

*Building search app
for public mailing lists in*

15 *minutes
with Elasticsearch*

Agenda

- 1. Who & Why*
- 2. Searching mailing lists*
- 3. How to do it*
- 4. Implementation details*
- 5. Challenges*
- 6. Give aways (free stuff inside!)*

Who am I?

Lukáš Vlček

*Senior Software Engineer
JBoss Community team
Red Hat, Czech Republic*



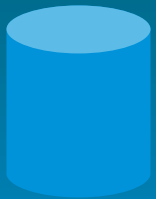
Why I am here?

*My mission is to **improve search**
for **JBoss.org** content.*

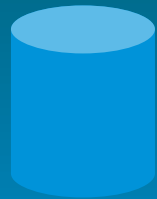
Why I am here?

*My mission is to **improve search**
for JBoss.org content.*

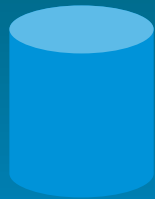
JBoss Community



Mailing lists



Blogs



Documentation



Forums



Project pages



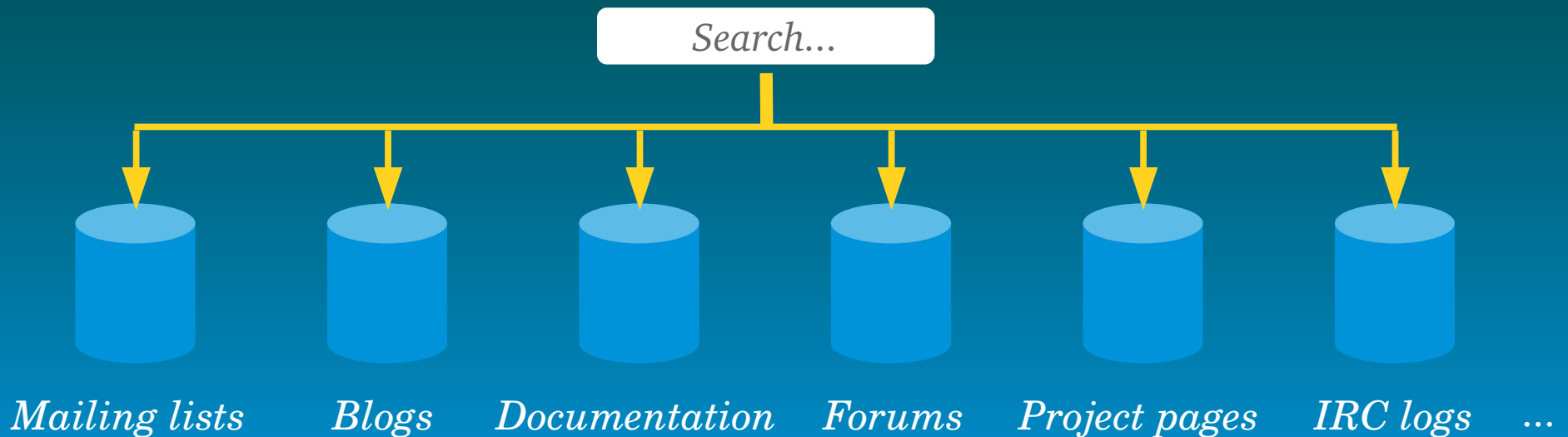
IRC logs

...

Why I am here?

*My mission is to **improve search** for JBoss.org content (and make it really rock!).*

JBoss Community



And we are starting with...



mailing lists
of JBoss community projects (†)

(†) <https://lists.jboss.org/mailman/listinfo/>

It is show time!

Searching mailing lists

Starring:

Front end side:

jQuery

Sammy.js & {{ mustache }}

Highcharts

Isotope

Back end side:

Elastic Search

▼ Projects

Sort results by: **Relevancy**

Instant Search

Order by: name | frequency

Rules (1968)	Gatein (361)	DNA (100)	JBoss OSGI (18)	JBoss I10n (1)
Infinispan (1139)	Netty (311)	ESB (63)	JSFUnit (11)	Guvnor (0)
JBPM (861)	JBoss WS (310)	mod_cluster (56)	JBoss TS (6)	JBossWS Metro (0)
Weld (732)	JBoss AS7 (289)	JOPR (44)	JBossWS CXF (3)	RichFaces (0)
Hibernate (671)	JBoss Tools (157)	Wise (39)	ModeShape (2)	Scribbling (0)
Seam (425)	Teiid (157)	JBoss Cluster (26)	EmbJOPR (1)	Tohu (0)
JBoss Cache (419)	Teiid designer (107)	Errai (19)	HornetQ (1)	

