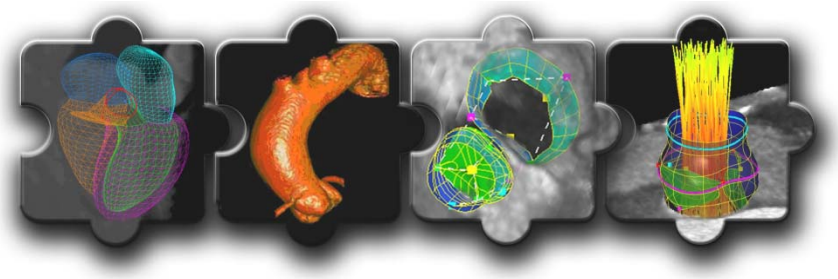
FP7-ICT-2009-4 (248421)
SeC
Sim-e-Child



Collaboration Project

Thematic Priority: ICT

Deliverable D3.2
Data Model Mapping Report

Due date of delivery: 31 October 2010
Actual submission date: 31 August 2011

Start date of project: 1 January 2010
Ending date: 30 June 2012

Partner responsible for this deliverable: Siemens Corporate Research (SCR)

SEVENTH FRAMEWORK PROGRAMME

Revision 2

Project co-funded by the European Commission within the FP7	
Dissemination level	
RE	Restricted to a group specified by the consortium

Document Classification

Title	Data Model Mapping Report
Deliverable	D3.2
Reporting Period	January 2010 - October 2010
Authors	Fabian Moerchen
Workpackage	WP3
Security	Restricted
Nature	Report
Keywords	

Document History

Name	Remark	Version	Date
Fabian Moerchen	First draft	0.1	01/11/2010
Fabian Moerchen	Submitted version	1.0	29/12/2010
Fabian Moerchen	First draft of revision	1.1	01/03/2011
Michael Suehling	Review comments and edits	1.2	11.07.2011
Michael Suehling	Updated data curation and data model mapping sections	1.3	29.08.2011
Fabian Moerchen	Minor edits	2.0	31.08.2011

Sim-e-Child Consortium

The partners in this project are:
01. Siemens AG (Siemens)
02. Lynkeus Srl (Lynkeus)
04. maat France (MAAT)
05. Technische Universität München (TUM)
06. I.R.C.C.S. Ospedale Pediatrico Bambino Gesù (OPBG)
07. Siemens Corporate Research, Inc. (SCR)
08. Johns Hopkins University (JHU)
10. American College of Cardiology Foundation (ACCF)
11. Siemens Program and System Engineering srl (PSE)

List of contributors

Name	Affiliation	Co-author of
Fabian Moerchen	SCR	
Benedetta Leonardi	OPBG	
Dumitru Ciubenco	PSE	
Michael Suehling	Siemens	

List of reviewers

Name	Affiliation
Michael Suehling	Siemens

Table of Contents

1. Introduction

The Sim-e-Child project requires a database of anonymized cases with medical images in specific formats and relevant clinical parameters. Such a database will allow the development of models and simulations and interpretation and validation thereof by clinicians. The data will be contributed by partners in the EU (OPBG) and the US (JHU and their partners in clinical trials and registries). The cases will thus have medical image data originating from different equipment and extracted from different PACS systems. The accompanying clinical data will be exported from different RIS and clinical registries. The differences in the data need to be abstracted into a common data model for use in Sim-e-Child. A third potential source of data is the integrated case database from the preceding Health-e-Child project.

1.1. Purpose of the Document

The purpose of this document is to describe the existing data at a high level and outlines a concept how to map the data into a common model and populate the Sim-e-Child database with anonymized electronic patient records.

1.2. Scope of the Document

The purpose of the data integration efforts is to directly serve the clinical aims of the Sim-e-Child project rather than to aim for a complete mapping of information that is not relevant for modelling, simulation and validation.

