

School of Computing and Engineering

Final Year Project

Abstract

*Project Title:*

In depth study of

RESTful Web Services

*Undertaken By:*

Khan Najam ul Asre (21303779)

*Submitted On:*

??????

*Supervisor:*

??????

*Second Marker:*

??????

Contents

[1 Introduction 4](#_Toc513329854)

[1.1 Background 4](#_Toc513329855)

[1.1.1 Monolithic Applications 4](#_Toc513329856)

[1.1.2 Information Technology 4](#_Toc513329857)

[1.1.3 Distributed Systems 5](#_Toc513329858)

[1.1.4 Rise of Computing Devices 5](#_Toc513329859)

[1.1.5 Heterogeneity 6](#_Toc513329860)

[1.1.6 System Agility 6](#_Toc513329861)

[1.1.7 Cloud Computing 6](#_Toc513329862)

[1.1.8 Service Oriented Architecture (SOA) 7](#_Toc513329863)

[1.1.9 Microservices Architecture (MSA) 7](#_Toc513329864)

[1.1.10 Architectural Attributes 8](#_Toc513329865)

[1.2 Motivation 9](#_Toc513329866)

[1.3 Aims and Objectives 10](#_Toc513329867)

[1.3.1 Aim 10](#_Toc513329868)

[1.3.2 Objectives 10](#_Toc513329869)

[2 REST – A Better Style for Architecting Modern Applications 11](#_Toc513329870)

[2.1 What is REST 11](#_Toc513329871)

[2.1.1 Misconceptions About Rest 11](#_Toc513329872)

[2.2 Why REST 12](#_Toc513329873)

[2.2.1 Interoperability 12](#_Toc513329874)

[2.2.2 Network-based API vs. Library-based API 12](#_Toc513329875)

[2.2.3 Devices 12](#_Toc513329876)

[2.2.4 The Cloud 13](#_Toc513329877)

[2.3 Why Distributed Systems Fail 13](#_Toc513329878)

[2.3.1 Requirements-Driven Architecture 13](#_Toc513329879)

[2.3.2 Fallacies of Distributed Systems 13](#_Toc513329880)

[2.4 How REST Mitigates Failures 13](#_Toc513329881)

[2.4.1 Constraint-driven Architecture 13](#_Toc513329882)

[2.5 REST Constraints 13](#_Toc513329883)

[2.5.1 Client-Server 13](#_Toc513329884)

[2.5.2 Stateless 14](#_Toc513329885)

[2.5.3 Cache 15](#_Toc513329886)

[2.5.4 Uniform Interface 17](#_Toc513329887)

[2.5.5 Layered System 18](#_Toc513329888)

[2.5.6 Code on Demand (OPTIONAL) 19](#_Toc513329889)

[2.6 Richardson’s Maturity Model 19](#_Toc513329890)

[2.6.1 Level 0: Swamp of POX 20](#_Toc513329891)

[2.6.2 Level 1: Resources 20](#_Toc513329892)

[2.6.3 Level 2: HTTP Verbs 20](#_Toc513329893)

[2.6.4 Level 3: Hypermedia Controls 20](#_Toc513329894)

[2.6.5 Levels ‘towards the REST’ not ‘of the REST’ 20](#_Toc513329895)

[3 Uniform Interface 21](#_Toc513329896)

[3.1 Identification of Resources 21](#_Toc513329897)

[3.1.1 Resources 21](#_Toc513329898)

[3.1.2 Resource Identifier 22](#_Toc513329899)

[3.2 Manipulation of Resources Through Representations 22](#_Toc513329900)

[3.3 Self-Descriptive Messages 23](#_Toc513329901)

[3.3.1 Describing Request Message 23](#_Toc513329902)

[3.3.2 Describing Response Message 24](#_Toc513329903)

[3.3.3 HTTP Status Codes 25](#_Toc513329904)

[3.3.4 HTTP Methods 26](#_Toc513329905)

[3.4 Hypermedia As The Engine Of Application State (HATEOAS) 29](#_Toc513329906)

[4 Implementation 31](#_Toc513329907)

# Introduction

## Background

### Monolithic Applications

Early computers had limited processing, storage and communication capabilities. Despite, they were too expensive and too large in size for individuals to own or maintain. Hence, they were used only within the large organisations. The software systems were monolithic where functionally distinct concepts, e.g. data persistence and retrieval, business logic, user interface, error handling and logging, were strongly interwoven without any clear boundaries or architectural separation. Business requirements would change rarely and hence the software systems. It was very common that a version of software remained useful for the business for years, so monolithic applications worked really well.



Figure ‎1‑1Monolithic Architecture

### Information Technology

Over the last few decades, computing technology has seen dramatic advancement. Computing devices are becoming smaller and smaller in physical size, cheaper in cost but growing in computational power, data storage capacity and communication capabilities. The introduction of the Internet provided a global communication infrastructure. These factors together gave birth to the Information Technology which involves “the development, maintenance and use of computer systems, software and networks for the processing and distribution of data”[[1]](#footnote-1).

### Distributed Systems

The powerful devices and global communication network infrastructure revolutionised the business information and management systems. It was now possible for organisations to have geographically isolated locations with “autonomous computers, connected through a network and distribution middleware, which enabled computers to coordinate their activities and to share the resources of the system, so that users perceive the system as a single, integrated computing facility.”[[2]](#footnote-2)



Figure ‎1‑2 Distributed Software Architecture

### Rise of Computing Devices

Once thought to be of the interest of large enterprises, computing technology has now become a household commodity. With the introduction of smart hand-held devices computing devices are now personal pocket-items. This has changed software requirements of the business organisations. The market has become competitive. The businesses have to reach vast customer-base across the globe.



Figure ‎1‑3 Computing/Smart devices

### Heterogeneity

Lots of devices and manufacturers means lots of operating platforms and lots of software development frameworks. This poses the challenge of interoperability. Today’s ideal software systems have to be platform independent and capable of communicating and working with systems built using various frameworks and running on various platforms. As organisation’s customer base grows so does the need for system interoperability, to ensure that business is able to reach customer owning different devices running on different platform.

### System Agility

To keep going alongside the competitors, organisations have to change their marketing strategy and product presentation quickly and continuously. This requires the software systems that are agile and responsive, that can be changed quickly with or without the need of redeployment; or support Continuous Integration and Delivery.

