Knowledge management services and closely related records management tasks

Current System Description – Reference D

December, 2015

Contents

1. Introduction

Sub-sections of Section 1 include the following.

- Section 1.1 ("Purpose and Scope") explains the purpose and scope of the SAD, and indicates what information is and is not included. This tells you if the information you're seeking is likely to be in this document.
- Section 1.2 ("How the Current System Description (CSD) Is Organized") explains the information that is found in each section of the CSD. This tells you what section(s) in this CSD are most likely to contain the information you seek.

1.1. Purpose and Scope

This document specifies the software architecture for the Knowledge Management and Document Management Services.

The definition given below provides the basic litmus test for what information is included in this CSD, and what information is relegated to downstream documentation.

What is software architecture?

The software architecture for a system[1] is the structure or structures of that system, which comprise software elements, the externally-visible properties of those elements, and the relationships among them [Bass 2003]. "Externally visible" properties refers to those assumptions other elements can make of an element, such as its provided services, performance characteristics, fault handling, shared resource usage, and so on.

Elements and relationships:

The software architecture first and foremost embodies information about how the elements relate to each other. This means that architecture specifically omits certain information about elements that does not pertain to their interaction. Thus, a software architecture is an *abstraction* of a system that suppresses details of elements that do not affect how they use, are used by, relate to, or interact with other elements. Elements interaction will be documented in this CSD accordingly. On the other hand, private details of elements—details having to do solely with internal implementation—are not architectural and will not be documented in the CSD.

Multiple structures:

The definition of software architecture makes it clear that systems can and do comprise more than one structure and that no one structure holds the irrefutable claim to being the architecture Thus, this CSD follows the principle that documenting a software architecture is a matter of documenting the relevant views and then documenting information that applies to more than one view.

1.2. How the Current System Description (CSD) Is Organized

This CSD describes the knowledge management and the document management services one by one following the below described structure:

- Part 1: For each service provides systems overview.
- Part 2: For each service explains solution background

2. Terminology Server

2.1 System Overview

ECDC TS (Terminology Server) is a software application that supports the creation, maintenance and dissemination of ECDC terminology and provides terminology services on concepts that are related to the activities of ECDC to both human users and software applications.

The backbone of the ECDC terminology system is ontology — an ordered set of concepts and their relations. This ontology is created and maintained externally and imported into ECDC TS. It is available for download from ECDC TS.

The terminology content of ECDC TS consists of sets of concepts (e.g. variables of an ECDC application or the terms of an external reference concept set, as e.g. ICD10) that are represented as categories of a value set. All categories in different value sets are mapped to the ECDC TS ontology. The value sets represent a certain subject field, or domain or terms used by a client ECDC application. Value sets are created externally and imported into ECDC TS in SKOS or ClaML formats. They can be edited by administrators either directly in ECDC TS or externally; updated value sets can be validated and imported into ECDC TS.

ECDC TS is primarily designed for use by ECDC experts, by health ICT application developers, administrators, by various medical professionals in EU member states, such as epidemiologists, public health experts, and others, including the general public. These users can search and view ECDC TS terminology content.

The terminology content in ECDC TS is maintained by designated ECDC TS content administrators that work with value sets assigned to them.

ECDC TS system administrators manage access rights to value sets, upload the ECDC TS ontology and perform other system administration tasks.

ECDC ICT system administrator performs user management in ECDC ICT Active Directory.

ECDC TS has two main types of users: human users and machine users. Machine users connect to ECDC TS

Web service and use its terminology content for other software applications.

