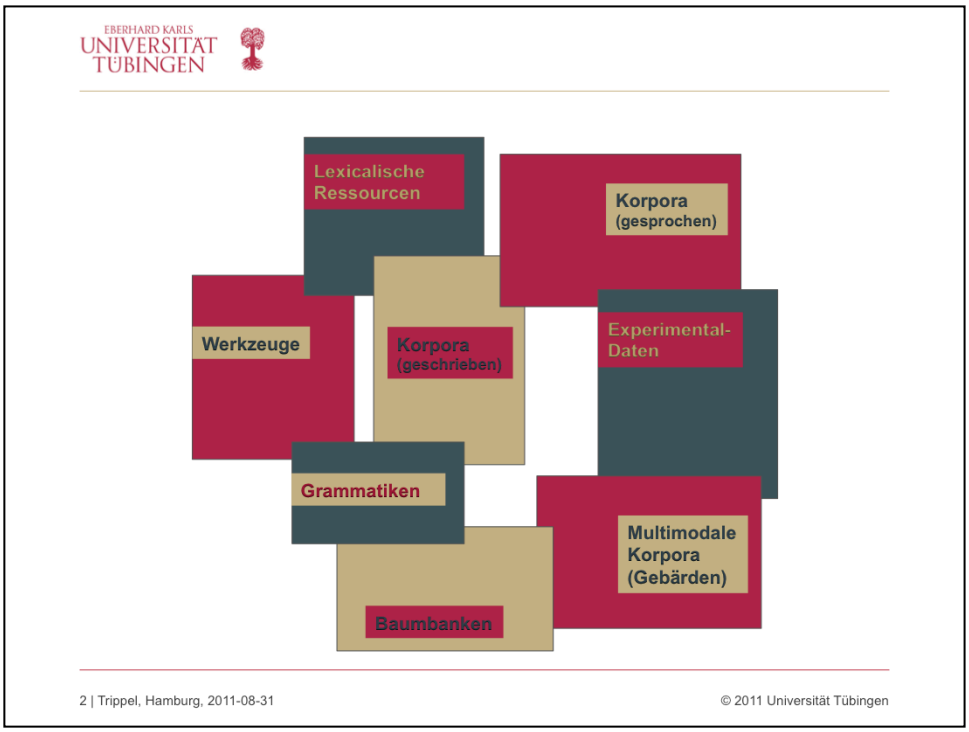
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Nachhaltigkeit
Linguistischer
Daten

Archivierung von Forschungsdaten – langweilig, unwichtig oder was?

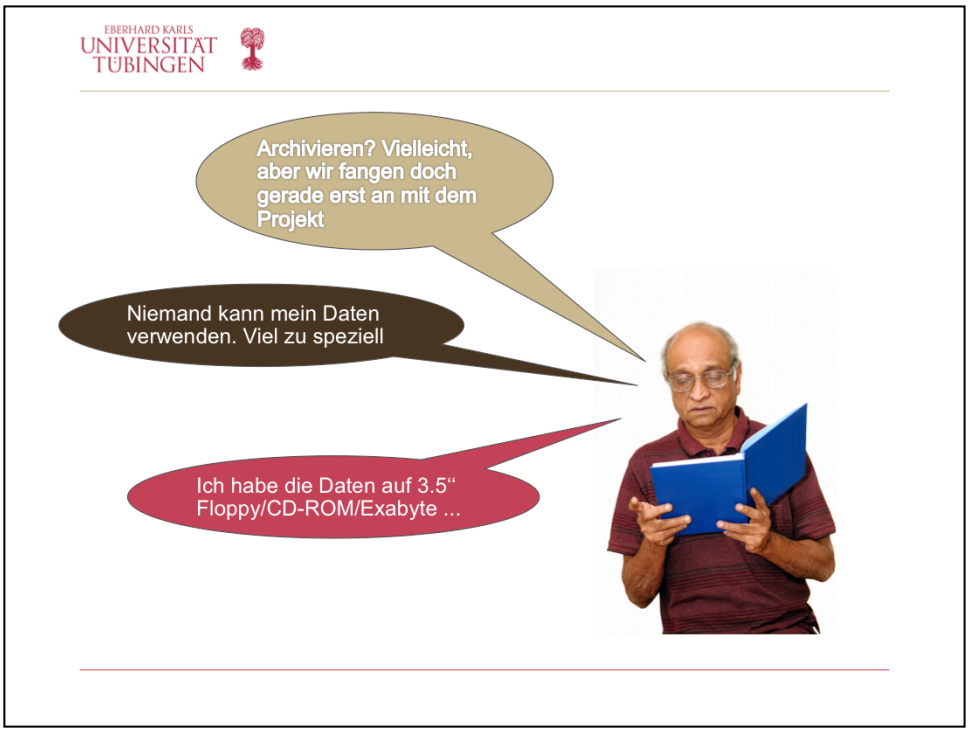
Thorsten Trippel
Computerlinguistik
Seminar für Sprachwissenschaft
Projekt: Zentrum für Nachhaltigkeit Linguistischer Daten



Sprachwissenschaftler machen Experimente, Aufnahmen, Annotationen, Grammatiken und verwenden diese Forschungsdaten für Unterricht, Qualifikationsarbeiten und Publikationen verwendet, teilweise über Jahre und Jahrzehnte. Wie können solche Ressourcen so lange verwendet werden? Können auch andere diese Daten verwenden. Teilweise werden Ressourcen in Drittmittelprojekten erzeugt und in anderen Projekten weiterverarbeitete und verwendet.

Aus dem Bereich der Lexika ist dies ein ganz übliches Verfahren und aus den vor-elektronischen-Zeiten wissen wir gut, wie das geht: man schreibt ein Buch und publiziert es, die Grunddaten werden auf Karteikarten gesammelt und mehr oder weniger oft durchgesehen und überarbeitet.

Heute haben wir wesentlich größere Datenmengen, die auf Computern gespeichert werden, wir erstellen sie, arbeiten damit und sind häufig ganz zufrieden damit, wie das läuft. Warum sollte man sich auch gedanken dazu machen, ob und wie solche linguistischen Daten längerfristig verfügbar bleiben?