### Cloud Computing

Traditionally businesses hosted their own on-premises computing infrastructure. For stable software systems this was feasible both financially and technically. But as the need grew for system agility companies start looking at maintenance and upgradation of on-site computing infrastructure as a continuous financial and technical pressure. This motivated the introduction of cloud computing where specialist organisation hosted and managed computing infrastructure which and be leased by the other business organisations. This shifted the responsibility of system maintenance and upgradation from consumer organisations to the cloud providers. Cloud offered kind of elastic resources that can grow or shrink on demand. The consumer organisations have to pay only what they consume. This is why a huge number of organisations have moved to cloud over last decade and process of migration to cloud still continuous. Although cloud offered a scalable infrastructure it does not come out of the box. The software architecture has to be cloud friendly to take full advantage of scalable cloud infrastructure.

### Service Oriented Architecture (SOA)

Unfortunately, monolithic applications lacked the architectural separation of concerns, therefore unable to become distributed systems. Monolithic software systems are mostly built for single platform using single framework. They are difficult, even impossible sometimes, to be changed or scaled. System designed to run on a single platform lack the interoperability, therefore reaching customer with devices running on different platform is not possible. Monolithic systems therefore failed to meet aforementioned challenges. They cannot take advantage of scalable cloud infrastructure. This lead the software architects to favour software systems composed of small, self-contained, independent and interoperable components rather than a giant monolithic system. Service oriented architecture (SOA) was one answer to such problems. SOA is a “system architecture in which application functions are built as components (services) that are loosely coupled and well-defined to support interoperability and to improve flexibility and reuse”[[3]](#footnote-3).



Figure ‎1‑4 Service Oriented Architecture

### Microservices Architecture (MSA)

“Microservices Architecture is basically Service Oriented Architecture done well”[[4]](#footnote-4). Microservices Architecture “is a method of developing software applications as a suit of independently deployable, small, modular services in which each service runs a unique process and communicates through a well-defined, lightweight mechanism to serve a business goal”[[5]](#footnote-5). Being able to meet almost all challenges of modern businesses, in the recent years MSA for many organisations has become a preferred architecture of creating enterprise applications. Martin Fowler notes Netflix, Amazon, eBay, Twitter, UK Government Digital Services and many other applications and websites have evolved from monolithic to MSA[[6]](#footnote-6).



Figure ‎1‑5 Microservices Architecture

### Architectural Attributes

Now that we have established some of the requirements of the modern software systems and challenges involved thereby, we sum up the properties or attributes of and ideal software architecture that supports distributed applications development in a way that they are easy to integrate and response to changes in the business quickly.

1. **Interoperability:** The application components should have ability to communicate and interoperate with each other regardless of the language tools and frameworks they are built with or platforms that they are running on. This will open up the vast possibility of integrating components built with different tools and technologies.
2. **Scalability:** The application should be able to scale itself should the need arise. This means that introduction of more infrastructural resources should proportionally enhance the system’s capacity to compute, communicate and response to the requests. This will ensure that businesses are capable of meeting quickly a sudden rise in demand.
3. **Evolvability:** The different component of the software system should be able to evolve independently. This will allow new features can be added quickly to keep up with the business’s market demands. Independent evolvability reduces the deployment and regression testing cost as only the changed components need to be tested and deployed.
4. **Visibility:** It should be possible that system performance and failures are visible without need for exposure of internal implementation details. This enables the addition of monitoring tools, load balancers and intelligent gateways without fearing the disclosure of proprietary estate.
5. **Reliability:** The system should be able to recover from full or partial failure and support graceful degradation should the need arise.
6. **Efficiency:** The system should be able to make efficient use of resources. It should be able to keep the server load to the minimum so the server component can manage their own resources efficiently.
7. **Performance:** Systems should be able to deliver responses to the requests quickly with minimum delay and improve overall perceived performance of the application.
8. **Manageability:** Systems should be easily manageable and should support the introduction of management tools.

## Motivation

Distributed Systems are advanced level of the architecture that started from Inter Process Communication (IPC). Different software vendors introduced, from time to time, different frameworks and tools. For example, Microsoft Component Object Model (COM) introduced in 1993 provided IPC and served as basis for some future Microsoft technologies. Java Remote Method Invocation (RMI) and Microsoft .Net Remoting next examples of technologies that worked provided Remote Procedure Call (RPC) capabilities thus influencing software systems to be composed of distributed processes. Enterprise Java Beans (EJB) and Microsoft Windows Communication Foundation (WCF) were further advancements that provided some level of scalability and facilitated development of Distributed Enterprise Systems. Almost all of these technologies were proprietary and hence could not inter-operate. SOAP based web services provided a unified standard that enabled interoperability between software sub systems built with different language tools/frameworks and running on different platform to communicate, coordinate and integrate in logically unified systems. SOAP based web services provided the interoperability but their very nature still had RPC mindset. Server publish service contract using Web Services Definition Language (WSDL). Clients shape themselves to conform to such service contracts. A change in service contract would mean modification and redeployment of all clients – a painstaking and expensive process that inhibited the system agility. None of these technologies elegantly and fully addressed the requirements and challenges mentioned in ‎1.1 above.

In 2000 Roy Thomas Fielding, in his PhD dissertation “Architectural Styles and the Design of Network-based Software Architectures”[[7]](#footnote-7) presented the concept Representational State Transfer (REST). Roy has been working on the WEB and the REST for over six years prior to the publication of his dissertation. Fielding was one of the main developers of HTTP Standard (RFC2616[[8]](#footnote-8)) and URI Specifications (RFC3986[[9]](#footnote-9)) and he developed REST to describe the architectural concepts behind the design of the Web. REST address all of the concerns mentioned above and provided standardised solutions to all of them. This is why RESTful services have now been recognised as generally the most useful methods to provide data-intensive services for web and mobile application development. A large number of business organisations have switched their application architectures to REST and rest are moving towards REST quickly. Therefore, there is strong technical as well as career motivation behind choosing this project to study RESTful architecture in depth and detail.

Next chapters of this report describe REST Architecture in detail and present a sample application architected in REST style to demonstrate the concepts.

## Aims and Objectives

### Aim

The aim of this project is to study and understand in greater detail the concept of RESTful  
architectural pattern of designing web services, thereby evaluating the RESTful services compared to  
other technologies like SOAP services and RPC.

### Objectives

1. To carry out in depth study of REST principles in order to understand how RESTful services differ from other options.
2. To build a data intensive web API software using REST architectural style to help  
   understanding and evaluation.
3. To gain a detailed understanding of the nature of HTTP protocol and how HTTP and RESTful  
   architectures are related to each other.
4. To make use of at least one HTTP client testing tool to gain a practical experience of how  
   HTTP (and thereby RESTful) communication works.
5. To understand the software project life cycle.

