

Czech Technical University in Prague  
Faculty of Electrical Engineering  
Department of Computer Science and Engineering



Master's Thesis

## Semantic Data Analysis and Visualization of User Interactions

*Bc. Michal Švácha*

Supervisor: Ing. Ivo Malý Ph.D.

Study Programme: Open Informatics

Field of Study: Software Engineering

May 23, 2016



## Aknowledgements

I would like to thank dearly to my cat, dog and goldfish I never had. You were a true inspiration to me. With you, I would have never made this document happen. Also, my PlayStation 3 proved to be an amazing tool to keep me sane while writing this thesis. Thank you SONY.



## Declaration

I hereby declare that I have completed this thesis independently and that I have listed all the literature and publications used.

I have no objection to usage of this work in compliance with the act §60 Zákon č. 121/2000Sb. (copyright law), and with the rights connected with the copyright act including the changes in the act.

In Prague on May 27, 2016

.....



# Abstract

The current status quo in data collection is to collect everything that is available. This approach has given rise to a trend in the last couple years - "Big Data". Data that is inconveniently large for processing, interpretation and inference. When a company decides to leverage big data, it usually ends up having too much information and no real business value. Each department uses different domain vocabularies and there is therefore no synchronization or understanding. This master's thesis is trying to take the opportunity of this disorder and connect the endpoints together while leveraging big data to provide an end-to-end solution.

The focal part of this thesis is the focus on user interactions - the collection of data from mobile applications. Before such collection can even happen, the interactions must be defined - the "what", "when" and "why". Starting with management, over to architecture and engineering to interpreting results as the destination - uniting all steps to form a bigger picture.

# Abstrakt

Momentální status quo ve sběru dat je sbírání a ukládání všeho, co je k dispozici. Tento přístup dal vzniknout trendu posledních let - "Velká data". Tedy data, která jsou nepohodlně objemná pro zpracování, výklad, a odvozování závěrů. Když se společnost rozhodne, že chce využít velká data, dopadne to většinou tak, že má příliš mnoho informací bez žádné reálné hodnoty. Každé oddělení používá vlastní doménové názvosloví a tudíž chybí synchronizace a porozumění. Tato diplomová práce se snaží využít této příležitosti neuspořádanosti pro spojení všech konců dohromady a vytvořit tak komplexní řešení.

Těžištěm práce je zaměření se na uživatelské interakce - sběr dat z mobilních aplikací. Než nějaký sběr vůbec nastane, je třeba mít interakce definované - tedy "co", "kdy" a "proč". Počínaje projektovým managementem, přes architekturu a softwarové inženýrství až k interpretaci výsledků jako konečným bodem - spojení všech kroků k vytvoření uceleného náhledu.





# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Work Environment . . . . .	3
1.1.1	Key Performance Indicators . . . . .	3
1.1.2	Software Development Life Cycle . . . . .	3
1.1.3	Agile Development Methodologies . . . . .	3
1.2	Personal Interviews . . . . .	3
1.3	Scope of Work . . . . .	4
<b>2</b>	<b>Analysis</b>	<b>5</b>
2.1	Product Opportunity Assessment . . . . .	5
2.2	Regulated environment constraints . . . . .	6
2.3	Current Workflow . . . . .	7
2.4	Existing Tools . . . . .	7
2.4.1	Google Analytics . . . . .	7
2.4.2	Google Tag Manager . . . . .	8
2.4.3	Tealium . . . . .	9
2.4.4	Fabric (formerly Crashlytics) . . . . .	10
2.4.5	App Pulse (formerly Pronq) . . . . .	10
2.4.6	Aptelligent (previously Crittercism) . . . . .	11
2.4.7	New Relic . . . . .	11
2.4.8	Apple . . . . .	11
2.5	Conclusion on tools . . . . .	13
<b>3</b>	<b>Design</b>	<b>15</b>
3.1	System Architecture . . . . .	15
3.2	Issue Tracker . . . . .	16
3.2.1	Knowledge structure . . . . .	16
3.2.1.1	Teams . . . . .	16
3.2.1.2	Projects . . . . .	16
3.2.1.3	Issues . . . . .	17
3.2.2	Knowledge Extraction . . . . .	17
3.2.2.1	DSL . . . . .	18
3.3	Semantic Data Manager . . . . .	18
3.4	Tracking Engine . . . . .	19
3.5	Tracked Device SDK . . . . .	19

<b>4</b>	<b>Implementation</b>	<b>21</b>
4.1	Deployment	21
4.2	Issue Tracker	22
4.2.1	API	22
4.2.1.1	JQL	23
4.2.1.2	Methods used	23
4.2.2	DSL	24
4.2.2.1	Examples	24
4.3	Semantic Data Manager	25
4.3.1	Spring Boot	26
4.3.2	Workflow	27
4.3.3	Technologies Used	27
4.3.3.1	H2 Database	27
4.3.3.2	Flyway	28
4.3.4	REST API	28
4.3.5	Code Deployment	28
4.3.6	User Interface	29
4.3.6.1	Application Flow	29
4.3.6.2	Technologies Used	29
4.4	Tracking Engine	29
4.5	Tracked Device SDK	30
4.5.1	Technologies Used	30
<b>5</b>	<b>Testing</b>	<b>31</b>
5.1	System Benchmarks	31
5.1.1	Tracking Engine	31
5.2	User Tests	31
5.2.1	Qualitative	31
5.2.2	Quantitative	31
<b>6</b>	<b>Conclusion</b>	<b>33</b>
6.1	Future development	33
6.2	Technical debt	33
6.3	Closing remarks	33
	<b>Bibliography</b>	<b>35</b>
<b>A</b>	<b>List of abbreviations</b>	<b>37</b>
<b>B</b>	<b>Instalační a uživatelská příručka</b>	<b>39</b>
<b>C</b>	<b>Obsah přiloženého CD</b>	<b>41</b>
<b>D</b>	<b>Pokyny a návody k formátování textu práce</b>	<b>43</b>
D.1	Vkládání obrázků	43
D.2	Kreslení obrázků	44
D.3	Tabulky	44

D.4	Odkazy v textu . . . . .	45
D.4.1	Odkazy na literaturu . . . . .	45
D.4.2	Odkazy na obrázky, tabulky a kapitoly . . . . .	47
D.5	Rovnice, centrovaná, číslovaná matematika . . . . .	47
D.6	Kódy programu . . . . .	48
D.7	Další poznámky . . . . .	48
D.7.1	České uvozovky . . . . .	48



# List of Figures

1.1	Real-World Example . . . . .	2
2.1	Balanced Approach to Waterfall and Agile Methodologies . . . . .	6
2.2	Google Analytics Dashboard . . . . .	8
2.3	Google Tag Manager Dashboard . . . . .	9
2.4	Apple App Analytics Dashboard . . . . .	12
3.1	Proposed Workflow . . . . .	15
4.1	AWS System Architecture . . . . .	21
4.2	JQL syntax . . . . .	23
4.3	Semantic Data Manager Deployment . . . . .	25
4.4	AWS Elastic Beanstalk Application Deployment . . . . .	26
4.5	Spring Boot architecture . . . . .	27
C.1	Seznam příloženého CD — příklad . . . . .	41
D.1	Popiska obrázku . . . . .	44



# List of Tables

3.1	Example Database Entries . . . . .	17
3.2	Extended Example Database Entries . . . . .	18
D.1	Ukázka tabulky . . . . .	44





# Chapter 1

## Introduction

In lean/agile product development, it is necessary to have formalized user feedback loops in place, to measure product performance against various (quantitative) metrics. Such feedback loops obtain the information and statistics regarding user engagement and interaction with the developed product. Are the users using it the way the creator imagined or did they find any other means of utilizing it? What stops the users from doing the task they intended to do? Do they get everything they need, and at the same time, does the creator get what was expected? In other words, are the dominant means of usage incentive compatible for the users?

Not only do the current solutions for gathering such user data for both quantitative and qualitative metrics work out-of-the-box with low-level semantics only (everything is a general activity on a general resource), but they also tend to run on somebody else's servers. What if the product developer is in a highly regulated market, such as pharmaceuticals, and has to own all their users' data? How can the current solutions' space be utilized and tweaked in order to fit such a schema?

### **Data Disconnectivity**

The most problematic issue in large corporations is the semantic disconnectivity of the data. This occurs when data is collected ad hoc without any set direction or goal to be achieved. It may physically be all there, however nobody knows what or how it should be connected in order for it to make sense and drive value. Do I have good data or do I just have petabytes of useless log trace? Am I gathering information on what the application was intended to do or am I only filling the database with irrelevant garbage information. Most importantly though, am I gathering the information I need in a consistent fashion that corresponds to the domain of the shareholders?

Let's draw an analogy here and let me illustrate the problem with a real-world example.

## Example

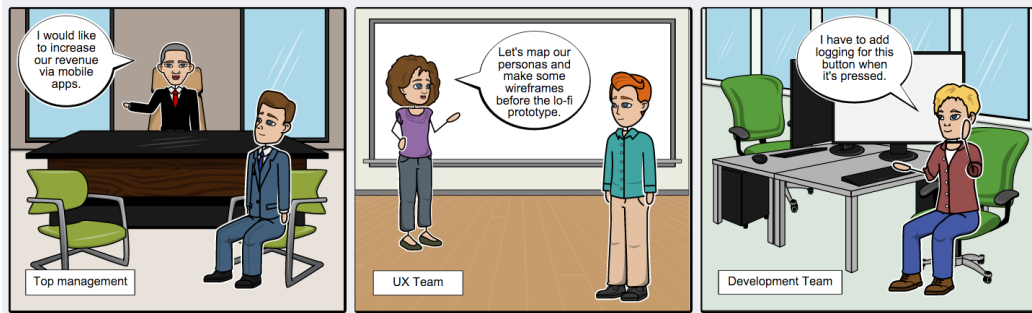


Figure 1.1: Real-World Example

This situation is mainly caused by:

1. Top management comes with a need to increase revenue via mobile devices.
2. Project management team takes over and breaks it down to user-stories, tasks etc.
3. Project management team asks the UX team, the mobile applications team and the backend team to perform their tasks in order to bring the product to life.
4. By the time product development is spread across three different teams, it is very much likely that each team will create their own jargon for every activity performed in the product.
  - The UX team is driven by the larger picture so they tend to refer to objects in a highly abstract way.
  - The backend team is driven by the inner processes happening between the database and the REST API, so they speak their precise technical language.
  - The mobile applications team is somewhere in-between, but never really aligned with either of the other teams.
5. When the product is delivered, the project management team has a really hard time translating the reported results into the top management's primary goals. Plus, the larger the teams are, the worse the situation actually gets.

