LUND UNIVERSITY

EASE
Embedded Applications Software Engineering

Automated Linking of Natural Language Software Artifacts

- A Research Overview

Markus Borg
Software Engineering Research Group
Dept. of Computer Science

About me

- PhD Student
  - Software Engineering Research Group, since Jan 2010
- Three years at ABB, Malmö
  - Process Automation
- Research interests
  - Requirements-Test Alignment
  - Traceability

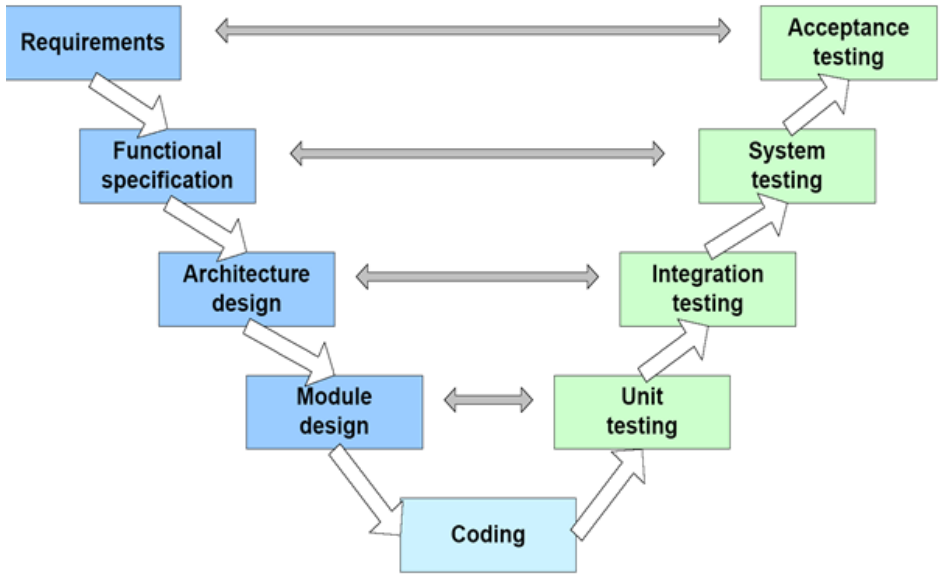
Outline of the Presentation

- Context and Motivation
- Information Retrieval (IR)
- IR-based Traceability Recovery
- A Research Overview

Context and Motivation

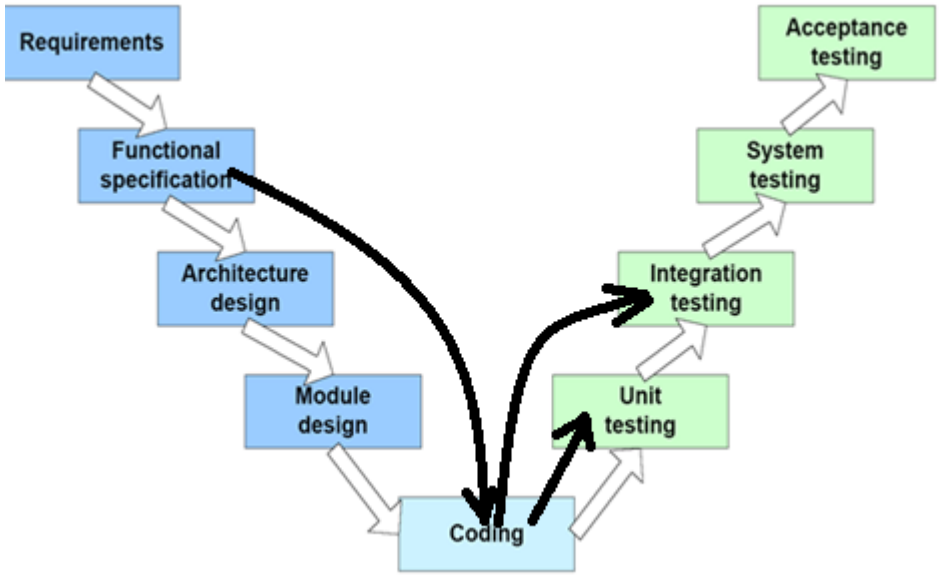
Large-scale Development

- Industrial development generated lots of information
  - Must be able to navigate this info space!