# REST – A Better Style for Architecting Modern Applications

## What is REST

In his dissertation, Fielding describes REST as:

“The name ‘Representational State Transfer’ is intended to evoke an image of how a well-designed Web application behaves: a network of web pages (a virtual state-machine), where the user progresses through the application by selecting links (state transitions), resulting in the next page (representing the next state of the application) being transferred to the user and rendered for their use.[[10]](#footnote-10)”

The description of REST above implies that first there was Web. Then it describes that how well-design system should behave. This also implies that REST is not a standard but is an architectural style. However, when we implement it we will use standards of course.[[11]](#footnote-11)

Although above description was given in context of web pages, principles of REST can be applied to any system. The idea is that client request a resource. The server returns the resource representation which is in a particular state. It also sends along controls that client can use the move the resource to next state. Figure ‎1‑1 depicts the process of Representational State Transfer. It begins with a task in New state. With this representation of task, server sends a control: Assign, which client can use to move the task to new state Assigned. An Assigned Task is another representation of the task. When an Assigned Task is requested by a client, server returns its presentation with tow controls: Complete and De-assign. Client can use Complete control to move task to Complete state or it can use De-Assign control to move it back to New state – an so on.



Figure ‎2‑1 State transfer

### Misconceptions About Rest

In literature it is often described that REST is not just another new way of calling remote procedure over the HTTP. This leads to the misconception that REST is any architecture that is not SOAP. This is not true. An architecture cannot be considered to be REST until it adheres to certain constraint. More on this later.

Almost of all the REST services use HTTP as protocol. This leads to another misconception that REST is HTTP. Theoretically this is not true. REST itself does not dictate any particular protocol. It leaves the selection of the standards and protocols with the implementer. The fact that almost all REST services on the Internet these days communicate over HTTP is no co-incidence. This is because Fielding who introduced REST was also one of the main developers of the HTTP Standard[[12]](#footnote-12). It is quite possible to write APIs that work on HTTP but still not RESTful.

Majority of the RESTful APIs use JSON (JavaScript Object Notation) to represent the resources. This lead to another misconception that an API that works with JSON is REST, which is not true. JSON is just one of many representations that can be used by a RESTful application. It is the media type that defines the representation. It can be JSON, XML, CSV or any custom format.

As said REST is an architectural style. The rest of this chapter explore that style and deriving constraints in detail, but first we will have a brief look at why we need REST.

## Why REST

The Chapter ‎1 already presents an abstract vision of what lead towards REST. Here we take a rather concrete account of some of key motivations to favour REST more.

### Interoperability

Interoperability refers to the ability to integrate various functional components built using different language tools and frameworks and running on different platforms. For example, a popular news website may have various elements that perform various functions. It may be providing search facility using Google integration, Advertisements provided by Ad Host integration, comments manged by integrating Disqus and sharing using integration of Facebook and Twitter. All these service providers need not be running on the same platform or built using single framework, therefore cannot make any assumptions about each other. Such wide integration requires a mechanism that is simple, consistent and reliable. This makes REST a good fit for the integration of heterogenous system.

### Network-based API vs. Library-based API

Rest comes up with the idea of network-based API rather than the Library-based API. Library-based API is mostly built using certain tools and framework that can run on a single or very few platforms. Network-based API on the other hand is not dependent upon a single platform or development framework. All it requires is to implement certain constraints that provide a standardised way of communication. If implemented correctly, such API offers unlimited interoperability between heterogeneous systems.

### Devices

The last decade saw a huge rise in smart devices. These devices may range from smartphone and tablets to in-car navigation to smart-tv and smart home air conditioning system. The businesses need to reach to the maximum devices. All devices need not be running web application. They have their own operating systems and native applications. In-car navigations system may request traffic data corresponding to GPS coordinates. The smart air conditioning controller may communicate with services provided by local meteorological office to maintain suitable in-house environment. Devices and services are provided by different vendors. They may be upgraded and evolve independent of each other and REST offers such independent evolvability.

### The Cloud

Rise in devices and increased interoperability poses another challenge to the software developers of systems that subscribe capabilities from other service providers and/or publish their own capabilities for other consumers. This opens up the possibility of dramatic rise in number of consumers. Over-subscription may push systems to the limits. To avoid such chaos, organisations tend to favour Cloud that provides elastic infrastructure that has ability to shrink or expand on demand, and they are only charged what they used. This provides financial benefits as well as scalability. But scalability of the cloud is not something out-of-the-box thing. “It is critical to build a scalable architecture in order to take advantage of a scalable infrastructure.[[13]](#footnote-13)” REST offers such scalable architecture.

## Why Distributed Systems Fail

Due to the rise in versatile smart devices, applications however small and simple tend to deviate from monolithic design and begin to resemble distributed systems. The Chapter ‎1 on Introduction describes in details the characteristics, requirements and challenges of modern software systems. It concludes that REST is a better approach for building modern distributed systems. Here we briefly discuss the key factors contributing to the failure of distributed systems.

### Requirements-Driven Architecture

Traditional RPC based architectures tend to take requirement driven approach. Business requirements are identified and software are designed to fulfil those requirements. When business requirements grow, design is grown to cover those new requirements. Such software designs are conceived and tested in a controlled environment and then deployed in the real environment. It is only after the deployment to the real environment that the limitations of the real environment are discovered that reduce the usability of the design, and eventually resulted in system failure. In requirements driven approach the architecture of the application is shaped to conform the business domain and result is an architecture that is tightly moulded within the shape of business.

### Fallacies of Distributed Systems

In 1994, at Sun Microsystems, Peter Deutsch identified seven assumptions that most of the architects of the distributed system tend to make. In 1997, James Gosling added another such fallacy[[14]](#footnote-14). Howard Dierking, a Pluralsight author, identifies yet another fallacy[[15]](#footnote-15). Unfortunately, these assumptions prove wrong in the long run, hence causing the system to fail. In literature these assumptions are known as Fallacies of Distributed Systems and are listed below (7 from Peter Deutsch, one from James Gosling and one from Dierking):

#### Network Reliability

Fallacy is that network is reliable. Obviously, this is not true. Power failure, hardware failure, people tripping over the cable – a whole lot of reasons to compromise network reliability.

#### Latency

Fallacy is that latency is zero. In one local area it might not be seen as a problem. But what if user is on other side of the globe? Even in case of local users, they might be using mobile devices with delayed response.