This isn't a problem of poor project management. The company can use the best project management tools on the market and have the smartest people working on their teams and still encounter this problem on a daily basis. This is a larger problem resulting from the lack of intersection between different domains (Management, UX, Development, DevOps).

## 1.1 Work Environment

The context of this work is a large company, operating in a regulated market, where products are developed across multiple departments varying not only in size and experience, but also in project management approach - traditional and agile methods. Both methods implemented in the organization have a common denominator: measured KPIs to report the status of the work.

### 1.1.1 Key Performance Indicators

Each department/team/person has very specific Key Performance Indicators (= KPIs) [19] that represent the value of their work. Even products have their KPIs - "How much value has this product brought?", "How much time is this product saving us?", etc.

### 1.1.2 Software Development Life Cycle

Software Development Life Cycle (= SDLC) is a standardized traditional and formalized approach to developing a software product for regulated markets. It enables companies to follow specific steps in order to get their software product certified for use in constrained environments. It is used not for the sake of bureaucracy, but to protect the customers and increase the level of traceability of a problem, should one ever occur.

### 1.1.3 Agile Development Methodologies

In the modern era of software development, it has become "cool" to not follow traditional waterfall models and to adopt agile methodologies, such as scrum or kanban. For many years these were found exclusively in start-ups. But because time and time again, research [5] has shown that agile methodologies do impact the level of productivity and innovation, larger companies are working on adopting them in their constrained context as well. This thesis operates in such a place, where agile methodologies are being deployed across the whole innovation department.

## 1.2 Personal Interviews

In order to find out what needs there are, I have conducted several interviews with leaders of various departments. These are their reactions to what bothers them about product monitoring during the development phase:

#### 1. Associate Director, Applied Technology

"The real problem I see is the fact that all the information I need is on somebody else's server. We can't store any sensitive, let alone confidential information *somewhere* with some random vendor. It's actually illegal in some countries. Unfortunately, sometimes sensitive data is exactly what we need to obtain from the applications to make an informed decision."

## 2. Associate Director, Mobile and Web

"Our needs for tracking KPIs are variable throughout time and unfortunately the current tools we use are quite inflexible. Because we are in a regulated market, each change that requires a new build of an application takes a longer period of time. And time *is* money."

## 3. Mobile Application Development Lead

"I have noticed that one of the biggest obstacles is how should we name what we measure. I have no vocabulary to help me during the development. The only thing we have is a robust Google Analytics toolkit that only allows us to gather low-level actions. We can log that a button was pressed, but what do we name such action? "Button pressed"?"

## 4. UX Lead

"Our team looks at the high-level needs. We are trying to make the user activities in an application as smooth as possible. When we design a low-fidelity prototype, we know what we want to measure. Having the opportunity to add a high-level KPI would help us a lot in order to gather feedback for our prototypes. We are not programmers, we don't know how to add it to the code, but I would love the idea of including what to measure along with the prototype."

# 1.3 Scope of Work

First, I will present the context of agile software development in a regulated market. Then I will take a look at the current tracking solutions being used in mobile development (the focus is on iOS development, but most of the tools are multiplatform solutions). I will examine and analyze their strengths and weaknesses.

Next I will propose a workflow to fit the needs of a larger company with multiple departments, operating in a regulated market. I will utilize existing tools and build on top of them in order to drive value without reinventing the wheel couple times.

Lastly, I will develop a working PoC (Proof of Concept), verify its functions through user testing and further discuss with the stakeholders whether or not it is the correct path to take in order to unite all departments and connect the dots in the data.

# Chapter 2

## Analysis

### 2.1 Product Opportunity Assessment

Product opportunity assessment is a helpful procedure to separate the wheat from the chaff. It points out the real pressing issues that are mission critical for the company.

1. *What problem are we trying to solve?*

Enabling alignment of business language and user interaction reporting is a key part to a success of a new product. No commercially successful platforms allow such alignment, not even any higher semantic analysis.

In a regulatory market it is vital to have an option of storing the user data on custom servers. No such option along with the higher-level analysis is currently available at the market.

There is no standardized toolkit and thus the lack of interchangeability between tools tends to end up with a vendor lock-in.

2. *For whom do we solve that problem?*

For all departments in the company that are participating in product development, top to bottom. From setting a business goal to defining the needs, while also helping the developers and the designers.

3. *How will we measure success?*

Having gathered data that is securely placed on custom servers. Data that is united and connected by shared domain vocabulary, ready to be analyzed and visualized in order to draw conclusions.

4. *What alternatives are out there?*

There is some proprietary software on the market, but none is flexible enough for the defined needs.

5. *Why now?*

The competitors [9] are really getting into the data-driven decisions. The intention is not to gather as much data as possible, but to generate valuable insights as soon as possible in order not to fall behind the competition.

6. *What factors are critical to success?*

Alignment of all departments participating in software development in a non-invasive fashion. It is critical to primarily focus on helping the departments to connect with each other.

The ability to run the whole solution on custom servers and have total control over all gathered data.

## 2.2 Regulated environment constraints

Regulated markets, especially pharmaceuticals have multiple rules that need to be carefully followed in order to be allowed to use their new IT products. Not only is important how much value does the final product bring to the end-user, but also how data storage is handled and how prone it is to exploitations and attacks.

From the development process point of view it is equally important to follow specific guidelines and processes during the development phase. Each step has to be carefully documented and approved by specific audit department. For that reason, most of the big pharmaceutical companies use the waterfall model implemented in SDLC. It enables companies to follow these specific steps in order to get their software product certified. This all hassle is not for the sake of bureaucracy, it is to protect the customers and increases the level of tracability of a problem, should one ever occur. After all, it is theirs and their customers' private data that goes on the server, so it is imperative that it is safe.

The limitation of this approach is the inflexibility of the waterfall model. The strict and rigorous requirements for the waterfall model are not allowing to go back half way through and re-define the objectives of a product. Therefore more and more large companies are trying to implement the best of both worlds [10].

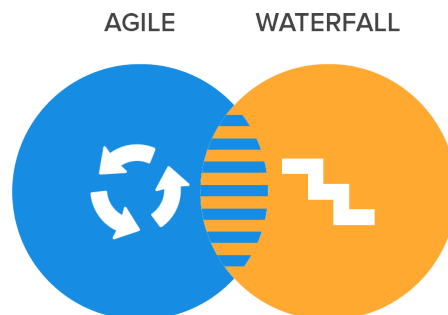


Figure 2.1: Balanced Approach to Waterfall and Agile Methodologies

In the context of my work, the approach is that prototypes, PoCs and internal beta versions of new software products are developed using various Agile Methodologies (Scrum and Kanban) and once their purpose is verified, the work done is considered as the starting point for the SDLC.

## 2.3 Current Workflow

Currently, the top management is in the USA, while the innovation department is in Prague, Czech Republic. Some members from the top management occasionally fly to Prague, but very irregularly and mostly for a quick check on the overall status of the department. Most of the communication happens remotely - via video conference calls, e-mails etc. Obviously, this produces a lot of noise in the data being transferred. Sometimes some parts are misheard or misunderstood, get lost and are causes for a larger problem in the future.

Not only there is a difference in the culture and language, but it clearly leads to a difference in the domain vocabularies used in everyday work. Tools for productive data exchange are implemented - such as:

- Issue tracker
- Internal Wiki-style website
- Enterprise chat

But the problem is, that there is no tool that unites them all to make sure that the domain vocabularies are the same throughout the whole development phase. If such tool existed, a one that could connect to all of the deployed knowledge systems, it would make communication of measured performance of a product a lot easier.

Let's take a look at the current tools available on the market that specialize in application performance measuring.

## 2.4 Existing Tools

I will now analyze and assess the tools available on the market. I'll focus primarily on the flexibility of the deployment and their analysis features. Anything else has a minor priority.

### 2.4.1 Google Analytics

Google Analytics is by far the most popular and widely used framework for monitoring user interactions in applications. It reports everything the developer wishes to. By default it does not report anything - the tool has to be activated during application start. Then each action needs to be hard-coded in the code. An action can be two things - the first one is after a certain user activity has happened - a press of a button, pulling down a list to refresh the data etc. One thing gets reported - "this event has happened". The second one is more open - an identifier is set to an item, let's say, a button. Whatever happens with that button gets reported, be it touch, swipe or anything else.

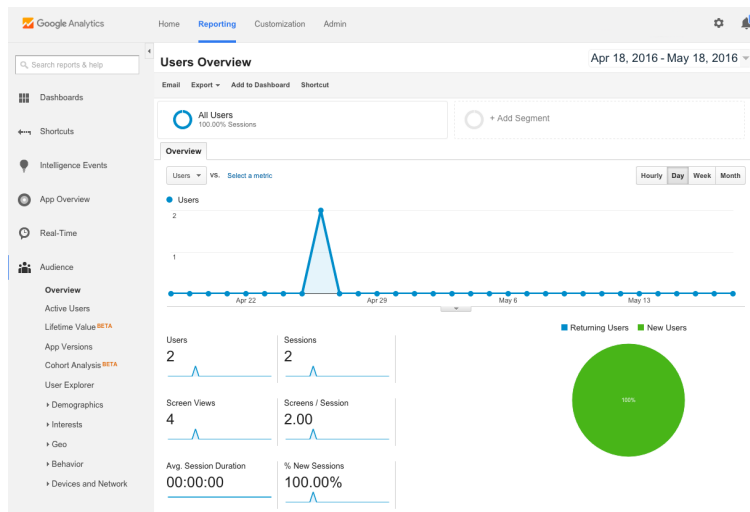


Figure 2.2: Google Analytics Dashboard

The dashboard website is very detailed and responsive. All data is nicely visualized in graphs and corresponds well with the whole GA ecosystem. Higher order semantics is remotely possible via definition of complex queries and dashboard setups. User management is very rich and enables forming multiple roles for different users. The main problem is, unfortunately, that all data goes to Google. There is no option of having the engine run on custom servers. All code is closed source and that simply wouldn't go through any risk assessments because Google might be sending the data to anywhere in the world.