Korpora im großen Maßstab werden noch nicht ganz so lange eingesetzt, aber diese Sammlungen von Texten mit zusätzlichen Informationen, z.B. zu Wortart oder Bedeutung oder Übersetzung, kennen wir, manchmal wissen wir auch schon, wieviel mühe es macht, so etwas über längere Zeiträume zu verwenden. Und so etwas, wie ein fertiges Korpus gibt es nicht, man hat immer noch Möglichkeiten, es zu vergrößern oder weitere Beschreibungsebenen hinzuzufügen. Und wenn man genügend Analysen damit gemacht hat, dann ist das Material so reich und so speziell für die eigenen Bedürfnisse bearbeitet, dass man sich kaum vorstellen möchte, dass es auf er Welt noch irgendeine andere Person gibt, die ähnlich detalierte Informationen auch nur begrenzt interessant findet.

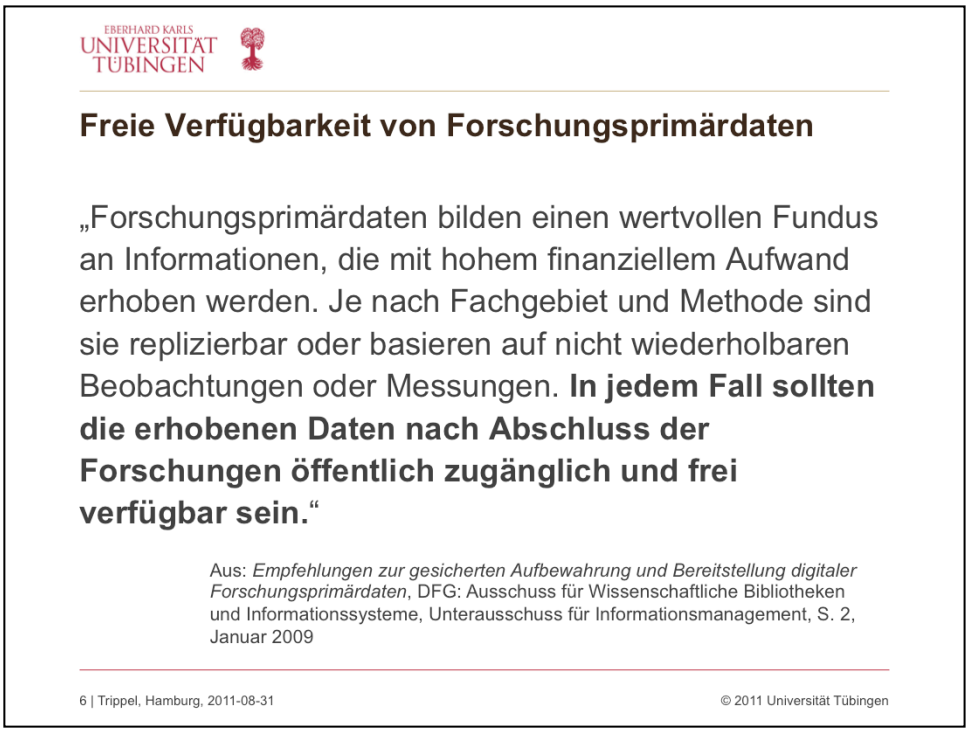
Außerdem: Die Daten werden auf Computern abgelegt, wenn man die jemand anderem geben will, kann man die Dateien einfach weitergeben. Und geht der Computer kaputt: Wohl dem der ein Backup gemacht hat. An diesem Beispiel mit den 3.5 Zoll Disketten oder Exabytes sehen Sie aber schon ein Problem: wer hat denn sowas heute schon noch? Und USB-Speichermedien sind nicht für ihre langfristige Zuverlässigkeit bekannt.

Wenn das so kompliziert ist mit dem Archivieren, warum sollte man sich dann überhaupt darum kümmern?

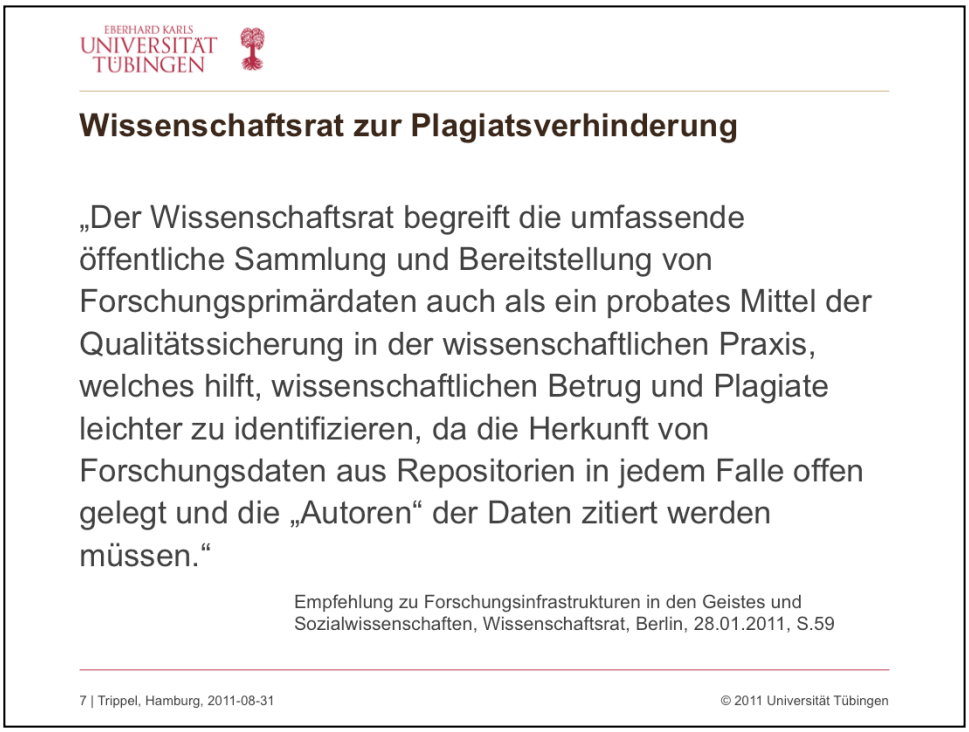
Für uns, für diejenigen, die sich intensiv damit auseinandersetzen, wie das passiert, ist die Frage natürlich einfach: Es ist interessant, anspruchsvoll, es ist nicht klar, wie man dafür sorgen kann, dass Video-Dateien in 10 Jahren noch lesbar – und zugänglich – sind. Kurz: Es ist wissenschaftlich spannend und faszinierend, wenn man seine Forschungsergebnisse nach Jahren ansehen kann, auf den zugrundeliegenden Daten neue Verfahren ausprobieren kann. Oder auch zum Vergleich ähnliche Daten nochmals erfassen kann, um Längsschnittstudien zu machen, die wirklich die unterschiedlichen Grunddaten erfassen. Z.B. Dialektale Veränderungen, Änderungen am Gesteninventar oder beim Ausdruck mittels Gebärden. Oder auch nur zu überprüfen, ob Ergebnisse, die man für Erklärungsbedürftig hält, unter neuen Ansätzen und Theorien nicht erklärbar sind.

