

# **Warum Intranet-Suchmaschinen unbrauchbar sind ...und was dagegen getan werden kann**

2013-07-03 – Corsin Decurtins



# My Experience with Intranet Search

Notes

I have seen a lot of Intranets and Intranet Search engines in my time  
Organizations that I worked for  
Customers where Netcetera did projects

To make it short: **All of them were useless!**

email archive



**34'203** results found for **email archive**

## **Management Meeting 2003-04-09**

We should have an **archive** for **emails** at Netcetera...

## **Re: Email Archive**

Hey guys, any news re the **email archive**? Cheers Peter

## **Email Archive Release Notes 2013-01-18**

We fixed some bugs in the **email archive**...

...

much, much later ... on page **534**

## **Mailstore**

Welcome to the **email archive** of Netcetera...



# Relevance

# My Experience with Intranet Search

## Notes

The problem is **relevance**.

If you were paranoid, you could think that the Intranet Search is really out to get you.

It almost seems like it is mocking you.

The results are (usually) absolutely correct, but still useless.

Does this ring a bell or is it just me?

# My Experience with Intranet Search

## Notes

The worst Intranet Search that I have ever seen was ... ours.

It is debatable if we even had one.

We actually had an Intranet Search at some time, but nobody used it really.

It suffered from the above mentioned problems as well as performance issues.

Google™

# What's the problem?

Google solved the search problem 15 years ago, on the Internet scale

Why are we still talking about Intranet Search?

If Google can solve it for the Internet, it should be a piece of cake to solve it in the Intranet.

To answer the question, you have to understand (at least at a very high level) how Internet search engines work, particularly with respect to relevance.