Data is accessible via REST API, but registration for it is required (it is not possible for a "normal" user to start using the REST API). Single user account is free, enterprise account is paid.

### 2.4.2 Google Tag Manager

Google Tag Manager is a tag management system that allows quick and easy update of tags and code snippets in a mobile application. It allows adding and updating AdWords, Google Analytics, etc. from the Tag Manager user interface instead of editing the code.

A tag is a snippet of code that sends information to Google. It is not necessary to wire it up in code - it works through configuration files from the admin user interface.

Tag Manager is deployed in conjunction with the Firebase SDK, with support for both Android and iOS. The container replaces all other manually-coded tags in the mobile application, including tags from AdWords, Google Analytics etc. It basically builds on top of Google Analytics to make the integration a little easier. It is aimed to be used by people focused mainly on marketing, thus the tool itself serves as an overall performance dashboard. It allows them to create complex tags in a short period of time. Unfortunately it works well only on the web - the tags have to be hard-coded in mobile applications.



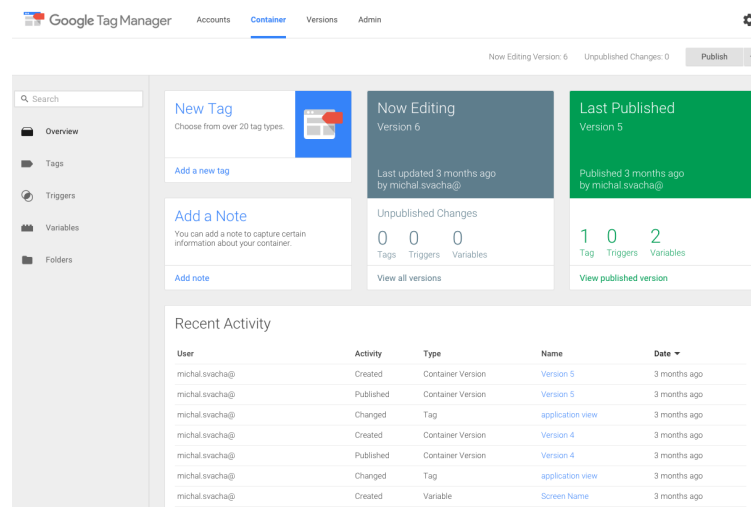


Figure 2.3: Google Tag Manager Dashboard

The dashboard is similar in style to the Google Analytics one. Tags can be created, modified and linked with a specific Google Analytics query. The same as for Google Analytics holds for Google Tag Manager - everything is closed source and runs exclusively on Google's servers.

### 2.4.3 Tealium

Tealium is a complex tool combining all marketing tools on the market (Criteo, Socialbakers, but also Google AdWords etc.). It does offer insights into usage and performance and it is very robust.

"Combining the leading enterprise tag management solution, an omnichannel customer segmentation and action engine, and a suite of rich data services, the Tealium CDP enables organizations to unlock customer data trapped in siloed marketing systems, build a unified customer view, and take real-time action." [7]

As of December 2015, Tealium successfully completed a Health Insurance Portability and Accountability Act (HIPAA) and Health Information Technology for Economic and Clinical Health Act (HITECH) attestation examination [8], making it compliant to store data in the cloud securely. The only limitation with this is, that it is only valid in the USA - it does not extend to countries such as China, Japan or Russia, where customer data have to be stored within the country's borders.

The tool itself is intended to be wired up and used as-is. Tealium has their own servers, their own front-end and no API. While for smaller to mid-sized companies it may be sufficient, larger companies require code customizations and integrations into their systems that are currently not provided by this tool.

#### 2.4.4 Fabric (formerly Crashlytics)

Fabric is a platform made by Twitter. Collected and fine-tuned I should say, as it comprises of multiple acquired start-up platforms. The focal point of this platform is a former crash reporting tool Crashlytics [2]. Aside from that Fabric is a full-featured developer platform used for a variety of tasks. It helps to overcome obstacles a developer may face - even the distribution of beta versions to testers, which has always been a problem for iOS developers.

"The Fabric platform is made of three modular kits that address some of the most common and pervasive challenges that all app developers face: stability, distribution, revenue and identity. It combines the services of Crashlytics, MoPub, Twitter and others to help build more stable apps, generate revenue through the world's largest mobile ad exchange and enable to tap into Twitter's sign-in systems and rich streams of real-time content for greater distribution and simpler identity. Installation takes minutes, and most features only require a few lines of code." [14]

Fabric has been taken it a step further and is not only about reporting crashes, but it also reports overall statistics, like Google. One really nice feature is a video visualization of user movements in the application. Formerly Appsee [1], now a UX part of Fabric, creates a video of steps users take in the application to give the developer a new perspective on how their application is actually used.

The source code of Fabric is closed source (except for some parts, like Fastlane [6]) and all of the statistics run on Twitter's servers. The Data is not accessible in any other way, than via Fabric's custom dashboards.

The whole platform is free to use. Registration of developers and applications is required along with the installation of a custom Fabric installation program which integrates the framework into existing projects.

#### 2.4.5 App Pulse (formerly Pronq)

Hewlett-Packard Enterprise AppPulse Mobile is a mobile app performance monitoring tool that tracks the real user experience of mobile applications.

App Pulse is very different from the previously listed tools in a way that it reports everything at all times. The usage is fairly low - tens of kilobytes per week, but it is very thorough. Screen time, actions, movements - it is all there in the report and no setup is required for the start. That obviously leads to gathering a lot of junk by default.

Contrary to iOS standards, App Pulse's SDK is distributed via a compiled library that has to be drag-and-dropped manually into the Xcode project. This is not a good practice, as Apple's updates are regular and oftentimes break a lot of old code, so keeping the SDK up to date manually is a lot of pain.

Because it works out-of-the-box and "automagically" tracks everything that happen in the application, the one big issue is the need to have a consistent naming of all views, labels, buttons etc. - as it does everything on its own, without hooking up the actionable items to the framework manually, it can be hard to determine which button was referred to in the

report. One really interesting feature is battery level monitoring, though it is disputable, as the user may have multiple applications running in the background.

Free 30-day trial allows up to two mobile applications, 25.000 active users per application (everything above is discarded, or offered for the standard monthly price) and access to the App Pulse community. The tools are closed source and all of the statistics runs on Hewlett-Packard Enterprise servers as it is a Software as a Service (= SaaS). No API is provided.

#### 2.4.6 Aptelligent (previously Crittercism)

Aptelligent doesn't stand out from previously mentioned tools - it runs on Aptelligent's own servers, has custom SDK and works seamlessly. While it does have some benchmarking that others don't, at the end of the day, Aptelligent again is the owner of your users' data. It seems very enterprise oriented - it enables 3rd party API integration into their system, which enables monitoring the performance of other APIs used in the application to really find what can be the bottleneck of the overall performance.

The tools are closed source and all of the statistics run on Aptelligent's servers. API for data retrieval is provided.

#### 2.4.7 New Relic

Other than the standard package (SDK, custom servers, periodical reports), New Relic has a nice alerting system - when a crash occurs, a web hook to ticketing system can be defined to streamline bug reports into the issue tracker.

Similar to Tealium, New Relic does stand out with the focus on data encryption:

"You have complete control over what, if any, sensitive information is sent to New Relic. We are unique among software analytics solutions. When you deploy our agent, by default, our security settings and regulatory compliance exceed industry standards, and all data is encrypted in transit." [11]

While this is a good start, it simply isn't compliant with the legislation of many countries, to store the users' data on a "random" cloud computer storage. The iOS SDK is distributed in a compiled binary file, having enormous 28,5MB. It is not open sourced. API for data retrieval is provided.

#### 2.4.8 Apple

The last tool I will mention isn't considered as a framework, but is important nonetheless. As companies fight for user data from mobile applications - such as Twitter, who even gives out their platform to everybody for free, naturally the platform owners (Apple and Google) strive for keeping all that precious data for themselves. Apple announced in 2015 at their Worldwide Developer Conference (= WWDC) a new iTunes Connect portal with a new feature - App Analytics.

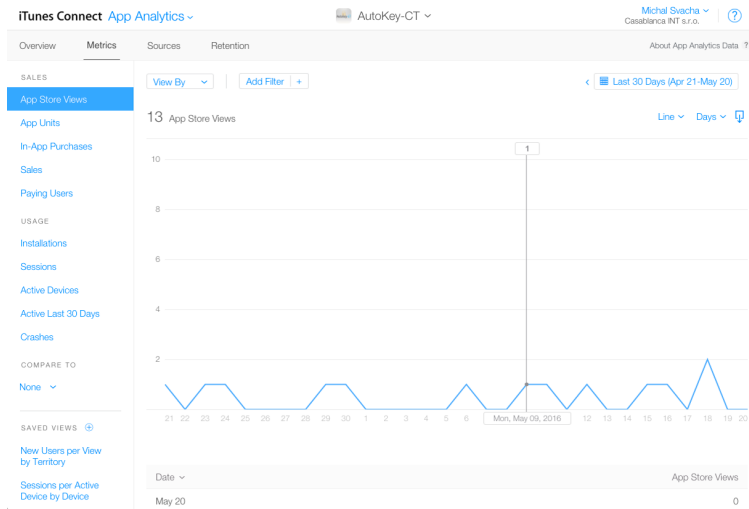


Figure 2.4: Apple App Analytics Dashboard

It is fairly thorough in means of usage, downloads, screen time etc., but the overall reporting seems very marketing oriented. Crashes do get reported, but they are not very detailed compared to Fabric's Crashlytics tool.