Traceability

- The ability to describe and follow the artifact life-cycle
  - Example: a use case is implemented by one or more classes that are tested by a set of test cases



Why Traceability?

- It is required or suggested by many standards
  - MIL-STD-498, IEEE/EIA 12207 (Military)
  - ISO/IEC 12207
  - DO178B, DO254 (Avionic)
  - EN50128 (Railways)
- Understand your system
- Change impact analysis
- Requirements traceability (forwards/backwards)
  - Good source for metrics
- Support software reuse

Traceability Can Exist Between

- Requirements and source code
- Source code and design
- Requirement and test cases
- Design and requirements
- Bug report and manual page
- Manual page to requirements
- ...

Traceability Challenges

- Maintaining traceability links during software evolution
  - Endless and error prone task
  - Information not updated or it is not there at all
  - Poor traceability contributes to project delays and failures

- State-of-the-practice tools do not provide sufficiently good support for traceability link generation and maintenance
  - Manually managed traceability matrix

Information Retrieval (IR)

What is Information Retrieval?

- **The process of actively seeking out information relevant to a topic of interest**

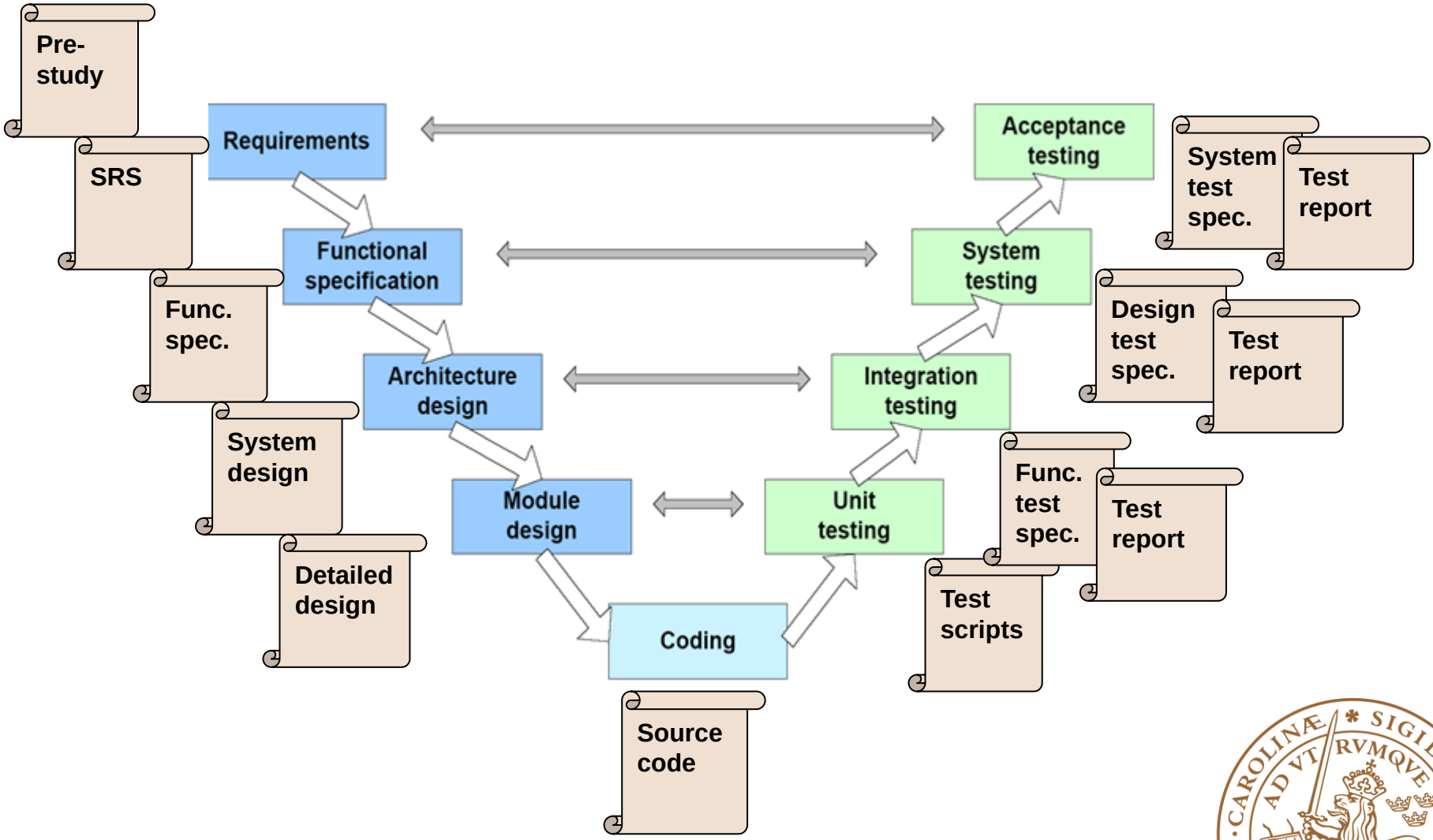
  (van Rijsbergen)