Das sagt ein begeisterter Ressourcensammler. Aber es gibt auch ganz profane Gründe:

Der Wissenschaftsrat, ein zentrales Gremium zur Beratung der Bundesregierung, hat sich mit Forschungsprimärdaten und deren Zugänglichmachung beschäftigt. Die Ergebnisse als Empfehlung wurden Ende Januar veröffentlicht. Und die wollen das.



Und die sagen auch warum: Wenn die Öffentlichkeit bezahlt hat, will sie auch darauf zugreifen können. Daher sollen die Daten öffentlich zugänglich und verfügbar sein



Und zu einem Zeitpunkt, als das Thema nur von der Fachöffentlichkeit wahrgenommen wurde, sah der Wissenschaftsrat bereits eine Anti-Plagiatsfunktion: "Der Wissenschaftsrat begreift die umfassende öffentliche Sammlung und Bereitstellung von Forschungsprimärdaten auch als ein probates Mittel der Qualitätssicherung in der wissenschaftlichen Praxis, welches hilft, wissenschaftlichen Betrug und Plagiate leichter zu identifizieren, da die Herkunft von Forschungsdaten aus Repositorien in jedem Falle offen gelegt und die „Autoren“ der Daten zitiert werden müssen."

Mittels der Archivierung von Primärdaten sollen also auch Betrug und gefälschten Studien vorgebeugt werden.

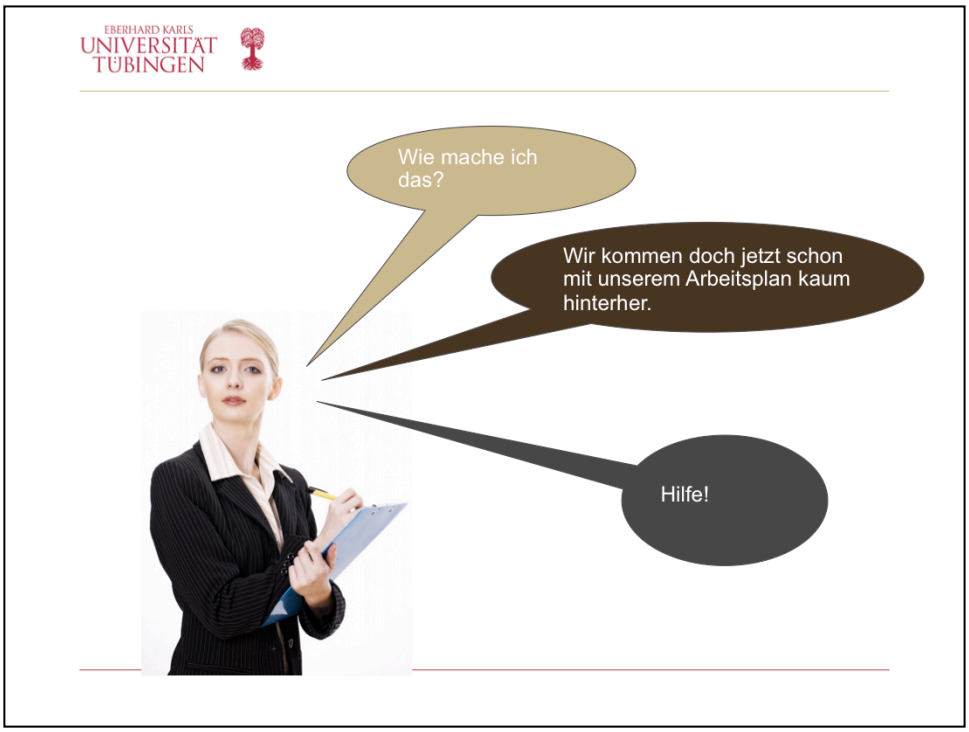
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Konsequenzen aus der Verfügbarkeit

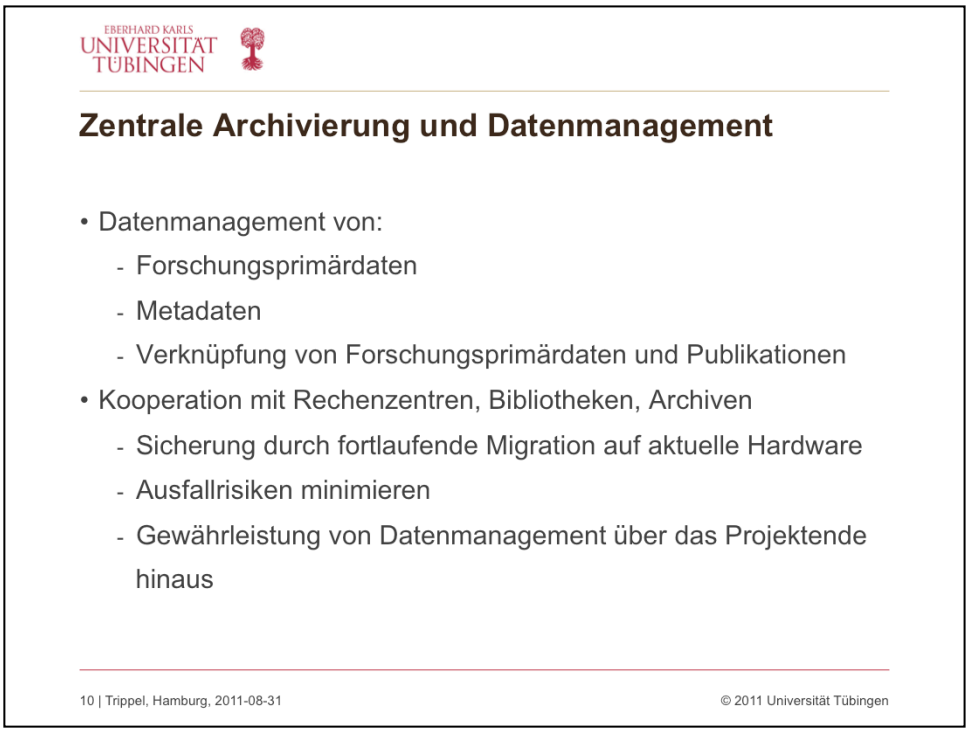
„Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für **zehn Jahre** aufbewahrt werden.“