One of the biggest obstacles every iOS developer faces is the fact that since iOS 5 [13] it is not possible to uniquely identify a device. Therefore it is virtually impossible to connect two user entries of the same user if the user deletes and redownloads the application. The only way it could be identified is through Apple's own analytics solution. However, the help tooltip for Installations says:

"The total number of app installations and redownloads. Installations doesn't include app updates."

Which looks like Apple disabled identification of the same device for themselves as well.

App Analytics in iTunes Connect is provided to every single developer automatically for free on the iTunes Connect website. There is no framework to be implemented, the statistics is gathered whether the developer likes it or not. Naturally Apple keeps all of the user data on their servers. Even though iTunes Connect has a JSON REST API, none of the sales or analytics data endpoints are exposed.

## 2.5 Conclusion on tools

### 1. Google Analytics

**PROS:** robustness, reliability, reputation

**CONS:** completely proprietary, very low-level

### 2. Google Tag Manager

**PROS:** abstraction above GA, flexibility

**CONS:** completely proprietary, hard initial setup

### 3. Tealium

**PROS:** versatility, omni-channel connectivity

**CONS:** completely proprietary

### 4. Fabric

**PROS:** versatility, community support

**CONS:** completely proprietary, required Twitter registration, data inaccessibility

### 5. App Pulse

**PROS:** automatic initial setup

**CONS:** completely proprietary, non-standard SDK distribution, data inaccessibility

### 6. Aptelligent

**PROS:** 3rd party tool integration

**CONS:** completely proprietary

### 7. New Relic

**PROS:** user data encryption

**CONS:** completely proprietary, non-standard SDK distribution

### 8. Apple

**PROS:** no installation or setup required

**CONS:** completely proprietary, data inaccessibility

None of the tools meet the requirements for deployment (custom servers) and therefore privacy as well for a global enterprise in regulated market. It is necessary to design an architecture that would fit the needs - run the whole solution on custom servers and enable the mash-up of various knowledge resources deployed across the company.



## Chapter 3

# Design

### 3.1 System Architecture

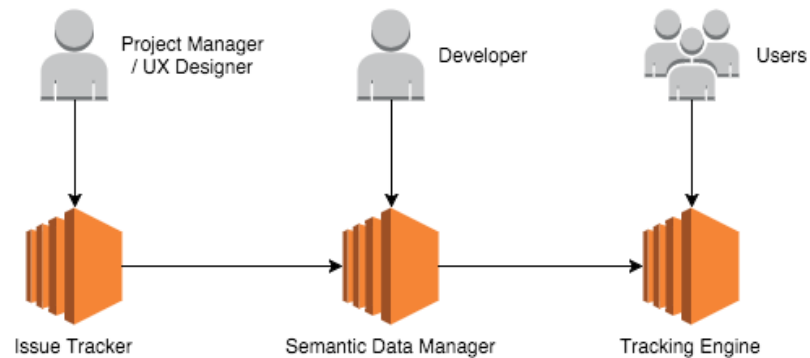


Figure 3.1: Proposed Workflow

#### Description

The workflow should be as linear as possible with respect to the natural flow of data. The central knowledge base is the issue tracker, because that's where the project management team defines objectives set by the top management. The UX team has specific needs for measuring certain tasks in the application so it is absolutely valid that they would add their needs in the issue tracker.

From then on, the domain vocabulary is created and stored for everybody to see. In order to measure the tasks performed by the developer (for example "implement login form") using the same domain vocabulary, something has to deliver a computer-readable version of the vocabulary. That is the task of the Semantic Data Manager - to extract the knowledge from the Issue Tracker and serve it to the developer so he/she can use it easily within the application.

The user data is then reported to the Tracking Engine using the same domain vocabulary that was defined at the beginning. That will enable alignment and easy reporting of the results.

## 3.2 Issue Tracker

Issue Tracker is a software deployed in a company for managing and tracking the work done by multiple employees across different departments. It serves the project management team to define projects, milestones and to break the work down to tasks and sub-tasks. Those are then assigned to specific people and marked as done when they're done and reviewed. It is considered a good practice that when a task/issue is created and assigned, it becomes immutable with respect to its name (and of course project label and ID). The detailed description and comments can be added, but in order to preserve some stability, the name should remain the same.

There are many solutions on the market, such as Bugzilla, JIRA or open-source Redmine. Most of the solutions provide a REST API for 3rd party integrations.

### 3.2.1 Knowledge structure

Every company uses a slightly different way of structuring their knowledge in their project management tools. In my situation, the structure is as follows:

- Teams
- Project
- Issues

Different tools have different naming conventions, but this structure is as generalized as possible so it is applicable to most of the tools on the market.

#### 3.2.1.1 Teams

Everybody is distributed into teams in various departments. While departments are important in the company structure, in the project management, teams are more relevant. Every team has its own Project Board (either Scrum or Kanban) that reflects the state of the team in time - how much work has been done, what will be done and what's currently being done.

#### 3.2.1.2 Projects

Each team works on their own projects. Often times, collaboration does occur, but the owner of the project is still the team. If other person collaborates with other team, it is reflected in his/her work statistics (storypoints, hours spent), but the ownership stays within the scope of the project and thus the team as well.



### 3.2.1.3 Issues

Issue is the most granular entity in the system. It defines a step or set of steps to be performed and is usually assigned to one person. Workflows for reviewing etc. vary from team to team. There is a possibility to break down the work on single issue (if it's especially large one) to sub-tasks, but those are not given IDs, so they are not unique per se. They only serve for clarity and visibility of work being done during the day. When all sub-tasks are finished, the whole issue is finished. That is the state reflected by the API.

## 3.2.2 Knowledge Extraction

The extraction of the knowledge will build on top of the premise that some parts of the data at certain point of time are immutable. In order to make it as bulletproof as possible, I will only assume immutability of the name of the issue, its automatically assigned identifier (ID) and, obviously, the project it belongs to. Everything else is considered fully mutable and is welcome to be changed in order to provide flexibility.

The idea behind the tracking itself in the same domain vocabulary is that the beginning of an issue/task/sub-task and the end is reported. Giving a direct feedback of success or failure of the implemented workflow.

### Example

An issue is created in the Issue Tracker - "Create form for user registration.". The Issue Tracker automatically assigns an identifier, for example "SA-01". The resulting database table should look like this:

Issue ID	Name	Type	...
SA-01	Form for user registration.	Start	...
SA-01	Form for user registration.	Finish	...
SA-01	Form for user registration.	Start	...
SA-01	Form for user registration.	Finish	...
SA-01	Form for user registration.	Start	...

Table 3.1: Example Database Entries

It is clear what is measured here<sup>1</sup> - the success of a registration of a new user. For each commenced registration, there is a "Start" entry. For each successfully finished registration, there is a "Finish" entry. There was not defined any other KPI to measure, so without storing any junk data, it is easy to calculate the percentage (2/3) of users that successfully registered versus that didn't (1/3).

<sup>1</sup>The "..." marks obvious missing columns, such as timestamp, device type etc.

### 3.2.2.1 DSL

In order to track more than just the beginning and the end of the workflow defined in the scope of the issue, it is necessary to give the stakeholders a way to define certain observable events to pay attention to. Usually, every issue has a field specifically for description, that contains some human-readable set of instructions. Why not piggyback on that and give it just enough structure to make it also computer-readable?

Let's assume it would be desired, to measure how many times the user clicks on the help button in the registration form. Then the database table would be extended to:

Issue ID	Name	Type	...
SA-01	Form for user registration.	Start	...
SA-01	Form for user registration.	Finish	...
SA-01	Help button pressed.	Action	...

Table 3.2: Extended Example Database Entries

Another type has appeared in the database table - "Action". Some measurable item, linked to the registration form itself. Neatly labeled with the relevant Issue ID, easy to find and in the same domain vocabulary as was intended.

## 3.3 Semantic Data Manager

This is the most crucial component of the whole solution - connecting the Issue Tracker and tracking capabilities. The main task is to obtain actionable items from the Issue Tracker and create computer-readable domain vocabularies for application engineers to include in their source code. The computer-readable format should in the form of a configuration file. Therefore, there must be a framework that parses such configuration file, and offers in-code assistance with implementing the tracking itself. It goes without saying that the usage should be as smooth as possible.

The produced configurations should be persisted for the sake of reproducibility. Though it should avoid storing any metadata (detailed measurement description) as it is subject to change (DSL used in the description). Only the unique issue identifier and its name should be persisted, in order to be able to trace back what was the configuration file meant for. Persisting anything else could lead to disconnection between data and that's exactly what I am trying to solve.

Connecting to the Issue Tracker will be handled via its REST API. It should automatically retrieve a list of all projects and their subsequent issues/tasks/sub-tasks that may or may not contain specific metadata regarding fine-grained measurement demands.

Semantic Data Manager will run as a stand-alone microservice and will provide a REST API for any kind of UI to be implemented on top of it. It can easily be a website, mobile application or integration into a 3rd party tool.

## 3.4 Tracking Engine

The Tracking Engine itself is a completely separate component from the Semantic Data Manager. It can easily be any existing solution - Google Analytics or Google Tag Manager. It should be a stand-alone service into which data is stored using the already obtained domain vocabularies.

For reasons listed in previous chapter, mainly in-house data storage, I will also develop this component and deploy it alongside the Semantic Data Manager. This gives me more flexibility in optimizing the storage and provides the company with a custom tailored solution.

## 3.5 Tracked Device SDK

The SDK has to have two layers:

1. Interpreter of the configuration file from the Semantic Data Manager - load the vocabularies and provide them during the development.
2. Interface for storing the events - should the Tracking Engine change, the only thing that would be necessary in the application code would be to download dependencies for it and implement the interface. This approach ensures that the in-code measuring always remains the same, only the underlying engine changes.