Detailed information about ECDC TS users is provided in the following table:

User type	User role	Usage scenario
Human users		ECDC experts, health ICT application developers, administrators, and various medical professionals in EU member states, such as epidemiologists, public health

		experts, and others, including the general public, can search and view ECDC CS terminology content
	ECDC TS content administrators	maintain terminology content of ECDC TS, working with value sets assigned to them
	ECDC TS system administrators	manage access rights to value sets, upload the ECDC TS ontology and perform other system administration tasks
Machine users	ECDC TS client applications	connect to ECDC TS using its web service and reuse its terminology content
	ECDC TS client application	Connect to web service of the ECDC Administrative application in order to upload Value Sets and Ontology

Table 1 Target users and their usage scenarious

2.2 Goals and Context

The main goal of the architecture is to:

- Support human users with ECDC specific terminology at various end-points at ECDC
- Serve terminology services to the various systems in the federated system environment at ECDC
- Support industry accepted data format standards for terminology exchange to allow external partners and organisations reusing of ECDC terminology data
- Use of standardised services to allow easy and flexible connection to Terminology Server

2.3 Significant Driving Requirements

The main significant driving requirements are as follows:

- **Data model extendibility**: Systems at ECDC evolve with the time and often they come up with new requirements to store additional terminological information that was not foreseen at the requirement analysis of the initial development of Terminology Server. Therefore data model of Terminology Server must be flexible and extendible upon such requests.
- **Query flexibility**: With more and more ECDC systems consuming terminological data from Terminology Servers, the querying possibilities must be very flexible in order for Terminology Server to be able to return required data.
- **Compatibility:** As it is foreseen that Terminology Server will be used not only by ECDC systems, but also external organisations, the data format for exchange must conform to industry accepted standards.

2.4 Solution Background

2.4.1   Architectural Approaches

The following list explains and details the architectural decisions on which the system is based upon:

- **Microsoft technologies**: ECDC as an organisation is using Microsoft technologies, support is provided only for Microsoft technologies based software, therefore following products have been chosen for system implementation/hosting:

- o (Development platform) .NET Framework;
- o (Database software) MS SQL Server;
- o (Development environment) Visual Studio;
- o (Web server) Internet Information Services (IIS);
- o (Server operating system) Windows Server;
- o (Single sign on) Active Directory Federation Service (ADFS);
- o (Background tasks) Windows Server Services;

- **Data format standards:** following industry accepted W3 standards, below are listed data formats that are used for data storage and exchange: The ECDC TS handles data in **SKOS** format. The import/export format of the data is **SKOS**, **ClaML**, however, some of the TS functions allow export data also in **CSV** format as simple listings, **OBO** format as OBO ontologies or as structured **DOC** documents. When exporting data in ClaML format, the user must be aware of conversion constraints. The ontology is represented in **OWL** and managed externally using the ontology editor Protege, and is converted to SKOS (using the conversion rules with their constraints) by import procedure of the TS before importing into the TS. The ontology is limited in complexity. Communication with the machine interface is using web **WSDL** based upon **SOAP** 1.1 and for queries **SPARQL** is allowed as well.

- **Data storage in XML format**: data are stored in XML format in the MS SQL database for performance reasons. It is formed out of individual concept blocks during the upload or editing of the data and then is returned as a concatenated piece of requested results. Metadata of the concepts are extracted and stored in other relational database tables for search and querying purposes as well as for displaying simple data in the user interface.

- JavaScript and JQuery usage for XML editing: due to the complexity and numerous restrictions in SKOS data format, editing of data is developed on the client side using JavaScript and JQuery library. This was the easiest way how to fulfil the editing within the browser. After the user has finished editing in the browser form, the XML then is formed on the client side and sent to server for further processing.

- Replication service of SQL database: Terminology Server is available both – inside ECDCNET domain and also externally inside ECDCDMZ domain. Due to security issues another instance of TS is deployed in ECDCDMZ domain with another database. Since there is no data modification allowed externally, the synchronisation of the two database is done using one-way replication of the SQL database.

2.4.2   Analysis Results

Software was developed by outsourced contractor who was selected based on the best offer. The software had listed requirements in terms of technologies used and functional requirements. The architectural solution was proposed and developed by the outsourced company.

2.4.3   Requirements Coverage

Requirements that are addressed by the architecture are following:

o TS shall handle terminology using the SKOS standard in order to allow exchange of data with external partners (inside and outside ECDC).

   SKOS data format was chosen for storage and exchange of data.

o TS shall allow to attach DCMI (Dublin Core Metadata Initiative) metadata to any conceptual element (ontology, concept, value set, and category).

   SKOS data format was extended with the inclusion of DCMI metadata standard elements.

o Web services shall be defined with WSDL. Web services communication shall be based on SOAP.