1.3. Abbreviations

CT	Computed Tomography
ETL	Extract, transform and load (ETL) is a process in database usage
HeC	Health-e-Child
LV	Left Ventricle
MRI	Magnetic Resonance Imaging
PACS	Picture Archiving and Communication System
RIS	Radiology Information System
RVO	Right Ventricular Overload
SeC	Sim-e-Child
ToF	Tetralogy of Fallot

2. Data Description

This section briefly describes the existing data models to be integrated and mapped into a common SeC data repository. For a detailed description of the different data protocols, we refer to the accompanying Deliverable D3.1, “Aligned Clinical Protocol and Assessment Report”.

2.1. Health-e-Child data model

The cardiology data from the Health-e-Child project was extracted from the databases for RVO which contains the ToF cases relevant for Sim-e-Child. The database has about 1500 variables. The database is very sparsely populated except for the first section with basic patient information and many quantitative parameters on the heart physiology such as right ventricular end diastolic volume is available for many patients. The Health-e-Child data was mainly used for the validation of the right-ventricular modelling as described in D5.1 “Health-e-Child Heart Models Clinical Validation Report” and is not considered for a common mapping since it lacks MRI data suited for the aortic modelling of Sim-e-Child.

2.2. Sim-e-Child data at JHU from COAST

The clinical data from the COAST trial was provided by JHU as PDF forms and spreadsheets with a data export from the database. The database has about 500 fields. While some sections such as interventions or adverse effects are obviously sparsely populated, most sections show a very high level of data quality.

2.3. Sim-e-Child data at JHU from GenTAC

The GenTAC data was assessed by the clinicians and deemed of limited use for the Sim-e-Child project. JHU has submitted an amendment to the IRB protocol to retrieve data and imaging on our patients with aneurysms. Once this data becomes available it can be analyzed for mapping purposes.

2.4. Sim-e-Child data at OBPG

The clinical data from OBPG was extracted from the Hospital and Radiology Information Systems (HIS, RIS) and from the PACS in the form of Excel documents and DICOM images, respectively.

3. Sim-e-Child Data Model Mapping

The clinical protocols and data models used for acquiring the data utilized within Sim-e-Child are documented in detail in Deliverable D3.1, “Aligned Clinical Protocol and Assessment Report”. Since each of the different data sources originates from a multi-centre clinical study or research project, each of the corresponding clinical protocols has been carefully designed and successfully been used over multiple sites and multiple years. For each data source, significant amounts of very specific information are necessary to guarantee the clinical goals of each study. To stay focused on the Sim-e-Child project goals, information irrelevant for the validation of the developed modelling and simulation will not be covered in the integrated Sim-e-Child data model (e.g. details regarding the catheterization and implants in the COAST trial). This means that a subset of the existing protocols will be used in Sim-e-Child and semantically mapped into the Web-based SciPort data base.

The main data sources that are being used to validate the anatomical modelling and, in particular, the simulation technologies developed in Sim-e-Child are the COAST trial data and OPBG data acquired during the course of the project. The Health-e-Child Tetralogy of Fallot data was mainly used to validate the right ventricular modelling as described earlier in Deliverable D5.1 “Health-e-Child Heart Models Clinical Validation Report”. Therefore, we focussed on the data communalities and mapping between the COAST and OPBG data.

For both relevant data sources, COAST and OPBG, the non-image data is available in structured Excel sheets independently developed in the past at the different sites. The image data is provided in standard DICOM format. Using these structured Excel sheets and DICOM data as input, we developed ingestion tools to automatically import and map the image and non-image data into the SciPort database.

The common SeC data model and mappings from the COAST and OPBG data have been developed in close cooperation with the clinical project partners. Figure 1 shows the common data fields and elements agreed upon to be used in SciPort and the corresponding mappings from the COAST and OPBG data sources. Subsets of data elements are only contained in one of the two data sources and therefore do not have a match in the other one.

The focus of the common data model is on basic patient information such as demographics and general disease information and, in particular, on MRI related information, especially related to the aorta which plays a central role in the project. The main goal is to provide an efficient access and search mechanism to select for example sub-group of patients that are suited for a particular validation study or experiment (e.g. find all patients that have an aortic root diameter larger than a particular value). Figure 2 illustrates the SciPort web-based graphical user interface to browse and search the data.