The whole company runs on iOS devices, so I will develop an iOS framework to be used while developing applications. While iOS may seem limited to only iPhones and iPads, the ecosystem actually enables the frameworks to be written for all Apple platforms - iOS, tvOS and Mac OS X.



## Chapter 4

# Implementation

### 4.1 Deployment

The whole solution is designed for and deployed in the Amazon Web Services (= AWS) cloud environment. It is a managed Virtual Private Cloud (= VPC) environment that is only accessible with company credentials and certificates.

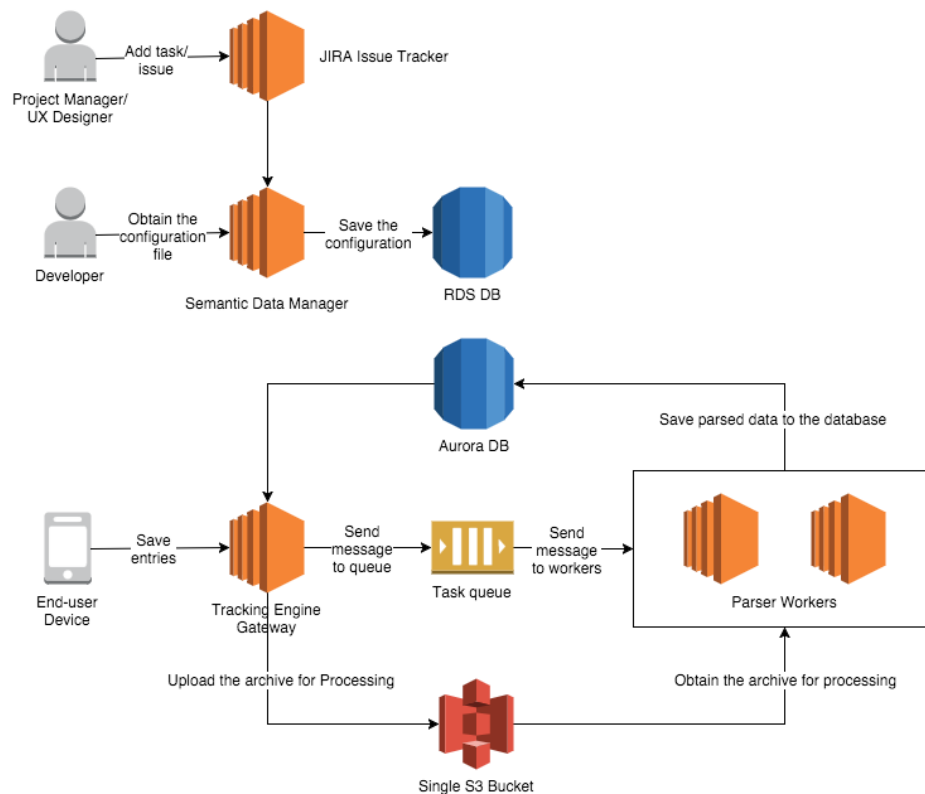


Figure 4.1: AWS System Architecture

## 4.2 Issue Tracker

The Issue Tracker used in my work environment is JIRA<sup>1</sup> by Atlassian. It is a standard project management tool providing bug tracking, issue tracking and many other functions. It is conveniently synergistic with other Atlassian tools such as Confluence (for documentation and wiki) and Bitbucket (formerly Stash), which is a server for version control (Mercurial and Git). The advantage is that all these three components are deployed in the AWS VPC environment. Thus, as a programmer I have fairly easy access to its internal APIs without being afraid to leak data where it shouldn't.

### 4.2.1 API

JIRA REST API is quite a mess, which is because Atlassian didn't develop all their products from scratch (Bitbucket was acquired) and it is still visible that the usage isn't seamless. In order to access Teams, Projects and Issues, two API endpoints have to be used:

1. JIRA REST API
2. JIRA REST AGILE

Both of them are APIs (= Application Programming Interface), but I will use their names to distinguish one from another. Both are similar, but also slightly different from each other.

To illustrate the subtle differences that drive any software engineer mad:

- When JIRA REST API is used to obtain the issues, the resulting array uses pagination, because there could be a lot of issues and loading them all at once could take a significant amount of time. In order to determine whether the array I have is final, a parameter "total" is present in the response. This parameter tells how many issues in total there are. In order to load the whole list, it is necessary to keep track how many there are, and how many are left on the stack.
- When JIRA REST AGILE is used to obtain the teams, the resulting array also uses pagination. In order to determine whether the array I have is final, a parameter called "isLast" is present in the response, having, surprisingly, a boolean value true/false. Obviously, when the value is false, one has to load the next page with the last index that came before.

There are plenty of these little surprises that are so easily breakable with any update of the whole system. I honestly do not know, why it isn't the top priority for Atlassian to unite their APIs.

What struck me most though, is the absence of OAuth or Token-based communication. Every query is done via basic auth. While for development it is fine as it allowed me to quickly prototype on top of the API without the need to develop a complex token manager, for production it is quite inconvenient. Even though SSL certificates are all valid and in place, it simply is a terrible architectural choice to not have a proper way to authenticate other applications using the APIs.

---

<sup>1</sup>No abbreviation here. It is short for GOJIRA, which is Godzilla in Japanese. Rumor has it that it is because the main competing product is Bugzilla.

#### 4.2.1.1 JQL

JQL stands for JIRA Query Language [12]. It enables the API user to query the JIRA knowledge graph and extract information.



Figure 4.2: JQL syntax

1. **Field** - Fields are different types of information in the system. JIRA fields include priority, fixVersion, issue type, etc.
2. **Operator** - Operators are the heart of the query. They relate the field to the value. Common operators include equals (=), not equals (!=), less than (<), etc.
3. **Value** - Values are the actual data in the query. They are usually the item for which we are looking.
4. **Keyword** - Keywords are specific words in the language that have special meaning. In this post we will be focused on AND and OR.

I used JQL in order to get all issues for a certain project:

---

["https://jiraURL/issues/search?jql=project=SAUI"](https://jiraURL/issues/search?jql=project=SAUI)

---

Here I used simple query to search all issues where the project is SAUI (= Semantic Analysis of User Interactions). All URL encoders handle the double "=" and it has never happened to me, that it would encode the parameters badly.

It can obviously be even more powerful, but I was glad it helped me easily get what I needed.

#### 4.2.1.2 Methods used

All communication is handled via HTTP GET and all responses are in JSON format. Cross-site request forgery (= CSRF/XSRF) token system is disabled.

1. To get all teams, method **/board** has to be called on JIRA REST AGILE.
2. To get all projects, method **/projects** has to be called on JIRA REST API.
3. To get all issues, method **/search** with JQL query has to be called on JIRA REST API.

Interestingly enough, even though, there are two API endpoints, the data is connected, so no further processing was necessary. It is important to note, that the responses are **very** verbose and it is possible to tell in the query to the server not to send some fields back.

### 4.2.2 DSL

After a discussion with a UX lead, I came up with a simple solution - add keyword "WATCH:" on a new line and describe what to observe. Parsing is done line by line where the code searches the line for "WATCH:" (case insensitive) and extracts whatever follows until the end of line or occurrence of another "WATCH:". In order not to make it complex, end of line is the end of any description, it does not carry over to the next line.

#### 4.2.2.1 Examples

Here are some examples how the parser for the DSL works. Validation of this technique will be covered in the Testing chapter.

##### Example 1 - Success

---

```
As a user, I want to be able to list all projects in the mobile app currently
    being tracked along with the number of tracked versions in the tracking system.
```

```
WATCH: Number projects expanded to the highest detail
WATCH: Filters used to extract information
```

---

This succeeds perfectly, as it parses everything without any hassle.

##### Example 2 - Success

---

```
As a user, I want to be able to list all projects in the mobile app currently
    being tracked along with the number of tracked versions in the tracking system.
```

```
Watch:    Number projects expanded to the highest detail
Watch:    Filters used to extract information
```

---

This succeeds too, because the search is case insensitive and after the search, white spaces are extracted, so the result is the same as in the previous example.

##### Example 3 - Semi-success

---

```
As a user, I want to be able to list all projects in the mobile app currently
    being tracked along with the number of tracked versions in the tracking system.
```

```
WATCH: Number projects expanded to the highest detail, Filters used to extract
    information
```

---

This is a semi-success, almost a failure, but it still yields all the information that the user wanted. It just isn't nicely separated and would need some changes. The error is visible to the programmer and is easy to fix.



**Example 4 - Failure**


---

As a user, I want to be able to list all projects in the mobile app currently being tracked along with the number of tracked versions in the tracking system. WATCH: Number projects expanded to the highest detail. WATCH: Filters used to extract information

---

This yields only one result - "Number projects expanded to the highest detail.". While it seems like it is a good solution, the fact that it seems that way is unfortunately the worst thing about it - because it is hard to discover that there is an error. The programmer sees one observable action item and doesn't see that some got lost during the process, because it was all on one line.

**4.3 Semantic Data Manager**

The central part of the project should be robust and reliable. For that reason I chose Java as the main technology. For convenience and standardization of the code-base I opted for Spring Boot framework to help me with bootstrapping the heavy work (scheduling, threading, persistence etc.).

