeere, Katzentapen, Moz, Schwarzbeere. [...] *Blame<sup>r</sup>* usw. Über *bl-* entwickelten sich die Formen mit *fl-*: *Flame<sup>r</sup>, Flambee<sup>r</sup>* usw.

**c) ElsWb, Bd.2, Sp. 189a, Z. 34 - 49**

**Bram**, Bräm, Brämer [Próm *Hi.*; Präm *Wh.*; Prámær *Weiler*; meist *Pl. -ə*] *Brombeerstrauch*; meist *Pl., weil sie auf den Äckern fortspinnen.* [...].

**d) LoWb, S.66a, Z.36 – 37:**

**Brombier** [brómbér ohne *Pl. Si.*] *m. Brombeere.* – mittelhochdeutsch brämber.

**3.1 Goethe-Wörterbuch Bd. 2, S. 903a, Z. 32 – 42**

**Brombeerart** Darstellung deutscher Brombeer-Arten mit Kupfern [vgl. *Brombeerwerk*] [...] **Brombeere als wildwachsende Frucht** [*Treufreund.*] Gibt's keine Heidelbeeren, Himbeeren, Mehlbeeren [...]

**4.1 Oekonomische Encyklopädie Krünitz Bd. 6, Sp. 779 ff.:**

**Brombeere, Brombeerstaude, Brombeerstrauch, Bremen, Bromen, Brummelbeere, Kratzbeere, Rahmbeere, Robotbeere, Rubetbeere, L. Roncea, Rubetum, Rubus, Fr. Ronce.** Dieses Pflanzengeschlecht setzt Hr. v. Linné unter die zwanzigmännervielweiberigen [...] **Rabets und Rubeten** [...] **Fuchs=Maulbeeren, Fr. Meure de renard,** [...]

Bereits an diesem einfachen und stark verkürzten Beispiel zeigt sich die Bandbreite des zur Verfügung stehenden Materials. Mit seiner konsequenten Auswertung kann ein qualitativer Sprung in den Möglichkeiten der semantischen Analyse und Annotierung, zumindest für bestimmte Bereiche, erreicht werden.

Darüber hinaus lässt sich das über die Vernetzung der Wörterbuchdaten erzeugte Informationsnetz jedoch auch noch zu weiteren Auswertungen und für andere Verfahren nutzen. Erzeugt man nun anhand all dieser vernetzten Wörterbücher ein Umkehrlexikon (Bedeutungsangabe wird zum Lemma und umgekehrt), ergeben sich neue Erkenntnisse und Einsichten in die Entwicklung und vor allem in die Beschreibung semantischer Felder in Raum (z.B. Dialektwörterbücher) und Zeit (z.B. Sprachstadienwörterbücher). Eine Analyse der Belegstellenangaben kann ferner zeigen, durch welche Quellen die Beschreibung im Wörterbuch determiniert ist. Insgesamt kann die Kenntnis der Beschreibungsmöglichkeiten semantischer Felder auf eine neue Basis gestellt werden.

Durch Zuhilfenahme der Semantic-Web-Technologien unter Einsatz von GermaNet werden all diese Ressourcen zu einem semantischen Erschließungstool. Dies kann von den in AP2 entwickelten Textdatenverarbeitungstools genutzt werden, z.B. um automatisch lexikalische Referenzen in einen Text zu integrieren, automatische Register zu erstellen, etc. Über die TextGrid-Middleware-Plattform aus AP3 werden sowohl einzelne Wörterbücher, als auch das GermaNet in den virtuellen Desktop für Textwissenschaftler integriert.

Konkrete Arbeitsschritte:

- Evaluierung und Reports über Ontologie-Software als Basis für eigene Module (s. AP 1)
- Weiter- bzw. Neuentwicklung von Tools zur Verknüpfung von Wörterbüchern untereinander, Vorarbeiten bestehen im Modul WB-Link (s. Eigenleistung U-Trier) sowie zur Verknüpfung von

Wörterbüchern und Primärquellen unter Verwendung von GermaNet; Abstimmung mit AP 2 (Linking-Tool)