#### Bandwidth

Fallacy is that bandwidth is infinite. Although we now have much greater bandwidth then we ever had, but it is still finite. This is particularly true in the case of mobile device, where even there is large bandwidth, user may be charged for the bandwidth they use.

#### Security

The fallacy is that network is secure. Not all networks secure by default. And then this is the most overlooked aspect in practice of software development.

#### Network Topology

The fallacy is that the network topology never changes. In the modern Internet world, this is absolutely untrue. Servers and intermediaries keep moving. DNS, IP address and URLs keep changing. Even the relative paths and query strings at the server keep changing. Topology also regularly changed when applications are scaled and more hardware is added.

#### Administration

The fallacy is that there is one administrator. Obviously, this is not true in the case of applications distributed over the network. For example, when a remote or third-party service fail, local administrators have now access or control to diagnose the problem.

#### Transport Cost

The fallacy is that the transport cost is zero. This is also overlooked aspect during the development of the application. It is assumed that setting up a hardware and network infrastructure has zero cost or at least is one off cost. While in fact such cost is regular. Things like maintenance needs, upgradation, load balancing, bandwidth cost need for scalability contribute to regular cost.

#### Heterogeneous Network

The fallacy is that all nodes on the network is same. This is also false. We have seen already that in the modern world of computing, network and devices are not some. They are heterogeneous, particularly with the rise of variety of mobile devices.

#### Complexity

This fallacy assume that consumers of our service have enough domain and context knowledge of our service such that they will use our services correctly. This is obviously not true. In today’s Internet world, publisher and consumer may not know each other or may not have proper level of technical support available.

## How REST Mitigates Failures

The forces identified by Fielding in his dissertation that influence the system behaviour are closely inline with the fallacies of distributed systems discussed above. REST is defined by taking such forces into consideration.

### Constraint-driven Architecture

Contrary to requirement driven design REST on the other hand advocates constraint-driven architecture. It begins with identifying the forces that could impact the system usability. It then defines and apply constraints on the architecture design such that the impact of those forces can be eliminated or at least minimised. REST, therefore, requires the software developers to shape business domain to fit the architectural style. The resulted product is a design that works with rather than against those forces.

## REST Constraints

Now when we have established that REST was designed to work with forces that impact system behaviour rather than against them, and also that REST advocates constraint-drive approach, it is time to look into what actually those constraints are that define the RESTful style. We will also see what influencing forces those constraint address and what architectural attributes (‎1.1.10 above) are achieved as a result of enforcing each of the constraints.

### Client-Server

This constraint defines that all communication between any two nodes within the distributed system is considered being between client and server. Client sends request, then server performs some processing and sends the response back to the client. The goal of this constraint is separation of concerns.



Figure ‎2‑2Client-Server

This constraint Addresses the following concerns:

* Security: the scope of network security is narrowed down to the connections between client and server.
* Administration: the scope of administration is narrow down the connections between client and server.
* Complexity: Client know about the server but server has no knowledge of client. This decreases the complexity.

This lead to achieve following benefits:

* Client Interoperability: Since server and clients are separated and server doesn’t need to know anything about the client, this means client running different platforms can work with server.
* Scalability: Since server is separate from the client, it is possible to spin up multiple instances of server which can be load-balanced. This allows system to scale easily.
* Evolvability: Separation of concern means no dependency between client and server, so nodes can evolve independently.

### Stateless

Stateless constraint defines that server should get all the state information to process the request along with the request itself and must not rely on any context information saved on the server. The stat is actually maintained on the client and it is the responsibility of client to send any state information along with the request that server may need to process the request. As is evident from Figure ‎2‑3, it allows a great flexibility to the process workflow. A client can contact any server for subsequent requests as the requests are self-contained with the state which would not be possible if the state was maintained in server.



Figure ‎2‑3 Stateless systems

This constraint addresses the following influencing forces:

* Network Topology: This constraint allows nodes to be added and removed from the network without any risk of state corruption as each request brings required state with itself.
* Network Reliability: Since the state is maintained on the client, the system can conveniently recover from any network errors by starting with last known good state at the client.
* Administration: This constrain simplifies server administration because the administration does not have to worry about the state.
* Complexity: Addition of new nodes is simplified because it does not require the complex state management.

This constrain bring about following benefits:

* Visibility: Since request contains the state, it is visible to server and any intermediaries. This visibility opens up possibility of introduction of controls like intelligent gateways that can make smart decision using the state information.
* Reliability: This is the most valuable benefit of statelessness. Since state exists at one point at any given time, client, server and network can be recovered in a deterministic way in case of failure. For example, if client fails, it can request the last know representation from the server and take it from there. If server fails, the client can update it with the current state in next request.
* Scalability: Since the server does not have to remember the state, new server nodes can be added at any point during the workflow.

### Cache

Cache constraint requires that responses should be explicitly labelled as cacheable or non-cacheable. The cache control declaration with messages opens up possibilities of multi-level caching including server caching, client caching and caching on any intermediary devices. For example, in case of web application, responses can be cached at browser, at server, at a proxy server or any devices that sits somewhere in the route.



Figure ‎2‑4 Cacheable Architecture

This constraint addresses the following influencing forces:

* Bandwidth: Due to caching, the request may be responded from a cache before it reaches server in which case it will use less network segment. In the case of client cache there is no bandwidth usage at all.
* Transport Cost: Again, cache reduces the total number of network requests thus reducing the transport cost.
* Latency: Caching can significantly reduce the latency by eliminating the need to make some request or serving the request from a cache that is closed to the client.

The constraint brings following benefits:

* Efficiency: Caching makes the application more efficient with respect to both latency and bandwidth.
* Scalability: Caching allows the introduction of more client by simply scaling the workload over the entire network rather than the server alone.
* Performance: By caching responses, performance can be dramatically improved both by responding from the node that is nearest to the client and by reducing the processing cycles server has to perform to fulfil the request.

### Uniform Interface

This constraint is the key differentiator between RESTful and other architectures. In the early days of Web consistency between the all nodes of the network (i.e. clients, servers and intermediaries etc.) was maintained by requiring them to use the common client-server implementation library called CERN libwww[[16]](#footnote-16). The designer of the Web soon realised that providing consistency by way of enforcing certain implementation which is tightly coupled to the platform was not a scalable solution. Therefore, uniform interface constraint was devise to allow web scale reliably and quickly.



