

# Kibana als Werkzeug zur Unterstützung der Metadatenkuration

Kolloquium Wissensinfrastruktur, 26.1.2018



# AGENDA

## 1) Motivation

Metadaten in OpenAIRE

## 2) Kibana

Indexierung, Suche & Analyse, Schwierigkeiten

## 3) Referenzen

# Motivation / Projekt OpenAIRE2020

- **promote open scholarship and improve discoverability and reusability of research publications**
- **OpenAIRE platform / technical infrastructure interconnecting collections of research outputs across Europe**
- **create workflows and services on top of this repository content**
- **assist in monitoring [EC funded] research outputs, collaboration with national funders**

[<https://www.openaire.eu/project-factsheets>]

# Motivation / OpenAIRE aggregiert heterogene Metadaten

<b>Plattformen:</b>	<b>OJS, DSpace, EPrints, ...</b>
<b>Länder:</b>	<b>Europa, Asien, Amerika, Afrika, Australien</b>
<b>Quellen:</b>	<b>Aggregator, Institut, Journal, ...</b>
<b>Publikationen:</b>	<b>Literatur, Forschungsdaten, Software, ...</b>
<b>Sprachen:</b>	<b>viele! betrifft Publikationen und Metadaten</b>
<b>Formate:</b>	<b>Dublin Core, DateCite, JATS, ...</b>

# Motivation / OpenAIRE-Portal offenbart Datenlücken

All

SEARCH

Publications | Research Data | Projects | People | Organizations | Data Providers

FUNDER	ACCESS MODE	PUBLICATION YEAR	DOCUMENT TYPE
<a href="#">European Commission</a> (241442)	<a href="#">Open Access</a> (22728299)	<a href="#">2015</a> (1851013)	<a href="#">Article</a> (13656953)
<a href="#">National Institutes o...</a> (163258)	<a href="#">Restricted</a> (270932)	<a href="#">2014</a> (1804975)	<a href="#">Unknown</a> (1937403)
<a href="#">National Science Foun...</a> (133115)	<a href="#">Closed Access</a> (132801)	<a href="#">2016</a> (1706341)	<a href="#">Preprint</a> (1798780)
<a href="#">Wellcome Trust</a> (53600)	<a href="#">not available</a> (33929)	<a href="#">2013</a> (1640377)	<a href="#">Research</a> (1345195)
<a href="#">Swiss National Scienc...</a> (50901)	<a href="#">Embargo</a> (6019)	<a href="#">2012</a> (1520597)	<a href="#">Doctoral thesis</a> (1230289)
<a href="#">View more</a> →		<a href="#">View more</a> →	<a href="#">View more</a> →

DOCUMENT LANGUAGE	DATA PROVIDER	COMMUNITIES
<a href="#">English</a> (11088760)	<a href="#">Europe PubMed Central</a> (4444873)	<a href="#">EGI Foundation</a> (23930)
<a href="#">Undetermined</a> (2069938)	<a href="#">DOAJ-Articles</a> (2971649)	<a href="#">FET FP7</a> (7587)
<a href="#">Japanese</a> (1945685)	<a href="#">JAIRO</a> (2537755)	<a href="#">FET H2020</a> (943)
<a href="#">Russian</a> (1604448)	<a href="#">arXiv.org e-Print Arc...</a> (1333969)	
<a href="#">Portuguese</a> (1181550)	<a href="#">CyberLeninka -</a>	
<a href="#">View more</a> →	<a href="#">Russia...</a> (1250514)	
	<a href="#">View more</a> →	

# Motivation / Datenanalyse fördert Datenqualität

**Analyse nativer Daten (nicht transformierter) hilft:**

- Fehler zu entdecken
- fruchtbare Inhalte zu entdecken
- Formate zu vergleichen
- ...