Aus:
Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten, DFG: Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme, Unterausschuss für Informationsmanagement, S. 2, Januar 2009, Hervorhebung hinzugefügt.

Die Deutsche Forschungsgemeinschaft geht sogar von Datenhaltungen für 10 Jahre aus, und gibt damit den Wissenschafltern den sprichwörtlichen schwarzen Peter zur Archivierung und Zugänglichmachung.

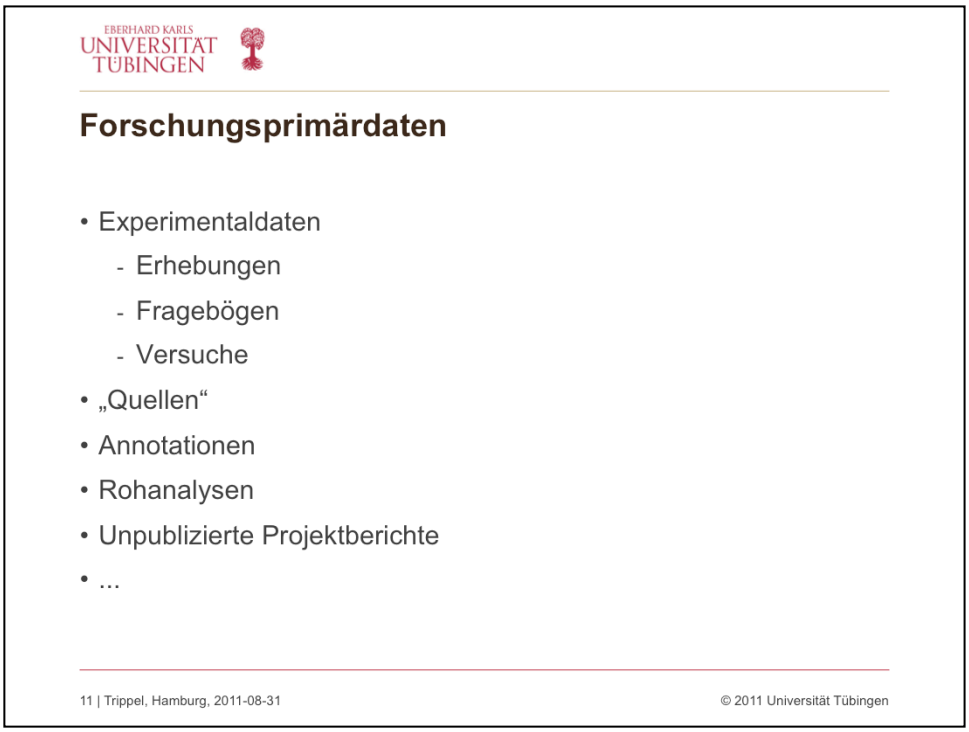


Der einzelne Wissenschaftler wird gerne damit alleine gelassen. Es wird einfach davon ausgegangen, es wird vorausgesetzt, dass er das tut, ohne dass es Anleitungen oder qualifizierte Hilfestellungen gäbe. Für Bücher gibt es normalerweise Bibliotheken mit Bibliothekaren, die große Bestände langfristig verwalten, die haben die Erfahrung, die katalogisieren, die wissen, wie das geht. In der Zwischenzeit gibt es aber eine Reihe von Projekten, die auch von den Geldgebern für Forschungsprojekte getragen werden, die sich darum bemühen, die Forschung im Bereich der Archivierung voranzutreiben und Infrastrukturen und Expertisen für die Archivierung aufzubauen