SciPort Data Field	SciPort common data terms	COAST data field	COAST data terms	OPBG data terms
Patient	ID	Demographics	Subject Number	Patient
	Gender		Gender	Gender
	Primary Indication	Baseline Subject Characteristics	Primary Indication	<Always: coarcation>
	Weight		Weight	Weight
	Height		Height	Height
Findings	Presence of coarcation	Subject Screening and Enrollment	Does the subject have native or recurrent coarctation	COA
	Presence of aortic aneurysms		Does the subject have aortic aneurysms	Does the subject have aortic aneurysms (yes, no)
Echo	Date	Post Implant Echo		Date
	Interval		Interval	Time between surgery and Echo
	Peak instantaneous Doppler gradient through region of coarctation v2 v1		Peak instantaneous Doppler gradient through region of coarctation v2 v1	
	Mean Doppler gradient through region of Coarctation		Mean Doppler gradient through region of Coarctation	
	Aortic Valve Peak gradient		Aortic Valve Peak gradient	Aortic valve pressure gradient
	LV ED diameter			DTD
	LV ES diameter			DTS
	EF			EF
MRI	Date	Site and Core Lab Cardiac MRI or CT Scan		Date of test performance
	Interval		Interval	Time between surgery and CMR
	Aortic root diameter			MR, aortic arch measurements, Aortic root, MAX(ant-post, trasv)
	Sino-tubular junction diameter			MR, aortic arch measurements, sinotubular junction, MAX(ant-post, trasv)
	Ascending aorta diameter		Ascending aortic diameter	MR, aortic arch measurements, ascending aorta, MAX(ant-post, trasv)
	Minimum transverse aortic arch diameter		Minimum transverse aortic arch diameter	
	Transverse arch diameter			MR, aortic arch measurements, posterior arch, MAX(ant-post, trasv)
	Isthmus diameter		Aortic Isthmus	MR, aortic arch measurements, isthmus, MAX(ant-post, trasv)
	Thoracoabdominal aorta diameter		Descending aorta at diaphragm	MR, aortic arch measurements, descending aorta, MAX(ant-post, trasv)
	Minimum stent diameter		Minimum stent diameter	
	Maximum stent diameter		Maximum stent diameter	
Interventions	Intervention performed (y/n)	Catheteriation Implant	first_implant ~ Was implantation of stent attempted	1 surgery (yes/no)
	Date of first intervention			1 surgery, DATE
	Type of first intervention		<Always: Catheteriation Implant>	1 surgery, TYPE
Re-interventions	Date of second intervention			2 surgery, DATE
	Type of second intervention			2 surgery, TYPE
	Category	Surgical Intervention	Category	
	Indication of hemodynamic dysfunction		Hemodynamic dysfunction ~ Indication	
	Indication of aortic aneurysm		Aortic aneurysm ~ Indication	
	Indication of restenosis (y/n)	Catheter Reintervention	Restenosis ~ Indication	
	Indication of hemodynamic dysfunction		Hemodynamic dysfunction ~ Indication	
	Indication of aortic aneurysm		Aortic aneurysm ~ Indication	
	Indication of in-stent restenosis (y/n)	Reintervention with Baloon Dilation or Additi	In stent restenosis	
		Reintervention with Covered CP Stent	In stent restenosis ~ Indication	
	Indication of aneurysm formation proximal (y/n)		Aneurysm formation Proximal ~ Indication	
	Indication of aneurysm formation distal (y/n)		Aneurysm formation Distal ~ Indication	
	Indication of aneurysm formation within (y/n)		Aneurysm formation Within ~ Indication	

Figure 1: SeC SciPort data model and semantic mappings of COAST and OPBG data sources.