Figure ‎2‑5 Uniform Interface

Uniform interface is the largest constraint of the REST as well as the most important constraint from implementation point of view, therefore the whole next chapter is dedicated to Uniform Interface. However, for completion, a brief definition is given here.

The goal of the Uniform Interface was to provide a single, generalised and platform independent technical interface for the nodes of the distributed system to communicate with each other.

The Uniform Interface has four elements or sub-constraints:

#### Identification of Resources

The client consumes server capabilities by interacting with the resources which are identified by resource identifiers. A resource identifier uniquely identifies a resource. In typical HTTP terms this is a URL, however in theory it can be anything that uniquely identifies a resource.

#### Manipulation Through Representation

The client does not directly interact with resources. It rather interacts with the representation of those resources. A representation is separate from resource. When a client holds the representation of a resource, it should also have enough information about the resource so it can update or delete the resource (if API allows it).

#### Self-descriptive Messages

Each message must include description of itself that is enough for client or server to understand and/or process the message. For example, if the representation of the resource uses JSON then headers must state that the representation is JSON.

#### Hypermedia as the Engine of Application State (HATEOAS)

This is the constraint that most of the so called ‘RESTful’ APIs fail to implement. This constraint provides that client and server should be truly decoupled by server having to generate links and sending with response so the client can progress through the workflow of the application using those links.

The uniform interface constraint addresses the following influencing forces:

* Network Topology: existence of uniform interface for components to communicate with each-other allows system components to be created, added, removed and evolved independently.
* Administration: Since the interface is uniform, generalised tools can be created to investigate, manage and optimised network which facilitates administration. We will use Postman to test/demonstrate our artefact.
* Heterogeneous Network: The uniform interface lifts the platform dependency facilitating the interoperability on heterogeneous networks.
* Network Reliability: consistent communication semantics facilitate client and server to reliably recover from failures.
* Complexity: The complexity of the network application is limited to the complexity of the uniform interface.

This constraint helps achieving following attributes:

* Visibility: Uniform interface means that a message has same meaning for every component of the system involved in processing it without the need of any extra information.
* Evolvability: Uniform interface enables the system components to be upgraded or completely replaced without compromising the system stability.

### Layered System

This constraint requires that RESTful architecture can comprise multiple layers. It further enforces that a component in one layer should have knowledge of only the components in the next layer and not beyond that. Obviously, the introduction of intermediary layer can add to latency. However, this can be mitigated by using the advantages offered by other constraints, for example, by introducing intermediate caches and load balances etc.

This constraint addresses the following influencing forces:

* Network Topology: Changing a component affects only the components in the immediate layer and not beyond that.
* Complexity: The fact that a component can only know about and interact with the components in the immediate limits the magnitude of complexity.
* Security: Layers can be introduced on trust boundaries to hide layers inside the boundary from the layers outside the boundary.

This constrain brings about following properties:

* Scalability: Layered system means that scalability gets extended beyond the local resources.
* Manageability: Layered system enables the administration scope to be reduced and favours isolated managing entities. Managers at one layer only know and manage that layer and not the layers beyond.

### Code on Demand (OPTIONAL)

This constraint has been described as optional in Fielding’s dissertation. What this constraint says is that the along with resource representation and metadata, the server can also send code to the client so that the client to extend its functionality. By optional it may also imply that if this constraint is implemented, the code should not be an essential for the client to make progress. In today’s web applications the server sends extensive code in the form of JavaScript, particularly since the rise of JavaScript libraries like jQuery, Angular and React etc. However, it is not very common in the APIs. With the introduction of frameworks like NodeJS which enable JavaScript execution on native platform, it can be reasonably speculated that future API may start to make use of this constraint.

## Richardson’s Maturity Model

The Richardson’s Maturity Model was developed my Leonard Richardson and has now become a scale to measure the maturity of an API. This model has got attention recently by books like Rest in Practice and authors like Martin Fowler[[17]](#footnote-17).



Figure ‎2‑6 Richardson's Maturity Model[[18]](#footnote-18)

This model defines four levels of the maturity of an API.

### Level 0: Swamp of POX

An API at level 0 just works with a swamp of Plain Old XML (POX). Such APIs usually have single endpoint and are mostly RPC style. HTTP is used just for the remote interaction and no other HTTP capabilities are used. A single HTTP verb, e.g. POST is used to get, add, edit or delete information. Different procedures are called through a single URL and plain XML propagates as requests and responses. SOAP services are one example that lives at level 0.

### Level 1: Resources

APIs at this level have the notion of resources. These APIs use URIs and each URI uniquely identifies a resource. However, at this level APIs still not use the HTTP verbs as specified in the standard.

### Level 2: HTTP Verbs

APIs at this level are already at the previous level. In addition, they make uses of HTTP verbs like GET, PUT, POST, DELETE and so on for the purposes specified in the HTTP standard. Also, the responses contain the correct HTTP status codes to indicate the status of the response.

### Level 3: Hypermedia Controls

The APIs at this level support the Hypermedia as the Engine of Application State (HATEOAS). Request to GET a resource receives the requested resource as well as links that drive the application state (More on this in chapter on Uniform Interface).

### Levels ‘towards the REST’ not ‘of the REST’

It is important to note here that levels of Richardson’s Maturity Models are steps toward the REST and not the levels of the REST. This means that the only API that qualify as RESTful is the one at the Level 3 already. Any API below this level theoretically is not RESTful. This is the reason most of the APIs, even the famous ones, that claim to be RESTful are not actually RESTful just because those do not implement HATEOAS.

# Uniform Interface

Uniform interface constraint requires that communicating components share one single technical interface. Uniform Interface is the most important key constraint that differentiates REST from other architectural styles. The major part of REST implementation goes to this constraint. The goal of this constraint is to decouple components that communicate with each other to form a system. The level of decoupling has to be sufficient for components to evolved independently and fully interoperate with each-other. This in turn enables systems to scale quickly. In contrary to library-based API where components maintain communication consistency by adhering to the contract specifically provided by certain library, the purpose of Uniform Interface was to apply more general software design principles to the communication between distributed components.

The use of interface to remove dependency between classes in object-oriented programming is not a new concept. Uniform Interface is same concept on a larger scale, not to remove dependency between classes, but between communicating components such as clients and servers. As said earlier, REST architectural style has the concept of network-based API rather than library-based API. The interface in case of a library has limited scope, is specific to that particular library and known thorough it’s documentation. The Uniform Interface constraint was developed to emphasise the use of some globally known standard that can be implemented on any platform using any language tools, so that nodes in the Internet sized systems, developed, managed, deployed and evolved independently, can communicate with each other using such globally standardised interface. Although REST is not tied to a specific protocol, HTTP was one protocol that was designed according to the REST principles. This is why almost all RESTful services use HTTP as global standard, and academically REST principles are elaborated using the HTTP as an example.

