



High quality, low maintenance content tagging @ ZEIT Online

Breno Faria, Christoph Goller

IntraFind Software AG

- ▶ Elasticsearch Partner (we also do consulting)
- ▶ Specialist for Information Retrieval and Text Analytics
- ▶ Founded 2000, 30 employees
- ▶ More than 850 customers mainly in Germany, Austria, and Switzerland
- ▶ **Lucene Committers:** B. Messer, C. Goller
- ▶ Independent Software Vendor, entirely self-financed
- ▶ **Products are a combination of Open Source Components and in-house Development**
 - ▶ High quality Linguistic Analyzers for most European Languages (**also available as Solr and Elasticsearch plugins**)
 - ▶ Named Entity Recognition
 - ▶ Text Classification
 - ▶ Tagging Service – extraction of semantic meta data

1. The ZEIT Online Project 2010 → tagging and making the archive searchable
2. Editorial Workflow @ ZEIT Online
3. Feedback from the Editors
4. Meeting the Expectations

- ▶ Die ZEIT is a weekly newspaper founded 1946, one of the most renowned in Germany
- ▶ ZEIT Online, the web edition, exists since 1996

- ▶ Die ZEIT is a weekly newspaper founded 1946, one of the most renowned in Germany
- ▶ ZEIT Online, the web edition, exists since 1996
- ▶ 2010 → organize entire archive based on semantic meta data and make it searchable

- ▶ Die ZEIT is a weekly newspaper founded 1946, one of the most renowned in Germany
- ▶ ZEIT Online, the web edition, exists since 1996
- ▶ 2010 → organize entire archive based on semantic meta data and make it searchable
 - ▶ Persons, locations and organizations mentioned

Chinas Präsident Xi Jinping hat sich für eine neue asiatische Sicherheitsstruktur ohne Einbindung der USA ausgesprochen. In die regionale Kooperation sollten Russland und der Iran miteinbezogen werden, forderte Xi während der Konferenz für Interaktion und Vertrauensbildung in Asien in Shanghai. Zuvor hatte er Russlands Präsident Wladimir Putin empfangen. Der war inmitten der Krise in der Ukraine nach China gereist, um über militärische und energiewirtschaftliche Zusammenarbeit zu beraten. Russland will China 38 Milliarden Kubikmeter Gas pro Jahr liefern – das entspricht derzeit rund einem Viertel des chinesischen Verbrauchs. Putin ist international unter Druck, weil er den wachsenden Einfluss prorussischer Gruppen in der Ostukraine tolerierte und die ukrainische Halbinsel Krim annektierte. EU und USA verhängten Sanktionen. Das Partnerschaftsbekenntnis aus China dürfte ihm gelegen kommen. In der Ostukraine kämpfen ukrainische Polizei und Militär mit den prorussischen Separatisten um die Kontrolle. Die Regionen Donezk und Luhansk hatten sich vorvergangenen Sonntag in international abgelehnten Referenden für unabhängig erklärt. Am Dienstag mobilisierte der ukrainische Milliardär Rinat Achmetow in einer öffentlich übertragenen Ansprache Zehntausende Gegner der Separatisten und rief zum friedlichen Protest auf. Während eines kurzen Warnstreiks rollten

- ▶ Die ZEIT is a weekly newspaper founded 1946, one of the most renowned in Germany
- ▶ ZEIT Online, the web edition, exists since 1996
- ▶ 2010 → organize entire archive based on semantic meta data and make it searchable
 - ▶ Persons, locations and organizations mentioned
 - ▶ Statistically significant keywords

Chinas Präsident Xi Jinping hat sich für eine neue asiatische Sicherheitsstruktur ohne Einbindung der USA ausgesprochen. In die regionale Kooperation sollten Russland und der Iran miteinbezogen werden, forderte Xi während der Konferenz für Interaktion und Vertrauensbildung in Asien in Shanghai. Zuvor hatte er Russlands Präsident Wladimir Putin empfangen. Der war inmitten der Krise in der Ukraine nach China gereist, um über militärische und energiewirtschaftliche Zusammenarbeit zu beraten. Russland will China 38 Milliarden Kubikmeter Gas pro Jahr liefern – das entspricht derzeit rund einem Viertel des chinesischen Verbrauchs. Putin ist international unter Druck, weil er den wachsenden Einfluss prorussischer Gruppen in der Ostukraine tolerierte und die ukrainische Halbinsel Krim annektierte. EU und USA verhängten Sanktionen. Das Partnerschaftsbekenntnis aus China dürfte ihm gelegen kommen. In der Ostukraine kämpfen ukrainische Polizei und Militär mit den prorussischen Separatisten um die Kontrolle. Die Regionen Donezk und Luhansk hatten sich vorvergangenen Sonntag in international abgelehnten Referenden für unabhängig erklärt. Am Dienstag mobilisierte der ukrainische Milliardär Rinat Achmetow in einer öffentlich übertragenen Ansprache Zehntausende Gegner der Separatisten und rief zum friedlichen Protest auf. Während eines kurzen Warnstreiks rollten

- ▶ Die ZEIT is a weekly newspaper founded 1946, one of the most renowned in Germany
- ▶ ZEIT Online, the web edition, exists since 1996
- ▶ 2010 → organize entire archive based on semantic meta data and make it searchable
 - ▶ Persons, locations and organizations mentioned
 - ▶ Statistically significant keywords
 - ▶ Classification into corresponding department

ZEIT  ONLINE

SUCHEN

START POLITIK WIRTSCHAFT GESELLSCHAFT KULTUR WISSEN DIGITAL STUDIUM KARRIERE REISEN MOBILITÄT SPORT HAMBURG **ZEITmagazin**

Start > Schlagworte > A

Anmelden | Registrieren

A | SCHLAGWORTREGISTER

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z #

Aachen
Abacha, Sani
Abbado, Claudio
Abbas, Ferhat
Abbas, Machmud
Abbas, Mahmud
Abchasien
Abdel Fattah al-Sissi

Angela Merkel
Angeln
Angerer, Nadine
Angerer, Tobias
Angkor Wat
Anglizismus
Angola
Animationsfilm