- Weiter- bzw. Neuentwicklung und Erprobung philologisch-linguistischer Verfahren zur Erstellung der Meta-Lemmaliste
- Anwendung von Lemmatisierungsverfahren auf die Wörterbuchmaterialien, z.B. auf Belegstellenzitate und Abgleich mit der Meta-Lemmaliste zur Verfeinerung der Ontologie
- Anwendung von umkehrlexikographischen Verfahren zur Verfeinerung der Ontologie
- Entwicklung geeigneter Metadata Application Profiles
- Kollaborative Werkzeuge, die benötigt werden: Vernetzungswerkzeuge, Auswertungswerkzeuge

#### Abhängigkeiten zu anderen APs

Die Reports über Ontologie-Software (s. AP 1) sind die Basis für die Vorhaben des AP 5. Im Hinblick auf die Weiterentwicklung von WB-Link ist eine Abstimmung mit AP 2 (Linking-Tool) notwendig, im Hinblick auf die Lemmatisierungsverfahren ebenfalls. Die in AP 4 erstellten Materialien des Campe-Wörterbuchs müssen in die Meta-Lemmalisten integriert werden. Die Schnittstellen zu AP 3 müssen gemeinsam entwickelt werden.

#### 4.6. AP 6: Projektmanagement und Öffentlichkeitsarbeit

<b>Nummer des Arbeitspaketes:</b>	<b>6</b>							
<b>Titel:</b>	<b>Projektmanagement und Öffentlichkeitsarbeit</b>							
<b>AP-Leiter</b>	SUB Göttingen							
<b>Start-Monat:</b>	Februar 2006 (M1) bis Januar 2008 (M36)							
<b>Partner (Akronym):</b>	SUB	DAASI	Saphor	U-Wür	IDS	U-Tri	TU-Da	FH-W
<b>Personen-Monate pro Partner:</b>	18	2	1	1	1	1	1	-
<b>Gesamtanzahl PM</b>	<b>25</b>							

#### Ziele (Kurzbeschreibung)

Das Ziel des Arbeitspaketes liegt zum ersten darin, sicher zu stellen, dass das Projekt die Ziele innerhalb des vorgegebenen Zeitrahmens und des Budgets erreicht. Dazu gehört auch die formale und administrative Leitung des Projektes.

Ein weiterer Schwerpunkt besteht darin, national und international über das Projekt und dessen Teilergebnisse zu informieren. Besonderer Wert wird dabei auf den Informationsaustausch mit der eigenen Community gelegt.

## **Beschreibung des Arbeitspaketes (Detail)**

### ***Projektmanagement***

Das Management des Projektes umfasst die administrative und organisatorische Durchführung des Projektes. Dazu gehört zum Beispiel die finanzielle Verwaltung der zentral eingesetzten Mittel und die Beantragung von Mitgliedschaften des Konsortiums in nationalen und internationalen Standardisierungsinitiativen (z.B. DIN, CEN, Unicode, DCMI etc.). Die Sicherstellung der termingerechten Abgabe und Einhaltung von Meilensteinen und Ergebnisberichten liegt ebenfalls im Verantwortungsbereich des Arbeitspaketes. Das Projektmanagement gewährt den reibungslosen Ablauf der internen und externen Kommunikation. Dies beinhaltet auch die Kommunikation zum Projektträger und die damit verbundenen formalen und inhaltlichen Aspekte des Projektes. Das Management wird darüber hinaus für die formal-rechtliche Etablierung des Konsortiums und für die Aufnahme z.B. weiterer Assoziierter Partner im Verlauf des Projektes und danach sorgen.

Die Organisation und Durchführung der Meetings (Agenda, relevante Dokumente, Protokoll etc.) der Partner für die Projektbesprechungen werden von dem Projektmanagement geleistet, so z.B. alle 6 Monate ein „face-to-face“ Meeting und dazwischen Video- und/oder Telefonkonferenzen.

Die Gewährleistung des wissenschaftlichen und technischen Austausches zu anderen Communities und hauptsächlich zu der Integrationsplattform wird ein Hauptschwerpunkt des Arbeitspaketes sein.