The Uniform Interface constraint is composed of four sub-constraints.

## Identification of Resources

This sub-constraint requires that individual resources are identified in the request. Two things here: Resources and Resource Identifiers.

### Resources

Resources are concepts and not the entities as rows in the database table. However, they are related to entities. Resources map concept to one or more entities over time. There can be many-to-many relationship between resources and entities. While entities are internal implementation details, resources are concepts known to the outer world. This isolation of concept and implementation provides the abstraction required for decoupling. The resources are conceptually separated from the representation that is sent to the client. A representation of resource can be XML, JSON, HTML or any custom format if API supports it. A resource does not map to an entity. It may map to more than one entities in more than one database or even representation returned by other services.

The example in Figure ‎3‑1, taken from our implementation, demonstrates this mapping of resources on internal entities. As we can see in this example that the resources Account, Profile and Credential all are mapped to an internal entity User. Conversely, the resource Task maps to internal entities User, Group and Task.



Figure ‎3‑1 Resources-to-entities relation/mapping

### Resource Identifier

Resource identifier is a piece of information that uniquely identify a resource. A resource identifier can be anything, however it must be an agreed upon standard. In the Internet world, RFC3986[[19]](#footnote-19) (URI: Uniform Resource Identifier) standard is the globally agreed standard, hence used by almost all of RESTful services. In REST the URI space is owned by the server. Since the resource is an abstract concept, the resource identifier should not be frequently changed. However, mapping between resource and internal entities can change. This ensures that when server evolves, any clients who has subscribed to the URI space do not break. As an example of URIs, following are URIs for the example presented in Figure ‎3‑1:

/api/accounts/{user-name}

/api/accounts/{user-name}/profile

/api/groups/{group-id}/tasks/{task-id}

/api/accounts/{user-name}/credentials

## Manipulation of Resources Through Representations

An identified resource can be returned in various formats such as HTML, XML, JSON, PNG and so on. These formats are different representations of the resource. This constraint defines that when a client has a representation of the resource, it can use this representation (partially or fully) along with any metadata to add, update or delete the resource. Figure ‎3‑2 is a screenshot of Postman (a generic HTTP client) taken from our implementation. We can see that the client has JSON representation of Task resource and it is using HTTP PUT method and send this representation to the resource URI to update the Task.

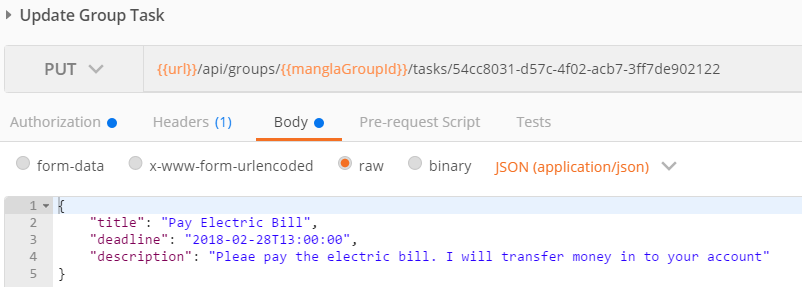


Figure ‎3‑2 Manipulating a resource through representation

## Self-Descriptive Messages

Both a request to server and response to client are messages. This constraint requires that the message should be self-descriptive. That means that the recipient of the message should receive enough information along with the message itself in order to understand the message and any operation that is required to perform. Put in other words, the messages should be state-less or context-less[[20]](#footnote-20). This means that the recipient should get all the information (description) with the message itself so it can understand the message and/or perform any required operation without the need of any contextual information from other than the message itself.

### Describing Request Message

Following is an example request message from our implementation. We see how it describes itself using HTTP standards.

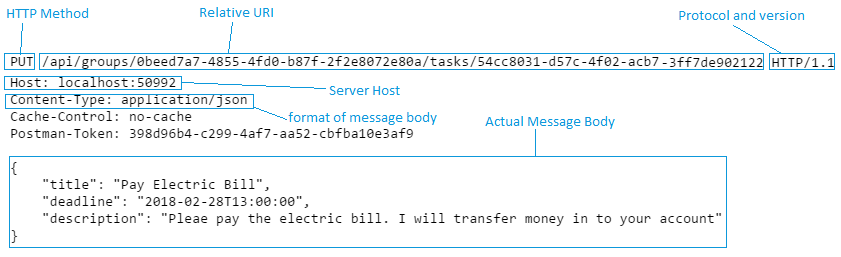


Figure ‎3‑3 Self-descriptive request message

* **Protocol and Version:** Tells the recipient that this message is sent using HTTP version 1.1 so should be interpreted using this protocol and version.
* **HTTP Method:** Tells the recipient that what to do with message. In this example it says PUT which means that if a resource exists at the given URI then update it using the resource representation in the message body, or insert the resource represented in the body at the give URI.
* **Relative URI:** This is the identifier that uniquely identifies the resource under question.
* Server Host: This HTTP header tells that this message is to the server hosted with this domain name.
* **Format of the message body:** This HTTP header tells the recipient that the message in the body is JSON representation of the resource.
* **Actual Message Body:** the message itself.

### Describing Response Message

Following is an example response message (Response to request given in Figure ‎3‑3). The example demonstrates how message is described using HTTP Standard.

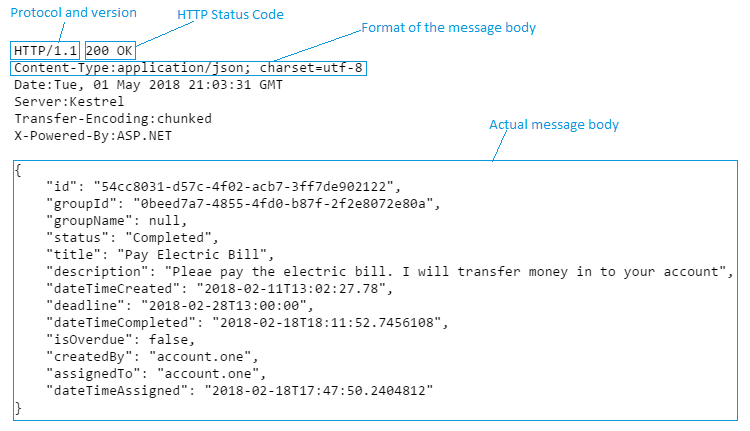


Figure ‎3‑4 Self-descriptive response message

* **Protocol and version:** Describes that this message should be interpreted as per HTTP 1.1 Standard
* **HTTP Status Code:** Describes that server was able to successfully fulfil the request (200 OK).
* Format of the message body: Describes that message body is JSON representation of the resource.
* **Actual message body:** Shows the “state” of the resource after update.