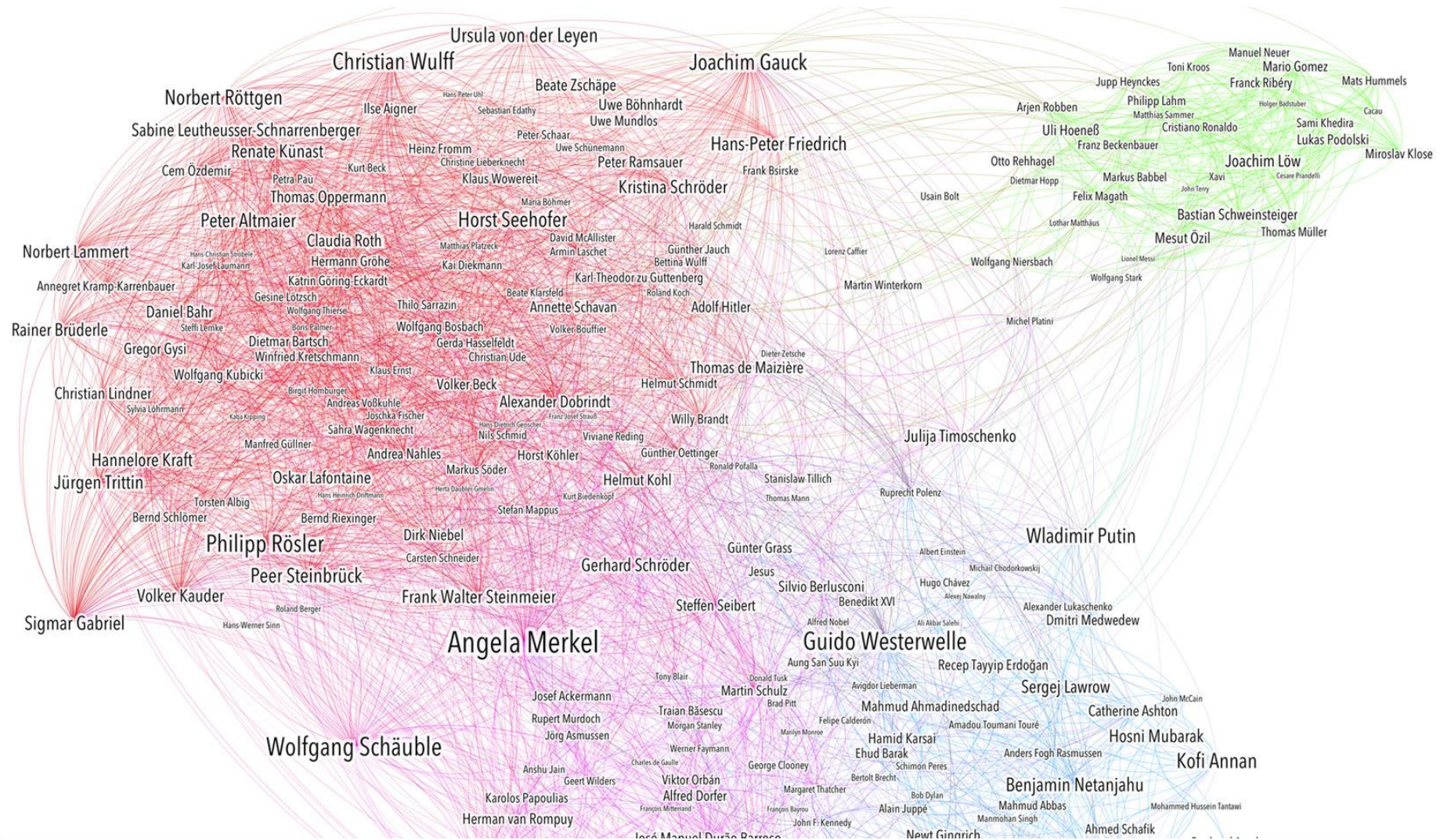
MEISTGELESEN

1. **ARMUT** Hightech gegen Flaschensammler
2. **EUROPAWAHLKAMPF** Wer ist hier der Kriegstreiber?
3. **UKRAINE-LIVE-BLOG** UN zählen schon 10.000 Flüchtlinge in der Ukraine
4. **UKRAINE-KRISE** Joschka weiß auch nicht so recht
5. **APP FLATASTIC** Klickt den Dreck weg!

MEISTKOMMENTIERT

1. **UKRAINE-LIVE-BLOG** UN befürchten Flüchtlingswelle aus der Ostukraine **(405)**
2. **NUTZTIERE** "Für ein Schinkenbrot werden Tiere einaesprert und verstümmelt" **(272)**

Amazingly, there is an API for accessing this tagged content! See developer.zeit.de



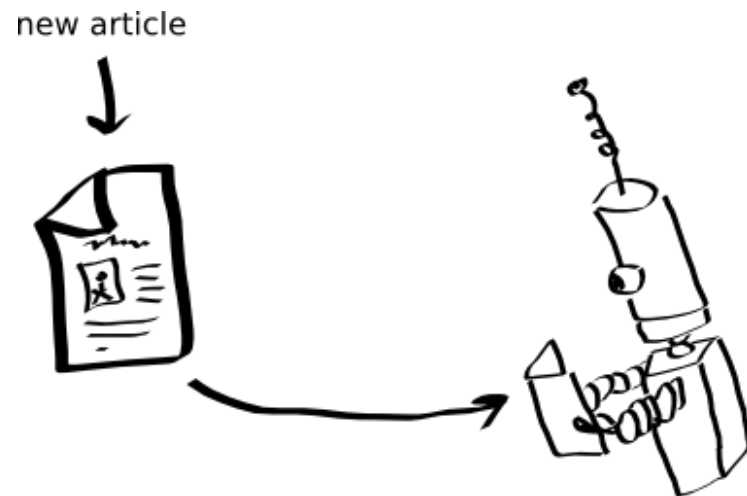
- ▶ Second step in the project was to integrate the content tagging system into the editorial workflow @ ZEIT Online

- ▶ Second step in the project was to integrate the content tagging system into the editorial workflow @ ZEIT Online