Im Namen des Konsortiums werden inhaltlich relevante Mitgliedschaften beantragt bei folgenden nationalen und internationalen Gremien und Standardisierungsinitiativen:

- TEI: Text Encoding Initiative
- DIN-Fördermitgliedschaft für zwei Arbeitskreise (NI29.01 Zeichensätze und NI22 Programmiersprachen)
- DCMI: Dublin Core Metadata Initiative

### ***Öffentlichkeitsarbeit***

Im Rahmen des Projekts TextGrid werden intensive PR-Maßnahmen für das Projekt mit nationaler und internationaler Wirkung in Zusammenarbeit mit dem Konsortium durchgeführt. Neben der allgemeinen Propagierung der Konzeption wird ein besonderer Wert auf die Information der eigenen Communities (Geisteswissenschaften und andere textbasierte Communities z.B. Wissenschaftsgeschichte der Mathematik) gelegt. Durch vielfältige PR-Maßnahmen und das Aufzeigen und Bereitstellen technischer Möglichkeiten der Realisierungen und Nutzung der Module sollen weitere Anwender aktiv gewonnen werden. Dazu wird auch die Qualität der Dienstleistungsmöglichkeiten für den Nutzer beitragen, die im Rahmen des Projektes angeboten werden wird.

### ***Workshops***

Es sind insgesamt 3 Workshops geplant, ein CEN/ISSS Workshop, um die Präsenz von TextGrid in internationaler Standardisierung zu sichern, und zwei TextGrid Workshops für Öffentlichkeitsarbeit und zur Information für TextGrid Nutzer.

Der geplante CEN/ISSS Workshop (CEN = Comité Européen de Normalisation; ISSS = Information Society Standardization System) ist ein wichtiges Mittel zur Standardisierung der im TextGrid definierten WSDL-

Schnittstellen und Datenaustauschformate in einem offenen, europaweiten Prozess. Es wird zur erhöhter Akzeptanz der Resultate und zu höherer Interoperabilität textwissenschaftlicher Dienste führen.

CEN/ISSS, das Information Society Standardization System der paneuropäischen Normierungsorganisation CEN, bietet den Workshop-Mechanismus als ein neutrales, allgemein akzeptiertes Forum zur Konsensbildung und zur Validierung der in Textgrid definierten Schnittstellen durch alle einschlägigen Gruppierungen in Europa und darüber hinaus. (s. auch

<http://www.cenorm.be/cenorm/businessdomains/technicalcommitteesworkshops/workshops/index.asp>)

Grid-Technologien leben von der Offenheit und Interoperabilität der realisierten Dienste. So wird sichergestellt, dass auch externe Partner ihre Dienste in TextGrid einbringen und problemlos integrieren können.

Von dem technischen Aspekt der Interoperabilität abgesehen ist der Workshop auch ein wichtiges Mittel der Öffentlichkeitsarbeit und der Bewusstseinschaffung (awareness activities) für die im TextGrid entwickelten Lösungen.

### *Schulungen*

Ein Trainings- und Ausbildungsprogramm für TextGrid soll die Community mit den neuen Möglichkeiten vertraut zu machen. Alle Konsortiumsmitglieder erklären sich bereit, an der Entwicklung und Durchführung entsprechender Schulungsmaßnahmen und Trainingsprogramme für die Anwender und Nachnutzer mitzuwirken, und diesbezügliche Aktivitäten werden im Rahmen der passenden Arbeitspakete AP3 und AP4 durchgeführt:

Zum einen wird technische Expertise zur Installation, Konfiguration und zum Betrieb von Textgrid-Komponenten vermittelt. Die hierfür notwendigen Arbeiten werden in AP 3 durchgeführt.

Der zweite Bereich der Schulungsmaßnahmen wendet sich an den Endanwender, der die Tools in seiner wissenschaftlichen Arbeit nutzen will. Diese Arbeiten werden in AP 4 durchgeführt, wobei zusätzliche Eigenleistungen der Universität Würzburg eingebracht werden, da die Entwicklung der Kursinhalte in das Programm der dortigen Aufbaustudiengänge übernommen und erste Tests mit Studierenden durchgeführt werden können.

Diese Arbeiten werden durch die Öffentlichkeitsarbeiten in AP 6 begleitet und gefördert. Die Konsortiumsmitglieder stellen selbstverständlich hierfür auch ihre Erfahrungen und Tools zur Verfügung, wie z.B. Erfahrung in der Entwicklung von Trainingsprogrammen (Tutorials), ein passendes Umfeld für den Test solcher Programme, auch unter Einsatz von Video-Conferencing (grenzüberschreitend) sowie die Option, die Trainingsmaterialien parallel auch in englischer Sprache bereitzustellen.