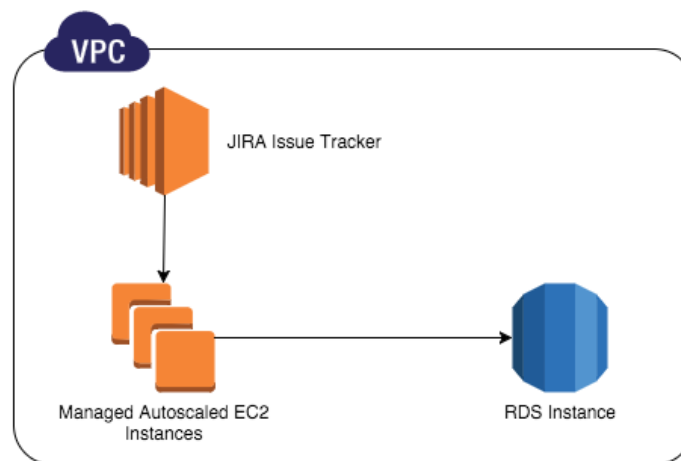


Figure 4.3: Semantic Data Manager Deployment

Semantic Data Manager is deployed in AWS Elastic Beanstalk (= EB), which is Amazon's Platform as a Service (= PaaS) for Java web applications. It reduces management complexity without restricting choice or control. All it takes is to upload the application, and Elastic Beanstalk automatically handles the details of capacity provisioning, load balancing, scaling, and application health monitoring. [3]

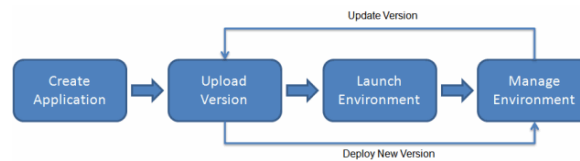


Figure 4.4: AWS Elastic Beanstalk Application Deployment

It is important that the EB application runs in the same availability zone of the VPC as JIRA, otherwise it wouldn't be reachable at all.

The database runs on managed Amazon RDS (Relational Database Service), which is fast, secure and scalable deployment of database engine. It manages backups, software patching, automatic failure detection, and recovery. The default database engine is MySQL.

### 4.3.1 Spring Boot

I first tried to use play2 framework for educational purposes, but I encountered too many obstacles deploying play2 application to AWS:

- It does not support WAR packaging.<sup>2</sup>
- It is not possible to run play2 packages (packaged by Activator tool) on Tomcat Server.
- It comes with its own Netty Server, which is really clumsy to set up in AWS environment.

All three combined resulted in inability to synchronize the play2 application on port 9000 and NGINX running on port 5000. Unfortunately Netty Server does not support compile-time port configuration and NGINX does not support running Activator to set up the port during run-time, so I had to drop the idea of using play2 as I was simply unable to deploy the application. After researching and discussing with my peers and coworkers I looked up Spring MVC and stumbled upon Spring Boot, also recommended by my classmate. I tried few sample apps and found out it supports WAR packaging, runs natively on Tomcat and comes with almost the same perks like play2. I was ready to give it a try.

"Spring Boot aims to make it easy to create Spring-powered, production-grade applications and services with minimum fuss. It takes an opinionated view of the Spring platform so that new and existing users can quickly get to the bits they need." [17]

Primary goals of Spring Boot are [16]:

- Provide a radically faster and widely accessible getting started experience for all Spring development.
- Be opinionated out of the box, but get out of the way quickly as requirements start to diverge from the defaults.

<sup>2</sup>There is an unofficial tool that packages the code in a WAR file, but it is not recommended for production environment. Being constrained by highly regulated market, something that already says that it is not production ready is an instant "No thanks".

- Provide a range of non-functional features that are common to large classes of projects (e.g. embedded servers, security, metrics, health checks, externalized configuration).
- Absolutely no code generation and no requirement for XML configuration.

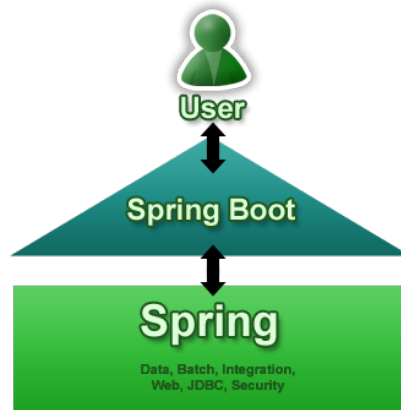


Figure 4.5: Spring Boot architecture

The advantages seemed to be strong, development environment was convenient (native support in IntelliJ IDEA) and after some validation with Amazon support regarding AWS EB deployment in our VPC environment, I was confident this would be a good choice.

### 4.3.2 Workflow

...

### 4.3.3 Technologies Used

In order to make the development as comfortable as possible, I used various open-source frameworks/technologies to help me out with things that would otherwise take a lot of time

#### 4.3.3.1 H2 Database

H2 is a Java SQL database that runs in-memory. It is great for fast prototyping and validating of design ideas before using production database (RDS). Due to the fact that the entire database is deleted after a server is restarted, it enabled me to quickly change the schema without the need to do tedious database wipes.

Another great thing is that it only needs to be defined in pom.xml and all other linking is provided out-of-the-box:

---

```
<dependency>
  <groupId>com.h2database</groupId>
  <artifactId>h2</artifactId>
</dependency>
```

---

Important note: as it wires up all connections automatically, it is absolutely vital not to have **any other** database engine loaded via Maven, otherwise it crashes on start-up of Tomcat.

#### 4.3.3.2 Flyway

Flyway is an open-source database migration tool. It aims to be clean and simple rather than robust and too complex. It uses 6 main commands:

1. Migrate - Migrates the schema to the latest version.
2. Clean - Drops all objects in the configured schemas.
3. Info - Prints the details and status information about all the migrations.
4. Validate - Validates the applied migrations against the available ones.
5. Baseline - Baselines an existing database, excluding all migrations upto and including baselineVersion.
6. Repair - Repairs the metadata table.

Since I didn't need to migrate the database, thanks to the flexibility of H2, I mainly used Flyway for cold start (empty database). I defined, what will be in the database after the server has started and Flyway filled it up for me.

#### 4.3.4 REST API

// provides nice API

#### 4.3.5 Code Deployment

The code is packaged by Maven and deployed as WAR file to AWS EB instance, running Tomcat 8 server. There is no need to configure anything with regards to basic networking - AWS EB is a back-to-back fully managed Platform as a Service (PaaS).

There are multiple things to consider when deploying to AWS EB, even though it seems "super-easy" in most instructional videos and ads:

1. Contrary to programmer's logic, one has to first create an "Application" and under that custom "Environments". In other words, Application is the main context, and environments are servers running in the same context, by default having the permission to communicate between each other.

2. In Environment setup, we can choose either a Web Server Environment or a Worker Environment. Web Server is the hub of any application and is used in both Semantic Data Manager and Tracking Engine. Workers are only used in Tracking Engine and will be explained later.
3. Chosen configuration was obviously Tomcat and I also opted for automatic load balancing and scaling.
4. Opting to automatically create an RDS instance along with the environment is a really bad design. Once an environment is terminated, so is the database instance which causes a serious data loss.
5. In order to access the servers via SSH, it is necessary to define a EC2 Security Group and assign it accordingly. Otherwise, all outside access is prohibited.
6. It is also crucial to wisely choose the instance type. I opted for m3.medium, because Java applications by themselves are quite demanding and I didn't want to risk being on the edge when strange errors occur because of insufficient memory capacity.
7. Permissions and roles are absolutely crucial when it is desired to connect the Environment with other AWS services, such as object storage (S3 = Simple Storage Service) or a messaging queue (SQS = Simple Queue Service).

The size of the instance is recommended for any application running JVM. The configuration is Intel Xeon E5-2670 v2 (Ivy Bridge), 4GB SSD storage and 3.75GB RAM. Any additional memory is handled via S3.

#### 4.3.6 User Interface

The component Semantic Data Manager provides REST API to be consumed by any kind of client capable of HTTP requests. Because my specialization are mobile applications, I chose to implement a mobile application for iOS as an administrating user interface for this component.

##### 4.3.6.1 Application Flow

SCREENSHOTS

##### 4.3.6.2 Technologies Used

Moya, Alamofire

## 4.4 Tracking Engine

Tracking Engine is also deployed in AWS EB but uses Aurora as database engine. Aurora is ... and has ... configuration, allowing to have really high traffic and maintain its speed and robustness.

Because the tracking engine was developed after the Semantic Data Manager, I opted to use Java again.

The code is deployed in the same fashion as the Semantic Data Manager.

## 4.5 Tracked Device SDK

TBD

### 4.5.1 Technologies Used

// Realm.IO // Alamofire // Moya // GZIP

# Chapter 5

## Testing

// black-box testing, UX testing

### 5.1 System Benchmarks

// methods used

#### 5.1.1 Tracking Engine

// json size, time to process

### 5.2 User Tests

#### 5.2.1 Qualitative

// test scenario + interview

#### 5.2.2 Quantitative

// test scenario benchmarking by dogfooding





## Chapter 6

# Conclusion

### 6.1 Future development

The current solution can be deployed in testing environment as an MVP (= Minimum Viable Product). In order for it to run in production environment, legal steps have to be carried out in order to be fully compliant.

### 6.2 Technical debt

Currently unknown

### 6.3 Closing remarks

TBD



# Bibliography

- [1] P. R. Appsee. Appsee partners with twitter.  
<https://www.appsee.com/press-releases/appsee-fabric-integration>, 8.12.2015.
- [2] W. Chang. Crashlytics is joining forces with twitter.  
<http://www.crashlytics.com/blog/crashlytics-is-joining-forces-with-twitter/>, 28.1.2013.
- [3] A. Documentation. What is aws elastic beanstalk?  
<http://docs.aws.amazon.com/elasticbeanstalk/latest/dg/Welcome.html>, 6.8.2013.
- [4] M. Haindl, Ľ. Kment, and P. Slavík. Virtual information systems. In *WSCG'2000 — Short communication papers*, pages 22–27. University of West Bohemia, Pilsen, 2000.
- [5] J. Highsmith and A. Cockburn. Agile software development: the business of innovation. *Computer*, 34(9):120–127, Sep 2001.
- [6] F. Krause. The easiest way to automate building and releasing your ios and android apps.  
<https://github.com/fastlane/fastlane>, up to date 22.5.2016.
- [7] B. Leonard. Tealium iq - the world's leading enterprise tag management solution.  
<http://tealium.com/products/tealium-iq-tag-management-system/>, 16.12.2015.
- [8] B. Leonard. Tealium's customer data platform achieves hipaa milestone.  
<http://tealium.com/press-releases/tealiums-customer-data-platform-achieves-hipaa-milestone>, 16.12.2015.
- [9] K. MacDonald. The data driven pharmacy today.  
<https://kitcheck.com/2015/06/data-driven-pharmacy-today/>, June.2015.
- [10] B. Matsugu. The limitations of linear methodology – waterfall or agile...or both?  
<http://www.blueprintsyst.com/the-limitations-of-linear-methodology-waterfall-or-agile>, 20.10.2014.
- [11] P. R. New Relic. People, process, privacy.  
<https://newrelic.com/why-new-relic/security>, up to date 22.5.2016.

