Process Modeling for Privacy - Conformant Biobanking: Case Studies on Modeling in UMLsec

Ralph Herkenhöner

Dept. for Computer Science, Christian-Albrechts-University, 24098 Kiel, Germany

Abstract. The continuing progress in research on human genetics is highly increasing the demand on large surveys of voluntary donors' data and biospecimens. By this new dimension of acquiring and providing data and biospecimens, a new quality of biobanking arose. Using automated data and biospecimens handling along with modern communication channels—such as the world wide web—assigns new challenges to protection of the donor's privacy . Within current discussions on privacy and data protection an emerging result is the need of auditing privacy and data protection within biobanks. For this purpose, finding a proper way for describing biobanks in terms of a data protection audit is a vital issue. This paper presents how modeling in UMLsec can improve the description of biobanks with the objective of performing a data protection audit. It demonstrates the use of UMLsec for describing security characteristics regarding data protection issues on the basis of two case studies.

1 Introduction

Research on human genetics was significantly advancing during the last decades. Primarily, this progress is due to the ability of fully sequencing the genotype of the human DNA. Complexity of acquisition and research has reached a scale that requires highly specialized acquisition and provision architectures. For this, a new generation of biobanks arises all over the world.

A biobank is storing biospecimens and sensitive medical information of voluntary donors. For research, this medical information is enriched by genetic information—so called genotypes—gained from the stored biospecimens. The possibility of directly combining medical, genetic, and identifying data demands an appropriate safeguard for protecting privacy issues of the donors.

In order to tighten biobank integrity and trust of the donors there need to be new attempts to audit biobanks regarding their soundness of privacy protection measures. For gaining such an audit, the biobank discloses its internal data and biospecimens management along with their appropriate protection measures to an external independent accredited entity.

For performing a data protection audit, it is necessary to describe how data protection is integrated into the processes of a biobank. However, a textually description of a biobank as a complex system faces the problem of being very difficult to be written in an understandable, complete and consistent way. Therefore, it is good practice to use

Herkenhöner R. (2008).
Process Modeling for Privacy - Conformant Biobanking: Case Studies on Modeling in UMLsec.
In *Proceedings of the 6th International Workshop on Security in Information Systems*, pages 3-12
DOI: 10.5220/0001732900030012

graphic modeling in order to get a better understanding of complex systems (e.g. software development and process management). Nevertheless, graphic modeling cannot replace textual description.

For modeling biobank processes with the objective of performing a data protection audit, the following requirements arise. An appropriate modeling language should:

- be easy to understand,
- require as little a-priori-knowledge as possible,
- allow a fairly complete description of all coherences on processes, and
- be described by a formal grammar.

Further, an appropriate modeling language should be able to describe:

- processes, roles, and their relations,
- data and control flow, and
- security characteristics regarding privacy and data protection issues.

A common language for graphical modeling is the *Unified Modeling Language (UML)* [1]. Originally, UML was intended to be used in the context of software engineering. As software engineering integrates process automation in real-world processes, UML unifies modeling techniques to describe program architectures, real-world processes, and real-world environments. Nowadays, UML is used in a wide set of modeling issues, including process management. Therefore, UML is considered to be a good candidate for modeling biobank processes.

Generally, processes, roles, and their relations can be modeled by *UML use case diagrams*. Further, data and control flow are modeled by *UML activity diagrams*. As UML does not support modeling of security characteristics by itself, an appropriate extension is necessary. As a prominent candidate, *UMLsec* meets this requirements.

This paper presents two case studies, demonstrating the use of UMLsec for description of biobanks with the objective of performing a data protection audit. In the next Section 2, a brief overview on related work is given. Afterwards, Section 3 introduces the UMLsec approach of modeling security requirements. Section 4 and 5 present case studies, modeling the *GENOMatch* and the *popgen* biobank in UMLsec. Finally, conclusion and outlook are given.

2 Related Work

UML was already used for business process modeling, as presented by *Kreische* [2]. Further, modeling security characteristics were introduced by *Jürjens*, extending UML to UMLsec [3]. Also, there are alternative approaches on process modeling. In the current version, the *Business Process Modeling Notation (BPMN)* [4] uses constructs similar to UML activity diagrams [5]. Further, the *Event-driven Process Chain (EPC)* introduced by *Keller et al.* [6] is using its own type of semi-formal modeling language.

Currently, there are several projects all over the world dealing with audit and standardizing issues of privacy and data protection in biobanks (e.g. by the USA National Cancer Institute [7], the UK Information Commissioner's Office [8], the German TMF [9] [10], and Swiss Academy of medical science (SAMW) [11]). Although, as all these projects focus on determining criteria for privacy and data protection, they do not consider description at all.