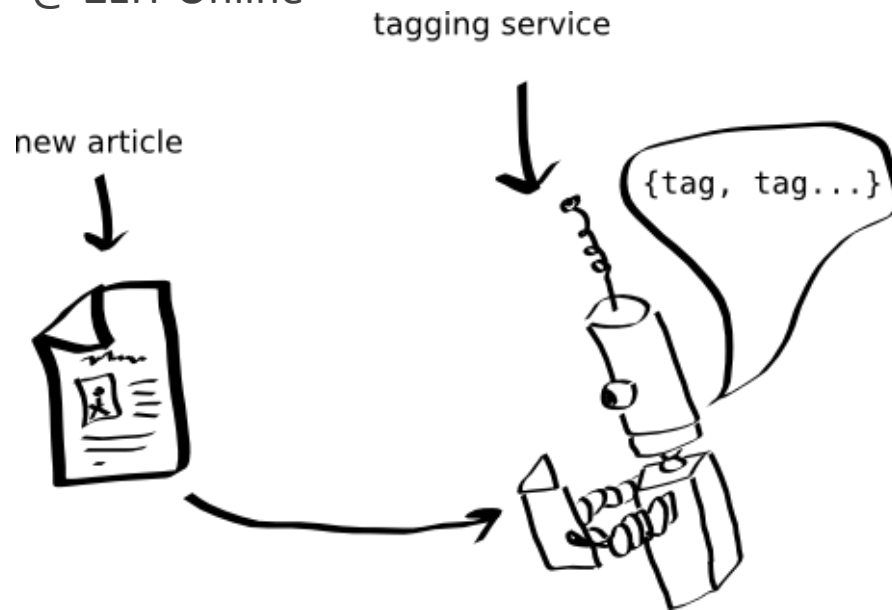
new article



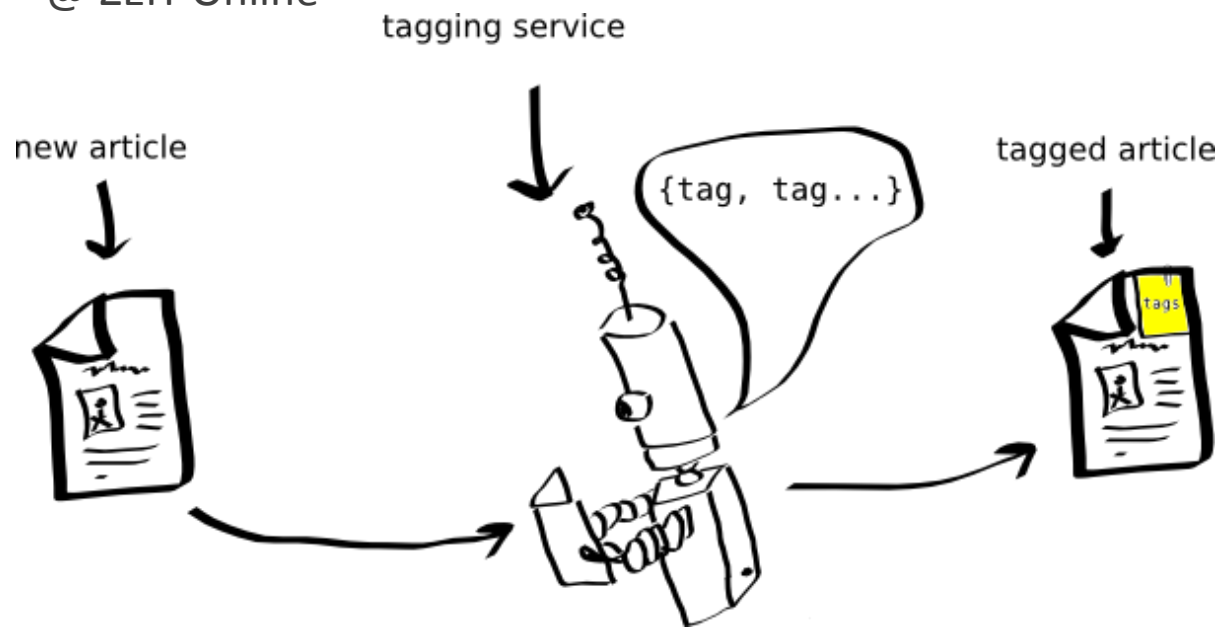
- ▶ Second step in the project was to integrate the content tagging system into the editorial workflow @ ZEIT Online



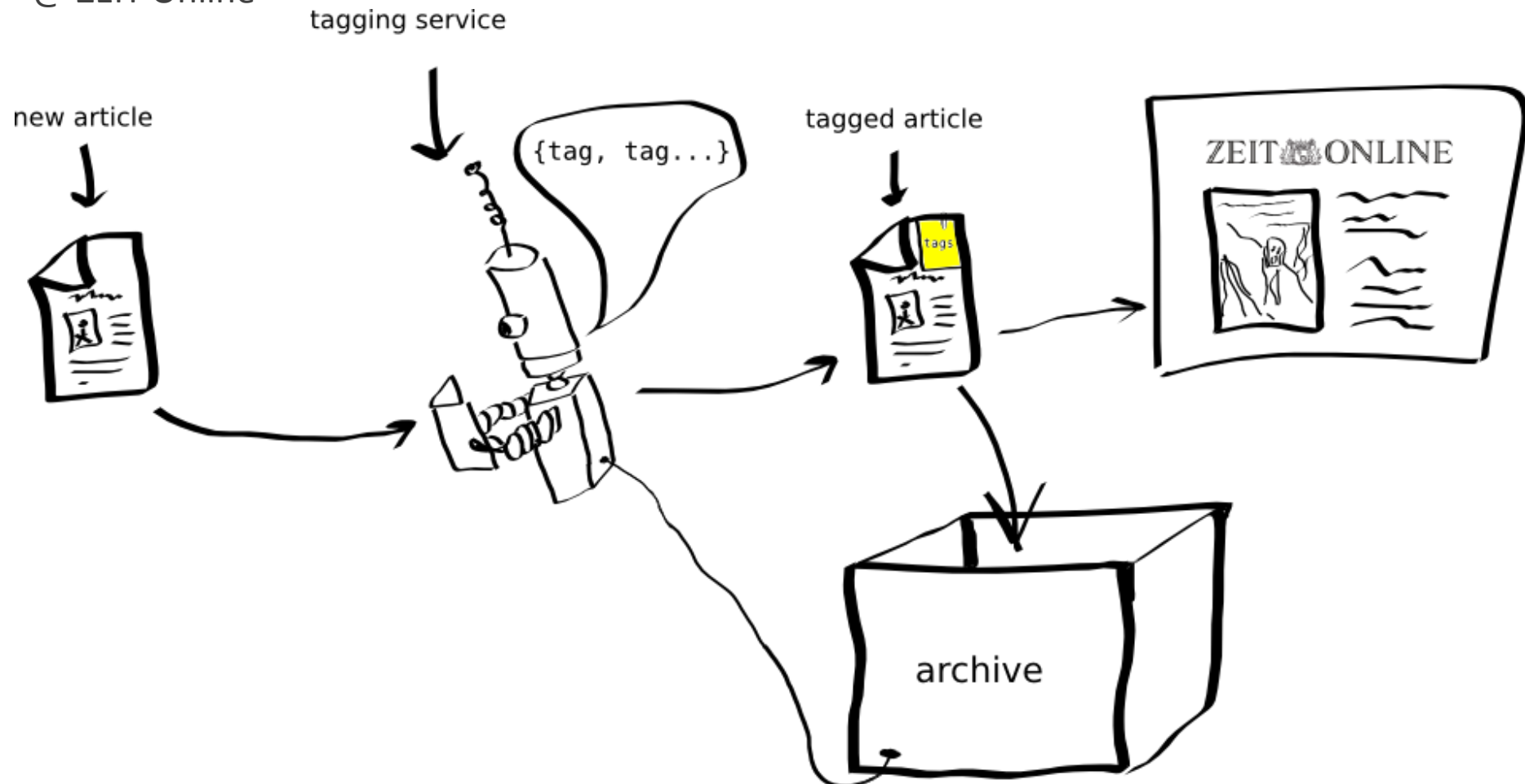
- ▶ Second step in the project was to integrate the content tagging system into the editorial workflow @ ZEIT Online



- ▶ Second step in the project was to integrate the content tagging system into the editorial workflow @ ZEIT Online



- ▶ Second step in the project was to integrate the content tagging system into the editorial workflow @ ZEIT Online



ZEIT ONLINE | UNTERNEHMEN

SUCHEN

START POLITIK **WIRTSCHAFT** GESELLSCHAFT KULTUR WISSEN DIGITAL STUDIUM KARRIERE REISEN MOBILITÄT SPORT HAMBURG **ZEITmagazin**

Start > Wirtschaft > Unternehmen > Schweiz: Credit Suisse zahlt Milliardenstrafe wegen Steuerbetrugs Anmelden | Registrieren

SCHWEIZ

Credit Suisse zahlt Milliardenstrafe wegen Steuerbetrugs

Die Schweizer Bank hat wegen Beihilfe zur Steuerhinterziehung in den USA eine Rekordstrafe akzeptiert und zeigt sich reuevoll. Die Justiz beklagt mangelnde Kooperation.