# Relevance in Internet Search

Reputation/Relevance/Rank

Pages are collected into sites

Basically, you count how many links point to a site

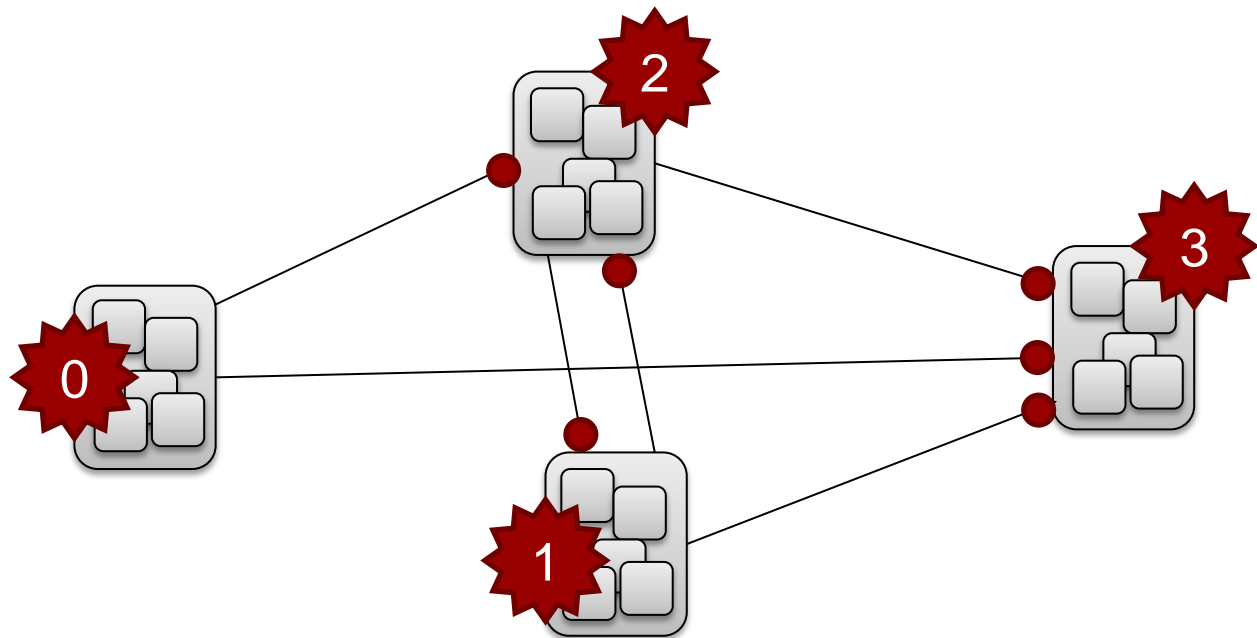
- Lots of links: important

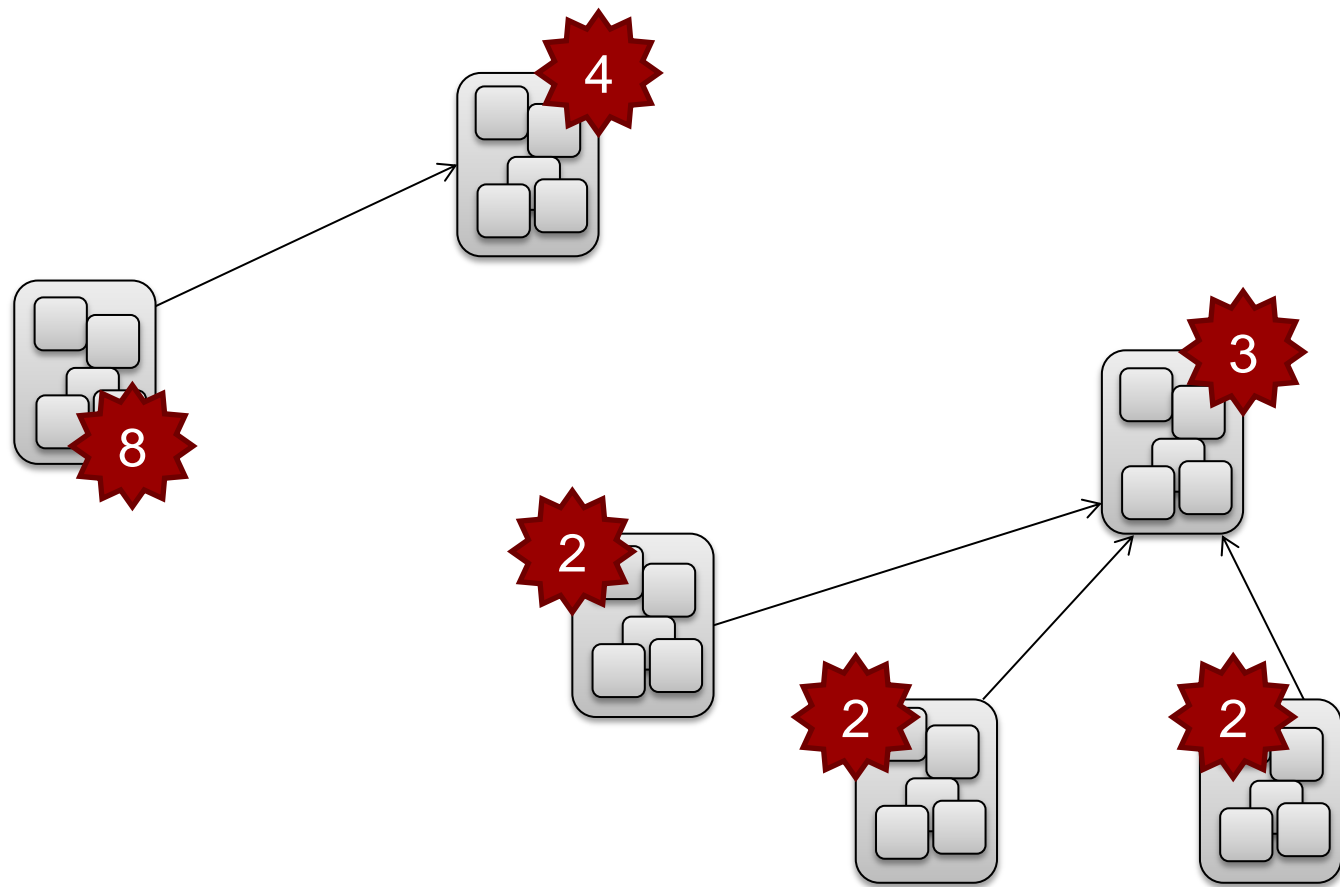
- Few links: not so important

Can be improved by looking at the reputation/relevance/rank of the source of the link

- Link from an important site -> very important

- Link from an unimportant site -> not so important





# Relevance in Internet Search

Main concepts:

- Aggregate pages into sites

- Every site has a relevance value

- Count links and calculate the relevance

- Combine relevance with the textual match of a query

There are other things of course (user profiling, social search, feedback loops, ...), but this is the core relevance metric.