3 Modeling Security Characteristics Regarding Data Protection

Focus of this paper is modeling of biobanks with respect to a data protection audit. For this purpose, it is necessary to describe how data protection is integrated into the processes of a biobank. As mentioned before, process modeling can be done in UML using use case and activity diagrams. Further, UMLsec enables modeling security characteristics using extension mechanisms of UML. In a data protection audit, the achieved degree of data protection is evaluated according to the presence or absence of security characteristics within the process flows. Therefore, an important question is whether modeling in UMLsec can help improving the description of biobanks for the objective of a data protection audit. Concerning the modeling of processes, roles, and their relations in an UMLsec-enriched use case diagram, we think this is fulfilled.

In the following, three security characteristics that can be described by UMLsec are examined, and subsequently their use is illustrated for two case studies.

3.1 Non-Repudiation

In general, non-repudiation is the property of assurance that no participant of an action can deny its participation. In UMLsec, this security characteristic is represented by the stereotype ≪provable≫. This stereotype extends an UML object with the property that certain use cases or activities inside the object are provable, and therefore undeniable. Concerning data protection in biobanks, this characteristic targets at the demand for transparency and traceability of data and biospecimen handling—commonly achieved by recording of all handling activities.

Formally, the stereotype ≪provable≫ requires three parameters, which are describing the use cases that must be provable, the prove and an adversary. In the following, these parameters were omitted within the diagrams due to the fact that all shown use cases are provable if the stereotype ≪provable≫ would be present. For this stereotype the adversary is always an insider threat having the same privileges as the involved actor.

3.2 Role-Based Access Control

Generally, role-based access consists in restricting access to systems or environments to authorized individuals. In UMLsec, this security characteristic is represented by the stereotype ≪rbac≫. This stereotype extends an UML object with the property that certain use cases or activities inside the object are restricted to certain actors only. Concerning data protection in biobanks these parameters address the requirement to limit access to data and biospecimens according to identity, duration, and amount.
Formally, the stereotype ≪rbac≫ requires parameters that describe the actors having access and the use cases or activities being accessed by the actors. In the following, these parameters were omitted due to their complexity.

3.3 Secured Communication

Secured communication is the concept of communicating only via secured links that fulfill requirements concerning confidentiality and integrity. In UMLsec, this security characteristic is represented by the stereotype ≪`secure links`≫. This stereotype extends an UML object with the property that interactions between certain use cases or activities inside the object and other use cases, activities or actors are using secured communication links only. For this, the following stereotypes are attached to the edges between the interacting objects (e.g. actor to use case, use case to use case):

- ≪`secrecy`≫ (communication is encrypted),
- ≪`integrity`≫ (communication is signed), and
- ≪`high`≫ (communication is signed an encrypted).

Concerning data protection in biobanks, this characteristic targets the need to keep control on biospecimen and information flow.

Formally, the stereotype ≪`secure links`≫ requires a parameter describing an adversary. Again, this parameter were omitted due to the fact, that the adversary is always an outsider threat, having no privileges within the biobank.

4 Case Study 1: The GENOMatch Biobank

In a first step of evaluation the GENOMatch biobank of Bayer Schering AG was modeled as a use case diagram. The evaluation targeted at determining whether modeling in UMLsec improves the description of biobanks with the objective of a data protection audit. For this, the report on the data protection audit of the GENOMatch biobank in 2003 provides a basis, as it specifies all actors along with their activities. A brief report on this data protection audit was published by the Independent State Center for Privacy Protection Schleswig-Holstein [14]. Excerpts of the full report are made public by Luttenberger et al. [12] [13]. In the following, the final process step within the pseudonymization—as defined by *Pfitzmann and Hansen* [15]—of the tubes storing the biospecimen—in the following *sample tubes*—is presented as an exemplar of this case study.

Fig. 1 illustrates an UMLsec-enriched use case diagram of the second step of the pseudonymization within the storing process of sample tubes. In this step, the sample tubes are relabeled before they are finally stored within the biobank.

There are two participating facilities in this process: the external *data custodian*—represented by the *SIM Center*, that is storing the pseudonym-links of the biospecimen tube labels—and the biobank it self—known as the *Central Sample Repository*. The Central Sample Repository is divided into three different areas of accountability. The first area is called *Safety Zone 1*. This area is responsible for the biospecimen transfer from the *Clinical Trial Site* to the biobank and for removing the identifier labels from the Clinical Trial Site—the first step of pseudonymization. The second area—*Safety Zone 2*—is liable for relabeling the sample tubes, which is the second step of pseudonymization. Storage and handling of the biospecimens for analysis and research is done within the third area—*Safety Zone 3*.