- [12] D. Radigan. Jql: the most flexible way to search jira.  
<http://blogs.atlassian.com/2013/01/jql-the-most-flexible-way-to-search-jira-14/>, 29.1.2013.
- [13] C. Rawson. ios 5 deprecates udid as identifier for developers, but it's not the end of the world.  
<http://www.engadget.com/2011/08/19/ios-5-deprecates-udid-as-identifier-for-developers-but-i>, 19.8.2011.
- [14] J. Seibert. Introducing fabric.  
<https://blog.twitter.com/2014/introducing-fabric-india>, 24.10.2014.
- [15] P. Slavík. Grammars and rewriting systems as models for graphical user interfaces. *Cognitive Systems*, 4(3/4):381–399, 1997.
- [16] Spring boot reference guide.  
<http://docs.spring.io/spring-boot/docs/current-SNAPSHOT/reference/htmlsingle/>, up to date 16.5.2016.
- [17] P. Webb. Spring boot – simplifying spring for everyone.  
<https://spring.io/blog/2013/08/06/spring-boot-simplifying-spring-for-everyone>, 6.8.2013.
- [18] CSTUG —  $\mathcal{C}\mathcal{S}\mathcal{T}\mathcal{E}\mathcal{X}$  Users Group — hlavní stránka.  
<http://www.cstug.cz/>, stav z 2.3.2009.
- [19] A. Weber and I. R. Thomas. Key performance indicators. *Measuring and Managing the Maintenance Function*, Ivara, 2005.
- [20] K336 Info — pokyny pro psaní diplomových prací.  
<https://info336.felk.cvut.cz/clanek.php?id=400>, stav ze 4.5.2009.
- [21] Knihovna Grafické skupiny.  
<http://www.cgg.cvut.cz/Bib/library/>, stav z 30.8.2001.
- [22] Grafický vektorový editor pro práce vhodný pro práci  $\mathcal{L}\mathcal{A}\mathcal{T}\mathcal{E}\mathcal{X}$ em.  
<http://tclab.kaist.ac.kr/ipe/>, stav z 4.5.2009.
- [23]  $\mathcal{L}\mathcal{A}\mathcal{T}\mathcal{E}\mathcal{X}$  — online manuál.  
<http://www.cstug.cz/latex/lm/frames.html>, stav ze 4.5.2009.
- [24] Wiki books  $\mathcal{L}\mathcal{A}\mathcal{T}\mathcal{E}\mathcal{X}$ .  
<http://en.wikibooks.org/wiki/LaTeX/>, stav z 3.4.2009.
- [25] J. Žára, B. Beneš, and P. Felkel. *Moderní počítačová grafika*. Computer Press s.r.o, Brno, 1st edition, 1998. In Czech.

# Appendix A

## List of abbreviations

<b>API</b>	Application Programming Interface
<b>AWS</b>	Amazon Web Services
<b>CSRF/XSRF</b>	Cross-site Request Forgery
<b>EB</b>	Elastic Beanstalk
<b>EC2</b>	Elastic Compute Cloud
<b>JSON</b>	JavaScript Object Notation
<b>KPI</b>	Key Performance Indicator
<b>MVP</b>	Minimum Viable Product
<b>PaaS</b>	Platform as a Service
<b>PoC</b>	Proof of Concept
<b>RDS</b>	Relational Database Service
<b>S3</b>	Simple Storage Service
<b>SaaS</b>	Software as a Service
<b>SDLC</b>	Software Development Life Cycle
<b>UX</b>	User Experience
<b>WWDC</b>	The Apple Worldwide Developers Conference



## Appendix B

# Instalační a uživatelská příručka

Tato příloha velmi žádoucí zejména u softwarových implementačních prací.





## Appendix C

# Obsah příloženého CD

Tato příloha je povinná pro každou práci. Každá práce musí totiž obsahovat příložené CD. Viz dále.

Může vypadat například takto. Váš seznam samozřejmě bude odpovídat typu vaší práce. (viz [20]):

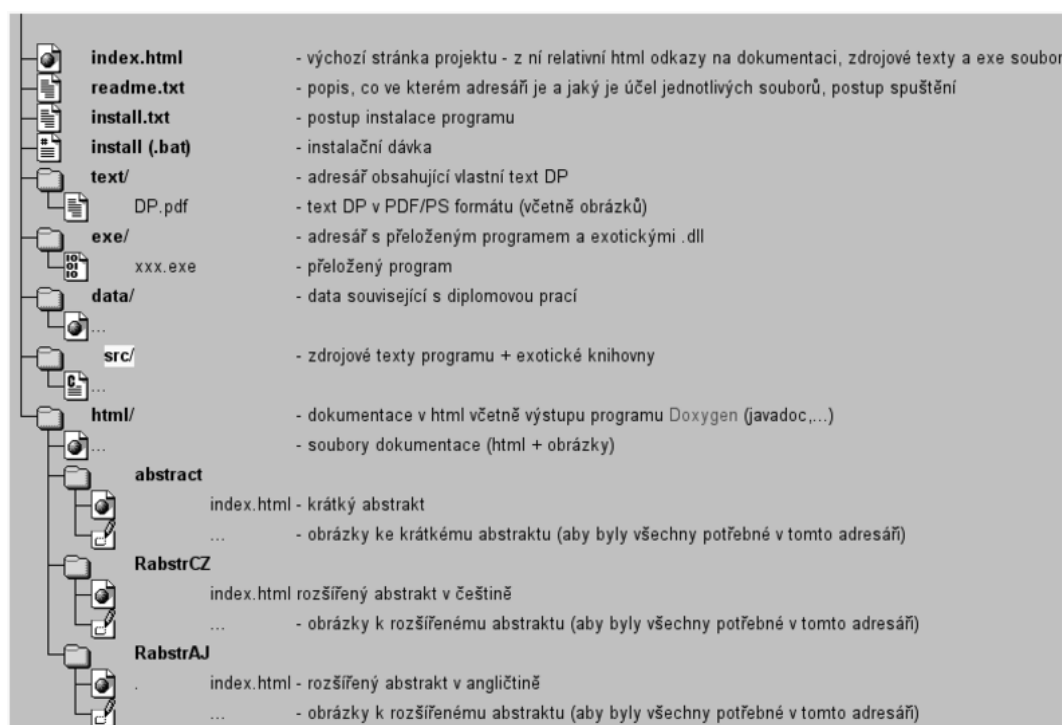


Figure C.1: Seznam příloženého CD — příklad

Na GNU/Linuxu si strukturu příloženého CD můžete snadno vyrobit příkazem:

```
$ tree . >tree.txt
```

Ve vzniklém souboru pak stačí pouze doplnit komentáře.

Z **README.TXT** (případně index.html apod.) musí být rovněž zřejmé, jak programy instalovat, spouštět a jaké požadavky mají tyto programy na hardware.

Adresář **text** musí obsahovat soubor s vlastním textem práce v PDF nebo PS formátu, který bude později použit pro prezentaci diplomové práce na WWW.

## Appendix D

# Pokyny a návody k formátování textu práce

**Tato příloha samozřejmě nebude součástí vaší práce. Slouží pouze jako příklad formátování textu.**

Používat se dají všechny příkazy systému  $\text{\LaTeX}$ . Existuje velké množství volně přístupné dokumentace, tutoriálů, příruček a dalších materiálů v elektronické podobě. Výchozím bodem, kromě Googlu, může být stránka CSTUG (Czech Tech Users Group) [18]. Tam najdete odkazy na další materiály. Většinou dostačující a přehledně organizovanou elektronikou dokumentaci najdete například na [23] nebo [24].

Existují i různé nadstavby nad systémy  $\text{\TeX}$  a  $\text{\LaTeX}$ , které výrazně usnadní psaní textu zejména začátečníkům. Velmi rozšířený v Linuxovém prostředí je systém Kile.

### D.1 Vkládání obrázků

Obrázky se umísťují do plovoucího prostředí **figure**. Každý obrázek by měl obsahovat **název** (`\caption`) a **návěští** (`\label`). Použití příkazu pro vložení obrázku `\includegraphics` je podmíněno aktivací (načtením) balíku `graphicx` příkazem `\usepackage{graphicx}`.

Budete-li zdrojový text zpracovávat pomocí programu `pdflatex`, očekávají se obrázky s příponou `*.pdf`<sup>1</sup>, použijete-li k formátování `latex`, očekávají se obrázky s příponou `*.eps`.<sup>2</sup>

Příklad vložení obrázku:

```
\begin{figure}[h]
\begin{center}
\includegraphics[width=5cm]{figures/LogoCVUT}
\caption{Popiska obrazku}
\label{fig:logo}
```

---

<sup>1</sup>`pdflatex` umí také formáty PNG a JPG.

<sup>2</sup>Vzájemnou konverzi mezi snad všemi typy obrázku včetně změn velikostí a dalších vymožeností vám může zajistit balík ImageMagick (<http://www.imagemagick.org/script/index.php>). Je dostupný pod Linuxem, Mac OS i MS Windows. Důležité jsou zejména příkazy `convert` a `identify`.