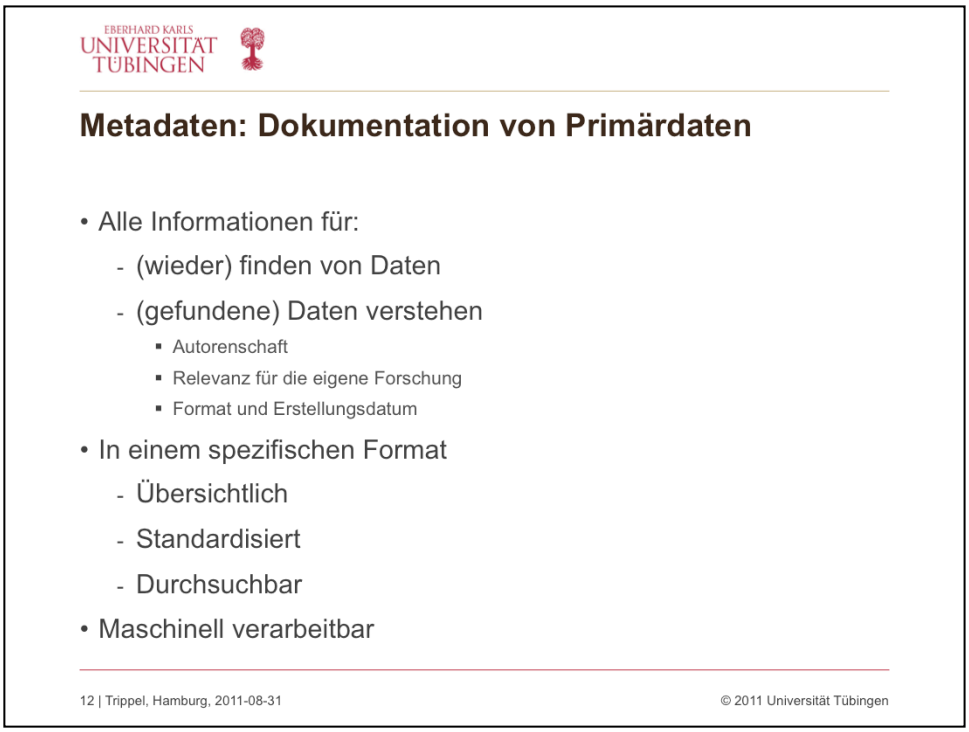


Als zentrale Anlaufpunkte etablieren sich Archive, die häufig mehr oder weniger eng mit etablierten Hochschuleinrichtungen wie Rechenzentren und Bibliotheken zusammenarbeiten. Dabei gibt es sowohl lokale Einrichtungen, etwa hier in Hamburg das Hamburger Zentrum für Sprachkorpora; hier am Institut für Gebärdensprache das Archiv, das meines Wissens das beste und größte Archiv für Gebärdensprachendaten in Deutschland ist; es gibt aber auch große Verbundprojekte und Einrichtungen, die im Bereich der Standardisierung und Archivierung zusammenarbeiten. Allen gemein ist: es werden Forschungsprimärdaten erfasst, mit beschreibenden Informationen. Es werden, genau wie bei Büchern und Bibliotheksnummern, Identifikatoren vergeben, so dass auf die Forschungsprimärdaten verwiesen werden kann.

Und weil bestimmte Probleme mit der Archivierung grundlegender Art sind, sorgt die Kooperation mit zentralen Infrastruktureinrichtungen wie Bibliotheken und Rechenzentren für ein Maximum an Sicherheit, sowohl was die Datenhaltung und Backups, angeht, als auch die Verfügbarkeit.



Worüber sprechen wir? Also über verschiedene Daten, aber auch über Aufnahnmen und andere Quellen mit ihren Annotationen und Rohanalysen. Natürlich ist es immer eine Frage, was Rohdaten sind und was bereits analysiert ist, aber lassen Sie mich das heute ein wenig lax angehen.

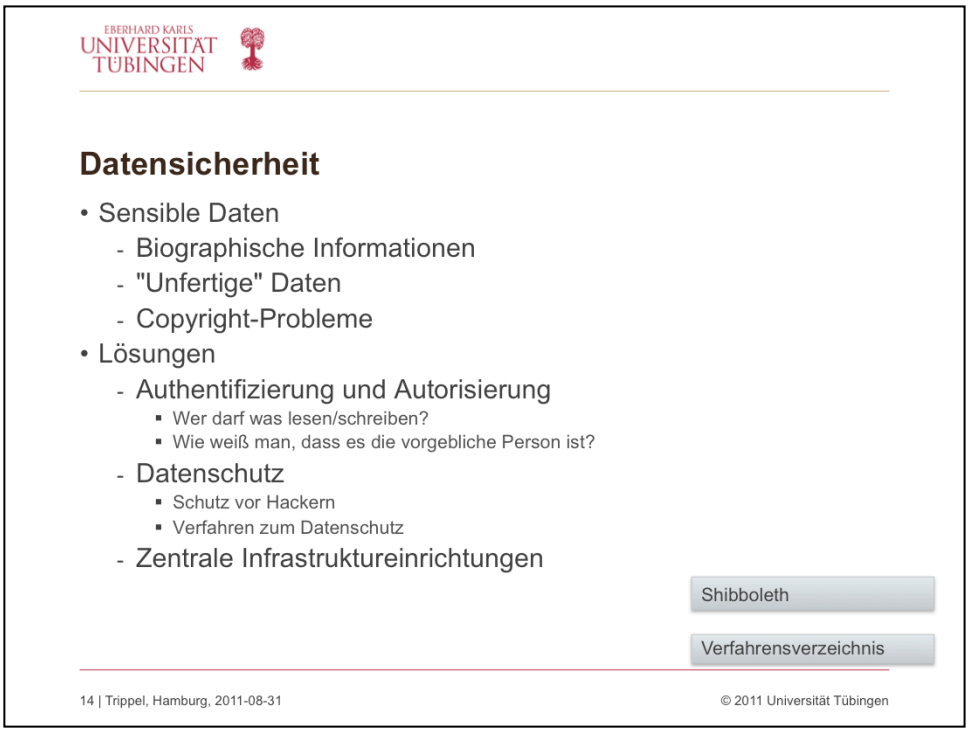


Zur Beschreibung dieser Primärdaten werden Metadaten verwendet. Metadaten können daher als die Informationen angesehen werden, die man benötigt, um Daten, also z.B. eine Datei auf einem Computer, wieder zu finden oder um zu verstehen, um was es sich bei einer Datei, die man gefunden hat, wirklich handelt.

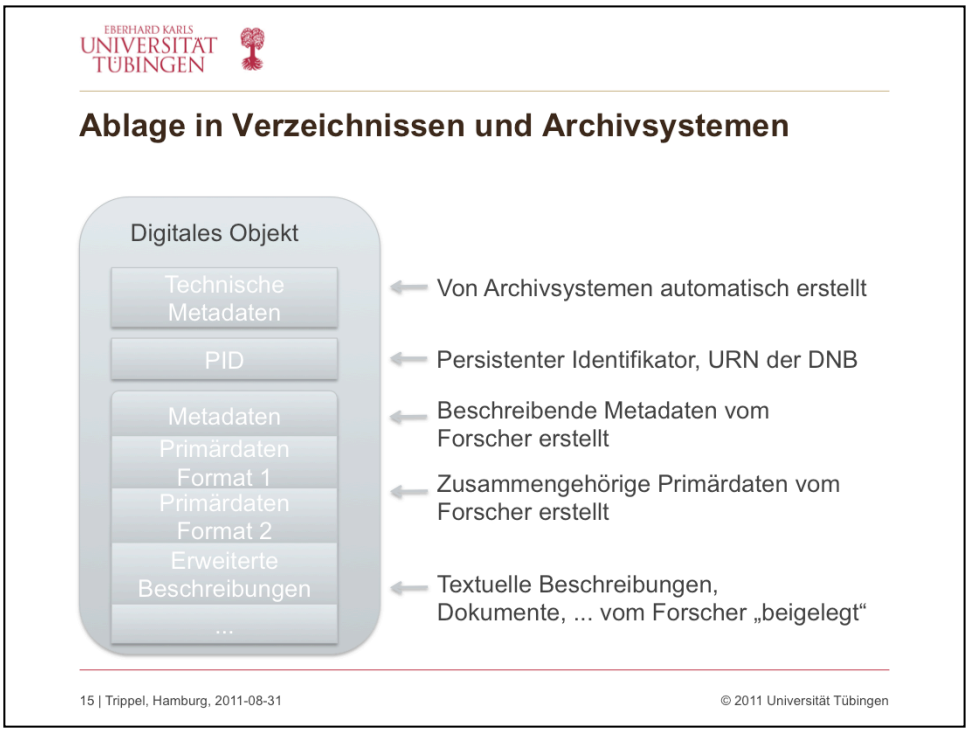
Dazu müssen diese Informationen in einem spezifischen Format vorliegen, das übersichtlich, standardisiert und durchsuchbar ist, so dass sowohl Suchmaschinen in einer maschinellen Verarbeitung darauf zugreifen können, als auch Menschen einen Eindruck davon erhalten, was diese Daten denn beinhalten. Beispiele für Metadaten sind etwa die bibliographischen Angaben in Publikationen, in Bibliotheken oder auch die Titelei.