20. Mai 2014 07:28 Uhr 4 Kommentare |

QUELLE ZEIT ONLINE, dpa, tst

SCHLAGWORTE Credit Suisse | Steuerhinterziehung | Schweiz | Bankgeheimnis | Ermittlung | Justizminister

NEU AUF ZEIT ONLINE

1. **THAILAND** Kabinett verabredet Geheimgespräch zur Krise
2. **UKRAINE-LIVE-BLOG** Merkel will Beziehung zu Russland verbessern
3. **UKRAINE-KRISE** Joschka weiß auch nicht so recht
4. **USA** China wehrt sich gegen Cyberspionage-Klage
5. **KOALITION** Wirtschaft gibt sich mit Rentenkompromiss nicht zufrieden

NEU IM RESSORT

1. **SCHWEIZ** Credit Suisse zahlt Milliardenstrafe wegen Steuerbetrugs
2. **INTERNATIONALER WÄHRUNGSFONDS** Besser Brücken bauen als Mütterrente zahlen
3. **GASSTREIT** Ukraine, der unwichtige Freund
4. **EUROPÄISCHE UNION** EU-Mythen im Check

It's not as simple as that

- ▶ Keywords will be visible to humans! → you cannot rely on a robot's good judgement and publish everything that comes out...

It's not as simple as that

- ▶ Keywords will be visible to humans! → you cannot rely on a robot's good judgement and publish everything that comes out...
- ▶ Ever heard of "inter-indexer consistency"? → it probably wouldn't work letting every editor choose freely

It's not as simple as that

- ▶ Keywords will be visible to humans! → you cannot rely on a robot's good judgement and publish everything that comes out...
- ▶ Ever heard of "inter-indexer consistency"? → it probably wouldn't work letting every editor choose freely

Solution:

- ▶ curated list of allowed keywords
- ▶ AND editor picks a subset of allowed keywords for the article

It's not as simple as that

- ▶ Keywords will be visible to humans! → you cannot rely on a robot's good judgement and publish everything that comes out...
- ▶ Ever heard of "inter-indexer consistency"? → it probably wouldn't work letting every editor choose freely

Solution:

- ▶ curated list of allowed keywords
- ▶ AND editor picks a subset of allowed keywords for the article

Curating the keyword list is expensive

... going through large lists of keyword candidates also

It's not as simple as that

- ▶ Keywords will be visible to humans! → you cannot rely on a robot's good judgement and publish everything that comes out...
- ▶ Ever heard of "inter-indexer consistency"? → it probably wouldn't work letting every editor choose freely

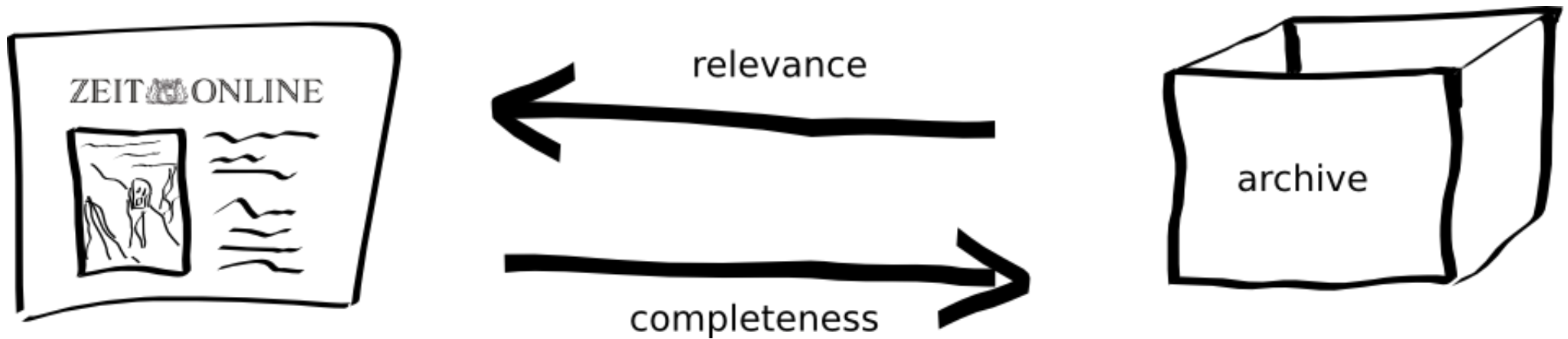
Solution:

- ▶ curated list of allowed keywords
- ▶ AND editor picks a subset of allowed keywords for the article

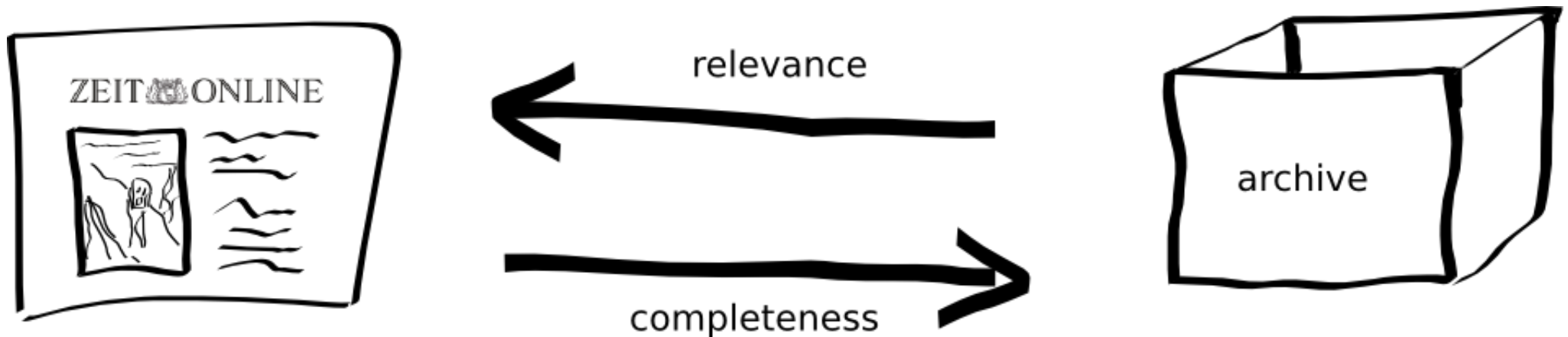
Curating the keyword list is expensive

... going through large lists of keyword candidates also ← **we want to solve this problem**

- ▶ Tradeoff: **relevance** vs. **completeness**

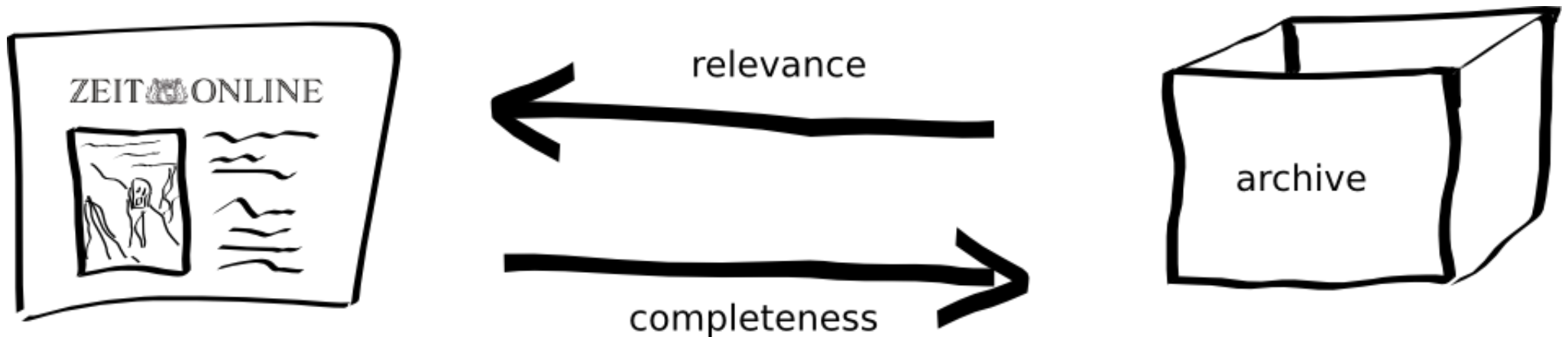


- ▶ Tradeoff: **relevance** vs. **completeness**



generic better than specific (*Stuxnet* vs. *Stuxnet-Virus*)
expand to similar keywords (*Prism* → *NSA*)
no 'stop-keywords' (e.g. *Angela Merkel*)
no out-of-context keywords
consider trends!