**damit lässt sich:**

- die OpenAIRE-Transformation verbessern
- den Datenquellen Rückmeldung geben

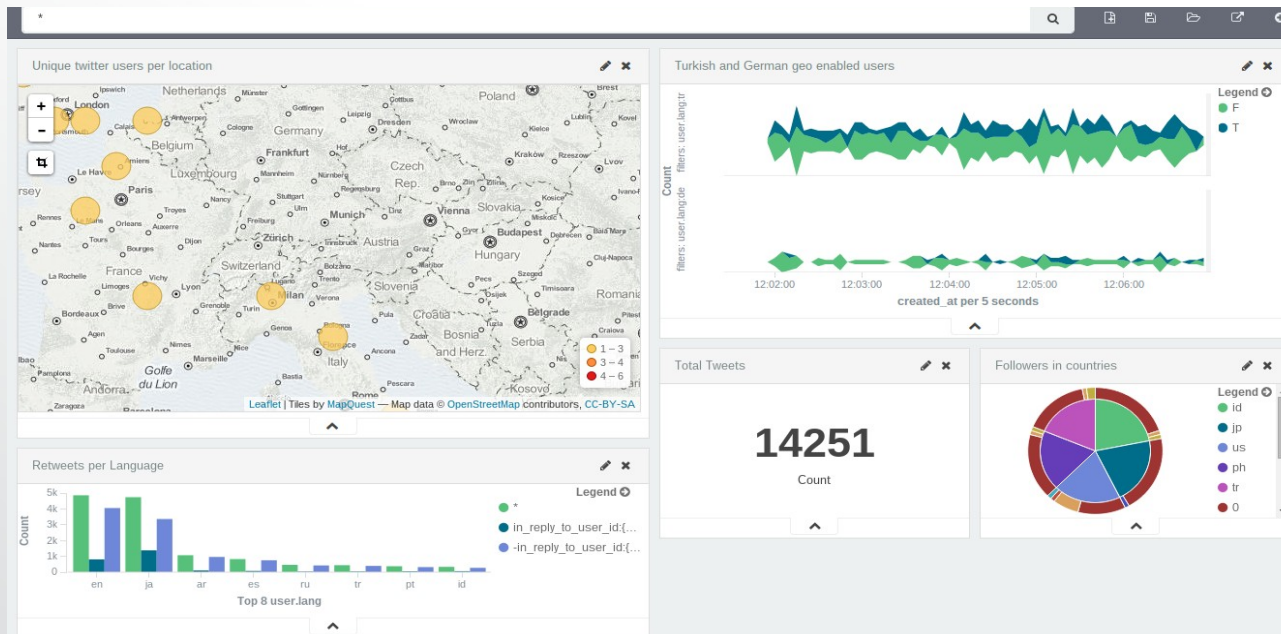
# Kibana / das OpenSource-Tool hilft bei Metadatenanalyse

Kibana erlaubt, Metadaten  
 ~ zu inspizieren & suchen  
 ~ zu evaluieren & visualisieren



Kibana basiert auf

~ Elasticsearch  
 ~ JSON

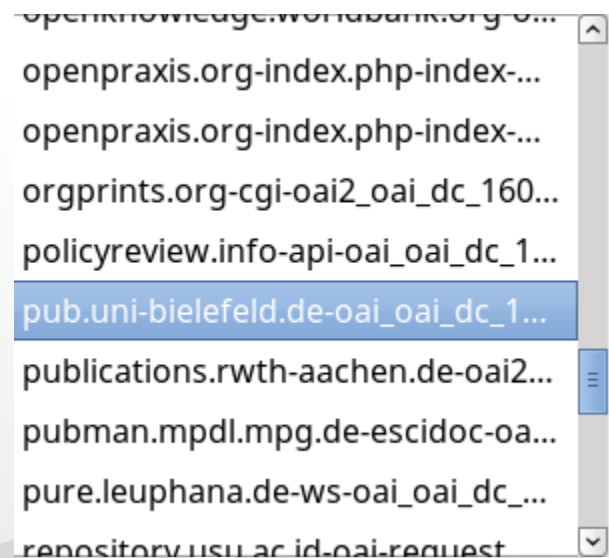


# Kibana Import / Indexierung von Metadaten via JSON

zum Import von Metadaten in Kibana werden

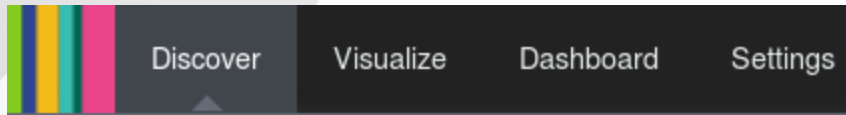
- Datenquellen 'geharvested'
- die Metadaten nach JSON transformiert (XML.toJSONObject)
- die Metadaten in Elasticsearch indexiert