Um eine Inselbildung zu vermeiden, wird großer Wert auf Interoperabilität der Datenzentren und deren verschiedenen Lösungen gelegt. Um z.B. Metadaten auszutauschen, werden Verfahren und Protokolle verwendet, die sich im Bibliothekswesen bewährt haben. Das Open Access Initiative Protocol for Metadata Harvesting (OAI-PMH) stellt so ein Protokoll dar, es erlaubt Metadaten auch in verschiedenen Formaten und unterschiedlichem Detailreichtum auszutauschen.

Um über die strukturierten Metadaten hinreichend gut suchen zu können, ist es wiederum notwendig, dass es gemeinsame Strukturen gibt. In der Entwicklung hat sich allerdings herausgestellt, dass die Beschreibungen von Büchern und anderen traditionellen Medien sich sehr stark davon unterscheidet, was man z.B. für die linguistische Feldforschung braucht, wo neben Audio- und Videoaufnahmen, Texten, Fragebögen, geographische Informationen, etc. beschrieben werden. Linguistische Software wiederum hat ebenso stark davon abweichende Beschreibungsebenen. Um möglichst viele Metadaten ähnlich zu gestalten, werden Beschreibungsebenen zu Blöcken zusammengefasst, die für verschiedene Ressourcentypen relevant sind. So kann etwa ein Autor, Ursprungsprojekt, der Ressourcenname als Einheit zusammengefasst für viele Ressourcentypen verwendet werden, diese gibt es meist auch für traditionelle Medien in Bibliotheken, wo solche Datentypen



Lassen Sie mich auf die beiden unteren Boxen, eingehen, die direkt mit der Datensicherheit und der Privatsphäre zu tun haben. Da wir im Bereich der Sprache teils mit sehr sensiblen Daten umgehen, müssen wir dafür auch Vorkehrungen treffen. Daten wie Korpora oder Experimentaldaten enthalten häufig Rückschlüsse auf biographische Informationen. Auch sind Daten, über die noch nicht publiziert wurde, häufig noch vertraulich, manchmal sind auch andere Eigentumsrechte betroffen. Der Besitzer darf natürlich darauf zugreifen, aber es ist auch notwendig dafür zu sorgen, dass nur berechtigte Personen zugreifen dürfen. Technische Verfahren wie das Shibboleth-System, was von vielen Universitäten unterstützt wird, erlauben dabei Nutzer zuverlässig zu erkennen, Archivsysteme müssen dann nur noch feststellen, ob diese Person auf Daten zugreifen darf. Zentrale Infrastrukturen wie Rechenzentren haben auch ein große Arsenal zur Hand, um die Daten technisch zu schützen, z.B. gegen Diebe und Hacker, für die Zugangsbeschränkung, also wer etwas lesen darf, müssen aber noch weitere Instrumentarien verwendet werden. Wir verwenden dafür Verfahrensverzeichnisse, wie sie auch von Behörden und ähnlichen Einrichtungen verwendetet werden, in denen detailiert geregelt wird, wer das Recht hat, anderen Zugansrechte zu geben und welche Voraussetzungen der zu erfüllen hat. Für einen Einzelforscher ist dieser Overhead zu groß, für Spezialeinrichtungen mit Verbindung zu Infrastruktureinrichtungen ist dies zu befältigen.



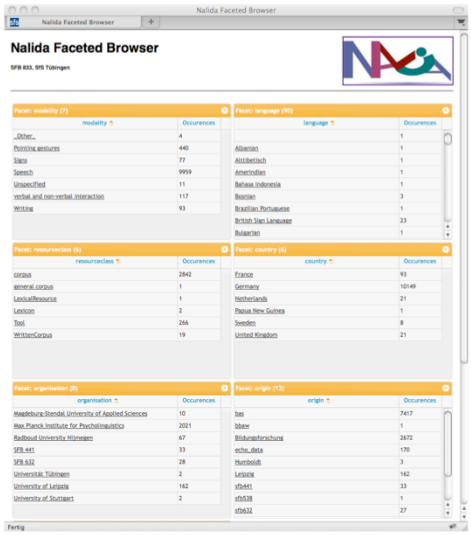
Forschungsprimärdaten und beschreibende Metadaten werden zusammen zu sogenannten digitalen Objekten zusammengefasst und im Archivsystem mit Bezug zueinander abgelegt werden. Wie in einer Bibliothek, in der ein neues Buch dadurch aufgenommen wird, dass die bibliographischen Informationen in den Katalog eingefügt werden, Bibliotheksspezifische Angaben in die Titelei eingestempelt werden und das Buch einem Standort zugewiesen wird, werden auch elektronische Primärdaten in ein Archivsystem dadurch aufgenommen, dass beschreibende Metadaten und die Primärdaten zusammen aufgenommen werden, technische Metadaten und ein Identifikationsmerkmal, ein Persistenter Identifikator, angelegt wird.

