



Qualitätsanalyse durch Clustern von Datenwerten

Pilot-Evaluation

20.12.2021, 13:30-17 Uhr















- 13:30 Einführung (45 min) (~ 40 min)
- 14:10 Training (35 min) (~ 20 min)
- 14:30 Analyse Teil 1 (55 min) (~ 30 min)
- 15:00 Analyse Teil 2 (45 min) (~ 20 min)
- 15:20 Pause (1 h) (~ 15 min)
- 15:35 Fazit zur Analyse (15 min) (~ 10 min)
- 15:45 Auswertung Teil 1 (20 min) (~ 15 min)
- 16:00 Auswertung Teil 2 (15 min) (~ 10 min)
- 16:10 Analyse Teil 3 (~ 20 min)
- 16:30 Abschlussbesprechung (10 min) (~ 30 min)



Ziele & Nutzen des Workshops

Für die Community

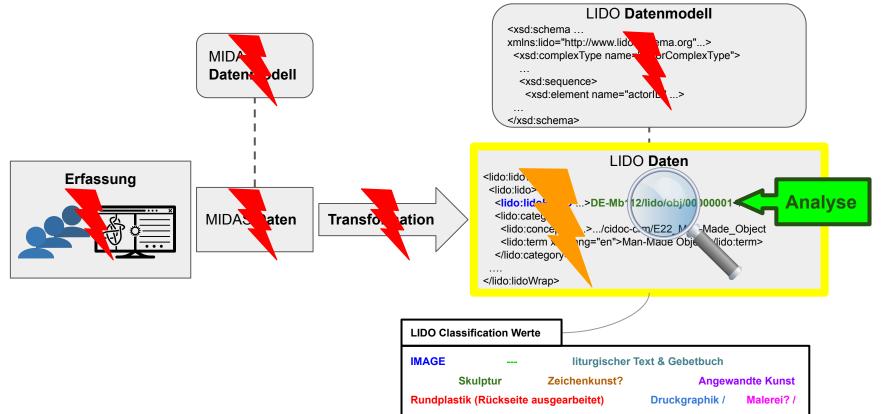
- Kennenlernen eines innovativen Tools zur explorativen Datenqualitätsanalyse
- Direkte Einflussnahme auf die zukünftige Verbesserung und Erweiterung des Tools

Für die Forschung & Entwicklung

- Evaluation des Tools in praxisnahem Szenario
- Erkennen von Stärken, Schwächen und neuen Anforderungen



Datenqualitätsprobleme und ihre Ursachen





Motivation

Beobachtung: Werte im selben Feld sind oft heterogen

IMAGE -- liturgischer Text & Gebetbuch Skulptur Zeichenkunst? Angewandte Kunst Rundplastik (Rückseite ausgearbeitet) Druckgraphik / Malerei? /

Wie verschafft man sich hier einen Überblick?

Was verrät die Heterogenität bzgl. Datenqualität?

Wo liegt die Ursache der Heterogenität?



Running Example: Classification

/lidoWrap/lido/descriptiveMetadata/objectClassificationWrap/classifica

classificationWrap (element)

Definition: A wrapper for classification information.

Sequence:

classification (0-unbounded)

Definition: Concepts used to categorize an object / work by grouping it together with others on the basis of similar characteristics. How to record: The category belongs to a systematic scheme (classification) which groups objects of similar characteristics according to uniform aspects. This grouping / classification may be done according to material, form, shape, function, region of origin, cultural context, or historical or stylistic period. In addition to this systematic grouping it may also be done according to organizational divisions within a museum (e.g., according to the collection structure of a museum). If the object / work is assigned to multiple classifications, repeat this element.Preferably taken from a published controlled vocabulary.

Extension (base lido:conceptComplexType)