Schwerpunkte der Tätigkeiten dieses Arbeitspaketes werden die Entwicklung einer detaillierten Konzeption zur Öffentlichkeitsarbeit und deren Umsetzung sein:

<i>Typ</i>	<i>Geplante Anzahl</i>	<i>Bemerkungen</i>
Informationsveranstaltungen	mehrere	Pressemitteilungen der Partner, Kurzvorträge

Workshops im Rahmen von CEN Veranstaltungen	2	Workshops auf europäischer Ebene unter Einbeziehung internationaler Grid-Experten: CEN/ISSS Workshop zur Standardisierung der WSDL-Schnittstellen + der Austauschformate
Homepage TextGrid www.text-grid.org		Aktuelle Informationen zum Stand des Projektes, Konsortium, Bereitstellung von Projektergebnissen, Workshop-Ergebnissen, Informationen zu Technologie-Fortschritten etc.
Logo(s) entwerfen	4-5 verschiedene	Zum Einbau in Werbematerialien mit unterschiedlicher Auflösung für Druck und Onlineauftritt
Schriften und Logos für Partner bereitstellen		Über Website
Mailinglisten, Newsletter	Verschiedene Mailinglisten (interne und externe Kommunikation), Newsletter alle 3 Monate	Für aktive Benachrichtigung der Interessenten über Newsletter (deutschsprachige und englischsprachige Version)
Werbung in internationalen Gremien, Standardisierungsinitiativen (z.B. DIN, TEI, W3C, Unicode, Dublin Core etc.)		Insbesondere bei den im Antrag genannten Gremien und Standardisierungsinitiativen
Vorträge auf ausgewählten Kongressen/Veranstaltungen		Vgl. Formulare aller Partner
Erstellung von Info-Materialien: Flyer, Broschüren		

**Abhängigkeiten zu anderen APs:**

Es besteht grundsätzlich eine starke Abhängigkeit zu allen anderen Arbeitspaketen, da die Ergebnisse von

dort direkt in die Öffentlichkeitsarbeit einfließen werden.

#### 4.7. Ressourcenplan: Zusammenfassung

Die folgende Übersicht zeigt nicht nur, in welchen Arbeitspaketen die einzelnen Partner ihre Arbeitsschwerpunkte setzen werden, sondern dokumentiert zugleich auch die für das Gesamtprojekt gültige Konzentration auf alle Aspekte der praktischen Nutzung und Anwendung von TextGrid. Die Vorbereitung bzw. Entwicklung von Community-spezifischen Applikationen in AP2 beansprucht immerhin mehr als ein Drittel der personellen Ressourcen. Aber auch für die Ankopplung des Projektes an die Integrationsplattform werden von den Partnern knapp 20% der zur Verfügung stehenden Personalkapazitäten eingesetzt. Der dafür betriebene Aufwand liegt sogar noch etwas über den Mitteln, die für das hinsichtlich Außenwirkung und wissenschaftlicher Wahrnehmung des Projektes so wichtige AP4 eingesetzt werden.

	AP 1	AP 2	AP 3	AP 4	AP 5	AP 6	Summe
<b>SUB</b>	5	10	17	11	8	18	<b>69</b>
<b>DAASI</b>	2	7	32	5	6	2	<b>54</b>
<b>Saphor</b>	1	37	6	9	-	1	<b>54</b>
<b>U-Wür</b>	2	3	-	21	6	1	<b>33</b>
<b>IDS</b>	5	14	3	5	5	1	<b>33</b>
<b>U-Trier</b>	2	26	3	12	10	1	<b>54</b>
<b>TU-Da</b>	5	24	3	-	-	1	<b>33</b>
<b>FH Wo</b>	2	2	4	4	-	-	<b>12</b>
<b>Summe</b>	<b>24</b>	<b>123</b>	<b>68</b>	<b>67</b>	<b>35</b>	<b>25</b>	<b>342</b>

#### 5. Verbundpartner (Konsortium)

Das vom Bundesministerium für Bildung und Forschung (nachfolgend „BMBF“ genannt) geförderte Projekt „TextGrid: Modulare Plattform für verteilte und kooperative wissenschaftliche Textdatenverarbeitung. Erstellung eines Community-Grids für die Geisteswissenschaften“ wird in