   .NET was chosen to develop WSDL web service with SOAP interface.

o TS shall have two types of interfaces: web-based User Interface (UI) for human users, and web services for system to system communication.

   Multilayer architecture was used to separate data level from UI level, allowing systems to communicate directly with the web service for data access.

o "TS shall provide three parallel views for browsing categories of a value set:

   • The category hierarchy in a tree view,

   • A list of the category attributes,

   • A graph representing the direct semantic network (directly related categories, concepts and relations).

   Generating image on the server side was chosen to create a graph representation that would work cross browsers/OS and would not require installation of any add-ons.

o All attributes and relations of the ontology/concepts/value sets/categories shall be searchable.

   In addition to storing data in XML format in SQL database, all the relations and attributes were extracted and stored in a relational database for search purposes.

o The ECDC Active Directory shall be used to manage TS users and their role in the system.

o TS shall allow the creation of new categories by administrator. TS shall allow the editing of existing categories by administrator.

   User access is determined using the Active Directory services.

o TS shall have a notification service via email.

   SMTP server is used to distribute email notifications. It is configurable in the service configuration file.

o "TS shall be based on the ECDC standard technology:

   • MS .NET framework

   • MS SQL Server 2005 for the database

   • *Required technologies were used.*

3 Talent Map – Enterprise Profile Management Tool

3.1 System Overview

The ECDC Talent Map based Profile Management Tool is a SharePoint application for managing professional profile information of ECDC staff. Education and professional experience information is stored in a structured way and all of the skills and knowledge of persons are mapped to competences in **Talent Map.** This structure allows managers (and other ECDC staff members) to find experts with different competences very easily.

Talent Map is a common shared resource that contains the structured set of professional competencies. There are hundreds of competences in this map, so it would be very difficult to choose competences from a simple list.

Talent Map has well designed **structure** to ease finding items. Each **competence** (which can be a *knowledge* or *skill*) is assigned to a **domain** to group similar competences.

Example: '*applied mathematics*' domain has the following ECDC relevant mathematical knowledge and skills:

- 'mathematical modelling and simulation',
- 'modelling in neurobiology',
- 'modelling the spread of diseases',

Domains are organized into a hierarchy where parent always means a wider area (for example 'applied mathematics' belongs to 'Other natural sciences relevant to ECDC' parent domain which is a top-level domain).

There is a special kind of competence named 'complex competence' where the subject of competence also needs to be defined. Specific areas define the subjects which the competence applies to. Best example is language knowledge: if you have language knowledge as mother tongue you need to specify the exact language, too. In this case the list of languages is stored as specific area in Talent Map and if you select language knowledge you also need to select the proper language. There are some cases when specific areas are also organized into a hierarchy, because competence can apply to a wider of narrower area.

3.2 Goals and Context

The main goal of the architecture is to:

- allow ECDC staff members to upload their existing Europass CV or create their profile manually;

- all the information is stored in structured and searchable way;

- profile information is shared with other applications through web services;

- connect to Active Directory and User Profile Service to query basic personal and organisational information about the users;

3.3 Significant Driving Requirements

The main significant driving requirements are as follows:

1. **Availability and supportability**: The target audience of TaMa tool is the whole ECDC staff, because every experts of ECDC should have the profile in this system. Because of the high number of users a web application is created in the standard, SharePoint environment.

2. **Integration with other systems**: There are a bunch of web services that make available Talent Map functions to other systems, too. Personal profiles can be also accessed through a specific web service. Documents can be attached to profiles from Document Management System (DMS).

3.4 Solution Background

3.4.1 Architecture

The following list will explain and detail the most significant architectural decisions on which the system is based upon:

- **Microsoft technologies**: ECDC as an organisation is using Microsoft technologies, making this a natural choice for EPPM. In light of this the system is based on / using:

  - Development platform:
    - Web application: ASP .NET 3.5 framework (C# with some JavaScript, JQuery scripts and Telerik controls)
    - Web services: .NET 4.0 framework, WCF (C#)
  - Development environment: Visual Studio 2010 Team System
  - Web environment: SharePoint (MOSS) 2010
  - Web server: Internet Information Services (IIS) 7;
  - Database: MS SQL Server 2008;
  - Database access: Database is accessed using MS Entity Framework. (Only web services have access to the database, application can access or modify data only through web services.)
  - Server operating system: Windows Server 2008;
  - Single sign on: Windows Authentication (Active Directory is also used to query user account and membership information);
  - TaMa (Talent Map) web service use simple HTTP binding without security checking, EPPM (Professional Profile) web service use 'wsHTTPBinding', caller windows account is authenticated.