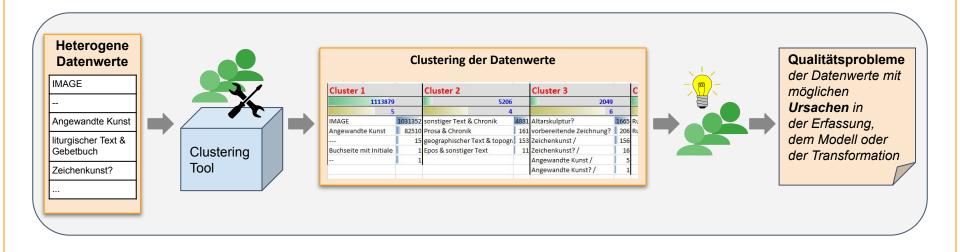
Attribute lido:type

Attribute lido:sortorder



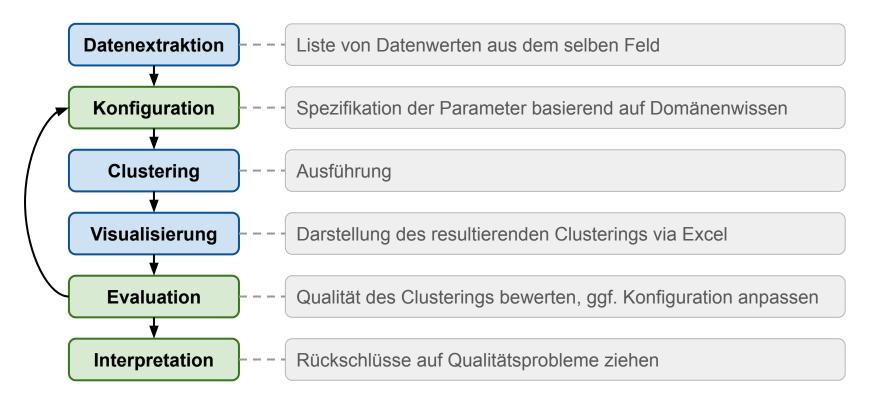


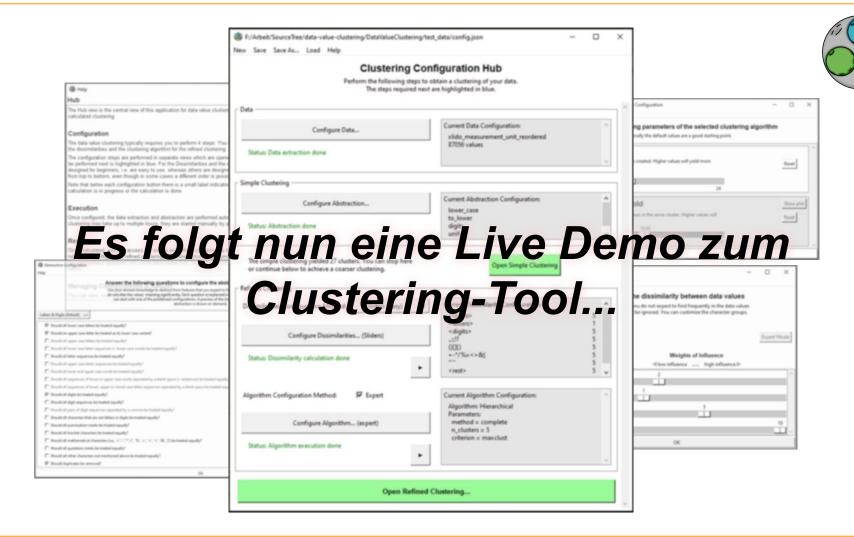
Das Clustering-Tool





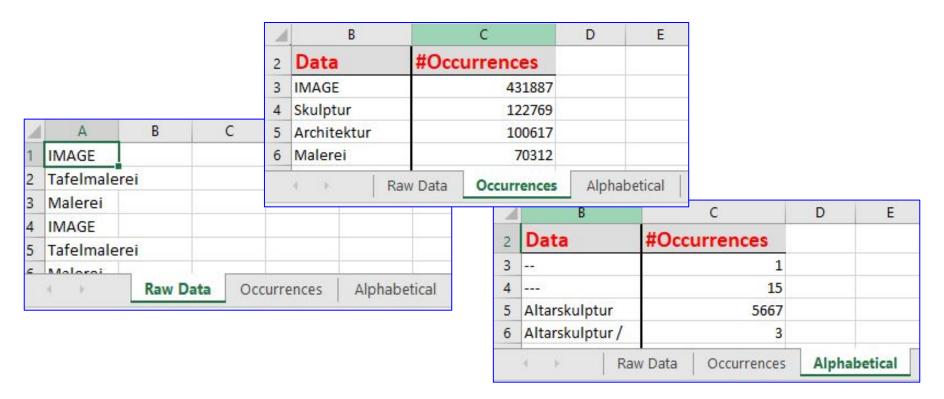
Workflow des Clustering-Ansatzes







Alternative Methode: Werteliste(n)







Ablauf



14:10 Hands Training (~ 20 min)

14:30 Analyse Teil 1 (~ 30 min)

15:00 Analyse Teil 2 (~ 20 min)

15:20 Pause (~ 15 min)

15:35 Fazit zur Analyse (~ 10 min)

15:45 Auswertung Teil 1 (~ 15 min)

16:00 Auswertung Teil 2 (~ 10 min)

16:10 Analyse Teil 3 (~ 20 min)

16:30 Abschlussbesprechung (~ 30 min)





Welche Daten betrachten wir?