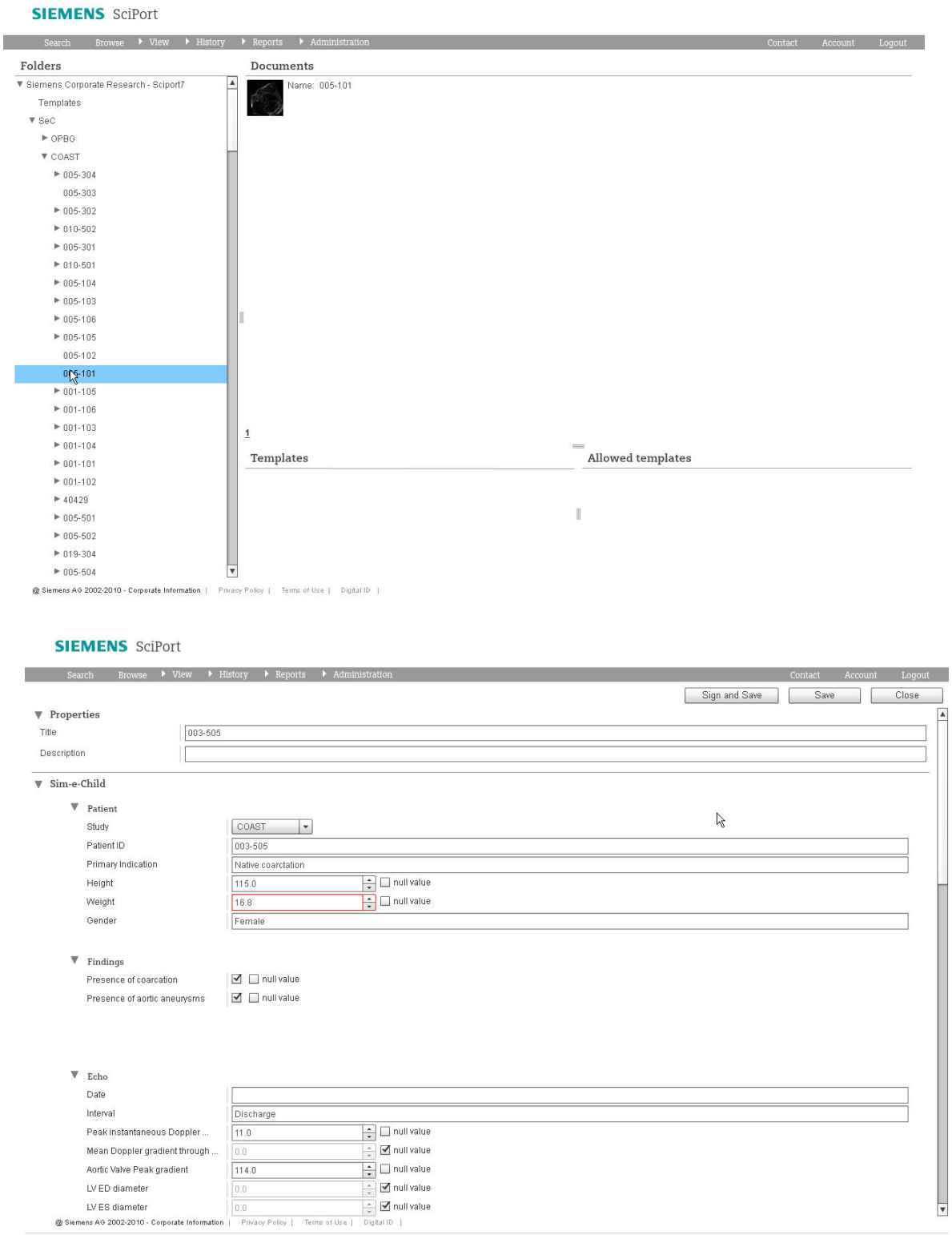


Figure 2: SciPort patient browser (top) and single-patient view (bottom).