- Document
  - Generic term for an information holder (book, test case description, article, wiki page, source code file, method, requirement, etc.)
- Basis for Internet search engines

Why Analyze Textual Information in SE?

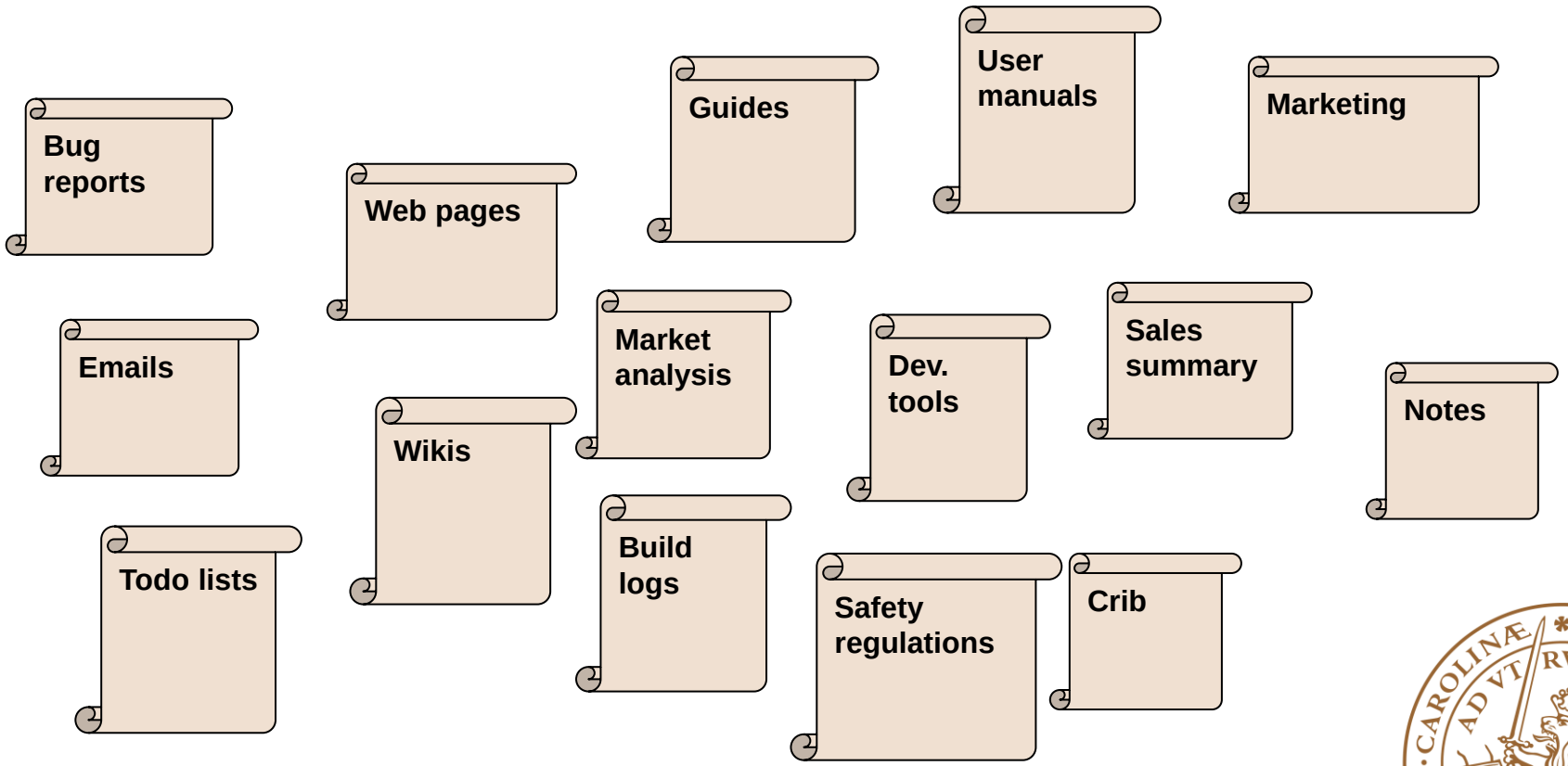
- Text is the common form of information representation
  - different abstraction levels
  - time dimension
- No predefined structure, grammar, vocabulary required
- Can be applied on legacy systems