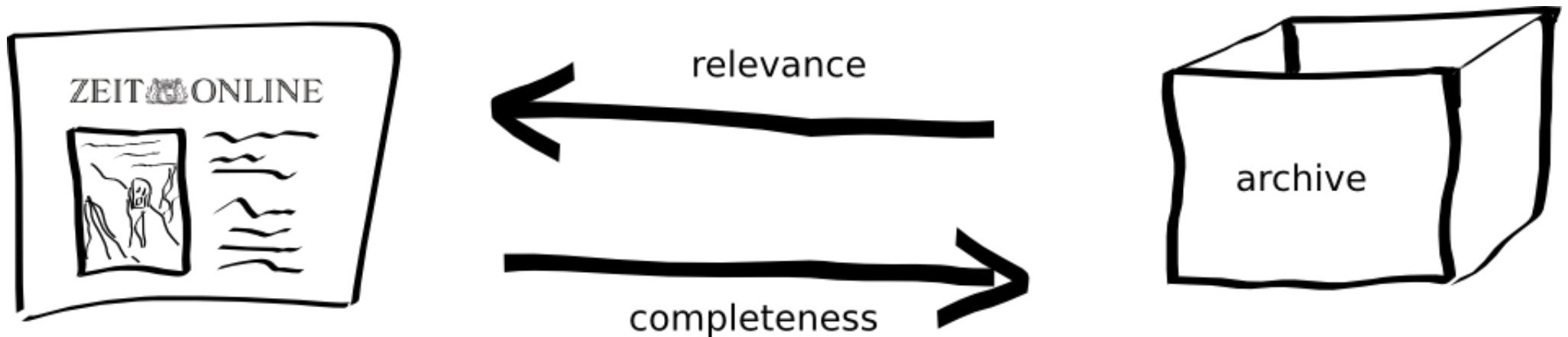
- ▶ Tradeoff: **relevance** vs. **completeness**



generic better than specific (*Stuxnet* vs. *Stuxnet-Virus*)
expand to similar keywords (*Prism* → *NSA*)
no 'stop-keywords' (e.g. *Angela Merkel*)
no out-of-context keywords
consider trends!

all possible keywords, don't miss anything!

- ▶ Tradeoff: **relevance** vs. **completeness**



generic better than specific (*Stuxnet* vs. *Stuxnet-Virus*)
expand to similar keywords (*Prism* → *NSA*)
no 'stop-keywords' (e.g. *Angela Merkel*)
no out-of-context keywords
consider trends!

all possible keywords, don't miss anything!

Oh, and please don't make us work more with your changes.

Provide a perfect ranking of keywords

Provide a perfect ranking of keywords

- ▶ This allows us to present only the relevant keywords to the editor

Provide a perfect ranking of keywords

- ▶ This allows us to present only the relevant keywords to the editor
- ▶ ... and we still have all possible keywords for the archive

Meeting the Expectations Baseline Scoring

First problem: how do we compare apples and bananas? (different sorts of entities and keywords)

Meeting the Expectations

Baseline Scoring

First problem: how do we compare apples and bananas? (different sorts of entities and keywords)

- ▶ We will compute the document hit count in the archive by searching for each tag found

Meeting the Expectations

Baseline Scoring

First problem: how do we compare apples and bananas? (different sorts of entities and keywords)

- ▶ We will compute the document hit count in the archive by searching for each tag found
- ▶ We can rely on our linguistic analyzers to account for different forms of the same tag: e.g. „Bundeswirtschaftsminister“ == „Bundesminister für Wirtschaft“

Meeting the Expectations

Baseline Scoring

First problem: how do we compare apples and bananas? (different sorts of entities and keywords)

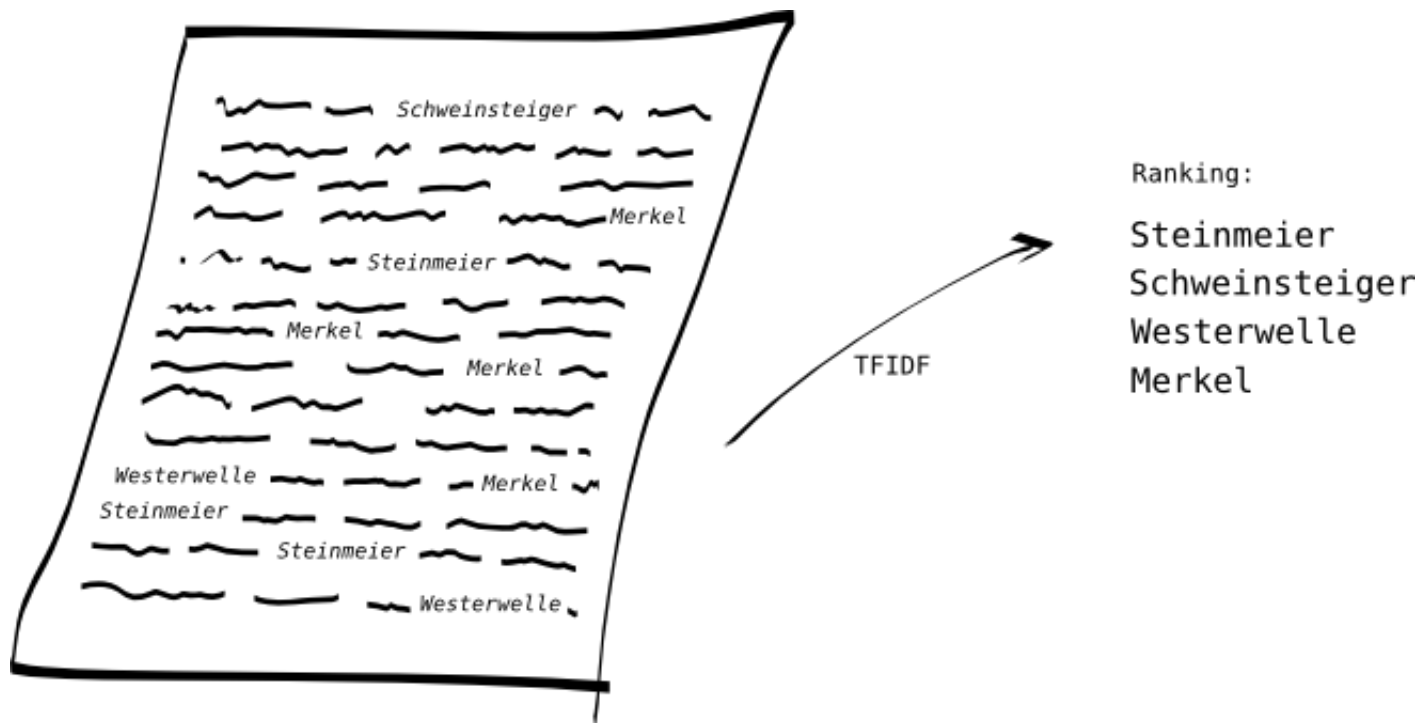
- ▶ We will compute the document hit count in the archive by searching for each tag found
- ▶ We can rely on our linguistic analyzers to account for different forms of the same tag: e.g. „Bundeswirtschaftsminister“ == „Bundesminister für Wirtschaft“
- ▶ Use a Lucene Similarity to compute the TFIDF of each tag