Form eines Verbundprojektes durchgeführt. Jeder der beteiligten Vertragspartner geht ein separates Rechtsverhältnis mit dem BMBF ein, das sich in einem eigenen Zuwendungsbescheid manifestiert. Folgende gleichberechtigte Partner sind in diesem Verbundprojekt aktiv:

#### **Institutionen an Universitäten:**

- Georg-August Universität Göttingen: Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB)
- Universität Trier: Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier (U-Trier)
- Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaften (TU-Da)
- Bayerische Julius-Maximilians-Universität Würzburg, Institut für deutsche Philologie, Kompetenzzentrum für EDV-Philologie (U-Wür)
- Institut für deutsche Sprache, Mannheim (IDS)
- Fachhochschule Worms (FH-Wo)

#### **Kleine und mittlere Unternehmen (KMU):**

- DAASI International GmbH (DAASI), Tübingen
- Saphor GmbH, Tübingen

### **5.1. Beschreibung der Partner**

#### **5.1.1. Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB)**

Die Staats- und Universitätsbibliothek Göttingen gehört zu den größten wissenschaftlichen Universalbibliotheken in Deutschland. Sie nimmt neben ihrer Kernaufgabe der Literatur- und Informationsversorgung für die Universität Göttingen im nationalen und internationalen Rahmen zahlreiche Aufgaben aus einem über klassische Bibliotheksarbeit weit hinausgehenden Bereich des Informationsmanagements und der Entwicklung von digitalen Bibliotheksdiensten wahr.

Mit der Einrichtung des Göttinger Digitalisierungszentrums (GDZ) verfolgt die SUB seit etwa sieben Jahren konsequent das Ziel der digitalen Aufbereitung ihrer Bestände und des Aufbaus von Kollektionen für die weltweite wissenschaftliche Nutzung.

In enger Kooperation mit führenden Institutionen des Informationswesens arbeitet die SUB Göttingen auch mit internationalen Partnern an der Lösung zentraler Fragen der weltweit vernetzten Informationsversorgung mit Schwerpunkten wie Standardisierung und der wissenschaftlichen Weberschließung mit.

#### **5.1.2. DAASI International GmbH (DAASI)**

Die DAASI International GmbH wurde Ende 2000 als Spin-Off der Universität Tübingen gegründet. DAASI will ihren Kunden – insbesondere Behörden und dem Forschungsumfeld – in Form von Beratung, Konzeption, Entwicklung, Implementierung und Schulung modernste Technologie in

benutzerfreundlichen Anwendungen zur Verfügung stellen. DAASI ist in nationalen und internationalen Arbeitsgruppen und Standardisierungsgremien aktiv, wie z.B. Global Grid Forum, IETF, ISOC, TERENA Arbeitsgruppen, Teletrust etc. Ein weiteres Merkmal der Firma sind ihre engen Beziehungen zu deutschen und internationalen Forschungseinrichtungen und Behörden, wie das Deutsche Forschungsnetz (DFN), das niederländische Forschungsnetz SURFNet, oder der Verbund europäischer Forschungsnetze TERENA. Die Mitarbeiter verfügen über langjährige Erfahrung in Projektmanagement, Technologieberatung, Open Source-Programmierung, Schulung und Öffentlichkeitsarbeit sowie im Verwalten von personenbezogenen Daten und in den sich hieraus ergebenden datenschutzrechtlichen Erfordernissen. Expertise liegt v.a. in den Bereichen IT-Security, Identity-, Informations- und Wissensmanagement sowie im Grid-Computing vor.

#### 5.1.3.Saphor GmbH

Die Saphor GmbH wurde 2002 in Tübingen als Nachfolgerin der seit 2001 aktiven Saphor GbR gegründet.

Schwerpunkte der Firmentätigkeit sind die Verarbeitung von Textdaten sowie – eng damit zusammenhängend – Dienstleistungen rund um SGML/XML:

- Datenkonvertierung, -aufbereitung und -management
- Entwicklung und Anpassung von Redaktionssystemen
- Elektronische Publikationen
- Daten- und Systemintegration/Web Services.