### HTTP Status Codes[[21]](#footnote-21)

When a client makes a request to the server and gets a response back, the client cannot make any assumption as to whether the request was successfully fulfilled or otherwise there had been a problem at the server. HTTP defines standard list of status codes. The server can and should send the appropriate status code with the response message in order for the requester to know what was happened to the request. As mentioned earlier, status codes are part of message description, hence conformance to the self-descriptive messages requires status codes being sent to the request originator.

HTTP Standard RFC2616 Standard defines a long list of status codes. Status code is a number with some associated meanings. RFC2616 defines various classes of status codes. The right-most digit of the status code indicates the class. Here are the classes defined by the HTTP Standard.

#### Informational 1xx

This class of status codes indicates provisional information response.

#### Successful 2xx

This class of status codes indicates that the client’s request was successfully received, understood and accepted.

#### Redirection 3xx

This class of status codes indicates that further action needs to be taken by the user agent in order to fulfil the request.

#### Client Error 4xx

This class of status code sis intended for cases where the client seems to have erred.

#### Server Error 5xx

This class of status codes indicate cases in which the server is aware that it has erred or is incapable of performing the request.

#### Status Codes Used

The following table lists the HTTP Status Codes used in our implementation

Table 1 HTTP Status Codes used in our implementation

|  |  |  |
| --- | --- | --- |
| Status Code | RFC2616/4918 Description | We used for |
| 200 OK | The request has succeeded. | Successful GET request  Successful HEAD request  Successful PUT request for update  Successful POST request for logon\* |
| 201 Created | The request has been fulfilled and resulted in a new resource being created. | Successful POST request for creation  Successful PUT request for creation |
| 204 No Content | The server has fulfilled the request but does not need to return an entity-body | Successful OPTIONS request  Successful DELTE request  Successful PUT request for change password\* |
| 400 Bad Request | The request could not be understood by the server due to malformed syntax. The client SHOULD NOT repeat the request without modifications. | Expected request message body missing  Malformed request message body |
| 401 Unauthorized | The request requires user authentication. | When request required a logged-on user, but no user is logged on. |
| 403 Forbidden | The server understood the request but is refusing to fulfil it. | When logged-on user is not authorised to a resource |
| 404 Not Found | The server has not found anything matching the Request-URI. | When request referred to a resource that did not exist. |
| 409 Conflict | The request could not be completed due to a conflict with the current state of the resource. | When request is not valid for the current state of the resource, for example, a user tries to assign a task that has already been completed |
| 422 Unprocessable Entity[[22]](#footnote-22) | the server understands the content type of the request entity (hence a 415 (Unsupported Media Type) status code is inappropriate), and the syntax of the request entity is correct (thus a 400 (Bad Request) status code is inappropriate) but was unable to process the contained instructions. | Validation failed on the request message |
| 500 Internal Server Error | The server encountered an unexpected condition which prevented it from fulfilling the request. | When unexpected exception occurs for which the code does not provide a handling |

### HTTP Methods

In the RPC world, when caller makes a call, the procedure being called is supposed to implement the functionality of the operation intended. In the REST style there is no concept of function or procedure, instead the user agent makes request for the resource identified by the URI the request was made to. When it comes to know what to do with the request, the request message brings with itself what is called HTTP method. It is this method that describes the operation to be performed by the server. Section 9 of HTTP Standard defines a set of methods. Section 9.1 t defines two important concepts associated with the HTTP methods.

#### Method Safety and Idempotency

##### Safe Method

A method is said to be safe if invoking that method does not cause any side effect. In simple words, safe method does not make any change to the state of the resource. However, it does not mean that server always returns the same response. When a client makes a request using a safe method, it can do so with the explicit knowledge that it is not going to make any change to state of the resource

##### Idempotent Method

A method is said to be idempotent when invoking the method many times has the same effect as if it was called once.

The table below summarises the safety and idempotency of commonly used HTTP methods.

Table 2 HTTP Methods' safety and idempotency

|  |  |  |
| --- | --- | --- |
| METHOD | Safe | Idempotent |
| OPTIONS | Yes | Yes |
| GET | Yes | Yes |
| HEAD | Yes | Yes |
| POST | No | No |
| PUT | No | Yes |
| PATCH | No | No |
| DELETE | No | Yes |

#### OPTIONS

This method requests the communication options available for the resource identified by the requested URI. Response of OPTIONS request usually does not contain a message body. It usually returns “Allow” header with the comma-separated list of HTTP methods available for the identified resource.

* This method should be safe, i.e. should not cause a change to the resource identified by the request-URI
* This method should be idempotent, i.e. same outcome of invoking multiple times should be same as invoking single time
* It should return 200 Ok, 204 No Contents for the found resource, or 404 Not Found if the resource at the request-URI was not found.
* It can return other appropriate status codes in cases of, for example, authentication/authorisation failures or malformed request etc.

#### GET

This method requests the transfer of the current representation of the resource identified by the URI. The GET method can be scoped to a single resource or collection resource. The response to a GET request contains the current representation of the resource in the message body. It may also return specific headers describing the message/metadata.

* This method should be safe, i.e. should not cause a change to the resource identified by the request-URI.
* This method should be idempotent, i.e. same outcome of invoking multiple times should be same as invoking single time
* It should return 200 Ok for the found resource, or 404 Not Found if the resource at the request-URI was not found.
* It can return other appropriate status codes in cases of, for example, authentication/authorisation failures or malformed request etc.

#### HEAD

This method requests the resource metadata. The response to the HEAD method should be identical to that of GET request except it MUST NOT contain the message body. The header/metadata should be exactly same as a GET request would return for the same resource. This method is used to obtain the resource metadata without transferring the resource representation.

* This method should be safe, i.e. should not cause a change to the resource identified by the request-URI.
* This method should be idempotent, i.e. same outcome of invoking multiple times should be same as invoking single time
* It should return 200 Ok for the found resource, or 404 Not Found if the resource at the request-URI was not found.
* It can return other appropriate status codes in cases of, for example, authentication/authorisation failures or malformed request etc.