Format: LIDO v1.0 (www.lido-schema.org/schema/v1.0/lido-v1.0-schema-listing.html)

Thema: Objekte der materiellen Kultur

Entstehung:

- Manuelle Erfassung in MIDAS
- Transformation zu LIDO

Felder:

- Training: "ShapeMeasurements"
- Analyse Teil 1 und 2: "ActorName" und "RepositoryName"
- Extra Teil 3: "RepositoryLocationName"

Pilot



ShapeMeasurements

/lidoWrap/lido/descriptiveMetadata/objectIdentificationWrap/objectMeasurementsWrap/objectMeasurementsSet/objectMeasurements/shapeMeasurements

shapeMeasurements (0-unbounded)

Definition: The shape of an object / work. Used for unusual shapes (e.g., an oval painting). How to record: Example values: oval, round, square, rectangular, and irregular.

Simple content

Extension (base lido:textComplexType)

Attribute lido:sortorder

LIDO Shape Measurements

 oval
 fünfeckig (unregelmäßig)
 Saalkirche? /
 Kopfbahnhof

 Eckhaus mit Turm
 24-zeilig
 Doppelstatue

 gerahmt
 5/ 8-Schluss
 gerahmt (Architekturrahmen)



ActorName

/lidoWrap/lido/descriptiveMetadata/eventWrap/eventSet/event/eventActor/actor/nRole/actor/nameActorSet/appellationValue

actorComplexType (complex type)

Definition: Contains identifying and indexing actor information.

How to record: Data values of the type attribute: person, corporation, family, group.

Sequence:

actorID (lido:identifierComplexType 0-unbounded)

nameActorSet (lido:appellationComplexType 1-unbounded)

Definition: A wrapper for name elements.

How to record: if there exists more than one name for a single actor, repeat Name Actor Set.

Notes: Indicates names, appellations, or other identifiers assigned to an individual, group of people, firm or other corporate body, or other entity.

nationalityActor (0-unbounded)

vitalDatesActor (lido:dateComplexType 0-1)

genderActor (<u>lido:textComplexType</u> 0-unbounded)

LIDO ActorName

unbekannt Rembrandt Harmensz, van Rijn Rembrandt

Jean (Berry, Herzog, 1, le Magnifique) ? Miro, Joan

Palant, Werner (Graf, 2) & Palant, Elverad von & Bergerhausen, Margarete von



RepositoryName (Teil 1)

/lidoWrap/lido/descriptiveMetadata/objectIdentificationWrap/repositoryWrap/repositorySet/repositoryName/legalBodyName/appellationValue

repositorySetComplexType (complex type)

Definition: Wrapper for designation and identification of the institution of custody and, possibly, indication of the exact location of the object. How to record: If there are several designations known, e.g., a current one and former ones (see: type attribute), repeat the element.Data values of the type attribute: current, former.

Sequence:

repositoryName (lido:legalBodyRefComplexType 0-1)

Definition: Unambiguous identification, designation and weblink of the institution of custody.

workID (0-unbounded)

repositoryLocation (lido:placeComplexType 0-1)



RepositoryName (Teil 2)

/lidoWrap/lido/descriptiveMetadata/objectIdentificationWrap/repositoryWrap/repositorySet/repositoryName/legalBodyName/appellationValue

legalBodyRefComplexType (complex type)

Definition: Reference information to a legal body.