SciPort models are extensible at runtime even if some of the data has already been imported. We expect the models to evolve while scientific experiments continue and as additional requirements come up. As shown in Figure 3, SciPort data fields can easily be adapted to

given needs through the graphical user interface. In addition, the automated ingestion tools are designed to be easily adaptable to include additional data fields for instance. SciPort models also support the attachment of arbitrary files including the Excel sheets with the complete extract of information from the originating studies. We hence have the flexibility to model variables in the SciPort model that are relevant either for scientific experiments or query and retrieval of the cases by physicians. The remaining secondary information will always be accessible via full text search in addition to the more focused search on the primary data fields.

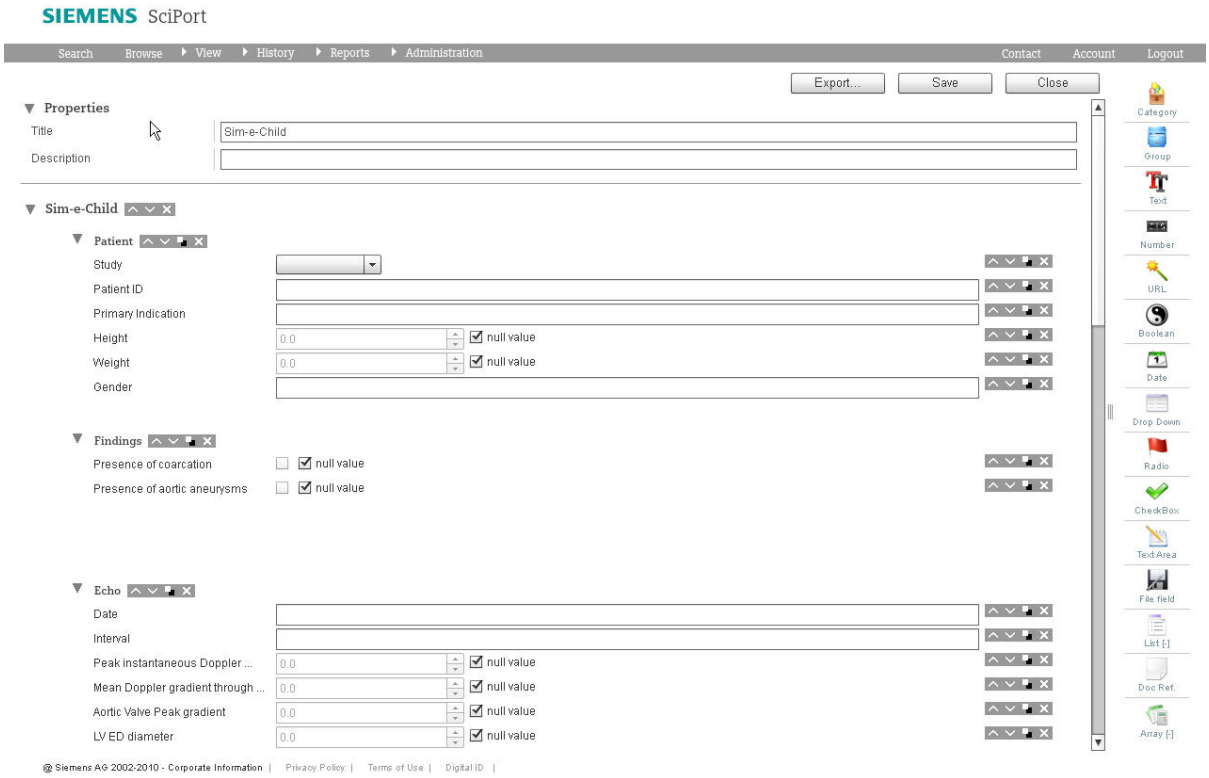


Figure 3: SciPort graphical user interface for easy definition and design of data field templates.

In the following, we describe additional tools developed to ensure highest data quality and privacy standards agreed upon within the project.