Textual Content Everywhere



Textual Content Everywhere

- **And also…**



IR-based Traceability Recovery

Main Idea and Assumption

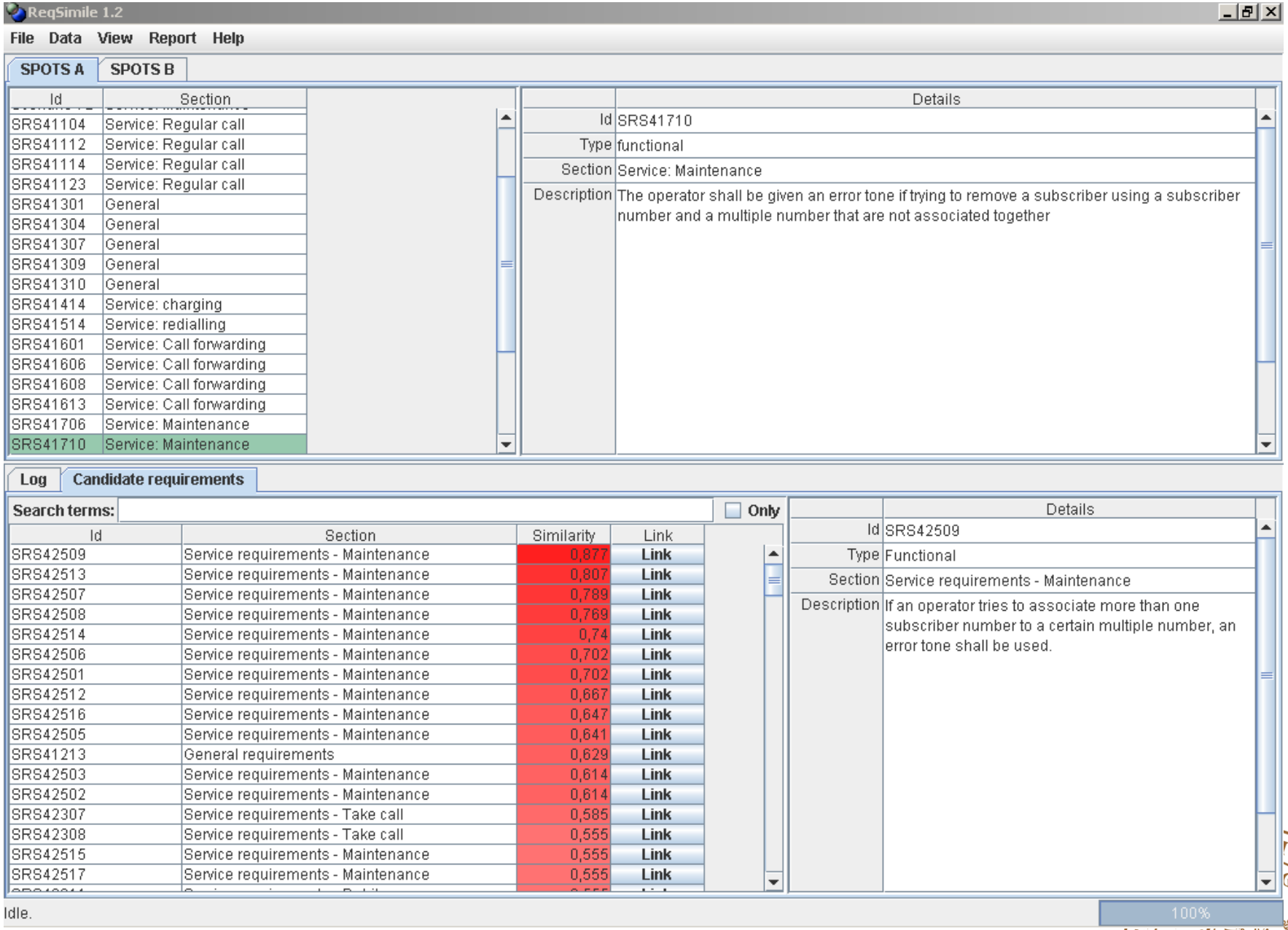
- Artifacts having a high textual similarity are good candidates to establish links between
- Developers use consistent naming in various artifacts