# Can we use this in the Intranet?

## Notes

Sites as collections of page do not work.

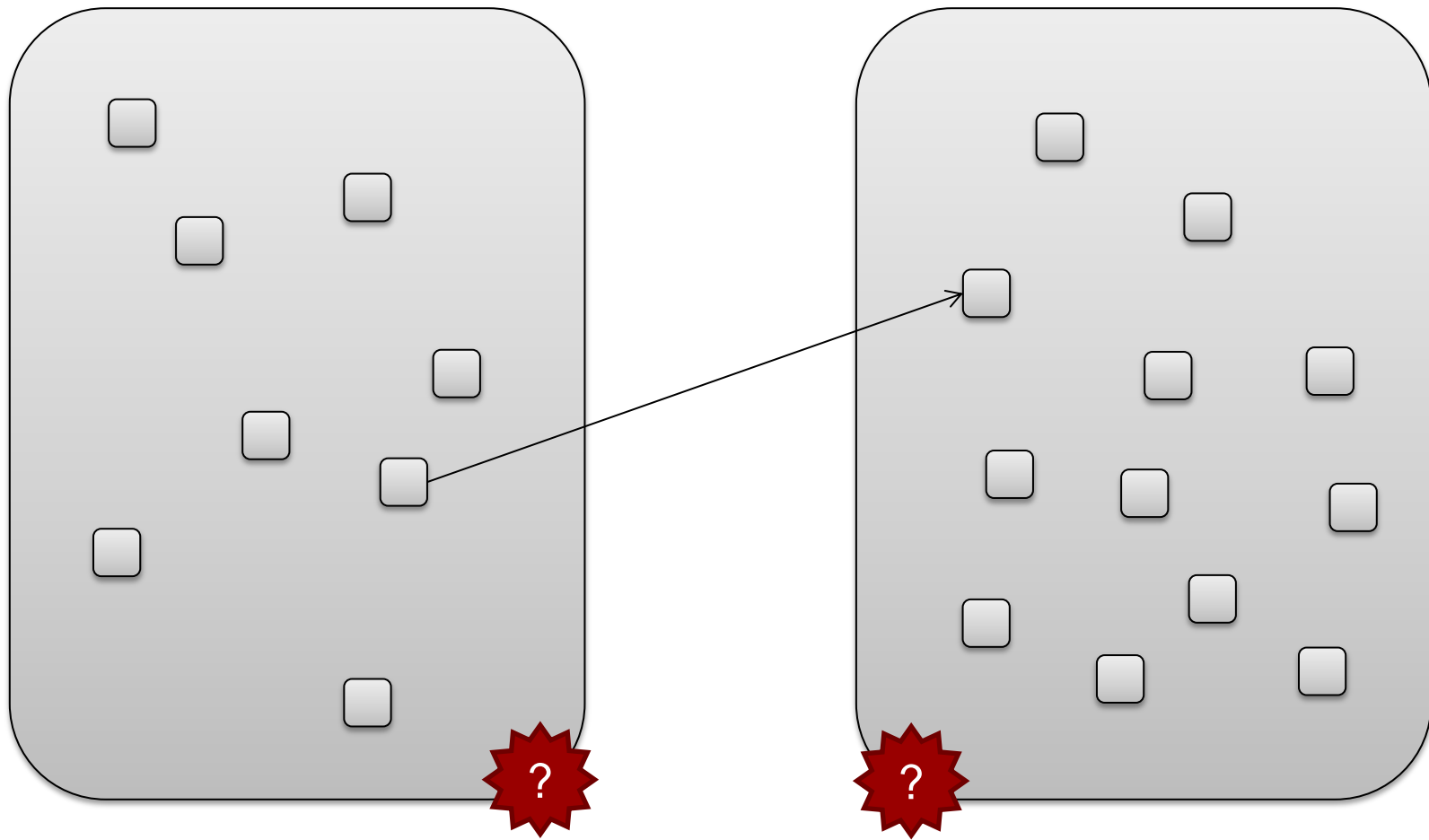
The Intranet contains much fewer links.