Mailing List

- Dev (5494)
- Users (2783)
- Announce (19)
- N/A (1)

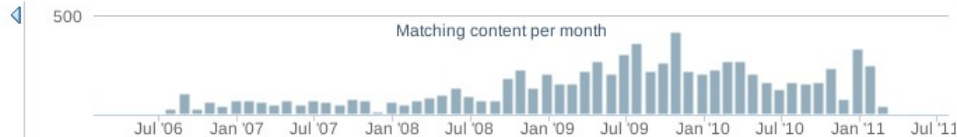
Top Authors

- Manik Surtani (527)
- Pete Muir (209)
- Mark Proctor (186)
- Emmanuel Bernard (178)
- Mircea Markus (166)
- Brian Stansberry (154)
- Dan Allen (153)
- Sanne Grinovero (138)
- Gavin King (122)
- Gavin King (114)
- Edson Tirelli (98)
- Max Rydahl Andersen (82)
- Steve Ebersole (81)
- Galder Zamarreño (80)
- Galder Zamarreno (72)

Time Frame

Quick time selection:

Last [week](#), [month](#), [quarter](#) or [year](#).



Total hits: 8297 for query **transaction manager** (took 49ms)

Page 1 of 830

[Drools Transaction Manager](#)

using **Transaction manager**: - **TRANSACTION** FOR JBOSS <property name="hibernate.transaction.mana ... ; - **TRANSACTION** FOR TOMCAT <property name="hibernate.transaction.manager_lookup_class" ... same **transaction manager** forboth application servers? Regards,Ram-- View this message in con ...

[rules](#) > [users](#) | 2011-01-31 13:07:56 | ramram <ramram858(at)gmail.com>

[Transaction Not Allowed](#)

s own **transactions**. If so, then you must use BMT (Bean **Managed Transaction**) and not CMT (Container M ... naged **Transactions**). CMT is used per default if you don't give any **transaction** attribute at all. Che ... er on **transactions** to learn more http://java.sun.com/javase/5/docs/tutorial/doc/[/url]. Also check t ...

[jbossws](#) > [users](#) | 2007-10-04 05:47:59 | oskar.carlstedt <do-not-reply(at)jboss.com>

[Drools Flow Persistence : How can I use Geronimo Transaction Manager Instead of Bitronix Transaction Manager.](#)

g the **Transaction Manager** in Drools persistence. I am working on OFBiz project an application in wh ... is a **transaction** conflict between them. Since I want to use OFBiz **transaction manager** i.e. Geronim ... ronix **Transaction Manager** to Geronimo **Transaction Manager**.BTW I am using Tomcat as an application ...

[rules](#) > [users](#) | 2009-11-03 05:48:26 | Pardeep.Ruhil(at)Intinfotech.com

[jbossws-2.0.1 released](#)

eptor **transaction**="Container">org.jboss.ejb.plugins.TxInterceptorCMT</interceptor> | ... eptor **transaction**="Container">org.jboss.ejb.plugins.CallValidationInterceptor</interceptor> ... eptor **transaction**="Container">org.jboss.ws.integration.jboss40.ServiceEndpointInterceptor</int ...

[jbossws](#) > [users](#) | 2007-10-12 20:01:33 | ptenn10 <do-not-reply(at)jboss.com>

Note: the web UI can be different once we go into production.

transaction manager

Search

Projects

Sort results by: Relevancy

Instant Search

Order by: name | frequency

Rules (1968)	Gatein (361)	DNA (100)	JBoss OSGI (18)	JBoss I10n (1)
Infinispan (1139)	Netty (311)	ESB (63)	JSFUnit (11)	Guvnor (0)
JBPM (861)	JBoss WS (310)	mod_cluster (56)	JBoss TS (6)	JBossWS Metro (0)
Weld (732)	JBoss AS7 (289)	JOPR (44)	JBossWS CXF (3)	RichFaces (0)
Hibernate (671)	JBoss Tools (157)	Wise (39)	ModeShape (2)	Scribbling (0)
Seam (425)	Teiid (157)	JBoss Cluster (26)	EmbJOPR (1)	Tohu (0)
JBoss Cache (419)	Teiid designer (107)	Errai (19)	HornetQ (1)	

Mailing List

- Dev (5494)
- Users (2783)
- Announce (19)
- N/A (1)