- **Web-based software**: To enable ECDC staff members to use the system efficiently, it was agreed to proceed with a web-based architecture allowing users to interact with the system using a web browser. Some Talent Map functions and profile data is also available using standard SOAP messaging.

- **Multitier architecture**: The system is divided into various parts, architecturally separating the responsibility into sub-parts:

  - Client browser: The frontend closest to the user working with Telerik AJAX controls to provide a seamless interaction with the system;
  - Web site: This solution is a collection of web parts that are deployed to a web site. The user frontend containing layered user interface and simple business logic related to the user interface;
  - Web service: The SOAP (WCF) frontend containing layered SOAP interface and the main business logic. BSC is not used.;
  - All of the information stored in Talent Map and the application i available only through web services, EPPM application also use these services to access data
  - Database: The application database containing both the Talent Map and the personal profile data in separated schemas;

3.4.2   Requirements Coverage

The EPPM is a complex system with many requirements that are fully described in the system documentation, this list here covers the main, high level requirements:

- Windows authentication should be used to avoid additional log on process to EPPM
- Access rights and organisational hierarchy is managed using AD information.
- A user can upload a Europass CV file (PDF or xml) to the system through a user interface in the application. The mentioned user interface allows editing of the parsed CV in order to fix mistakes of the automated parsing due to the very various content details of CVs.
- Profile data can be exported to an Excel file and data can be imported from Excel file through a user interface in the application.
- The Talent Map (TaMa) administrator can update the Talent Map from the application and he can list the changes of Talent Map during a given time interval.
- Users have the possibility to attach documents to their profiles from Document Management System (DMS). Application should show these documents in a new window.
- Users have the possibility to attach external links to their profiles, application should show these links in a new window.
- Users can share their profiles with different groups using the organisational hierarchy.
- The application should support searching based on competences or specific areas. Templates (roles) can be also defined for searching using a group of competences (specific areas) and setting the weight of each competence within the role. Weights and the lengths of experience with competences are used when the application calculates matching score during the search. It can be also indicate if one competence is mandatory for a role.

3.4.3　Information Sources

The system is using various information sources for metadata, or other data input:

- **ECDC Terminology Server**: Talent Map elements (domains, competences, specific areas and their relationships) are manually synchronised with the Terminology Server.
- **Active Directory**: Basic user data (name, email, organisational unit) coming from AD and organisational hierarchy also built on based AD information.
- **Document Management System (DMS)**: Users can attach documents from DMS (using Document ID) and TaMa application can show these documents using DMS services.

4 ECDC Enterprise Search

4.1 System Overview

ECDC Enterprise Search is a search system that allows users to search in a multitude of ECDC systems. The system is set up to crawl as many systems as possible and to index both – their content and metadata. It also provides advanced filtering possibilities by different type of metadata, supports linguistic search mechanisms and automatically extracts keywords from the crawled content.

The system is targeted to ECDC staff and is available only within ECDC network. The reason for this is sensitive data that is crawled and made available for users to see.

4.2 Goals and Context

The goal of the system is to provide aggregated results of all the systems in one single place. It also needs to be as easy as possible to integrate with existing systems at ECDC that are mainly developed using .NET and SharePoint technologies. The enterprise search must provide advanced semantic and linguistic features that will make information more easily discoverable and will differentiate it from the standard search.

4.3 Significant Driving Requirements

Speed: It had to be fast for usability, 3 seconds for typical search was set as a maximum search duration target.

Extendibility: The system must be able to accommodate new systems to search in as easy as possible, involving minimum development at a later stage.