**Kibana kann viele Indizes  
enthalten, und individuell  
oder zusammen analysieren**



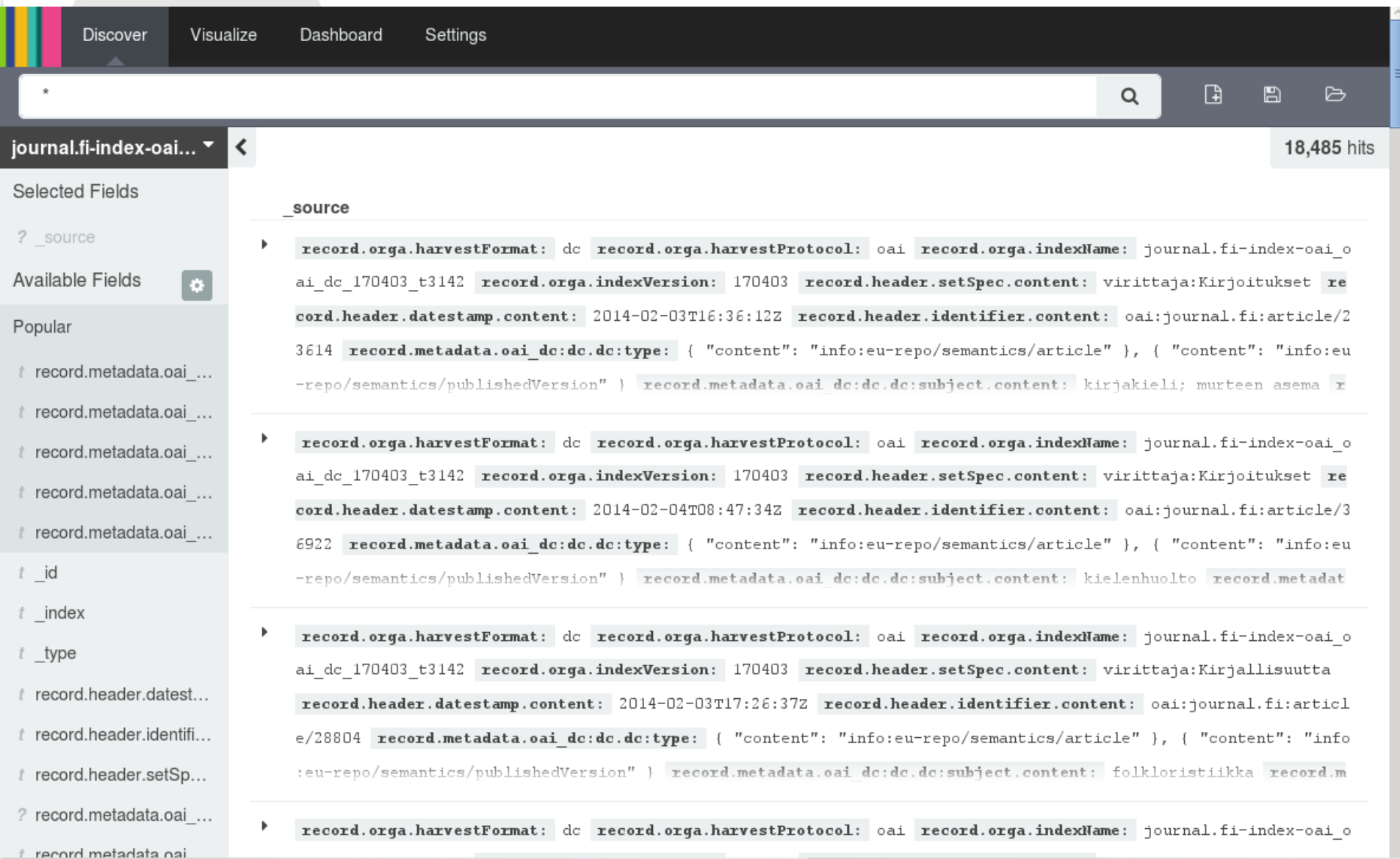


# Kibana Modi / Kibana bietet 4 Sichten



- Discover:** zeigt Datensätze als Liste/Tabelle
- Visualize:** visualisiert Daten mit Graphen, Karten, Tabellen, ...
- Dashboard:** stellt Visualisierungen bereit
- Settings:** fügt Indizes hinzu, zeigt Indexierungsdetails

# Kibana Discover / Discover-Sicht durchsucht Datensätze




The screenshot shows the Kibana Discover interface. At the top, there are navigation tabs: Discover, Visualize, Dashboard, and Settings. Below the navigation is a search bar with a magnifying glass icon and a search button. The search results are displayed in a list view, showing the source of the data and the metadata for each record. The left sidebar contains a 'Selected Fields' section with a search icon, an 'Available Fields' section with a gear icon, and a 'Popular' section with a list of fields.