- Challenges
  - Synonyms
  - Testers, developers, support engineers etc.

IR-based Traceability Tools

- Search once per document
  - Every document is used once as search query in the total document space

- Replace the degree of similarity between two documents with the probability of existence of a traceability link


ReqSimile 1.2
File Data View Report Help
SPOTS A
SPOTS B
Id
Section
SRS41104 Service: Regular call
SRS41112 Service: Regular call
SRS41114 Service: Regular call
SRS41123 Service: Regular call
SRS41301 General
SRS41304 General
SRS41307 General
SRS41309 General
SRS41310 General
SRS41414 Service: charging
SRS41514 Service: redialling
SRS41601 Service: Call forwarding
SRS41606 Service: Call forwarding
SRS41608 Service: Call forwarding
SRS41613 Service: Call forwarding
SRS41706 Service: Maintenance
SRS41710 Service: Maintenance
Details
Id SRS41710
Type functional
Section Service: Maintenance
Description The operator shall be given an error tone if trying to remove a subscriber using a subscriber number and a multiple number that are not associated together
Log
Candidate requirements
Search terms:
Only
Id
Section
Similarity
Link
SRS42509 Service requirements - Maintenance 0,877 Link
SRS42513 Service requirements - Maintenance 0,807 Link
SRS42507 Service requirements - Maintenance 0,789 Link
SRS42508 Service requirements - Maintenance 0,769 Link
SRS42514 Service requirements - Maintenance 0,74 Link
SRS42506 Service requirements - Maintenance 0,702 Link
SRS42501 Service requirements - Maintenance 0,702 Link
SRS42512 Service requirements - Maintenance 0,667 Link
SRS42516 Service requirements - Maintenance 0,647 Link
SRS42505 Service requirements - Maintenance 0,641 Link
SRS41213 General requirements 0,629 Link
SRS42503 Service requirements - Maintenance 0,614 Link
SRS42502 Service requirements - Maintenance 0,614 Link
SRS42307 Service requirements - Take call 0,585 Link
SRS42308 Service requirements - Take call 0,555 Link
SRS42515 Service requirements - Maintenance 0,555 Link
SRS42517 Service requirements - Maintenance 0,555 Link
Details
Id SRS42509
Type Functional
Section Service requirements - Maintenance
Description If an operator tries to associate more than one subscriber number to a certain multiple number, an error tone shall be used.
Idle.
100%