Usability: System must provide advanced features that enhance user experience while searching, e.g. find using synonyms, provide advanced filtering options by automatically extracted keywords etc.

4.4 Solution Background

4.4.1 Architecture

ECDC Enterprise Search is developed using Microsoft Technologies due to many constraints – not only ECDC Backoffice provides support only to Microsoft based systems, but also majority of the systems in ECDC that the Enterprise Search would have to index are built using Microsoft Technologies.

The basis for ECDC Enterprise Search is FAST Search for SharePoint 2010. The FAST ESP product was developed by the Norwegian company Fast Search & Transfer AS. In 2008 Microsoft acquired the company and the product was more adapted to the Microsoft product line in the future. FAST is deployed in addition to the standard SharePoint installation. FAST Search component allows out of the box crawling and indexing of SharePoint sites as well as other kinds of sources. It provides many additional features in comparison to the default SharePoint search and allows customization/additional development to extend its features.

Microsoft Visual Studio was used to develop various Web Parts for the ECDC Enterprise Search that would customize the search behaviour on the client side and translate user queries to advanced FQL language queries.

Microsoft .NET (C#) is used to develop web parts and web services that would act as a middle layer between complex web sites and crawling process of the FAST Search.

4.4.2   Requirements Coverage

Following are major functional and non-functional requirements for ECDC Enterprise Search:

- **Simple Search in SharePoint 2007/2010 system**

  FAST Search is able to search SharePoint systems out of the box. Necessary metadata elements need to be configured properly according to user needs.

- **Search in .NET web application**

  Enterprise Search must be able to search .NET based web applications in ECDC. Appropriate middle layer web service is often needed to be developed to achieve this.

- **Search in file system**

  Enterprise Search must be able to search file system (shared drives). This functionality comes out of the box. Appropriate configuration steps to set this up are needed.

- **Ensure security**

  Users should be able to search and retrieve only the information where they have appropriate security rights. For SharePoint and File System this is provided out of the box.

- **Look & feel must correspond to ECDC guidelines**

  ECDC master page for SharePoint applications must be used / adapted for Enterprise Search purposes.

- **Performance must be acceptable for end user**

  Typical query must run under 3 seconds. Appropriate caching for the relevant web parts must be done in FAST Search.

- **A user can search using synonyms**

  A query must be rewritten during the search process, dictionary of synonyms must be synchronised with ECDC Terminology Server and used for synonym searching.

- **A user can graphically navigate through the result set using navigation tree of ECDC Terminology Server**

  When user filters result set by topic, the topic must be retrieved from ECDC Terminology Server and shown in the graph with its directly related terms. The user then is able to select any topic in the direct neighbourhood of the selected term and change the filtering of the results.

- **Topic keywords must be automatically extracted during the indexing process**

  A custom keyword extractor needs to be built to extract keywords that come from ECDC Terminology Server automatically during the indexing process. This ensures that even if the users didn't tag all their documents/files, the results will still be filterable by topics from ECDC Terminology Server.

5 ECDC Document Management System

5.1 System Overview

The Document Management System (DMS) is a tool that controls the lifecycle of documents in the organization; how they are created, reviewed, published, and consumed, and how they are ultimately disposed of or retained. Document collaboration and adjoining process automation can greatly improve the organization's response to the ever-changing environment by bringing down the lead times needed to produce meaningful artefacts as vehicles of communication and exchange.

5.2 Goals and Context

The primary goal of a document management system within ECDC is to capture, preserve, maintain and evolve, within the relevant context, the compendium of collaborative effort; to adapt to the needs of business throughout the lifecycle of a recorded manuscript, may it be a vessel of factual information or a chain of events thereof. The DMS at ECDC is a software application should fulfil these goals and those that emerge from the spatial or temporal dimensions of business activity or processes.

5.3 Significant Driving Requirements

- The system must be simple in use
- The system must be responsive in its transitions
- The system must be transactional in its operations
- The System must be context-sensitive in its presentation
- The system must be consistent in its interfaces
- The system must be compliant to best practices
- The system must be security trimmed in all its areas
- The system must be able to scale with growth

5.3.1 Requirements Coverage

Functional Requirements

No	Requirement	Implementation
1	The system shall be web based which allows documents uploading, storing, archiving, searching, and retrieving. The system shall be able to store different document types and workflows.	SharePoint 2010 out of the box provides these functionalities.
2	The security/ access rights for the Document Management System shall be defined and implemented.	SharePoint 2010 out of the box fully covers this requirement.