Figure D.1: Popiska obrázku

DTD	construction	elimination
	$\text{in1} A B \text{ a:sum } A \ B$ $\text{in1} A B \text{ b:sum } A \ B$	$\text{case}([\_ :A]a)([\_ :B]a)ab:A$ $\text{case}([\_ :A]b)([\_ :B]b)ba:B$
+	$\text{do\_reg:A} \rightarrow \text{reg } A$	$\text{undo\_reg:reg } A \rightarrow A$
*, ?	the same like   and + with <code>empty_el:empty</code>	the same like   and + with <code>empty_el:empty</code>
$R(a,b)$	$\text{make\_R:A} \rightarrow B \rightarrow R$	$a: R \rightarrow A$ $b: R \rightarrow B$

Table D.1: Ukázka tabulky

```
\end{center}
\end{figure}
```

## D.2 Kreslení obrázků

Zřejmě každý z vás má nějaký oblíbený nástroj pro tvorbu obrázků. Jde jen o to, abyste dokázali obrázek uložit v požadovaném formátu nebo jej do něj konvertovat (viz předchozí kapitola). Je zřejmě vhodné kreslit obrázky vektorově. Celkem oblíbený, na ovládání celkem jednoduchý a přitom dostatečně mocný je například program Inkscape.

Zde stojí za to upozornit na kreslicí programe Ipe [22], který dokáže do obrázku vkládat komentáře přímo v latexovském formátu (vzroce, stejné fonty atd.). Podobné věci umí na Linuxové platformě nástroj Xfig.

Za pozornost ještě stojí schopnost editoru Ipe importovat obrázek (jpg nebo bitmap) a krelit do něj latexovské popisky a komentáře. Výsledek pak umí exportovat přímo do pdf.

## D.3 Tabulky

Existuje více způsobů, jak sázet tabulky. Například je možno použít prostředí `table`, které je velmi podobné prostředí `figure`.

Zdrojový text tabulky D.1 vypadá takto:

```

\begin{table}
\begin{center}
\begin{tabular}{|c|l|l|l|}
\hline
\textbf{DTD} & \textbf{construction} & \textbf{elimination} & \\
\hline
 $\mid$  &  $\verb+in1|A|B$  a:sum A B+ &  $\verb+case([_:A]a)([_:B]a)ab:A+\backslash\backslash$  &  $\verb+in1|A|B$  b:sum A B+ &  $\verb+case([_:A]b)([_:B]b)ba:B+\backslash\backslash$  \\
\hline
 $\&\&$  &  $\verb+do\_reg:A \rightarrow reg A+\&\verb+undo\_reg:reg A \rightarrow A+\backslash\backslash$  \\
\hline
 $\ast,?\$$  & the same like  $\mid$  and  $\&\&$  & the same like  $\mid$  and  $\&\&\backslash\backslash$  & with  $\verb+empty\_el:empty+$  & with  $\verb+empty\_el:empty+\backslash\backslash$  \\
\hline
 $R(a,b)$  &  $\verb+make\_R:A\rightarrow B\rightarrow R+$  &  $\verb+a: R \rightarrow A+\backslash\backslash$  &  $\verb+b: R \rightarrow B+\backslash\backslash$  \\
\hline
\end{tabular}
\end{center}
\caption{Ukázka tabulky}
\label{tab:tab1}
\end{table}
\begin{table}

```

## D.4 Odkazy v textu

### D.4.1 Odkazy na literaturu

Jsou realizovány příkazem `\cite{odkaz}`.

Seznam literatury je dobré zapsat do samostatného souboru a ten pak zpracovat programem bibtex (viz soubor `reference.bib`). Zdrojový soubor pro bibtex vypadá například takto:

```

@Article{Chen01,
  author   = "Yong-Sheng Chen and Yi-Ping Hung and Chiou-Shann Fuh",
  title    = "Fast Block Matching Algorithm Based on
              the Winner-Update Strategy",
  journal  = "IEEE Transactions On Image Processing",
  pages    = "1212--1222",
  volume   = 10,
  number   = 8,
  year     = 2001,
}

```

```

@Misc{latexdocweb,

```

```

author = "",
title = "{\LaTeX} --- online manuál",
note = "\verb|http://www.cstug.cz/latex/lm/frames.html|",
year = "",
}
...

```

**Pozor:** Sazba názvů odkazů je dána BibTeX stylem (`\bibliographystyle{abbrv}`). BibTeX tedy obvykle vysází velké pouze počáteční písmeno z názvu zdroje, ostatní písmena zůstanou malá bez ohledu na to, jak je napíšete. Přesněji řečeno, styl může zvolit pro každý typ publikace jiné konverze. Pro časopisecké články třeba výše uvedené, jiné pro monografie (u nich často bývá naopak velikost písmen zachována).

Pokud chcete BibTeXu napovědět, která písmena nechat bez konverzí (viz `title = "{\LaTeX} --- online manuál"` v předchozím příkladu), je nutné příslušné písmeno (zde celé makro) uzavřít do složených závorek. Pro přehlednost je proto vhodné celé parametry uzavírat do uvozovek (`author = "..."`), nikoliv do složených závorek.

Odkazy na literaturu ve zdrojovém textu se pak zapisují:

```

Podívejte se na \cite{Chen01},
další detaily najdete na \cite{latexdocweb}

```

Vazbu mezi soubory `*.tex` a `*.bib` zajistíte příkazem `\bibliography{}` v souboru `*.tex`. V našem případě tedy zdrojový dokument `thesis.tex` obsahuje příkaz `\bibliography{reference}`.

Zpracování zdrojového textu s odkazy se provede postupným voláním programů `pdflatex <soubor>` (případně `latex <soubor>`), `bibtex <soubor>` a opět `pdflatex <soubor>`.<sup>3</sup>

Níže uvedený příklad je převzat z dříve existujících pokynů studentům, kteří dělají svou diplomovou nebo bakalářskou práci v Grafické skupině.<sup>4</sup> Zde se praví:

```

...
j) Seznam literatury a dalších použitých pramenů, odkazy na WWW stránky, ...
Pozor na to, že na veškeré uvedené prameny se musíte v textu práce
odkazovat -- [1].
Pramen, na který neodkazujete, vypadá, že jste ho vlastně nepotřebovali
a je uveden jen do počtu. Příklad citace knihy [1], článku v časopise [2],
statí ve sborníku [3] a html odkazu [4]:
[1] J. Žára, B. Beneš;, and P. Felkel.
    Moderní počítačová grafika. Computer Press s.r.o, Brno, 1 edition, 1998.
    (in Czech).

```

<sup>3</sup>První volání `pdflatex` vytvoří soubor s koncovkou `*.aux`, který je vstupem pro program `bibtex`, pak je potřeba znovu zavolat program `pdflatex` (`latex`), který tentokrát zpracuje soubory s příponami `.aux` a `.tex`. Informaci o případných nevyřešených odkazech (cross-reference) vidíte přímo při zpracovávání zdrojového souboru příkazem `pdflatex`. Program `pdflatex` (`latex`) lze volat vícekrát, pokud stále vidíte nevyřešené závislosti.

<sup>4</sup>Několikrát jsem byl upozorněn, že web s těmito pokyny byl zrušen, proto jej zde přímo necituji. Nicméně příklad sám o sobě dokumentuje obecně přijímaný konsensus ohledně citací v bakalářských a diplomových pracích na KP.

- [2] P. Slavík. Grammars and Rewriting Systems as Models for Graphical User Interfaces. *Cognitive Systems*, 4(4--3):381--399, 1997.
- [3] M. Haindl, Š. Kment, and P. Slavík. Virtual Information Systems. In *WSCG'2000 -- Short communication papers*, pages 22--27, Pilsen, 2000. University of West Bohemia.
- [4] Knihovna grafické skupiny katedry počítačů:  
<http://www.cgg.cvut.cz/Bib/library/>

... abychom výše citované odkazy skutečně našli v (automaticky generovaném) seznamu literatury tohoto textu, musíme je nyní alespoň jednou citovat: Kniha [25], článek v časopisu [15], příspěvek na konferenci [4], [www odkaz](#) [21].

#### D.4.2 Odkazy na obrázky, tabulky a kapitoly

- Označení místa v textu, na které chcete později čtenáře práce odkázat, se provede příkazem `\label{navesti}`. Lze použít v prostředích `figure` a `table`, ale též za názvem kapitoly nebo podkapitoly.
- Na návěští se odkážeme příkazem `\ref{navesti}` nebo `\pageref{navesti}`.

### D.5 Rovnice, centrováná, číslovaná matematika

Jednoduchý matematický výraz zapsaný přímo do textu se vysází pomocí prostředí `math`, resp. zkrácený zápis pomocí uzavření textu rovnice mezi znaky `$`.

Kód `$ S = \pi * r^2 $` bude vysázen takto:  $S = \pi * r^2$ .

Pokud chcete nečíslované rovnice, ale umístěné centrovane na samostatné řádky, pak lze použít prostředí `displaymath`, resp. zkrácený zápis pomocí uzavření textu rovnice mezi znaky `$$`. Zdrojový kód: `$$ S = \pi * r^2 $$` bude pak vysázen takto:

$$S = \pi * r^2$$

Chcete-li mít rovnice číslované, je třeba použít prostředí `equation`. Kód:

```
\begin{equation}
  S = \pi * r^2
\end{equation}
```

```
\begin{equation}
  V = \pi * r^3
\end{equation}
```

je potom vysázen takto:

$$S = \pi * r^2 \tag{D.1}$$

$$V = \pi * r^3 \tag{D.2}$$

## D.6 Kódy programu

Chceme-li vysázet například část zdrojového kódu programu (bez formátování), hodí se prostředí *verbatim*:

```

(* nickname2 *)
Lego> Refine in1
      (do_reg (nickname1 h));
Refine by in1 (do_reg (nickname1 h))
?4 : pcddata
?5 : pcddata
      (* surname2 *)
Lego> Refine surname1 h;
Refine by surname1 h
?5 : pcddata
      (* email2 *)
Lego> Refine undo_reg (email1 h);
Refine by undo_reg (email1 h)
*** QED ***
```

## D.7 Další poznámky

### D.7.1 České uvozovky

V souboru `k336_thesis_macros.tex` je příkaz `\uv{}` pro sázení českých uvozovek. „Text uzavřený do českých uvozovek.“