Files, emails, tickets contain very few links.

You do not have enough data (links, sites, files, users, requests) to make the algorithms really work.

Relevance algorithms from the Internet do not work in the Intranet.

This explains the behavior that you see in Intranet Search: absolutely correct, but still useless search results.



email archive



**34'203** results found for **email archive**

## **Management Meeting 2003-04-09**

We should have an **archive** for **emails** at Netcetera...

## **Re: Email Archive**

Hey guys, any news re the **email archive**? Cheers Peter

## **Email Archive Release Notes 2013-01-18**

We fixed some bugs in the **email archive**...

...

much, much later ... on page **534**

## **Mailstore**

Welcome to the **email archive** of Netcetera...



# Can we use this in the Intranet?

## Notes

To the search engine, everything looks the same.

Everything has the same importance.

Meeting minutes, emails, log files, release notes, personal notes, contracts, dashboards, reports, official announcements from the management, home pages – they all look exactly the same



# Relevance in the **Intranet**

# Relevance in the Intranet

## Notes

So what do we do know?

Relevance mechanisms from the Internet do not work.

Are there alternatives for the Intranet?

Can we turn the weaknesses of the Intranet into strenghts?

Let's have a look at some examples:

## Convention

[...]/doc/architecture/old/software-architecture.pdf

[...]/doc/architecture/software-architecture.pdf

# Convention

## Notes

*"For old and outdated versions of documents, create a folder '**old**' and put those documents in there."*

### Interpretation for Search:

Documents in a folder 'old' are less important than other documents.

**May 4, 2013**

[...]/doc/legal/contract-20130504.pdf

[...]/doc/legal/contract-20130626.pdf

**June 26, 2013**

*"Documents can be 'versioned' with a date stamp. The date stamp has the format YYYYMMDD."*

## **Interpretation for Search:**

The "date" of a document can sometimes be extracted from the file name of the document. Newer documents are more important than old documents.

## **Project Identifier**

g:/projects/sbb-005-7/doc/specification.pdf

g:/projects/sbb-032-2/doc/specification.pdf

## **Project Identifier**

*"Project documents are put in a folder 'g:/projects/<projectId>/'."*

*"The project identifier consists of the identifier for the customer (three letters) a sequence number and a check digit."*

## **Interpretation for Search:**

A project with a higher sequence number and the same customer is more recent and thus more relevant.



## Project Identifier

g:/projects/sbb-005-7/doc/specification.pdf

g:/projects/sbb-032-2/doc/specification.pdf

## Project Identifier

### Infostore:

Project Identifier	Project Status
sbb-005-7	<b>running</b>
sbb-032-2	<b>closed</b>

*"We have various structured meta-data available in Infostore, our company database. This information can be linked to the documents that are being indexed. The structured meta-data can be taken into account for the ranking."*

## **Interpretation for Search:**

A file that belongs to a closed project is less important than a file that belongs to a running project.

# Relevance in the **Intranet**

# Relevance in the Intranet

## Notes

Internet mechanisms and algorithms to not work.

We have something that Google does not have.

Companies have a lot of structured information about their business.

**People**

**Departments**

**Teams**

**Solutions**

**Divisions**

**Projects**

**Business Domains**

**Product Lines**

# Structured Knowledge

## Notes

You might use different terms and concepts in your organizations.

We have a lot of information about these concepts.

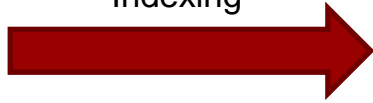
These are the concepts that people think in when they search.