journal.fi-index-oai... 18,485 hits

Selected Fields

? \_source

Available Fields 

Popular

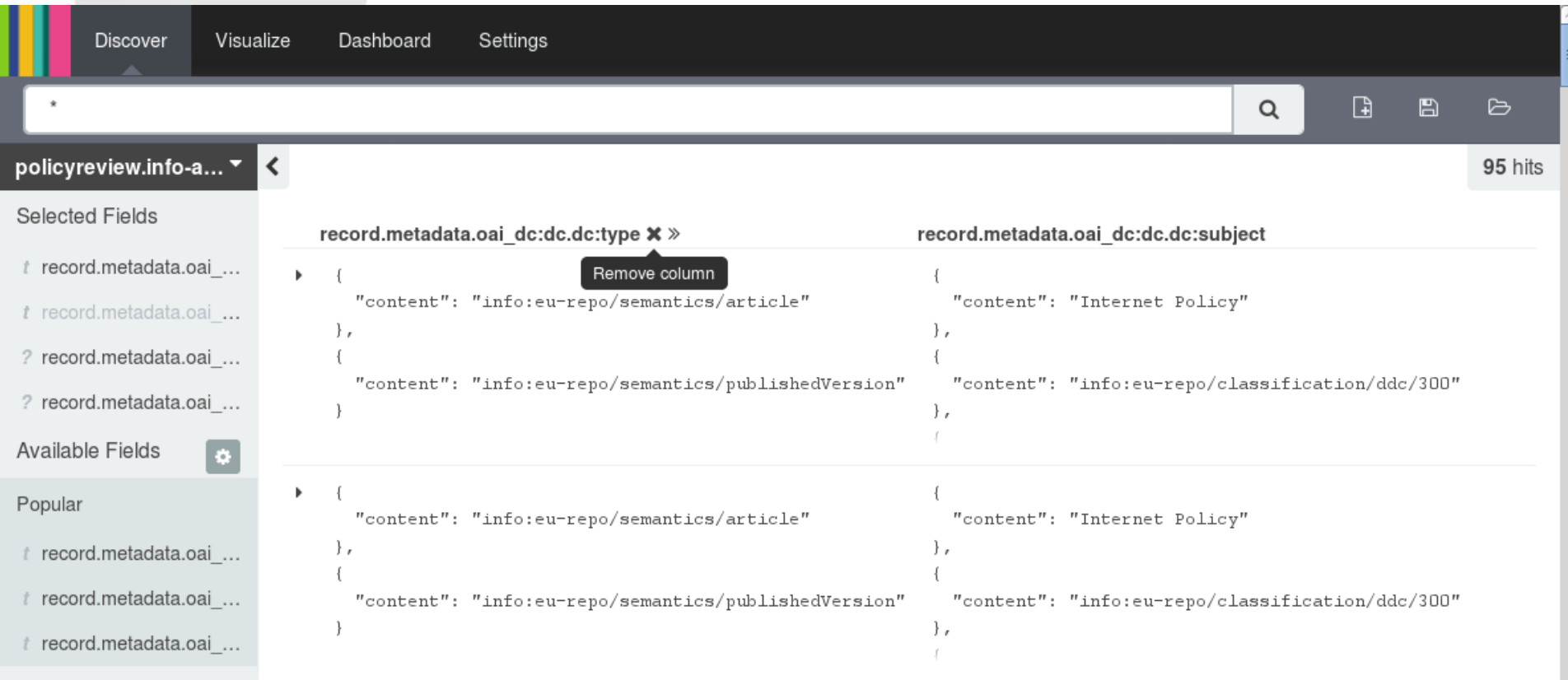
- f record.metadata.oai\_...
- f record.metadata.oai\_...
- f record.metadata.oai\_...
- f record.metadata.oai\_...
- f record.metadata.oai\_...
- f \_id
- f \_index
- f \_type
- f record.header.datest...
- f record.header.identifi...
- f record.header.setSp...
- ? record.metadata.oai\_...
- f record.metadata.oai\_...

**\_source**

- record.org.harvestFormat:** dc **record.org.harvestProtocol:** oai **record.org.indexName:** journal.fi-index-oai\_oai\_dc\_170403\_t3142 **record.org.indexVersion:** 170403 **record.header.setSpec.content:** virittaja:Kirjoitukset **record.header.datestamp.content:** 2014-02-03T16:36:12Z **record.header.identifier.content:** oai:journal.fi:article/23614 **record.metadata.oai\_dc:dc:dc:type:** { "content": "info:eu-repo/semantics/article" }, { "content": "info:eu-repo/semantics/publishedVersion" } **record.metadata.oai\_dc:dc:dc:subject.content:** kirjakieli; murteen asema r
- record.org.harvestFormat:** dc **record.org.harvestProtocol:** oai **record.org.indexName:** journal.fi-index-oai\_oai\_dc\_170403\_t3142 **record.org.indexVersion:** 170403 **record.header.setSpec.content:** virittaja:Kirjoitukset **record.header.datestamp.content:** 2014-02-04T08:47:34Z **record.header.identifier.content:** oai:journal.fi:article/36922 **record.metadata.oai\_dc:dc:dc:type:** { "content": "info:eu-repo/semantics/article" }, { "content": "info:eu-repo/semantics/publishedVersion" } **record.metadata.oai\_dc:dc:dc:subject.content:** kielenhuolto **record.metadat**
- record.org.harvestFormat:** dc **record.org.harvestProtocol:** oai **record.org.indexName:** journal.fi-index-oai\_oai\_dc\_170403\_t3142 **record.org.indexVersion:** 170403 **record.header.setSpec.content:** virittaja:Kirjallisuutta **record.header.datestamp.content:** 2014-02-03T17:26:37Z **record.header.identifier.content:** oai:journal.fi:articl e/28804 **record.metadata.oai\_dc:dc:dc:type:** { "content": "info:eu-repo/semantics/article" }, { "content": "info:eu-repo/semantics/publishedVersion" } **record.metadata.oai\_dc:dc:dc:subject.content:** folkloristiikka **record.m**
- record.org.harvestFormat:** dc **record.org.harvestProtocol:** oai **record.org.indexName:** journal.fi-index-oai\_o



# Kibana Discover / Felder lassen sich im Überblick betrachten



Discover Visualize Dashboard Settings

policyreview.info-a... 95 hits

record.metadata.oai_dc:dc.type <span>✕</span> »	record.metadata.oai_dc:dc.subject
<pre>{   "content": "info:eu-repo/semantics/article" }, {   "content": "info:eu-repo/semantics/publishedVersion" }</pre>	<pre>{   "content": "Internet Policy" }, {   "content": "info:eu-repo/classification/ddc/300" }</pre>
<pre>{   "content": "info:eu-repo/semantics/article" }, {   "content": "info:eu-repo/semantics/publishedVersion" }</pre>	<pre>{   "content": "Internet Policy" }, {   "content": "info:eu-repo/classification/ddc/300" }</pre>

Selected Fields

- record.metadata.oai\_...
- record.metadata.oai\_...
- record.metadata.oai\_...
- record.metadata.oai\_...