Meeting the Expectations

Baseline Scoring

First problem: how do we compare apples and bananas? (different sorts of entities and keywords)

- ▶ We will compute the document hit count in the archive by searching for each tag found
- ▶ We can rely on our linguistic analyzers to account for different forms of the same tag: e.g. „Bundeswirtschaftsminister“ == „Bundesminister für Wirtschaft“
- ▶ Use a Lucene Similarity to compute the TFIDF of each tag

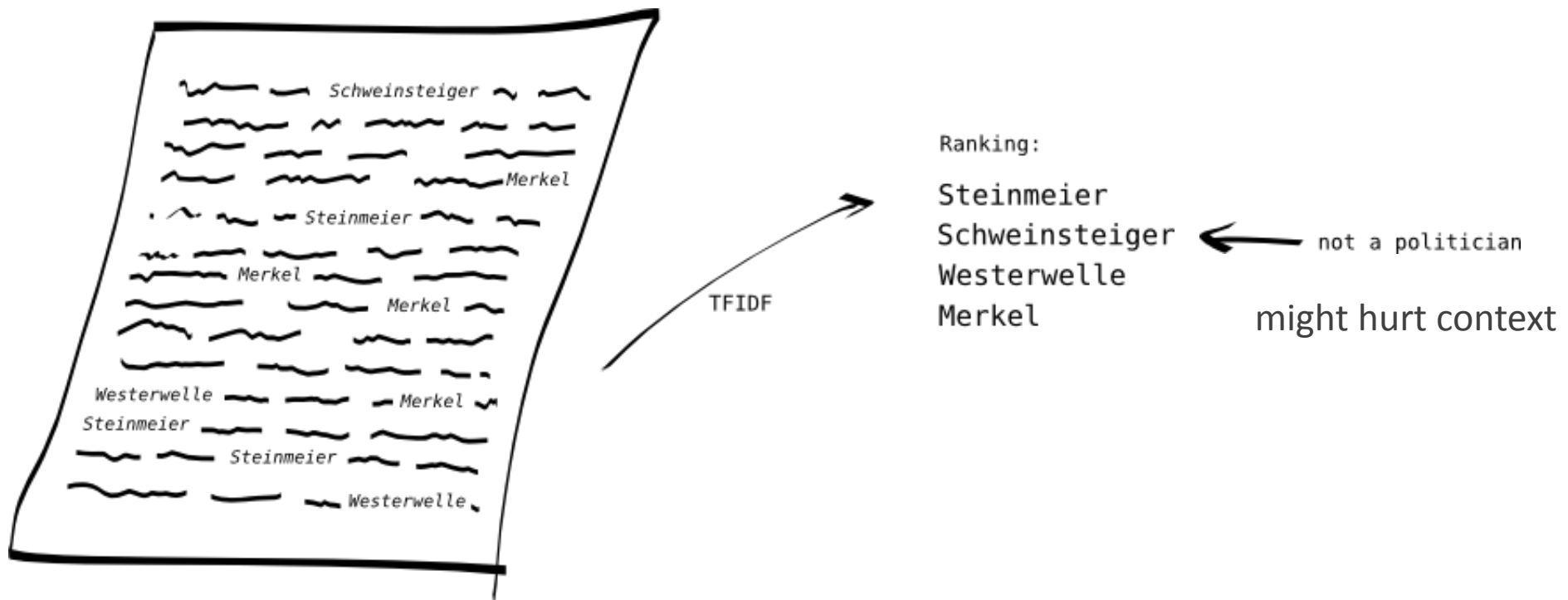


Meeting the Expectations

Baseline Scoring

First problem: how do we compare apples and bananas? (different sorts of entities and keywords)

- ▶ We will compute the document hit count in the archive by searching for each tag found
- ▶ We can rely on our linguistic analyzers to account for different forms of the same tag: e.g. „Bundeswirtschaftsminister“ == „Bundesminister für Wirtschaft“
- ▶ Use a Lucene Similarity to compute the TFIDF of each tag



Meeting the Expectations

Context Scoring

Idea: compare the document with other documents containing a particular tag

Meeting the Expectations

Context Scoring

Idea: compare the document with other documents containing a particular tag

- ▶ compute typical contexts of tag

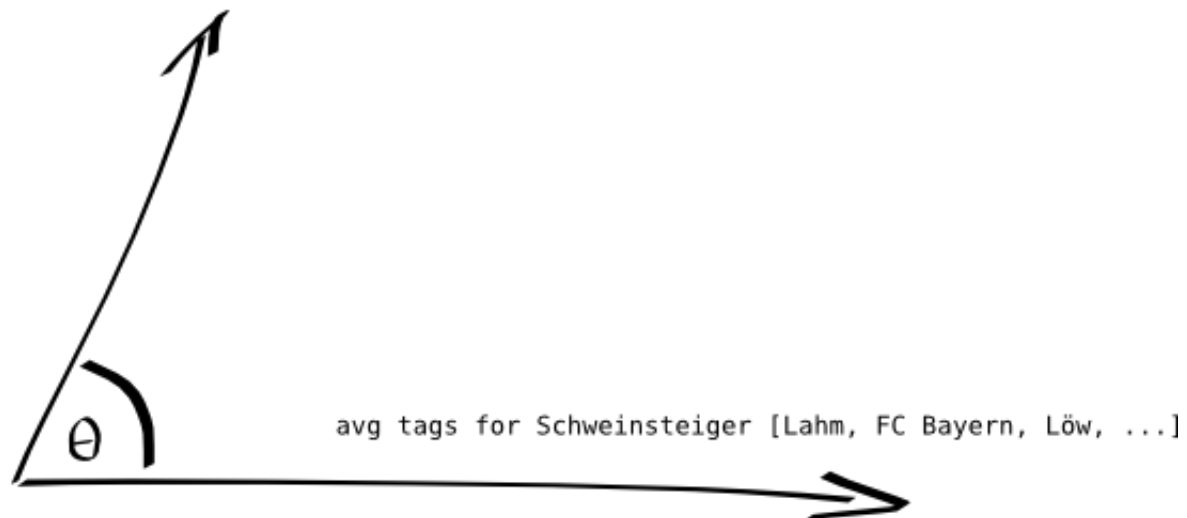
Meeting the Expectations

Context Scoring

Idea: compare the document with other documents containing a particular tag

- ▶ compute typical contexts of tag
- ▶ these contexts are a kind of prototypical document for all documents containing the keyword

found tags [Schweinsteiger, Steinmeier, Merkel, Westerwelle, ...]