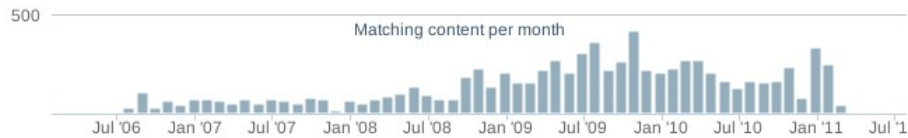
Top Authors

- Manik Surtani (527)
- Pete Muir (209)
- Mark Proctor (186)
- Emmanuel Bernard (178)
- Mircea Markus (166)
- Brian Stansberry (154)
- Dan Allen (153)
- Sanne Grinovero (138)
- Gavin King (122)
- Gavin King (114)
- Edson Tirelli (98)
- Max Rydahl Andersen (82)
- Steve Ebersole (81)
- Galder Zamarreño (80)
- Galder Zamarreno (72)

Time Frame

Quick time selection:

Last week, month, quarter or year.



Total hits: 8297 for query transaction manager (took 49ms)

Page 1 of 830

[Drools Transaction Manager](#)

using Transaction manager: - TRANSACTION FOR JBOSS <property name="hibernate.transaction.mana ... ; - TRANSACTION FOR TOMCAT <property name="hibernate.transaction.manager_lookup_class" ... same transaction manager forboth application servers? Regards,Ram-- View this message in con ...

[rules](#) > [users](#) | 2011-01-31 13:07:56 | ramram <ramram858(at)gmail.com>

[Transaction Not Allowed](#)

s own transactions. If so, then you must use BMT (Bean Managed Transaction) and not CMT (Container M ... naged Transactions). CMT is used per default if you don't give any transaction attribute at all. Che ... er on transactions to learn more http://java.sun.com/javase/5/docs/tutorial/doc/[/url]. Also check t ...

[jbossws](#) > [users](#) | 2007-10-04 05:47:59 | oskar.carlstedt <do-not-reply(at)jboss.com>

[Drools Flow Persistence : How can I use Geronimo Transaction Manager Instead of Bitronix Transaction Manager.](#)

g the Transaction Manager in Drools persistence. I am working on OFBiz project an application in wh ... is a transaction conflict between them. Since I want to use OFBiz transaction manager i.e. Geronim ... ronix Transaction Manager to Geronimo Transaction Manager.BTW I am using Tomcat as an application ...

[rules](#) > [users](#) | 2009-11-03 05:48:26 | Pardeep.Ruhil(at)Intinfotech.com

[jbossws-2.0.1 released](#)

eptor transaction="Container">org.jboss.ejb.plugins.TxInterceptorCMT</interceptor> | ... eptor transaction="Container">org.jboss.ejb.plugins.CallValidationInterceptor</interceptor> ... eptor transaction="Container">org.jboss.ws.integration.jboss40.ServiceEndpointInterceptor</int ...

[jbossws](#) > [users](#) | 2007-10-12 20:01:33 | ptenn10 <do-not-reply(at)jboss.com>

Facets →

Note: the web UI can be different once we go into production.

transaction manager

Search

Projects

Sort results by: Relevancy

Instant Search

Order by: name | frequency

Rules (1968)	Gatein (361)	DNA (100)	JBoss OSGI (18)	JBoss I10n (1)
Infinispan (1139)	Netty (311)	ESB (63)	JSFUnit (11)	Guvnor (0)
JBPM (861)	JBoss WS (310)	mod_cluster (56)	JBoss TS (6)	JBossWS Metro (0)
Weld (732)	JBoss AS7 (289)	JOPR (44)	JBossWS CXF (3)	RichFaces (0)
Hibernate (671)	JBoss Tools (157)	Wise (39)	ModeShape (2)	Scribbling (0)
Seam (425)	Teiid (157)	JBoss Cluster (26)	EmbJOPR (1)	Tohu (0)
JBoss Cache (419)	Teiid designer (107)	Errai (19)	HornetQ (1)	

Metadata List

- Dev (5494)
- Users (2783)
- Announce (19)
- N/A (1)

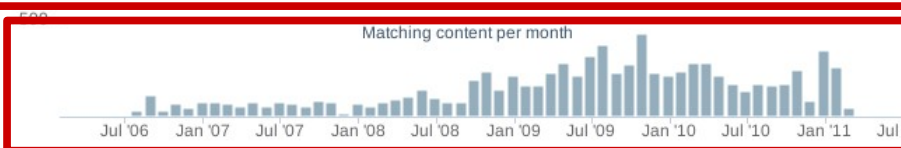
Top Authors

- Manik Surtani (527)
- Pete Muir (209)
- Mark Proctor (186)
- Emmanuel Bernard (178)
- Mircea Markus (166)
- Brian Stansberry (154)
- Dan Allen (153)
- Sanne Grinovero (138)
- Gavin King (122)
- Gavin King (114)
- Edson Tirelli (98)
- Max Rydahl Andersen (82)
- Steve Ebersole (81)
- Galder Zamarreño (80)
- Galder Zamarreno (72)

Time Frame

Quick time selection:

Last [week](#), [month](#), [quarter](#) or [year](#).