Available Fields ⚙️

Popular

- record.metadata.oai\_...
- record.metadata.oai\_...
- record.metadata.oai\_...

# Kibana Discover / gesucht wird nicht in orig. Feldwerten sondern im Index

**Suche wird beeinflusst durch**

- **Typisierung der Feldwerte**  
(interpretiert als String, Zahl, Datum, ...)
- **eventuelle Analyse der Feldwerte**  
(zerlegt bei Satz-/Leerzeichen, Kleinschreibung)



# Kibana Settings / Settings-Sicht zeigt je Index Indexierungsdetails der Felder

The screenshot shows the Kibana Settings interface. The top navigation bar includes 'Discover', 'Visualize', 'Dashboard', and 'Settings'. Below it, there are sub-navigation options: 'Indices', 'Advanced', 'Objects', and 'About'. The left sidebar shows a list of index patterns, with 'pub.uni-bielefeld.de-oai\_oai\_dc\_160712' selected. The main content area displays the index name, a star icon, a refresh icon, and a delete icon. Below the index name, there is a description: 'This page lists every field in the pub.uni-bielefeld.de-oai\_oai\_dc\_160712 index and the field's associated core type as recorded by Elasticsearch. While this list allows you to view the core type of each field, changing field types must be done using Elasticsearch's Mapping API'. A table below shows the fields (28) and scripted fields (0). The table has columns for name, type, format, analyzed, indexed, and controls. The visible rows are:

name	type	format	analyzed	indexed	controls
_source	_source				
record.metadata.oai_dc:dc.xmlns:xsi	string			✓	
record.metadata.oai_dc:dc.xmlns:dc	string			✓	
_type	string			✓	
record.metadata.oai_dc:dc.date.content	string			✓	
record.metadata.oai_dc:dc.relation.content	string			✓	

# Kibana Discover / in nicht-analysierten Feldern muß Suche genau sein

record.metadata.oai\_dc:dc:dc:creator.content:"Wolf, Sebastian"

pub.uni-bielefeld.d... 131 hits

Selected Fields

Available Fields

Popular

```
record.metadata.oai_dc:dc:dc:creator
{
  "content": "Summann, Friedrich"
},
{
  "content": "Wolf, Sebastian"
}
```

record.metadata.oai\_dc:dc:dc:creator.content:"Wolf"

pub.uni-bielefeld.d... 0 hits

No results found 😞

record.metadata.oai\_dc:dc:dc:creator.content:"Wolf, sebastian"

pub.uni-bielefeld.d... 0 hits

No results found 😞

# Discover / gesuchte Werte lassen sich verschieden beschreiben

## Suchoptionen:

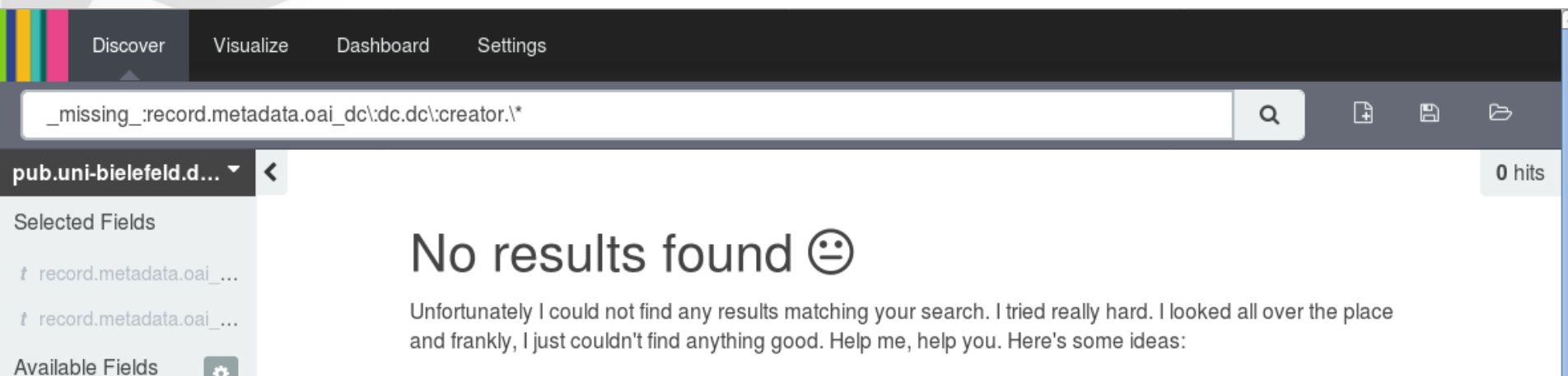
<b>Term:</b>	<b>hun</b>	<b>(ungarisch)</b>
<b>Phrase:</b>	<b>"..."</b>	
<b>Wildcard:</b>	<b>urn\:issn\:*</b>	
<b>regulärer Ausdruck:</b>	<b>/(ger ita)/</b>	<b>(deutsch oder italienisch)</b>