We can give people to possibility to filter using these concepts.

We can judge the relevance of artifacts based on these concepts or attributes of these concepts.



Indexing



Name	Foo Contract
URL	file:///g:/projects/nca-351-3/doc/legal/contract-20130615.pdf
Type	File
Content Type	application/pdf
Author	4711
Date	2013-07-01 14:07
Content	The parties named in this contract...

When we index a resource, e.g. a document, we basically turn the document into a list of parameters. Every parameter has a key and a value.



Name	Foo Contract
URL	file:///g:/projects/nca-351-3/doc/legal/contract-20130615.pdf
Type	File
Content Type	application/pdf
Author	4711
Date	2013-07-01 14:07
Content	The parties named in this contract...

Name	Foo Contract
URL	file:///g:/projects/nca-351-3/doc/legal/contract-20130615.pdf
Type	File
Content Type	application/pdf
Author	4711
<b>Date</b>	<b>2013-06-15 00:00</b> <del>2013-07-01 14:07</del>
Content	The parties named in this contract...
<b>Project</b>	<b>nca-351-3</b>

We can analyze the parameters for conventions that we use in the organization. Based on the data that we can extract from the conventions, we update existing parameters or add new parameters.

Name	Foo Contract
URL	file:///g:/projects/nca-351-3/doc/legal/contract-20130615.pdf
Type	File
Content Type	application/pdf
Author	4711
Date	2013-06-15 00:00
Content	The parties named in this contract...
Project	nca-351-3

Name	Foo Contract
URL	file:///g:/projects/nca-351-3/doc/legal/contract-20130615.pdf
Type	File
Content Type	application/pdf
Author	<b>Mike Franz</b> 4711
Date	2013-06-15 00:00
Content	The parties named in this contract...
Project	nca-351-3
Project Status	<b>closed</b>

# Enrichment using Structured Meta-Data

## Notes

We can look up additional meta-data for the parameter list from a structured data source using some of the existing parameters.

For example we can look up the name of a person by the user id or the project status for a given project identifier.

Name	Foo Contract
URL	file:///g:/projects/nca-351-3/doc/legal/contract-20130615.pdf
Type	<b>File</b> 1.0
Content Type	<b>application/pdf</b> 1.25
Author	Mike Franz
Date	2013-06-15 00:00
Content	The parties named in this contract...
Project	nca-351-3
Project Status	<b>closed</b> 0.25

Name	Foo Contract
URL	file:///g:/projects/nca-351-3/doc/legal/contract-20130615.pdf
Type	<b>File</b> 1.0
Content Type	<b>application/pdf</b> 1.25
Author	Mike Franz
Date	2013-06-15 00:00
Content	The parties named in this contract...
Project	nca-351-3
Project Status	<b>closed</b> 0.25
Ranking Factor	$1 * 1.25 * 0.25 = 0.3125$



# Ranking value calculation

## Notes

At the end of the indexing and enrichment process, we can calculate a ranking value for each document, based on some of the fields.

The value is a multiplicative value, so 1.0 is the default. Factors  $> 1.0$  increase the rank of the document, values  $< 1.0$  decrease the value.

At the end, we have a computed ranking value for each document.

# Idea / Concept

Enrich index entries by

- analyzing them for **conventions**

- linking** index entries to the **relevant concepts** of your organization

- loading** more data from **structured databases**

Calculate **relevance value** from the enriched index entry

Rank search results by a combination of **search term matching** and **relevance value**

Allow users to **filter** the search results by different parameters

**People** Type

**Departments**

**Teams**

Content Type

**Solutions**

**Divisions**

Container

**Projects**


**Business Domains**

**Product Lines**


File Size

Date

[Plaza](#) [Infostore](#) [Zimbra](#) [Mailstore](#) [JIRA](#) [Extranet](#) [Titra](#) [Tickets](#) [More ▾](#)

Plaza Search 

netcetera | Plaza



[Search Tipps](#)