Total hits: 8297 for query **transaction manager** (took 49ms) Page 1 of 830

Drools Transaction Manager

using **Transaction manager**: - **TRANSACTION** FOR JBOSS <property name="hibernate.transaction.mana ... ; - **TRANSACTION** FOR TOMCAT <property name="hibernate.transaction.manager_lookup_class" ... same **transaction manager** forboth application servers? Regards,Ram-- View this message in con ...
[rules](#) > [users](#) | 2011-01-31 13:07:56 | ramram <ramram858(at)gmail.com>

Transaction Not Allowed

s own **transactions**. If so, then you must use BMT (Bean **Managed Transaction**) and not CMT (Container M ... naged **Transactions**). CMT is used per default if you don't give any **transaction** attribute at all. Che ... er on **transactions** to learn more http://java.sun.com/javase/5/docs/tutorial/doc/[/url]. Also check t ...
[jbossws](#) > [users](#) | 2007-10-04 05:47:59 | oskar.carlstedt <do-not-reply(at)jboss.com>

Drools Flow Persistence : How can I use Geronimo Transaction Manager Instead of Bitronix Transaction Manager.

g the **Transaction Manager** in Drools persistence. I am working on OFBiz project an application in wh ... is a **transaction** conflict between them. Since I want to use OFBiz **transaction manager** i.e. Geronim ... ronix **Transaction Manager** to Geronimo **Transaction Manager**.BTW I am using Tomcat as an application ...
[rules](#) > [users](#) | 2009-11-03 05:48:26 | Pardeep.Ruhil(at)Intinfotech.com

jbossws-2.0.1 released

eptor **transaction**="Container">org.jboss.ejb.plugins.TxInterceptorCMT</interceptor> | ... eptor **transaction**="Container">org.jboss.ejb.plugins.CallValidationInterceptor</interceptor> ... eptor **transaction**="Container">org.jboss.ws.integration.jboss40.ServiceEndpointInterceptor</int ...
[jbossws](#) > [users](#) | 2007-10-12 20:01:33 | ptenn10 <do-not-reply(at)jboss.com>

← Facets

Note: the web UI can be different once we go into production.

Search query in detail

```
{  
  from : 0,  
  query : { ... },  
  fields : [ ... ],  
  highlight : { ... },  
  facets : { ... }  
}
```

Full example: <https://gist.github.com/1012051>

query : {...}

query_string : { query : "jboss server" }

Full example: <https://gist.github.com/1012051>

query : {...}

```
filtered : {  
  query : { query_string : { query : "jboss server" } },  
  filter: {  
    and : [  
      { range : { date : { from : "2007-07-25", to : "2010-12-16" } } },  
      { terms : { _index : ["weld"] } },  
      { terms : { mail_list : ["dev"] } },  
      { terms : { from.not_analyzed : [  
        "Galder Zamarreno <galder.zamarreno@redhat.com>",  
        "Pete Muir <pmuir@redhat.com>" ] } }  
    ] }  
}
```

Full example: <https://gist.github.com/1012051>

fields : [...]

"date", "document_url", "from", "mail_list",
"message_id", "message_snippet", "subject",
"subject_original", "to", "in_reply_to", "references"

Full example: <https://gist.github.com/1012051>

highlight : {...}

```
pre_tags : ["<span class='hlt'>"],
post_tags : ["</span>"],
fields : {
  first_text_message : { number_of_fragments : 3 },
  first_html_message : { number_of_fragments : 3 },
  message_attachments : { number_of_fragments : 3 },
  subject : { number_of_fragments : 0 }
}
```

Full example: <https://gist.github.com/1012051>

facets : {...}

histogram : {...},
projects : {...},
listType : {...},
author : {...},
author_filter : {...}