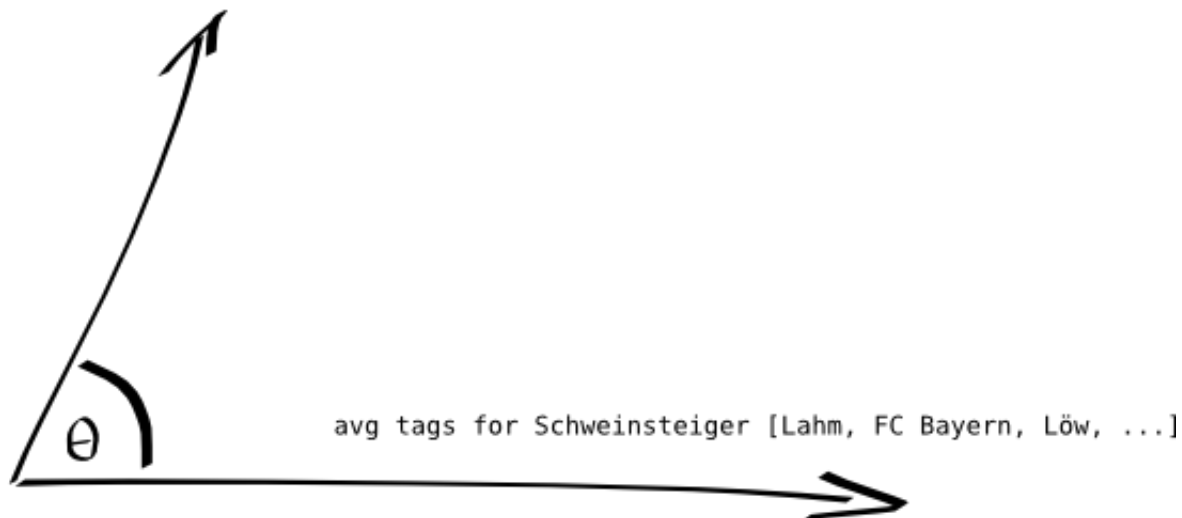
Meeting the Expectations

Context Scoring

Idea: compare the document with other documents containing a particular tag

- ▶ compute typical contexts of tag
- ▶ these contexts are a kind of prototypical document for all documents containing the keyword
- ▶ we compare the current context with this prototypical context, i.e. we compute a similarity

found tags [Schweinsteiger, Steinmeier, Merkel, Westerwelle, ...]



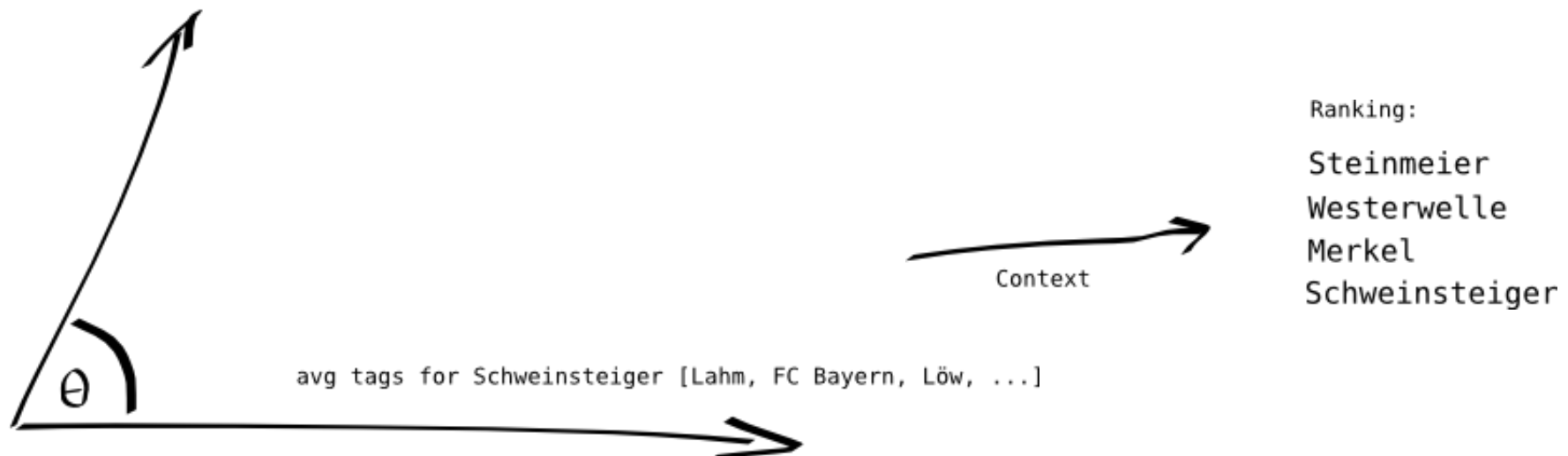
Meeting the Expectations

Context Scoring

Idea: compare the document with other documents containing a particular tag

- ▶ compute typical contexts of tag
- ▶ these contexts are a kind of prototypical document for all documents containing the keyword
- ▶ we compare the current context with this prototypical context, i.e. we compute a similarity

found tags [Schweinsteiger, Steinmeier, Merkel, Westerwelle, ...]



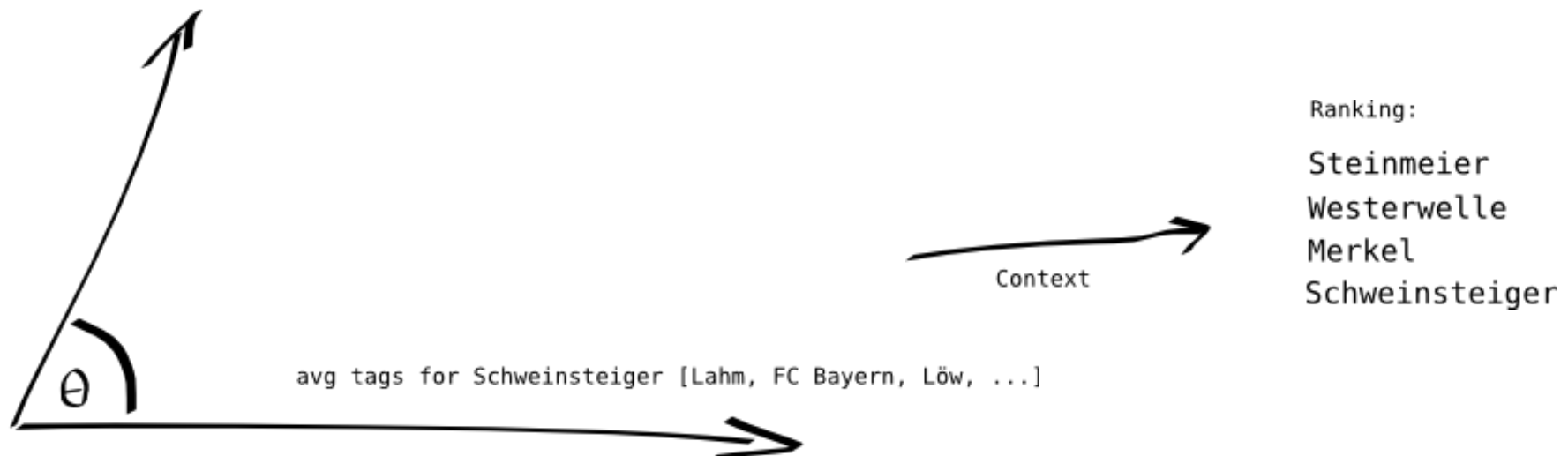
Meeting the Expectations

Context Scoring

Idea: compare the document with other documents containing a particular tag

- ▶ compute typical contexts of tag
- ▶ these contexts are a kind of prototypical document for all documents containing the keyword
- ▶ we compare the current context with this prototypical context, i.e. we compute a similarity

found tags [Schweinsteiger, Steinmeier, Merkel, Westerwelle, ...]