#### 5.1.4.Kompetenzzentrum für EDV-Philologie an der Universität Würzburg (U-Wür)

Das Kompetenzzentrum für EDV-Philologie an der Universität Würzburg wurde zum WS2003/2004 neu eingerichtet mit dem Ziel, die technischen Grundlagen für die seit Jahrzehnten am Institut für deutsche Philologie wie an der Philosophischen Fakultät II betriebene philologische Datenverarbeitung zu sichern und die Anwendungskompetenz im Bereich Philologie und Datenverarbeitung zu verstärken und auszubauen. Im Kompetenzzentrum sind in der Lehre die Studienangebote der Fakultät im Bereich von Philologie und EDV zusammengefasst und es betreut die dafür zur Verfügung stehenden Computer-Pools aus dem CIP-Programm. In der Forschung dient es als Brücke zu den EDV-gestützt arbeitenden philologischen Forschungsprojekten (z. B. der historisch-kritischen Jean-Paul-Edition, dem Sprachatlas Unterfranken, der Bayerischen Dialektdatenbank 'Bay-Dat', dem 'Repertorium der Sängsprüche und Meisterlieder des 12. bis 18. Jahrhunderts sowie kontrastiver Textkorpora) in der Fakultät und den Nachbarfakultäten. Ferner werden hier die enormen Bestände an qualitativ hochwertigen historischen Editionen aus zwei Würzburger Forschergruppen und einem Sonderforschungsbereich archiviert und gepflegt.

#### 5.1.5.Institut für deutsche Sprache (IDS)

Das Institut für Deutsche Sprache (IDS) ist das außeruniversitäre Forschungsinstitut zur wissenschaftlichen Erforschung und Dokumentation der deutschen Gegenwartssprache in ihrem gegenwärtigen Gebrauch und in ihrer neueren Geschichte. Als vom Bund und vom Land Baden-Württemberg finanzierte Forschungseinrichtung ist das IDS Mitglied der Leibniz-Gemeinschaft. Am IDS werden Projekte bearbeitet, die in ihrer Art und Größe nicht auch an germanistischen Universitätsinstituten durchgeführt werden könnten.

Einen Schwerpunkt in der Arbeit des IDS nehmen der Aufbau und die Pflege elektronischer Textkorpora sowie die Entwicklung von Methoden zu deren Erschließung und Analyse ein. Derzeit umfassen die seit 30 Jahren aufgebauten IDS-Textkorpora insgesamt knapp zwei Milliarden laufende Textwörter aus einer großen Zahl von Zeitungstexten, belletristischen, wissenschaftlichen und populärwissenschaftlichen Texten sowie einer breiten Palette weiterer Textarten. Das IDS verfügt damit über die weltweit größte elektronische Textsammlung mit geschriebenen deutschsprachigen Texten aus der Gegenwart und der jüngeren Vergangenheit. Die Korpora werden kontinuierlich weiter ausgebaut. Linguistisch nutzbar werden diese Korpora über die am IDS entwickelten automatischen Verfahren der Korpuserschließung (wie beispielsweise der Kookkurrenzanalyse) und insbesondere über das öffentlich zugängliche Korpus-Recherche- und Analyse-System des IDS, COSMAS II.

#### 5.1.6. Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier (U-Trier)

Das Kompetenzzentrum "Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften" an der Universität Trier wurde 1998 eingerichtet und hat sich inzwischen als feste Institution etabliert. Es wird gestützt von der Universität Trier, der Universitätsbibliothek und dem Universitätsrechenzentrum, der Akademie der Wissenschaften, der Literatur Mainz, und dem Zentrum für Wissenschaftliches Elektronisches Publizieren