# Kibana Discover / Suchen lassen sich auch komplexer gestalten

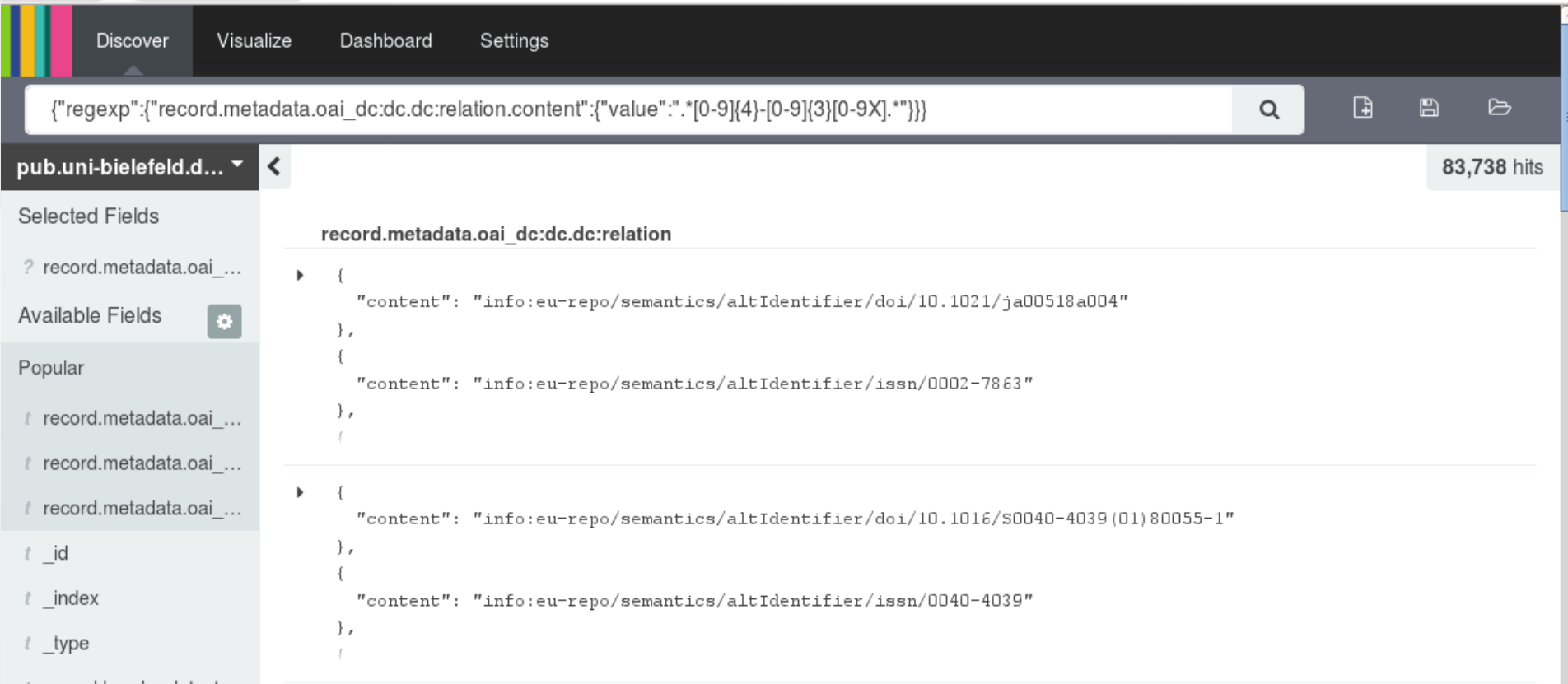
bestimmtes Feld:	<code>"dc:title":..., dc\:title:...</code>
fehlende Felder:	<code>_missing_:(dc\:creator OR dc\:contributor)</code>
existierende Felder:	<code>_exists_:"dc:creator" or _exists_:"dc:contributor"</code>
logischer Ausdruck:	AND, OR, NOT
JSON:	<code>{"wildcard":{"record.header.setSpec":"j?urnal:?"}}</code>

# Kibana Discover / fehlen wichtige Felder bzw. Feldwerte?



The screenshot shows the Kibana Discover interface. The search bar contains the query: `_missing_:record.metadata.oai_dc\dc.dc\creator.*`. The search results area displays "0 hits" and a message: "No results found 😞". Below the message, it says: "Unfortunately I could not find any results matching your search. I tried really hard. I looked all over the place and frankly, I just couldn't find anything good. Help me, help you. Here's some ideas:". The left sidebar shows "Selected Fields" with two entries: `record.metadata.oai_...` and "Available Fields" with a settings icon.