Wie sehen solche Metadaten aus? Hier habe ich einen Katalog linguistischer Ressourcen mitgebracht, Sie sehen verschiedene Gruppierungen für allgemein informationen, alles zur Erstellung und Dokumentation. Jeweils unter diesen Überschriften findet man dann die Angaben, die einem einen guten Überblick über die Daten gewähren.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faceted Browser

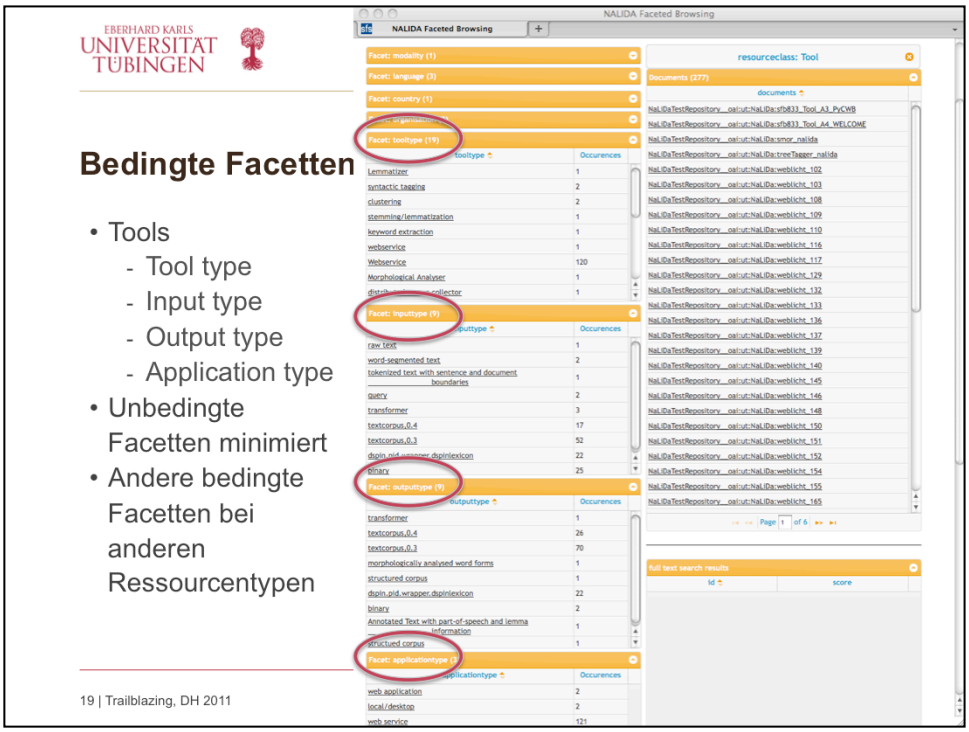
- Basierend auf OAI-PMH-Feeds
- Verwendung fachbezogener Facetten
  - z.T. vom Ressourcentyp gesteuert
  - Allgemeine Facetten für Suchraumbeschränkung
- Verwendung des Komponentenmodells für Metadaten



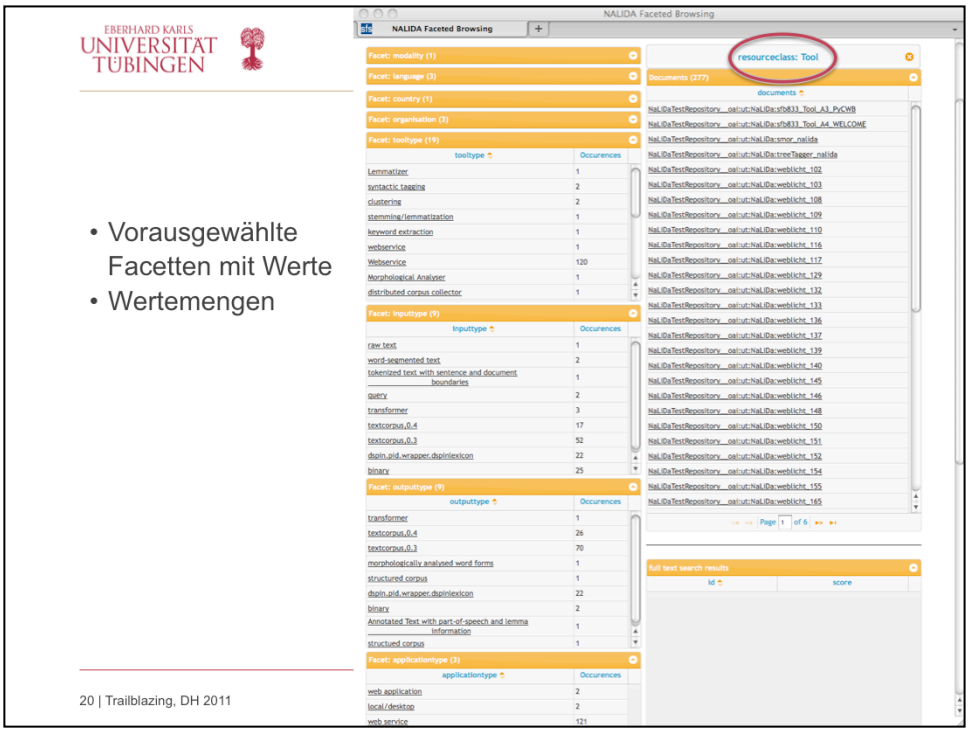
Diese Informationen können aber auch für die Suche nach Ressourcen verwendet werden. Hier zum Beispiel haben wir ein Suchinterface für Sprachressourcen, die auf solchen Ressourcen aufbauen. Abhängig vom Typ einer Ressource werden unterschiedliche Unterkategorien verwendet, die man als Facetten bezeichnet

Unbedingte Facetten sind diejenigen, die unabhängig von weiteren Einschränkungen verwendet werden, z.B. Modalität, Sprache, Organisation, etc. Im Allgemeinen können alle Facetten, die in allen Ressourcen verwendet werden, so benutzt werden, in der Praxis muss man aber eine Vorauswahl treffen, um die Übersichtlichkeit zu gewährleisten.

Die Zahl in jeder Zeile gibt an, wie viele Ressourcen mit den aufgeführten Eigenschaften im Suchraum vorhanden sind.