#### POST

This method tells the server that it should accept the enclosed representation as the subordinate of the resource identified by the request-URI. It is usually used to insert a resource to the resource collection. The PUT request must contain the representation of subordinate resource in its message body. The PUT request should not dictate the server the URI of the resource and server may decide the URI itself.

* This method is not safe as it can create a new subordinate resource to the resource identified by the request-URI hence changing its state.
* This method is not idempotent as calling multiple times inserts new resource at each call.
* It should return 201 Created if successful, 404 not found if resource at request-URI not found.
* It can return other appropriate status codes in cases of, for example, authentication/authorisation failures or malformed request etc.
* It should preferably return the representation of created resource.
* It should return the URI of newly created resource as a HTTP header “Location”.

#### PUT

This requests that the enclosed representation should be stored in the supplied request-URI. This means that PUT request must contain the representation in its message body as well as should dictate the URI of this representation. The server should create the resource at the give URI if it did not already exist or replace the existing resource at the URI with the supplied representation. This method is usually used for full resource update.

* This method is not safe as this can replace the existing resource which is essentially a change in state.
* This method is idempotent as invoking it more than one time has no more effect than the first invocation.
* It should return 200 OK, 204 No Content for update and 201 Created if no resource already existed at the request-URI.
* It can return other appropriate status codes in cases of, for example, authentication/authorisation failures or malformed request etc.
* It can optionally return the representation of updated/created resource in which case it must not return 204 No Content.

#### PATCH

This method is used to make partial updates to an existing resource. This means that it should send a representation that can be used to identify the requested changes to the existing resource state. We used the JSON Patch Specification[[23]](#footnote-23) to format our patch document.

* This method is not safe as it causes the change in state of the resource identified by the request-URI.
* This method is not idempotent as, for example, the patch document might dictate “Add” to a subordinate of the resource identified by the request-URI.
* This method should return 200 OK or 204 No Content. It should return 404 Not Found if the request-URI does not indicate an existing resource.
* It can return other appropriate status codes in cases of, for example, authentication/authorisation failures or malformed request etc

#### DELETE

This method requests that the server deletes the resource identified by the request-URI. Although it can be used for both single and collection resources, using it for collection is considered to be catastrophic and hence is not advisable.

* This method is not safe and deleting a resource essentially means change in resource.
* This method is idempotent as it has no effect after first invocation.
* It should return 200 OK, or 404 Not Found if the no resource at the request-URI was found
* It can return other appropriate status codes in cases of, for example, authentication/authorisation failures or malformed request etc.

## Hypermedia As The Engine Of Application State (HATEOAS)

This is key constraint that distinguishes the RESTful systems from the other architectural styles. What this constrain means is that how a client should interact with the application is through the hypermedia controls dynamically generated and provided by the server. The HTML standard already provides means for server to provide the controls necessary to transition the state. The most common is HTML anchor <a …> tag, that has a href and rel attribute. The href provides the link to next transition while rel attribute tells the client about the relation this link has with the currently represented resource.

HATOAS does not stop there. At it’s sophisticated level, it even provides controls to shape up that data the client has to submit to the server HTML <form...> element is an example. The server can dynamically generate the input elements and client can compose those values in to the request message.

However most of the representation other than HTML do not provide as such a standard for the hypermedia controls. The API designer can come up with a reasonable and comprehensible custom format for the links.

This constraint offers a great independence and loose coupling between client and server. The idea is that the client does not have to know or know minimum about how to consume or use the API. The server provides such information on the fly. All the client has to do is just to follow those dynamic links. This allows the server control over the application workflow that it can change any time without the fear or breaking clients. In an ideal situation, the client should only know about the root document. Then this document has the initial links to other documents.

Having server in charge of controlling the control also means that server can control things like whether a logged-on user has the access to certain resource. It then adds or removes the further links to depending upon current user’s access privileges. In other words, HATEOAS allows for a self-documenting API.

Despite being the core constraint of the REST, this is the constraint most of APIs fail to implement. Having not implemented this constraint does not mean that API is not good, however it is important to understand the importance of this constraint and benefits it brings to the REST.

# Implementation

1. https://www.merriam-webster.com/dictionary/information%20technology [↑](#footnote-ref-1)
2. http://www0.cs.ucl.ac.uk/staff/ucacwxe/lectures/ds98-99/dsee3.pdf [↑](#footnote-ref-2)
3. Service-Oriented Architecture (SOA) Compass: Business Value, Planning, and Enterprise Roadmap Bieberstein, Bose, Fiammante, Jones, & Shah, 2006 [↑](#footnote-ref-3)
4. https://www.pluralsight.com/courses/microservices-architecture [↑](#footnote-ref-4)
5. https://smartbear.com/learn/api-design/what-are-microservices/ [↑](#footnote-ref-5)
6. https://martinfowler.com/articles/microservices.html [↑](#footnote-ref-6)
7. https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding\_dissertation.pdf [↑](#footnote-ref-7)
8. https://tools.ietf.org/html/rfc2616 [↑](#footnote-ref-8)
9. https://tools.ietf.org/html/rfc3986 [↑](#footnote-ref-9)
10. https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding\_dissertation.pdf [↑](#footnote-ref-10)
11. https://app.pluralsight.com/library/courses/asp-dot-net-core-restful-api-building/table-of-contents [↑](#footnote-ref-11)
12. https://tools.ietf.org/html/rfc2616 [↑](#footnote-ref-12)
13. https://media.amazonwebservices.com/AWS\_Cloud\_Best\_Practices.pdf [↑](#footnote-ref-13)
14. http://java.sys-con.com/node/38665 [↑](#footnote-ref-14)
15. https://app.pluralsight.com/library/courses/rest-fundamentals [↑](#footnote-ref-15)
16. https://www.w3.org/Library/ [↑](#footnote-ref-16)
17. https://martinfowler.com/articles/richardsonMaturityModel.html [↑](#footnote-ref-17)
18. https://martinfowler.com/articles/richardsonMaturityModel.html [↑](#footnote-ref-18)
19. https://tools.ietf.org/html/rfc3986 [↑](#footnote-ref-19)
20. http://exyus.com/articles/rest-the-short-version/ [↑](#footnote-ref-20)
21. https://tools.ietf.org/html/rfc2616#section-9 (page 57 section 10) [↑](#footnote-ref-21)
22. https://tools.ietf.org/html/rfc4918 [↑](#footnote-ref-22)
23. https://tools.ietf.org/html/rfc6902 [↑](#footnote-ref-23)