Everything

Wiki

Files

Emails

Issues

People

Companies

Projects

Resources

Books & Magazines

Include Archive

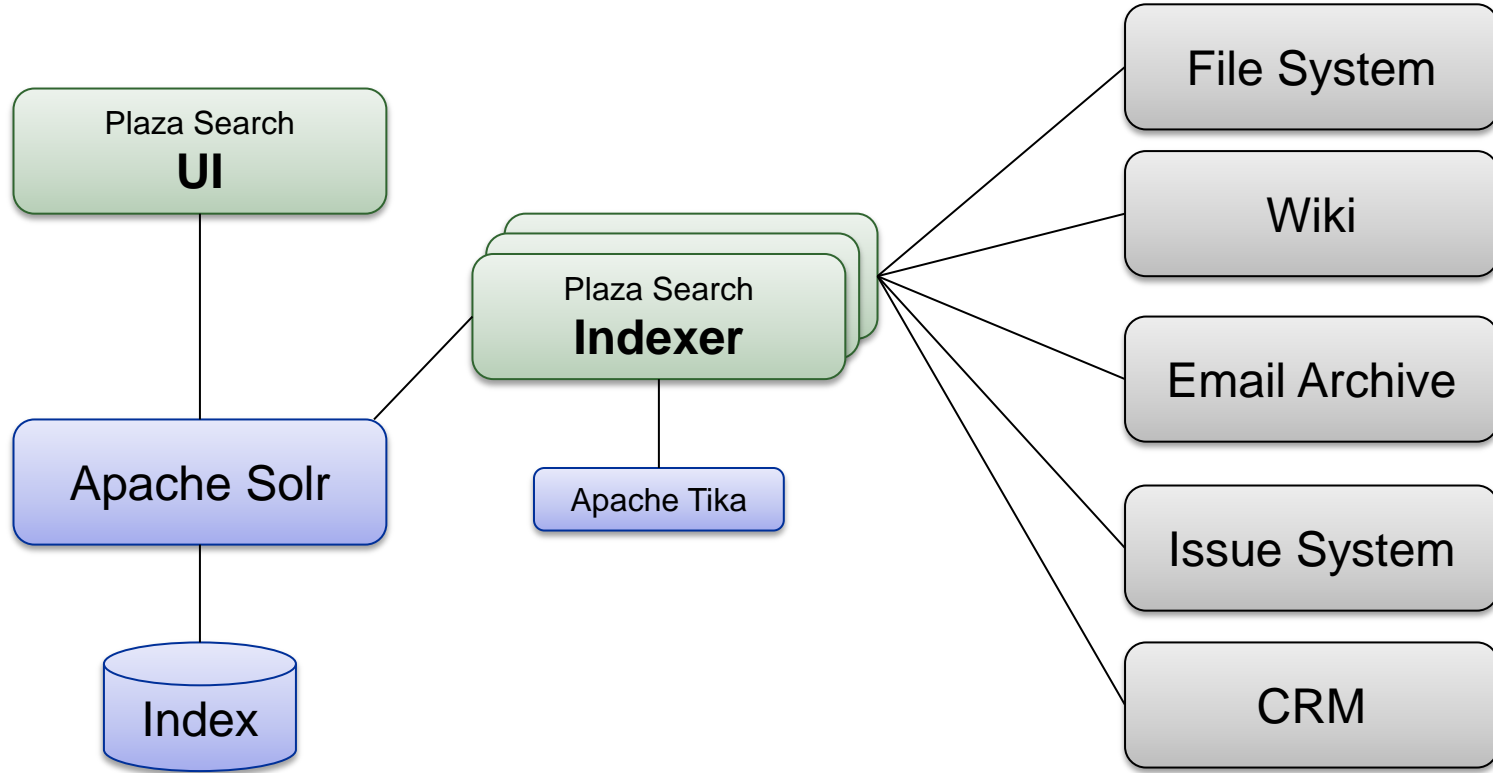
# Who's calling?

Got a phone call from a number that you do not know? If the number is in Infostore, Plaza Search can find it. Just type the phone number (or a part of it) into the search field.