Abhängig von der Ressourcenart werden auch bedingte Facetten definiert. Für Textkorpora sind z.B. Genre-Informationen zur Auswahl relevant, für Computerprogramme, wie hier dargestellt, ist der Input und Output wichtiger. Auch hier müssen Fachleute eine Vorauswahl treffen.



Note also that after the selection of facets, the facets and the selected value is still communicated to a user and the user may deselect these facet. Additionally, at least below a threshold, the documents that match to the facet values are listed. These lists can be rather long, hence a threshold should be selected based on human expertise.

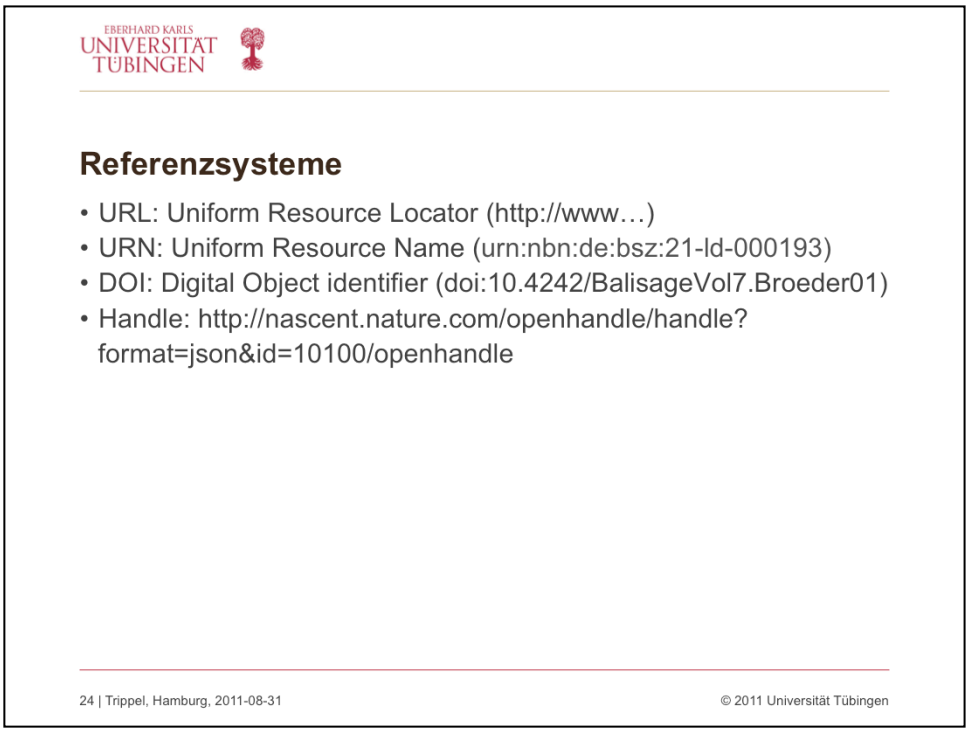
Im Idealfall erhält man nach der Auswahl der für einen interessanten Facetten nur eine kleine Anzahl von Dokumenten, hier z.B. habe ich nach einem Lemmatizer aus Stuttgart für geschriebene Ressourcen gesucht, in unserem Suchraum bleibt nur der TreeTagger übrig.

Und damit hat man dann zugriff auf den vollständigen Katalogeintrag.

Hier nochmal ein Katalogeintrag, wie man ihn als Ergebnis erhält. Wenn man also so einen Eintrag erhält und dadurch eine Ressource erhält, was kann man damit machen? Man kann diese Ressource in Publikationen und Analysen zitieren.

Der Wissenschaftsrat sagt dazu: Der Wissenschaftsrat fordert insbesondere die Forschungsförderer auf, Anreize zu schaffen, um qualitativ hochwertige Daten zu archivieren und langfristig zu sichern. Zu diesem Zweck sollten Referenz- und entsprechende Zitationsmöglichkeiten für Datensätze aufgebaut werden. *Persistent identifier* (PI) bzw. *digital object identifier* (DOI) erlauben eine eindeutige Identifizierung und Zitierbarkeit digital hinterlegter Daten selbst dann, wenn sie ihre Speicherorte (in der Regel referenziert über den *uniform resource locator*, URL) wechseln.

Da elektornische Primärdaten anders aufgebaut und anders archiviert werden, müssen sie auch anders als in Bibliotheken referenziert werden.



Der Vorteil von elektronischen Ressourcen ist, dass man sie eindeutig und einfach referenzieren kann. Im Gegensatz zu Büchern, deren bibliographischen Angaben man braucht, um sie im Bibliothekskatalot zu finden, wodurch man ihren Standort finden kann.

In der Welt des Internets kennen wir die URL, die leider nicht sehr persistent ist: Webserver ändern sich. Um das Problem zu lösen, gibt es verschiedene Serviceprovider, die dauerhafte identifikatoren vergeben, ändert sich dann ein Server, muss das nur bei diesem Serviceprovider geändert werden. Die Deutschen Nationalbibliothek etwa verwendet das URN-System, ursrpünglich vom Buch kommend, DOI werden insbesondere von Verlagen verwendet, Handles sind ein System, bei dem der Identifikator auch gleichzeitig eine URL ist: Ein click und man wird zum Standort der richtigen Ressource weitergeleitet.

URNs, DOI und Handle kann man angeben, um auf Ressourcen dauerhaft zu verweisen, ohne Datumsangabe oder ähnliches, weil sie immer eindeutig zu genau einer Ressource gehören.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Was will NaLiDa

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Archivierung von Forschungsdaten – spannend, wichtig und einfach zu bewerkstelligen...

- Motivation
  - Gefordert von Forschungsförderern
  - Neue Fragestellungen
- Datenzentren für den Overhead
- Metadaten zum Finden und gefunden werden
- Verweise über PIDs
- Informationen beim NaLiDa-Projekt
  - nalida@sfs.uni-tuebingen.de
  - http://www.sfs.uni-tuebingen.de/nalida

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Fotos

- Man reading Image: Arvind Balaraman / FreeDigitalPhotos.net
- Business Woman Taking Notes Image: Michal Marcol / FreeDigitalPhotos.net