3	The Document Management System shall allow users to view/update metadata fields.	The architecture of SharePoint 2010 allows integrating metadata in easily.
4	The Document Management System shall allow users to create documents based on templates stored and provided by the system	SharePoint 2010 and Office allows storing templates and creating documents based on them
5	The Document Management System shall provide the capabilities for documents versioning.	SharePoint 2010 libraries have this feature available.
6	The Document Management System shall allow users to view all documents created and / or uploaded, as well as the ones shared with them, by clicking on the navigation buttons available in My Documents window.	The system has several views of documents available for the users (My documents, Documents per Unit…)
7	The Document Management System shall allow users to Check documents In and Out.	SharePoint 2010 libraries have this feature available.
8	The Document Management System shall allow the user to share documents within ECDC.	By default all the documents are available for all DMS users.
9	The Document Management System shall allow the user to filter the documents.	The search System and the different filters allow users to filter documents.
10	The system shall provide the possibility of integration with different ECDC systems.	Custom developed DMS Services extension allows other systems to query DMS.
11	The Document Management System shall be integrated with Terminology Server.	

12	DMS is using SharePoint 2010 MMS Services which are synchronized with the Terminology Server	The synchronization is out of the scope of this project
13	The Document Management System shall host the Chrono module application.	The DMS is ready for it.
14	The Document Management System shall provide usage reports and statistics.	SharePoint 2010 out of the box can provide usage reports and statistics.
15	The Document Management System will have archiving functionalities, transforming documents into records based on the set retention period.	SharePoint 2010 has reports management module.
16	The Document Management System shall comply with ICT standards, e.g. content standards and services like core metadata and TS provided terminology.	All the DMS Document Types are based on the base Document Type provided by ICT.

6. Appendices

6.1 Appendix: Glossary and Acronym List

Term	Definition
software architecture	The structure or structures of that system, which comprise software elements, the externally visible properties of those elements, and the relationships among them [Bass 2003]. "Externally visible" properties refer to those assumptions other elements can make of an element, such as its provided services, performance characteristics, fault handling, shared resource usage, and so on.
view	A representation of a whole system from the perspective of a related set of concerns [IEEE 1471]. A representation of a particular type of software architectural elements that occur in a system, their properties, and the relations among them. A view conforms to a defining viewpoint.
view packet	The smallest package of architectural documentation that could usefully be given to a stakeholder. The documentation of a view is composed of one or more view packets.
viewpoint	A specification of the conventions for constructing and using a view; a pattern or template from which to develop individual views by establishing the purposes and audience for a view, and the techniques for its creation and analysis [IEEE 1471]. Identifies the set of concerns to be addressed, and identifies the modelling techniques, evaluation techniques, consistency checking techniques, etc., used by any conforming view.
Acronym	**Definition**
API	Application Programming Interface; Application Program Interface; Application Programme Interface
ATAM	Architecture Trade-off Analysis Method
CMM	Capability Maturity Model
CMMI	Capability Maturity Model Integration
CORBA	Common object request broker architecture
COTS	Commercial-Off-The-Shelf
EPIC	Evolutionary Process for Integrating COTS-Based Systems
IEEE	Institute of Electrical and Electronics Engineers
KPA	Key Process Area
OO	Object Oriented
ORB	Object Request Broker
OS	Operating System
QAW	Quality Attribute Workshop
RUP	Rational Unified Process
SAD	Software Architecture Document
SDE	Software Development Environment
SEE	Software Engineering Environment
SEI	Software Engineering Institute Systems Engineering & Integration Software End Item
SEPG	Software Engineering Process Group

SLOC	Source Lines of Code
SW-CMM	Capability Maturity Model for Software
CMMI-SW	Capability Maturity Model Integrated - includes Software Engineering
UML	Unified modelling Language