Evaluating Tools

- **Recall**

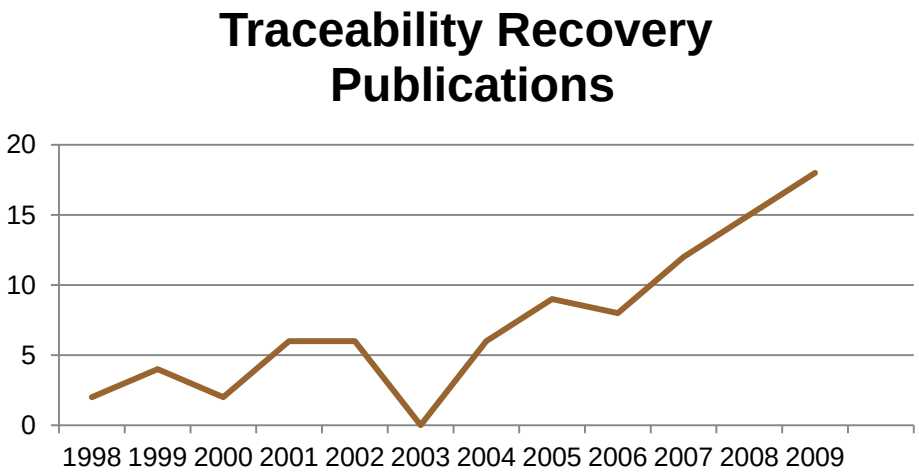
$$\frac{\text{\# Relevant documents retrieved}}{\text{\# Relevant documents}}$$

- **Precision**

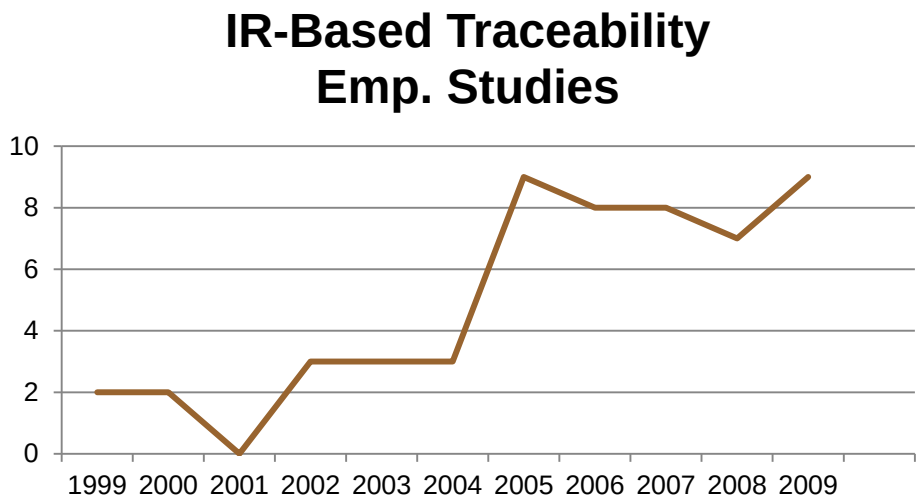
$$\frac{\text{\# Relevant documents retrieved}}{\text{\# Retrieved documents}}$$