Full example: <https://gist.github.com/1012051>

facets.histogram : {...}

date_histogram : { field : "date", interval : "month" }

Full example: <https://gist.github.com/1012051>

facets.author : {...}

```
terms : { field : "from.not_analyzed", size : 15 },
facet_filter : {
  and : [
    { query : { query_string : { query : "jboss server" } } },
    { range : { date : { from : "2007-07-25", to : "2010-12-16" } } },
    { terms : { _index : ["weld"] } },
    { terms : { mail_list : ["dev"] } }
  ]
},
global : true
```

Full example: <https://gist.github.com/1012051>

facets.author_filter : {...}

```
terms : { field : "from.not_analyzed", size : 15 },
facet_filter : {
  and : [
    { query : { query_string : { query : "jboss server" } } },
    { terms : { from.not_analyzed : [
      "Galder Zamarreno <galder.zamarreno@redhat.com>",
      "Pete Muir <pmuir@redhat.com>" ] } }
  ],
  { range : { date : { from : "2007-07-25", to : "2010-12-16" } } },
  { terms : { _index : ["weld"] } },
  { terms : { mail_list : ["dev"] } }
],
},
global : true
```

Full example: <https://gist.github.com/1012051>

Document Detail Preview

The screenshot shows a 'Document preview' window with a 'Conversation thread' section. The thread contains several entries, each with a 'dna>dev' header and a subject line: 'Non-XA compliant resource participation in distributed transactions'. The headers are color-coded: green for John P. A. Verhaeg, blue for Randall Hauch, and yellow for Mark Little. Below the thread is a yellow-highlighted header block with the following information:

Subject: [Re: \[dna-dev\] Non-XA compliant resource participation in distributed transactions](#)
Date: 2008-05-15 13:21:56
From: Randall Hauch <rhauch(at)redhat.com>
Project: [dna>dev](#)

The message body contains the following text:

Right. Now I guess I have a question, tho, regarding the implementation of federation connectors to systems (e.g., file systems in particular) that don't have any natural support for transactions. Is implementing the connector as if it is XA-compliant a bad thing to do?

On May 15, 2008, at 3:59 AM, Mark Little wrote:

- > As I said in an earlier email, there are significant differences
- > between compensating transactions and ACID transactions (of which XA
- > is just one possible implementation). You shouldn't offer
- > compensating transactions opaquely to users if they register
- > multiple 1PC-aware resources because the semantics are different.
- > Using them without understanding the implications can be a PITA at
- > best and cause significant data inconsistencies at worse
- > (compensations can fail or take an arbitrary amount of time to
- > resolve). ACID transactions are a pretty good default. If users have
- > multiple 1PC-aware resources then it may be in their best interests
- > to convert them to 2PC rather than be lulled into some false sense
- > of security that somehow a compensating transaction will give them
- > the same quarantees.

A 'Close' button is visible at the bottom right of the preview window.

Search query in detail

```
{  
  fields : [ ... ],  
  query : { ... },  
  filter : { ... },  
  highlight : { ... }  
}
```

Full example: <https://gist.github.com/1012096>

query : {...} & *filter* : {...}

```
query : {  
  bool : {  
    should : [{  
      query_string : { query : _userQuery_ }  
    }],  
    must : { term : { message_id : _documentId_ } }  
  }  
},  
filter : { ids : { type : "mail", values : [ _documentId_ ] } }
```

Full example: <https://gist.github.com/1012096>

highlight : {...}

```
highlight : {  
  pre_tags : [ "<span class='phlt'>" ],  
  post_tags : [ "</span>" ],  
  number_of_fragments : 0,  
  fields : {  
    subject_original : { },  
    first_text_message : { },  
    first_html_message : { },  
    message_attachments : { number_of_fragments : 3 }  
  }  
}
```

Full example: <https://gist.github.com/1012096>

Document Detail Preview

Document preview

Conversation thread: Header references only from dna>dev

dna>dev Non-XA compliant resource participation in distributed transactions by John P. A. Verhaeg on 2008-05-15 13:21:56
dna>dev Non-XA compliant resource participation in distributed transactions by Randall Hauch on 2008-05-15 13:21:56
dna>dev Non-XA compliant resource participation in distributed transactions by John P. A. Verhaeg on 2008-05-15 13:21:56
dna>dev Non-XA compliant resource participation in distributed transactions by Randall Hauch on 2008-05-15 13:21:56
dna>dev Non-XA compliant resource participation in distributed transactions by John P. A. Verhaeg on 2008-05-15 13:21:56
dna>dev Non-XA compliant resource participation in distributed transactions by Mark Little on 2008-05-15 13:21:56
dna>dev Non-XA compliant resource participation in distributed transactions by Mark Little on 2008-05-15 13:21:56
dna>dev Non-XA compliant resource participation in distributed transactions by Randall Hauch on 2008-05-15 13:21:56
dna>dev Non-XA compliant resource participation in distributed transactions by Mark Little on 2008-05-15 13:21:56