3.1. De-identification and data integration process

The process of de-identification and integration of patient records and images into the Sim-e-Child database is illustrated in Figure 4. The SciPort DICOM Anonymization Tool is provided to the physicians to process the clinical images. The physicians configure the tool to remove all DICOM headers that are not compliant with the privacy policies of the project. The patient identifiers are replaced with a unique study identifier. The identifiers are unique even over multiple runs of the same tool at different points in time as long as the same anonymization settings are used. The output folder contains the de-identified images that are then transferred to the researchers and to the database server using secure file uploads. In a separate folder that is not transferred the tool saves a text file that preserves the mapping from real patient identifier to study identifier. This file is to be kept private but would allow the physician to go back to the original clinical record if needed. Also this file is used to de-identify

any non-image data in the structured Excel spreadsheets. Automated import scripts are then applied to import and map the data into the SciPort data base.

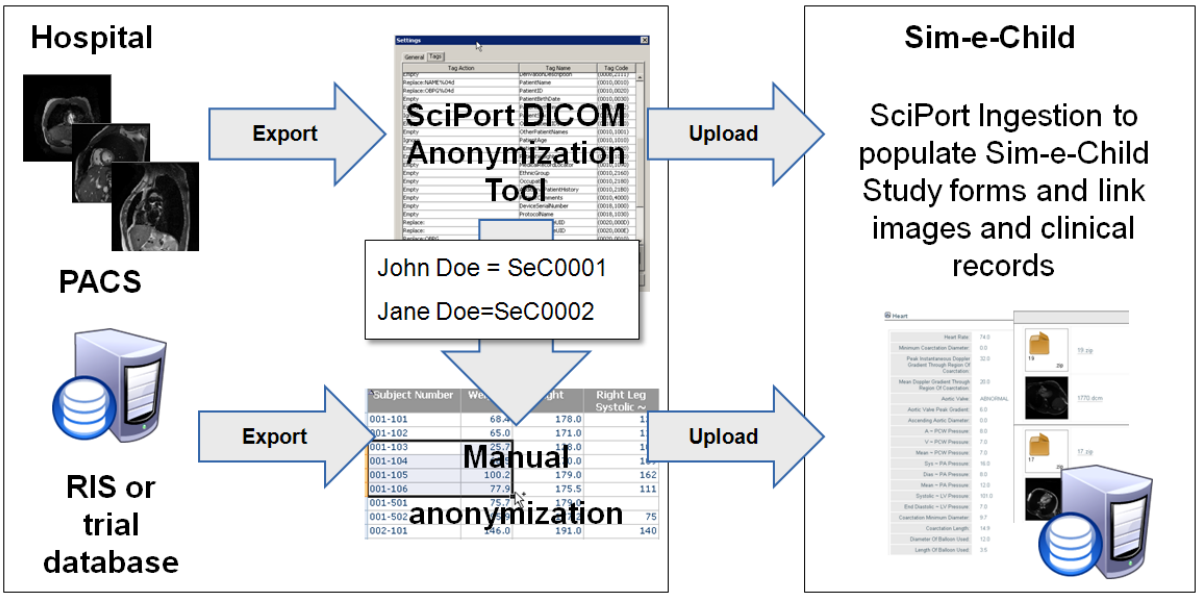


Figure 4: Data anonymization and integration process.

3.2. Data curation

In general we expect much of the data to be imported into the Sim-e-Child database to be of high quality because it has already undergone quality assurance processes in the originating databases such as the COAST or OPBG databases. Nevertheless, clinical database data entry errors may occur. Also since data from the US and EU are integrated there is a chance that different units of conversion of different numerical punctuations are used by accident.

The data curation process is based on the best practices learned in the Health-e-Child project. The SciPort database was extended with tools to generate reports that list missing and suspicious values in the combined data per data model field and offer references to quickly access the documents with the missing values to review them. The reports highlight suspicious values if they are outliers within the distribution of values from all documents.