A Research Overview

Publication Statistics

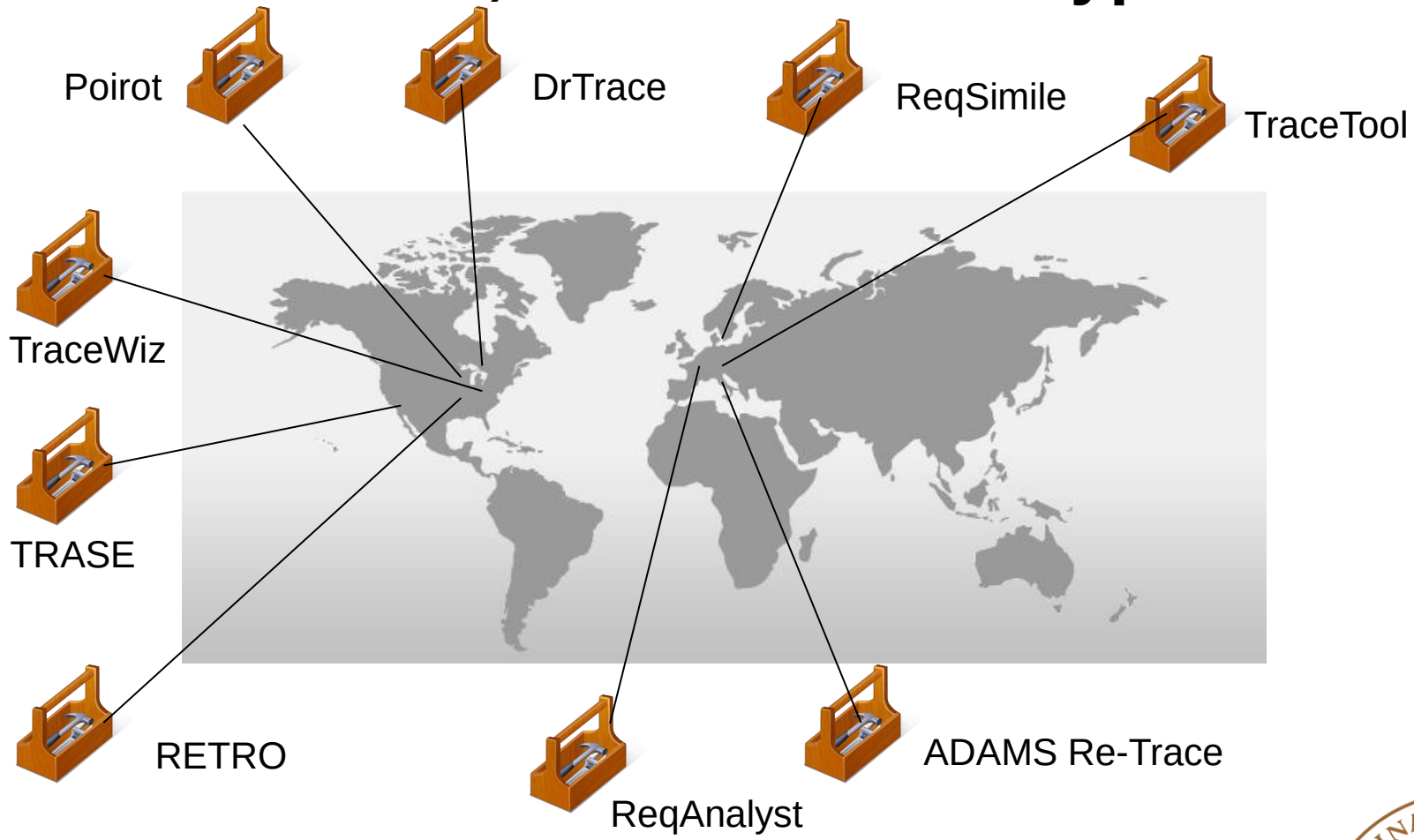


(Web of Science)



(Intermediate results, Mapping study)

Tool Overview, Research Prototypes



Case study, Antoniol et al. 2002

- LEDA 3.4
  - C++ abstract types library
  - 208 classes - 95 KLOC
  - Language: English
  - 88 manual pages
  - Traceability matrix reconstructed by hand
    - Manual pages to code: 124 correct links

Case study, De Lucia et al. 2008

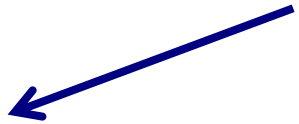
- EasyClinic
  - Developed by final year master students
  - Language: Italian
  - 30 use cases, 20 sequence diagrams, 63 test cases, and 37 code classes
  - Traceability matrix provided by the original developers: 1005 correct links