Subject: Re: [dna-dev] Non-XA compliant resource participation in distributed transactions
Date: 2008-05-15 13:21:56
From: Randall Hauch <rhauch(at)redhat.com>
Project: dna>dev

Right. Now I guess I have a question, tho, regarding the implementation of federation connectors to systems (e.g., file systems in particular) that don't have any natural support for transactions. Is implementing the connector as if it is XA-compliant a bad thing to do?

On May 15, 2008, at 3:59 AM, Mark Little wrote:

- > As I said in an earlier email, there are significant differences
- > between compensating transactions and ACID transactions (of which XA
- > is just one possible implementation). You shouldn't offer
- > compensating transactions opaquely to users if they register
- > multiple 1PC-aware resources because the semantics are different.
- > Using them without understanding the implications can be a PITA at
- > best and cause significant data inconsistencies at worse
- > (compensations can fail or take an arbitrary amount of time to
- > resolve). ACID transactions are a pretty good default. If users have
- > multiple 1PC-aware resources then it may be in their best interests
- > to convert them to 2PC rather than be lulled into some false sense
- > of security that somehow a compensating transaction will give them
- > the same quarantees.

Close

Do you mean federated within the same VM or across VMs? Mark. On 15 May 2008, at 14:21, Randall Ha

← *Threads*

```
{from : 0, size : count,  
  fields : [ ... ],  
  script_fields : { millis : { script : "doc['date'].date.millis" }  
  },  
  query : {  
    constant_score : {  
      filter : {  
        or : {  
          filters : [  
            { terms : { references : references } },  
            { terms : { message_id_original : references } }  
          ]}}}}}}}
```

Full example: <https://gist.github.com/1012115>

Challenges

Parsing and Indexing emails

Mail Subject normalization

Mail conversations

Quoted content

Author identification

...

Challenges

1) Parsing and indexing emails

Emails can have very complex hierarchical structure with different content types in it. They can be even malformed (typically SPAM emails).

Apache **Mime4J** and **jsoup** can do a very good job.

ES mapping example: <https://gist.github.com/1011913>

Challenges

2) Mail subject normalization

Emails in public mailing lists contain specific patterns in subject that should be handled carefully. Probably you do not want to index them as a part of the mail subject at all.

Challenges

2) Mail subject normalization

Typically something like the following:

- [\[infinispan-dev\]](#) Feedback on Infinispan patch
- [Re: \[infinispan-dev\]](#) Feedback on Infinispan patch
- [Re: \[hibernate-dev\] \[infinispan-dev\]](#) Feedback on Infi...

Challenges

2) Mail subject normalization

But it can be very funny too:

[jbossws-users] [JBossWS] – Re:
[jbossws-users] [JBossWS] - "
[rules-dev] ;

Challenges

2) Mail subject normalization

Or quite complex:

[infinispan-dev] [Fwd: [Fwd: Sometimes TCP responses not getting through on localhost]]

[hibernate-dev] [Fwd: [redhat.com #1341720] [Fwd: Re: Unable to checkout core/trunk/core]]

Challenges

3) Identification of conversation threads

Probably the two biggest challenges are **Thread hijacking** and **Non-linear threads**.

Probably the two most straightforward approaches are using mail headers (<in-reply-to>,<references>) and mail subjects.

Challenges

4) Quoted content

Should the quoted content be included?

(I think it should be included...)

Challenges

5) *Author identification*

Email value is not always enough.

```
" 이희승 (Trustin Lee)" <trustin@gmail.com>  
    Trustin Lee <tlee@redhat.com>  
"Trustin Lee ( 이희승 )" <trustin@gmail.com>  
Trustin Lee ( 이희승 ) <trustin@gmail.com>  
    Trustin Lee <trustin@gmail.com>  
Trustin Lee <trustin@gleamynode.net>  
    이희승 <trustin@gmail.com>
```

Want to try yourself?

You are welcome!

Some parts of the search application to be soon available on GitHub for anybody to index and search mbox files.

And one more thing...

Don't go home emptyhanded.

BigDesk

A tiny monitoring tool for Elastic Search.

<https://github.com/lukas-vlcek/bigdesk>

Thank you!

lvlcek@redhat.com
lukas.vlcek@gmail.com