Sequence:

legalBodyID (lido:identifierComplexType 0-unbounded)

legalBodyName (<u>lido:appellationComplexType</u> 0-unbounded)

Definition: Appellation of the institution or person.

legalBodyWeblink (lido:webResourceComplexType 0-unbounded)

LIDO RepositoryName

? (Paris, Prinz) (private Sammlung) (Paris)

3. Register

Ägyptisches Museum (Kairo)

Apotheke

zwischen Ruhrortsbrücke und Eisenbahnhafen

Universitätsmuseum für Kunst und Kulturgeschichte (Marburg)

Zur Krone

Sankt Lorenz ob Katsch

X





RepositoryLocationName (Teil 1)

/lidoWrap/lido/descriptiveMetadata/objectIdentificationWrap/repositorySet/repositoryLocation/namePlaceSet/appellationValue

repositorySetComplexType (complex type)

Definition: Wrapper for designation and identification of the institution of custody and, possibly, indication of the exact location of the object. How to record: If there are several designations known, e.g., a current one and former ones (see: type attribute), repeat the element.Data values of the type attribute: current, former.

Sequence:

repositoryName (lido:legalBodyRefComplexType 0-1)

workID (0-unbounded)

repositoryLocation (lido:placeComplexType 0-1)

Definition: Location of the object, especially relevant for architecture and archaeological sites.





RepositoryLocationName (Teil 2)

/lidoWrap/lido/descriptiveMetadata/objectIdentificationWrap/repositoryWrap/repositoryLocation/namePlaceSet/appellationValue

placeComplexType (complex type)

Definition: Structured element for place information

Sequence:

placeID (lido:identifierComplexType 0-unbounded)

namePlaceSet (lido:appellationComplexType 0-unbounded)

Definition: The name of the geographic place.

How to record: If there are different names of the same place, e.g. today's and historical names, repeat this element.

gml (0-unbounded)

partOfPlace (lido:placeComplexType 0-unbounded)

placeClassification (0-unbounded)

LIDO RepositoryLocationName

Marburg Goslar, Hoher Weg 1 Wien - 1. Bezirk (Wien), Graben

Lorch am Rhein & Lorch Schöntal (Hohenlohekreis)

Berlin - Charlottenburg (Berlin), Witzlebenstraße 4-5

Mainz, Marktplatz Rodgau - Nieder-Roden Mans, Le







- 13:30 Einführung (~ 40 min)
- > 14:10 Training (~ 20 min)
 - 14:30 Analyse Teil 1 (~ 30 min)
 - 15:00 Analyse Teil 2 (~ 20 min)
 - 15:20 Pause (~ 15 min)
 - 15:35 Fazit zur Analyse (~ 10 min)
 - 15:45 Auswertung Teil 1 (~ 15 min)
 - 16:00 Auswertung Teil 2 (~ 10 min)
 - 16:10 Analyse Teil 3 (~ 20 min)
 - 16:30 Abschlussbesprechung (~ 30 min)



Aktuelle Aufgabe: Download und Alias

Laden sie sich das Clustering-Tool herunter:

tinyurl.com/konda-dvc



- Denken sie sich ein Alias aus und notieren sie es lokal
- Tragen sie das Alias hier ein:

tinyurl.com/konda



5 Minuten Pause im Gange...

Aktuelle Aufgabe: Alias Ordner checken

- Bei Google abmelden oder Privates Fenster nutzen
- Unterordner in Drive checken:
 - tinyurl.com/konda-dvcworkshop
 - Anleitung (PDF)
 - Zuteilung zu einer Gruppe
 - Anweisungen wann welche Daten mit welcher Methode analysiert werden sollen usw.
 - Links zu Fragebögen und LIDO Dokumentation
 - Werteliste (XLSX)
 - Später kommt hinzu:
 - Ihre beiden ausgefüllten Fragebögen (PDF)
 - Zwei zu bewertende ausgefüllte Fragebögen einer anderen Person (PDF)



Aktuelle Aufgabe: Training

- Entpacken sie die Datei DataValueClustering.zip
- Führen sie die Anwendung DataValueClustering.exe aus



- Machen sie sich mit dem Clustering-Tool vertraut
- Nutzen sie die Feldwerte ShapeMeasurements
- Durchlaufen sie alle Konfigurationsphasen
- Betrachten sie das resultierende Simple und Refined Clustering
- Stellen sie jegliche Fragen im Plenum





- 13:30 Einführung (~ 40 min)
- 14:10 Training (~ 20 min)
- > 14:30 Analyse Teil 1 (~ 30 min)
 - 15:00 Analyse Teil 2 (~ 20 min)
 - 15:20 Pause (~ 15 min)
 - 15:35 Fazit zur Analyse (~ 10 min)
 - 15:45 Auswertung Teil 1 (~ 15 min)
 - 16:00 Auswertung Teil 2 (~ 10 min)
 - 16:10 Analyse Teil 3 (~ 20 min)
 - 16:30 Abschlussbesprechung (~ 30 min)