# Kibana Discover / sind z.B. ISSNs gegeben?



The screenshot shows the Kibana Discover interface. The top navigation bar includes 'Discover', 'Visualize', 'Dashboard', and 'Settings'. The search bar contains the query: `{"regex":"record.metadata.oai_dc:dc.dc.relation.content":{"value":".*[0-9]{4}-[0-9]{3}[0-9X].*"}]}`. The left sidebar shows 'Selected Fields' and 'Available Fields'. The main content area displays the search results for the query, showing a list of records with their content values.

pub.uni-bielefeld.d... 83,738 hits

Selected Fields

Available Fields

Popular

record.metadata.oai\_dc:dc.dc.relation

```
{
  "content": "info:eu-repo/semantics/altIdentifier/doi/10.1021/ja00518a004"
},
{
  "content": "info:eu-repo/semantics/altIdentifier/issn/0002-7863"
},
{
  "content": "info:eu-repo/semantics/altIdentifier/doi/10.1016/S0040-4039(01)80055-1"
},
{
  "content": "info:eu-repo/semantics/altIdentifier/issn/0040-4039"
}
```

# Kibana Visualize / Kibana bietet diverse Diagramme/Graphen

## Create a new visualization

Step 1



### Area chart

Great for stacked timelines in which the total of all series is more important than comparing any two or more series. Less useful for assessing the relative change of unrelated data points as changes in a series lower down the stack will have a difficult to gauge effect on the series above it.



### Data table

The data table provides a detailed breakdown, in tabular format, of the results of a composed aggregation. Tip, a data table is available from many other charts by clicking grey bar at the bottom of the chart.



### Line chart

Often the best chart for high density time series. Great for comparing one series to another. Be careful with sparse sets as the connection between points can be misleading.



### Markdown widget

Useful for displaying explanations or instructions for dashboards.



### Metric

One big number for all of your one big number needs. Perfect for show a count of hits, or the exact average a numeric field.



### Pie chart

Pie charts are ideal for displaying the parts of some whole. For example, sales percentages by department. Pro Tip: Pie charts are best used sparingly, and with no more than 7 slices per pie.



### Tile map

Your source for geographic maps. Requires an elasticsearch geo\_point field. More specifically, a field that is mapped as type:geo\_point with latitude and longitude coordinates.



### Vertical bar chart

The goto chart for oh-so-many needs. Great for time and non-time data. Stacked or grouped, exact numbers or percentages. If you are not sure which chart your need, you could do worse than to start here.

# Kibana Visualize / Bucket Aggregation zeigt Feldwerte im Überblick

Discover Visualize Dashboard Settings

pub.uni-bielefeld.de-oai\_oai\_dc\_160712

Data Options

### metrics

Metric Count

+ Add metrics

### buckets

Split Rows

Aggregation

Terms

Field

record.metadata.oai\_dc:dc:language.content

Order Top Size 5

Order By metric: Count

Advanced

+ Add sub-buckets

Top 5 record.metadata.oai_dc:dc:language.content	Count
eng	114,923
deu	84,539
spa	616
fra	572
hun	264

Export: [Raw](#) [Formatted](#)

# Kibana Visualize / welche Publikationstypen kommen vor?

Discover Visualize Dashboard Settings

pub.uni-bielefeld.de-oai\_oai\_dc\_160712

Data Options

**metrics**

Metric Count

+ Add metrics

**buckets**

Split Rows

Aggregation

Terms

Field

record.metadata.oai\_dc:dc.dc:type.content

Order Size

Top 100

Order By

metric: Count

Advanced



+ Add sub-buckets

Top 100 record.metadata.oai\_dc:dc.dc:type.content

	Count
text	201,046
doc-type:article	81,636
info:eu-repo/semantics/article	81,636
doc-type:doctoralThesis	39,104
info:eu-repo/semantics/doctoralThesis	39,104
doc-type:bookPart	28,358
info:eu-repo/semantics/bookPart	28,358
doc-type:conferenceObject	21,923
info:eu-repo/semantics/conferenceObject	21,923
doc-type:workingPaper	16,607
info:eu-repo/semantics/workingPaper	16,607
doc-type:other	4,077
doc-type:book	3,342
info:eu-repo/semantics/book	3,342
info:eu-repo/semantics/conferenceAbstract	3,201
doc-type:report	3,076
info:eu-repo/semantics/report	3,076