In particular, as shown in Figure 5, the data quality reporting module highlights

- Missing values for required data model fields: Any field in the Sim-e-Child data model can be marked as a required field. Running the data quality report enables the users to quickly identify all documents where this field is not populated. The users can open these document in edit mode and fill in the missing value.
- Suspicious values representing statistical outliers: The values of a numerical fields are compared across all documents. Running the data quality report enables the users to quickly identify values that are above (below) the mean value plus (minus) 3 standard deviations and view the documents to see this in the context of the other data.
- Data violating constraints: Any field in the Sim-e-Child data model can be constrainted to a specific value range using minimum and maximum values or for text fields to a specific length. Running the data quality report enables the users to quickly identify all documents where constraints are violated and correct them.

As indicated in Figure 5, the data quality reporting can easily be triggered by choosing the command “Analyze” in the “Reports” menu. The user can also select a template from a drop-down menu to be used for the analysis. Figure 5 displays results of a synthetic test case to verify the data curation functionality. The left column shows a list of document fields that have at least one suspicious value for a required field. Upon selection of such a document field, all documents with the indicated problems are listed in the table at the bottom for easy access and review by the user.

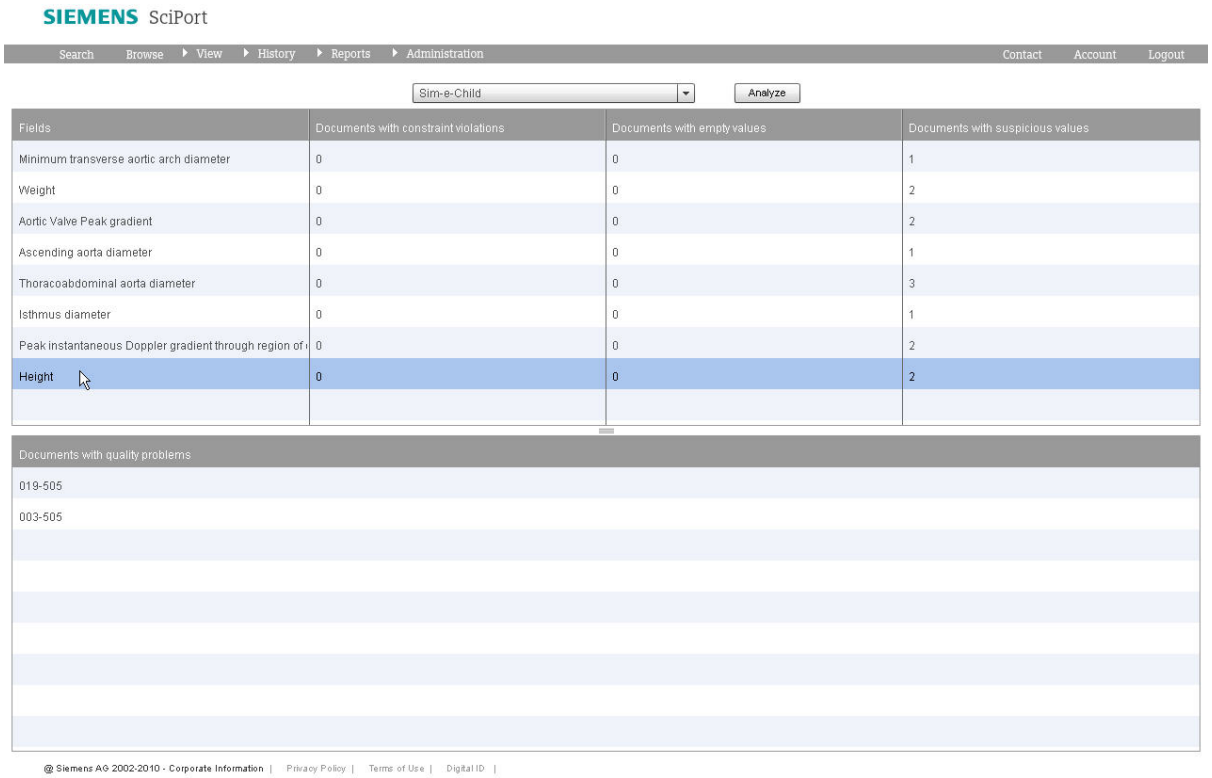


Figure 5: SciPort data curation functionality.