Es folgt nun eine Live Demo zu den







Aktuelle Aufgabe: Analyse Teil 1

Gruppe A

Gruppe A1: nutzt jetzt Clustering-Tool für ActorName

Gruppe A2: nutzt jetzt Werteliste für RepositoryName

Gruppe B

Gruppe B1: nutzt jetzt Clustering-Tool für RepositoryName

Gruppe B2: nutzt jetzt Werteliste für ActorName

Details: siehe Anleitung im eigenen Ordner







- 13:30 Einführung (~ 40 min)
- 14:10 Hands Training (~ 20 min)
- 14:30 Analyse Teil 1 (~ 30 min)
- > 15:00 Analyse Teil 2 (~ 20 min)
 - 15:20 Pause (~ 15 min)
 - 15:35 Fazit zur Analyse (~ 10 min)
 - 15:45 Auswertung Teil 1 (~ 15 min)
 - 16:00 Auswertung Teil 2 (~ 10 min)
 - 16:10 Analyse Teil 3 (~ 20 min)
 - 16:30 Abschlussbesprechung (~ 30 min)



Aktuelle Aufgabe: Analyse Teil 2

Gruppe A

Gruppe A1: nutzt jetzt Werteliste für RepositoryName

Gruppe A2: nutzt jetzt Clustering-Tool für ActorName

Gruppe B

Gruppe B1: nutzt jetzt Werteliste für ActorName

Gruppe B2: nutzt jetzt Clustering-Tool für RepositoryName

Details: siehe Anleitung im eigenen Ordner







- 13:30 Einführung (~ 40 min)
- 14:10 / Training (~ 20 min)
- 14:30 Analyse Teil 1 (~ 30 min)
- 15:00 Analyse Teil 2 (~ 20 min)

> 15:20 Pause (~ 15 min)

- 15:35 Fazit zur Analyse (~ 10 min)
- 15:45 Auswertung Teil 1 (~ 15 min)
- 16:00 Auswertung Teil 2 (~ 10 min)
- 16:10 Analyse Teil 3 (~ 20 min)
- 16:30 Abschlussbesprechung (~ 30 min)





- 13:30 Einführung (~ 40 min)
- 14:10 Hands Training (~ 20 min)
- 14:30 Analyse Teil 1 (~ 30 min)
- 15:00 Analyse Teil 2 (~ 20 min)
- 15:20 Pause (~ 15 min)
- > 15:35 Fazit zur Analyse (~ 10 min)
 - 15:45 Auswertung Teil 1 (~ 15 min)
 - 16:00 Auswertung Teil 2 (~ 10 min)
 - 16:10 Analyse Teil 3 (~ 20 min)
 - 16:30 Abschlussbesprechung (~ 30 min)



Aktuelle Aufgabe: Fazit zur Analyse

Füllen sie den Fragebogen zum Fazit zur Analyse aus



Details: siehe Anleitung im eigenen Ordner





- 13:30 Einführung (~ 40 min)
- 14:10 Hands Training (~ 20 min)
- 14:30 Analyse Teil 1 (~ 30 min)
- 15:00 Analyse Teil 2 (~ 20 min)
- 15:20 Pause (~ 15 min)
- 15:35 Fazit zur Analyse (~ 10 min)
- > 15:45 Auswertung Teil 1 (~ 15 min)
 - 16:00 Auswertung Teil 2 (~ 10 min)
 - 16:10 Analyse Teil 3 (~ 20 min)
 - 16:30 Abschlussbesprechung (~ 30 min)



Aktuelle Aufgabe: Auswertung Teil 1

Gruppen

- **Gruppen A1 und B2:** werten Fragebogen zu ActorName aus
- Gruppen A2 und B1: werten Fragebogen zu RepositoryName aus



Vorgehen

- Offnen sie den ausgefüllten Fragebogen der anderen Person (PDF)
 - o Die Person hat das Datenfeld mit der anderen Methode analysiert als sie
- Öffnen sie ggf. ihren eigenen Fragebogen (PDF)
- Lesen sie die von der anderen Person beschriebenen Probleme und Ursachen
- Füllen sie dabei den Fragebogen zur Auswertung aus