We can use the same method to expand our tags with related keywords!

Meeting the Expectations Trend Scoring

But what if the mention of "Schweinsteiger" is not incidental? Maybe it's world cup time?

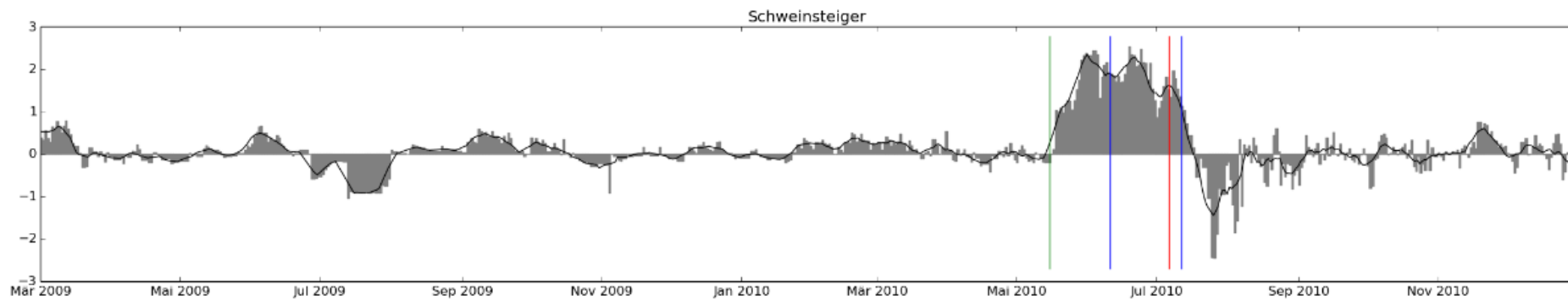
But what if the mention of "Schweinsteiger" is not incidental? Maybe it's world cup time?

- ▶ In our case, trend is a measure of variation of hit counts in a timespan
- ▶ We can compute trends from our archive, by counting hits in different timespans

Meeting the Expectations Trend Scoring

But what if the mention of "Schweinsteiger" is not incidental? Maybe it's world cup time?

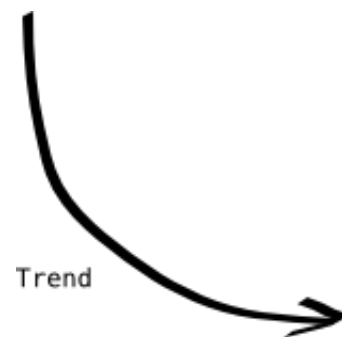
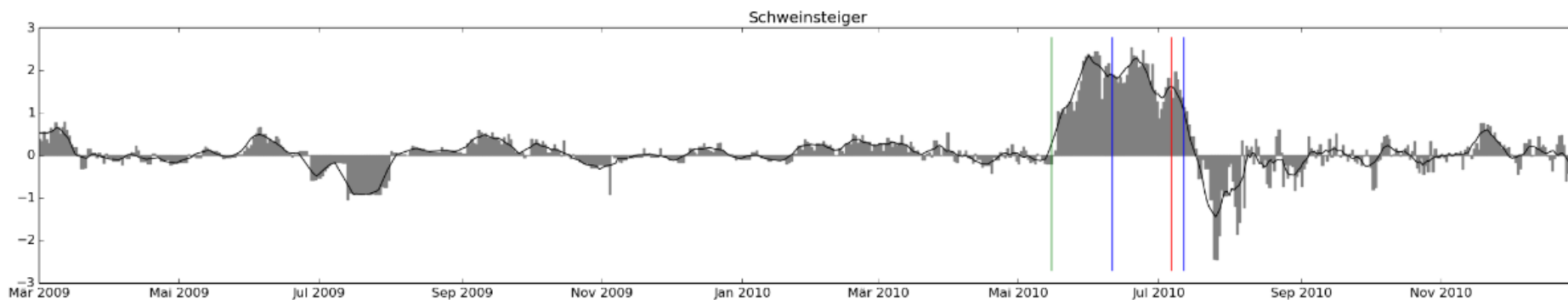
- ▶ In our case, trend is a measure of variation of hit counts in a timespan
- ▶ We can compute trends from our archive, by counting hits in different timespans



Meeting the Expectations Trend Scoring

But what if the mention of "Schweinsteiger" is not incidental? Maybe it's world cup time?

- ▶ In our case, trend is a measure of variation of hit counts in a timespan
- ▶ We can compute trends from our archive, by counting hits in different timespans



Ranking:

Schweinsteiger
Steinmeier
Westerwelle
Merkel

Meeting the Expectations Consolidating Scores



Meeting the Expectations Consolidating Scores



We combine the scores by

1. Individually scaling them onto the same interval

Meeting the Expectations Consolidating Scores



We combine the scores by

1. Individually scaling them onto the same interval
2. Multiplying each one by a weight

Meeting the Expectations Consolidating Scores



We combine the scores by

1. Individually scaling them onto the same interval
2. Multiplying each one by a weight
3. Summing up and again scaling the result

Meeting the Expectations Consolidating Scores



We combine the scores by

1. Individually scaling them onto the same interval
2. Multiplying each one by a weight
3. Summing up and again scaling the result

There's a lot to configure, and there is no such thing as the perfect configuration

Meeting the Expectations Consolidating Scores



We combine the scores by

1. Individually scaling them onto the same interval
2. Multiplying each one by a weight
3. Summing up and again scaling the result

There's a lot to configure, and there is no such thing as the perfect configuration

ZEIT Online has the freedom to fine-tune the ranking

- ▶ Requirements of an editorial office on a tagging system are complex
- ▶ Tradeoff between relevance and completeness of tags
- ▶ You need both. We can solve this problem the same way information retrieval systems have → ranking
- ▶ There is a lot one can do to enrich tags only by looking at a representative archive

Thanks for Listening

Thanks to Ron Drongowski and the ZEIT Online team!

Breno Faria (@brealbfar) & Christoph Goller (@ChGoller)

Phone: +49 89 3090446-0

Fax: +49 89 3090446-29

Email: {christoph.goller,breno.faria}@intrafind.de

Web: www.intrafind.de

IntraFind Software AG

Landsberger Straße 368

80687 München

Germany

The persons graph and most screen-shots are copyright material of ZEIT Online.