5800

[Learn more](#)

# Architecture



# Architecture

## Notes

Based on Apache Solr (and other components)

Apache Solr takes care of the text-search aspect

We certainly do not want to build this ourselves

We configure it with company-specific information (more about this later)

We implement the concepts that we talked about before

# A few numbers

Live since:	May 2012
Contains data since:	1996
Releases:	~ 25
Users:	~ 250
Indexed Resources:	~ 3'000'000
Index Database Size:	~ 75 GByte
Searches per Day:	~ 500-2'000
Core Team Size:	~ 1
QA Team Size:	~ 250
Effort:	~ 1-2 hours/week

[Search Tipps](#)11993 results for **decurtins****Everything**[Wiki](#)[Files](#)[Emails](#)[Issues](#)[People](#)[Companies](#)[Projects](#)[Resources](#)[Books & Magazines](#)**Include Archive****All Types**[Email \(9539\)](#)[Spreadsheet \(911\)](#)[Text \(419\)](#)[PDF \(367\)](#)[more ▼](#)**All Dates****Corsin Decurtins (CD)**

CTO

Netcetera AG

Person

[corsin.decurtins@netcetera.com](mailto:corsin.decurtins@netcetera.com)[+41 44 247 79 32 \(5932\)](tel:+41442477932)[+41 76 207 86 70 \(10000\)](tel:+41762078670)

Online



ZRH park eg



corsin

[Calendar](#) | [Infostore \(Person\)](#) | [Infostore \(Employee\)](#) | [RPL](#) | [Wiki](#)**Books and Magazines Decurtins**

Person

[Infostore](#)**Corsin Decurtins**<https://plaza.netcetera.com/wiki/display/~corsin/Corsin+Decurtins>2013-06-20 09:32 | 9.5 KB | [Wiki](#) | [HTML](#) | [Corsin Decurtins](#)

Profilecorsin About Mel grew up in Chur, Switzerland and studied Computer Science at ETH Zurich. In 2000 I joined Netcetera for a six month internship as part of my master's degree. But somehow I got stuck and just never left After finishing my master studies at ETH I continued to work part-time at Netcetera as software engineer and part-time as research assistant in the Global Information Systems group at ETH Zurich. My research focused on model-based approaches and infrastructure for infor...

[Space: Corsin Decurtins](#)



# Search for a person

Notes

Search terms are the name of a person

People come before documents and wiki pages

Netcetera employees come before other people

[Search Tipps](#)

5 results for plaza

Active Filters: **Project** [Reset Filters](#)[Everything](#)[Wiki](#)[Files](#)[Emails](#)[Issues](#)[People](#)[Companies](#)**[Projects](#)**[Resources](#)[Books & Magazines](#)**[Include Archive](#)****[All Types](#)****[All Dates](#)**

### nca-351-3 Plaza development

<https://plaza.netcetera.com/wiki/display/nca3114>

Project

Collaboration and communication platform for all the Netcetera companies, external partners and clients.

Customer	Netcetera AG	Status	permanent
Company	NCA	Type	build
Department		End Date	2012-12-31
Keywords	Plaza		
PL	Gorazd Titizov (GT)	TC	Panche Chavkovski (PCA)
PSB Contact	Slobodan Perchuklieski (SP)	AO Contact	Corsin Decurtins (CD)
Sales			

[Mailstore](#) | [Infostore](#) | [ISR](#) | [RPL](#) | [Tech-ISR](#) | [Wiki](#) | [Issues](#) | [Filesystem](#) | [Code Repository](#)

### nca-311-4 Plaza

<https://plaza.netcetera.com/wiki/display/nca3114/>

Project

Collaboration and communication platform for all the Netcetera companies, external partners and clients.