Die Tätigkeiten des Kompetenzzentrums zielen auf die allgemeine Zugänglichkeit und größtmögliche Verbreitung und Nutzung geisteswissenschaftlicher Grundlagenwerke sowie die nachhaltige Sicherung der Textdaten durch das Erstellen elektronischer Versionen und deren plattformabhängige, anwendungsneutrale, an internationalen Standards orientierte Kodierung. Sie reichen konkret von der Organisation der Erfassung bereits vorhandener Grundlagenwerke bis zur Beratung bei der Konzeption von elektronischen Publikationen neuer Werke. Das Dienstleistungsangebot des Kompetenzzentrums umfasst die zuverlässige Datenerfassung, die Konvertierung bestehender Altdaten, die Dokumentanalyse, die SGML/XML-konforme Auszeichnung, die Publikation auf CD-ROM, im Internet und im Buch (als Neu- oder Nachdruck), die Ausarbeitung vernetzter Strukturen, die Unterstützung effizienter Suchstrategien sowie die Entwicklung geeigneter Arbeitsumgebungen, auch für räumlich verteilte Projekte mit mehreren Arbeitsstellen. Die Arbeit erfolgt in enger Zusammenarbeit mit den Kooperationspartnern auf nationaler und internationaler Ebene und im Austausch mit den führenden Institutionen auf dem Gebiet der Textkodierung (Text Encoding Initiative, Unicode-Konsortium).

#### 5.1.7. Institut für Sprach- und Literaturwissenschaft, TU Darmstadt (TU-Da)

Prof. Dr. Fotakis Jannidis lehrt Neuere Deutsche Literaturwissenschaft am Institut für Sprach- und Literaturwissenschaft der TU Darmstadt, wo er zur Zeit an der Entwicklung eines MA „Computerphilologie, Korpuslinguistik“ mitwirkt. Er ist Herausgeber des „Jahrbuchs für Computerphilologie“ (mit Karl Eibl, Georg Braungart, seit 1997 online, seit 1999 auch mentis Verlag, Paderborn) sowie der Hybridedition „Der junge Goethe im Kontext seiner Zeit“ (mit Karl Eibl, Marianne Willems, Frankfurt: Insel 1999 mit CD-ROM). Er leitet die Kommission für editorische EDV-Anwendung in der Arbeitsgemeinschaft für germanistische Edition (Ekage). 2001-2002 war er Mitglied des TEI-Councils. Als Mitglied einer Arbeitsgruppe hat er das Portal [www.lirez.de](http://www.lirez.de) für Online-Rezensionen in der deutschen Literaturwissenschaft mitinitiiert. Laufende Projekte befassen sich unter anderem mit der Entwicklung von Online-Redationssystemen, Portalen, und Digitalisierung.

#### 5.1.8. Fachhochschule Worms (FH-Wo)

Die FH Worms wurde am 1.9.1996 als selbstständige Fachhochschule etabliert. Sie hat aktuell etwa 2500 Studierende in drei Fachbereichen, Wirtschaftswissenschaften (mit den Studiengängen Internationale Betriebswirtschaft, European Business Management / Handelsmanagement und Steuerwesen), Touristik / Verkehrswesen sowie Informatik und Telekommunikation mit den Studiengängen Informatik und Telekommunikation.

Die FH Worms und speziell deren Fachbereich Informatik und Telekommunikation trägt dazu bei, die Grid- und Web Service-Kompetenz innerhalb des Konsortiums auch im akademischen Umfeld signifikant zu vertiefen und zu verbreitern. Der Bereich Telekommunikation des Fachbereichs Informatik und Telekommunikation der Fachhochschule vertritt digitale Telekommunikationsnetze in Lehre und Forschung in ihrer gesamten Einsatzbreite vom lokalen Bereich bis hin zu weltweiten Netzwerken allgemeinsten Topologie. Dies schließt u. a. die folgenden Aspekte mit ein:

- physikalischer Aufbau und Signalübermittlung,
- Verfahren zur (gesicherten) Übertragung digitaler Daten,
- Datenstrukturen und Algorithmen der automatisierten Kontrollsysteme,
- Telekommunikationsdienste und die entsprechenden Benutzerschnittstellen,
- (verteilte) Anwendungen und die technischen, wirtschaftlichen sowie sozialen Aspekte ihres Einsatzes im privaten wie kommerziellen Umfeld,
- Multimedia-Anwendungen.

#### 5.2. Zusammenarbeit mit Dritten, assoziierte Partnerschaften

Das Konsortium sieht vor, assoziierte Partnerschaften in solchen Fällen einzurichten, wo sich inhaltliche Berührungspunkte mit interessanten und interessierten Institutionen ergeben, ohne dass diese als Zuwendungsempfänger im Rahmen der Projektförderung auftreten. Aus Sicht des Konsortiums sind vor allem Vereinbarungen zur Weiterentwicklung des Leistungsumfangs der im Grid angebotenen Dienste reizvoll.