Details: siehe Anleitung im eigenen Ordner





- 13:30 Einführung (~ 40 min)
- 14:10 / Training (~ 20 min)
- 14:30 Analyse Teil 1 (~ 30 min)
- 15:00 Analyse Teil 2 (~ 20 min)
- 15:20 Pause (~ 15 min)
- 15:35 Fazit zur Analyse (~ 10 min)
- 15:45 Auswertung Teil 1 (~ 15 min)
- > 16:00 Auswertung Teil 2 (~ 10 min)
 - 16:10 Analyse Teil 3 (~ 20 min)
 - 16:30 Abschlussbesprechung (~ 30 min)



Aktuelle Aufgabe: Auswertung Teil 2

Gruppen

- **Gruppen A1 und B2:** werten Fragebogen zu RepositoryName aus
- Gruppen A2 und B1: werten Fragebogen zu ActorName aus



Vorgehen

- Offnen sie den ausgefüllten Fragebogen der anderen Person (PDF)
 - o Die Person hat das Datenfeld mit der anderen Methode analysiert als sie
- Öffnen sie ggf. ihren eigenen Fragebogen (PDF)
- Lesen sie die von der anderen Person beschriebenen Probleme und Ursachen
- Füllen sie dabei den Fragebogen zur Auswertung aus

Details: siehe Anleitung im eigenen Ordner





- 13:30 Einführung (~ 40 min)
- 14:10 Hands Training (~ 20 min)
- 14:30 Analyse Teil 1 (~ 30 min)
- 15:00 Analyse Teil 2 (~ 20 min)
- 15:20 Pause (~ 15 min)
- 15:35 Fazit zur Analyse (~ 10 min)
- 15:45 Auswertung Teil 1 (~ 15 min)
- 16:00 Auswertung Teil 2 (~ 10 min)
- > 16:10 Analyse Teil 3 (~ 20 min)
 - 16:30 Abschlussbesprechung (~ 30 min)

Aktuelle Aufgabe: Analyse Teil 3

Pilot



Alle nutzen jetzt Clustering-Tool für RepositoryLocationName



Details: siehe Anleitung im eigenen Ordner







- 13:30 Einführung (~ 40 min)
- 14:10 Hands Training (~ 20 min)
- 14:30 Analyse Teil 1 (~ 30 min)
- 15:00 Analyse Teil 2 (~ 20 min)
- 15:20 Pause (~ 15 min)
- 15:35 Fazit zur Analyse (~ 10 min)
- 15:45 Auswertung Teil 1 (~ 15 min)
- 16:00 Auswertung Teil 2 (~ 10 min)
- 16:10 Analyse Teil 3 (~ 20 min)
- > 16:30 Abschlussbesprechung (~ 30 min)



Was passiert als nächstes?

Anhand der heute gesammelten Daten wollen wir u.A. folgende Forschungsfragen beantworten:

- Mit welcher Methode können *mehr* Datenprobleme und mehr mögliche Ursachen in der Erfassung, dem Datenmodell und der Datentransformation identifiziert werden?
- Mit welcher Methode können Probleme schneller identifiziert werden?
- Die Nützlichkeit welcher Methode wird als Höher wahrgenommen?
- Welche Arten von Problemen werden mit den Methoden häufig identifiziert? Gibt es Unterschiede?
- Wo liegen Stärken und Schwächen des Clustering-Tools?
- Was können wir am Tool verbessern?
- > Tool besser an Anforderungen der Community anpassen
- Forschungspapier schreiben



Was passiert als nächstes?

Anhand der heute gesammelten Daten wollen wir u.A. folgende Forschungsfragen beantworten:

Mit w dem

Mit \

Unsere Publikationen sind auf Zenodo verfügbar:

zenodo.org/communities/konda-project/

Die

Wel

• Wo

Was

- > Tool besse
- Forschungs

Weiteres Interesse an der Auswertung, Weiterentwicklungen usw. gern per E-Mail äußern

> arno.kesper@uni-marburg.de viola.wenz@uni-marburg.de konda@uni-marburg.de

der Erfassung,

Interschiede?





Feedback zum Design der Evaluation

Was war unverständlich, verwirrend, störend, unnötig, zu schnell, sonst irgendwie **problematisch** an Folgendem?

- Zeitplan
- Einführung
- Demos
- Aufgaben
- Fragebögen
- Clustering-Tool
- ..



Fragen, Anmerkungen, Kritik, Verbesserungsvorschläge, ...?



Vielen Dank für die Teilnahme!