Case study, De Lucia et al. 2008

- Lessons learned
  - The tool reduces the time spent by the software engineer
  - In general, the tool reduces tracing errors
  - Ability and Experience are influencing factors
  - The tool helps to reduce the gap between high and low ability subjects
  - The performances of the IR method is an influencing factor

Other Applications in Software Engineering

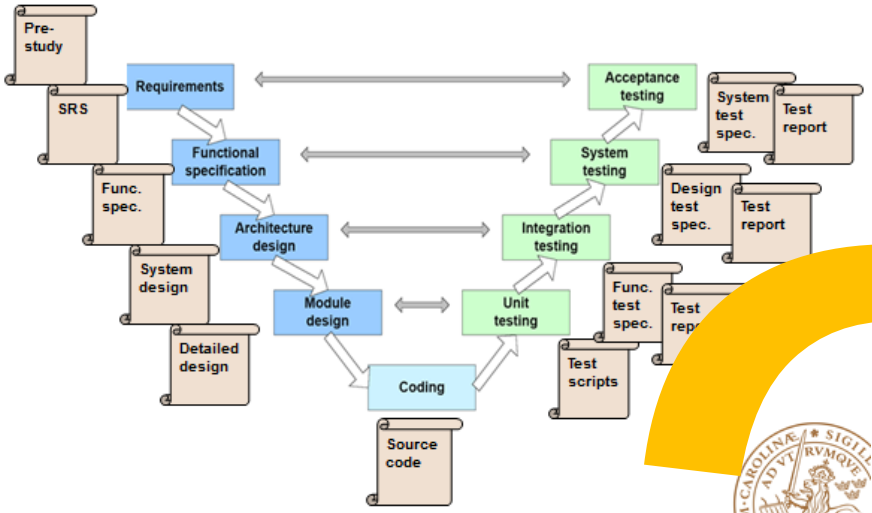
- Concept location (early step of impact analysis)
- Structural analysis
  - Coupling / Cohesion
  - Detection of code clones
- Bug triage
  - Duplicate detection
  - Developer recommendation
  - Automatic severity assignment
- And more…



Current Research Status

- Promising results
- Still few case studies on real industrial data, much focus on traceability to source code
  - Easier to get access to
- Stubborn hunt for recall and precision
  - Less focus on Return on Investment etc.
- The field matures
  - Recent publications from Daimler, IBM, Microsoft

Textual Content Everywhere



What is Information Retrieval?

- **The process of actively seeking out information relevant to a topic of interest**

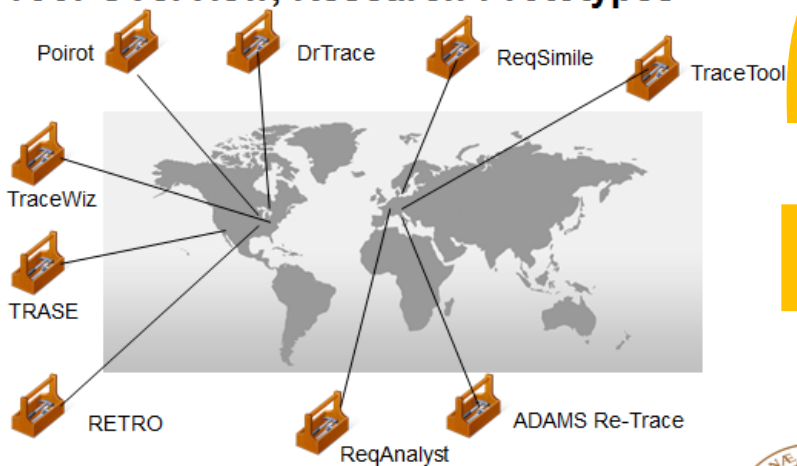
  (van Rijsbergen)

Typically it refers to the automatic (rather than manual) ...eval of documents

...ent" is the generic term for an information holder (...apter, article, webpage, class body, method, req... ...nt page, etc.)

- Bas... ...ternet search engines

Tool Overview, Research Prototypes



LEDA (Antoniol et al. 2002)