Customer	Netcetera AG	Status	closed
Company	NCA	Type	internal
Department	SWE	End Date	2012-10-01
Keywords	Plaza		
PL	Gorazd Titizov (GT)	TC	Zbarko Lezanovski (ZL)

**Archive**

# Search for a project

**Notes**

Running project comes first, closed project comes afterwards

## 7.1 Logging

2013-06-17 13:46 | 14.99 KB | Wiki | HTML | Kevin Seidler, Martin Jäger

3 Introduction NOTE: we might mainly refer to Girders and write here as little as possible. **Logging** in General Rules for **Log** Messages There are a few points to be obeyed when making **log** messages: Performance. **Logging** is usually not a performance issue, except if **logs** statements are generated a few thousand times a second or the generation of the **log** message itself is expensive (e.g. complex toString operations). Production use. **Logging** is not only something to be activated during the de

## Space: Thomas

## logging

2013-03-19 11:18 | 3.23 KB | Wiki | HTML | Rolf Koch, Wolfgang Habicht

[illegible]

Space: swc-042-2

## Issue Logging

2013-05-07 09:24 | 2.67 KB | File | HTML | Tobias Trusch

Activity Issue **Logging** (1) Start issue right for various processes. Correctly Subsequent Issue Identification Description Description - create new case (if not), if not already there - check issue description (if something unclear, please gather further information) - check customer's chosen priority Responsible to

## Books & Magazines

Include Archive

Project identifier or name

Name or shortname of an author

more

# Search for a generic term: 'logging'

Notes

First result is the Themas page on logging

Themas is our internal manual, best practices collection, guidelines collection, ...

Basically our Hitchhikers Guide to the Galaxy



## Corsin Decurtins (CD)

CTO

Netcetera AG

Person



[corsin.decurtins@netcetera.com](mailto:corsin.decurtins@netcetera.com)



+41 44 247 79 32 (5932)



+41 78 227 88 75 (100000)



Online



ZRH park eg



corsin

This Week							Next Week							W29	W30
M	T	W	T	F	S	S	M	T	W	T	F	S	S		

[Calendar](#) | [Infostore \(Person\)](#) | [Infostore \(Employee\)](#) | [RPL](#) | [Wiki](#)

# Enrichment of index entries in the UI

## Notes

Enrichment of entries can also happen at the UI level

Data that is not used for searching/filtering, but might still be relevant

Complex data that does not fit into an index

Dynamic data that changes too quickly to be indexed

## nca-362-6 Netcetera Javascript Dev Infrastructure

<https://plaza.netcetera.com/wiki/display/nca3626/>

Project

The goal of the project is the implementation and maintenance of development infrastructure for the engineering of web applications with Javascript, CSS and HTML for Netcetera project teams. This infrastructure includes things like IDEs, build systems, quality assessment tools, technology stack.

Customer	Netcetera AG	Status	setup
Company	NCA	Type	internal
Department		End Date	
Keywords	Javascript, HTML, CSS, LESS		
PL	Corsin Decurtins (CD)	TC	Maja Trajanoska (MTA)
PSB Contact	Ramon Grunder (RG)	AO Contact	Marcel Stör (MSR)
Sales			

[Mailstore](#) | [Infostore](#) | [ISR](#) | [RPL](#) | [Tech-ISR](#) | [Wiki](#) | [Issues](#) | [Filesystem](#) | [Code Repository](#)



# Enrichment of index entries in the UI

## Notes

Links to different representations of the object are very important  
Allows users to navigate to important views

goto

# Navigation Use Case

## Notes

Search Engine has to be fast

We support the **goto** keyword, for immediate redirect

For resource with more than one URL, you can specify the URL you want to go to.

goto jira

goto cal cd

goto mailarchive nca-351-3

# Personalization

# Personalization

## Notes

We do not support personalization yet

Potential is HUGE

We know a lot of relevant information about the users

Job Profile (Project Manager, Developer, Marketing, Manager, Accountant)

Projects that the person is involved in

When the person joined Netcetera (Newbie vs. Dinosaur)

...

# Summary

Intranet Search **can** deliver useful results

**Teach** your Intranet Search engine **about your company**

Use **structured data** to improve relevance ranking

Give users the possibility to **filter by meaningful concepts**

**Personalization** has a huge potential

**Investing in a good Intranet Search pays off**

# Contact



**Corsin Decurtins**

[corsin.decurtins@netcetera.com](mailto:corsin.decurtins@netcetera.com)

+41 44 297 55 55

@corsin