# Kibana Visualize / welche Publikationstypen weichen vom Vokabular ab?

pub.uni-bielefeld.de-oai\_oai\_dc\_160712


Data Options  

### metrics

Metric Count



### buckets

Split Rows 

Aggregation

Terms

Field

record.metadata.oai\_dc:dc.dc:type.content

Order Top Size 100

Order By metric: Count

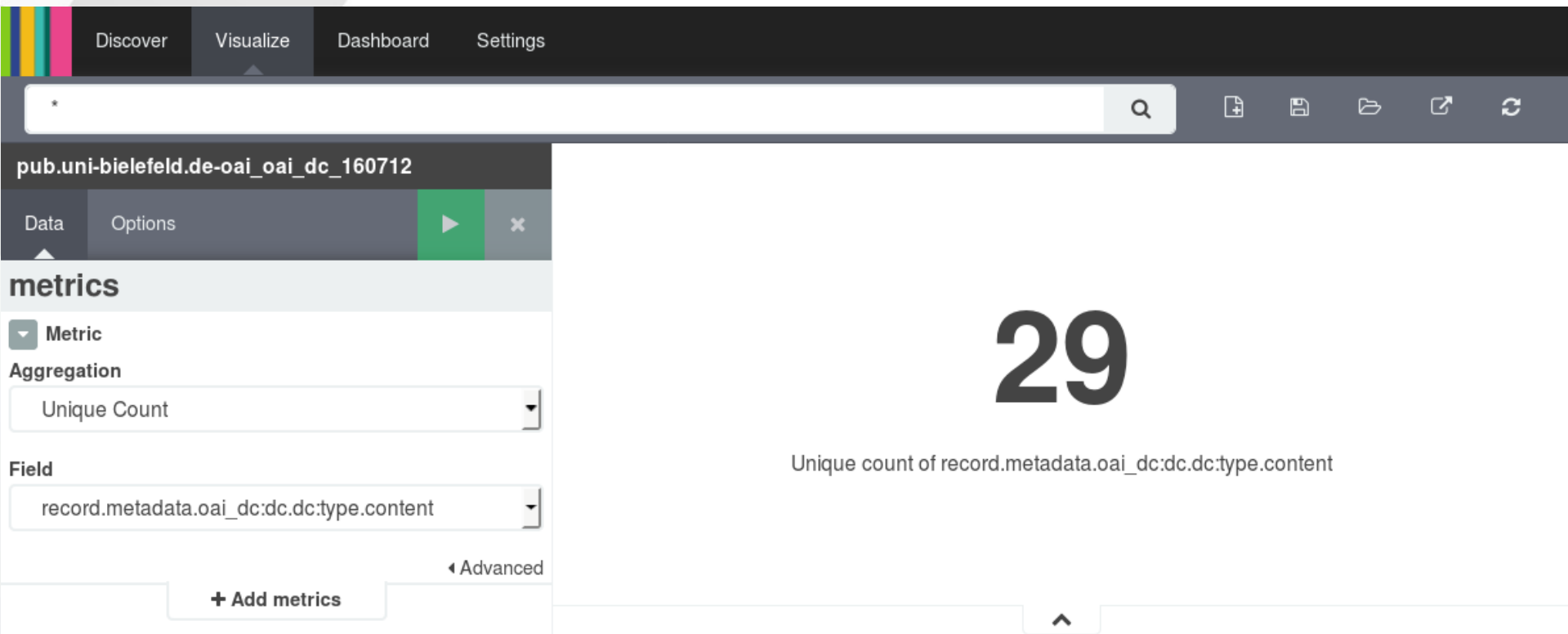
Exclude Pattern

info\eu-repo\semanticsV.\*

### Top 100 record.metadata.oai\_dc:dc.dc:type.content

	Count
text	201,046
doc-type:article	81,636
doc-type:doctoralThesis	39,104
doc-type:bookPart	28,358
doc-type:conferenceObject	21,923
doc-type:workingPaper	16,607
doc-type:other	4,077
doc-type:book	3,342
doc-type:report	3,076
doc-type:masterThesis	1,384
doc-type:contributionToPeriodical	748
doc-type:preprint	506
doc-type:bachelorThesis	285

# Kibana Visualize / wieviele Publikationstypen kommen vor?



The screenshot shows the Kibana Visualize interface. The top navigation bar includes 'Discover', 'Visualize', 'Dashboard', and 'Settings'. The 'Visualize' tab is active. The main visualization area displays a large number '29' representing the unique count of record.metadata.oai\_dc:dc.dc:type.content. The left sidebar shows the configuration for this visualization, including the 'metrics' section with 'Unique Count' aggregation and the field 'record.metadata.oai\_dc:dc.dc:type.content'. The URL in the address bar is 'pub.uni-bielefeld.de-oai\_oai\_dc\_160712'.

pub.uni-bielefeld.de-oai\_oai\_dc\_160712

Data Options

**metrics**

Metric

Aggregation

Unique Count

Field

record.metadata.oai\_dc:dc.dc:type.content

Advanced

+ Add metrics

29

Unique count of record.metadata.oai\_dc:dc.dc:type.content



# Fazit / zur Analyse von Metadaten ist Kibana sehr nützlich!

gewöhnungsbedürftig ist (noch):

- Indexierung komplex, via JSON
  - Komplexität erschwert Bedienung
  - GUI reagiert zuweilen etwas störrisch
  - reguläre Ausdrücke undurchsichtig
  - mehrfache Felder schlechter abfragbar
- selten nötig
  - erlaubt aber auch viel
  - JSON-Queries nutzen
  - Indexierung anpassen

```
? record.metadata.oai_dc:dc.dc:creator {
  "content": "Kaufmann, Franz-Xaver"
}
"content": "Rendtorff, Trutz"
}
```

Objects in arrays are not well supported.

## Kibana

- **Produkt:** <https://www.elastic.co/products/kibana>
- **Tutorial:** <https://www.elastic.co/guide/en/kibana/current/getting-started.html>
- **Tutorial:** <https://www.timroes.de/2015/02/07/kibana-4-tutorial-part-1-introduction/>

## OpenAIRE

- **Portal:** <https://www.openaire.eu>
- **Suche:** <https://www.openaire.eu/search/find?keyword=>