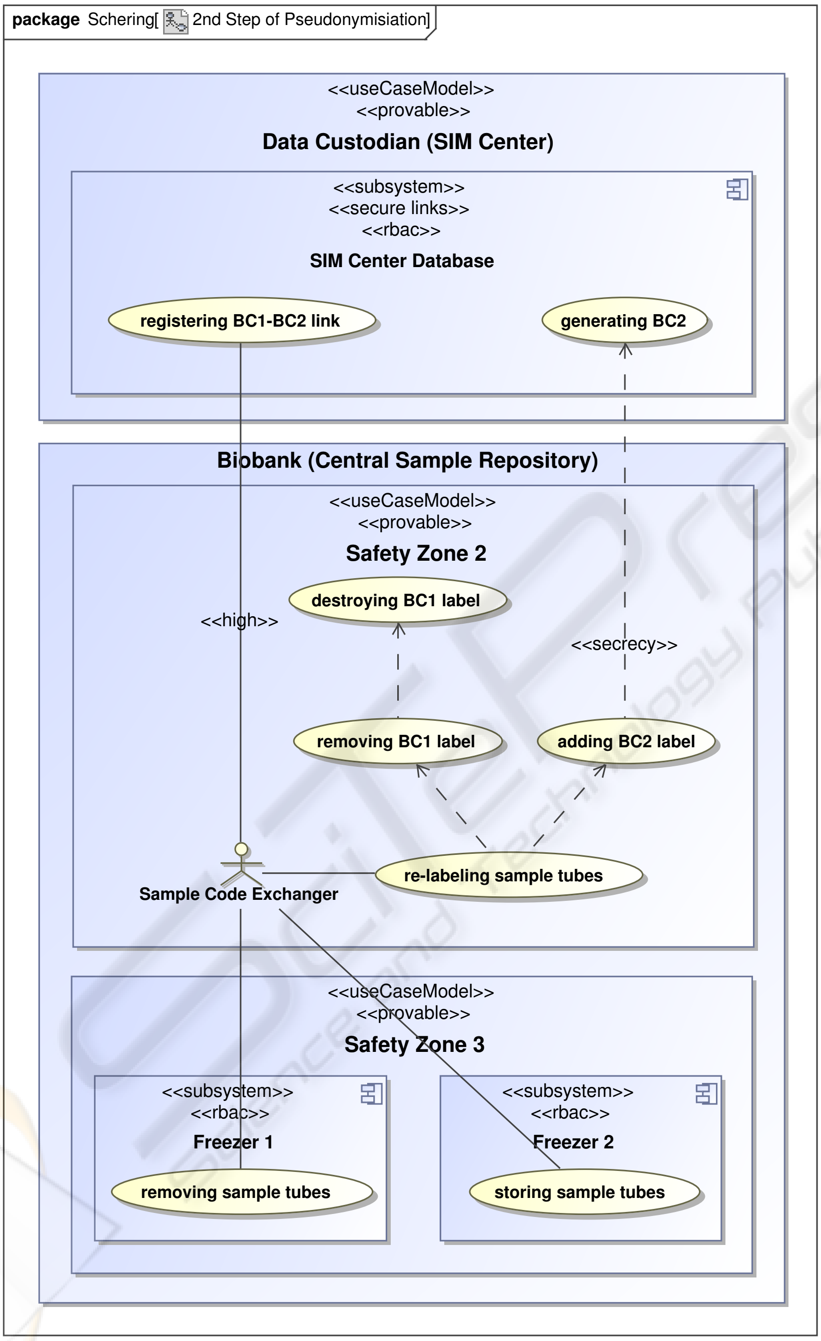

Fig. 1. UMLsec-enriched use case diagram of the second step of pseudonymization at GENO-Match (Safety Zone 1 is outside of this diagram.).

As the SIM Center is a fully automatic and PET[1]-protected database, the only actor in this use case is the *Sample Code Exchanger* within the biobank. He is responsible for relabeling the sample tubes, which is done in Safety Zone 2.

Relabeling the sample tubes includes the following activities: getting the sample tubes from Freezer 1 in Safety Zone 3, getting a new pseudonym from the SIM Center, saving the link between the new and the exchanged pseudonym at the SIM Center, and storing the relabled sample tubes in Freezer 3 in Safety Zone 3.

All these activities are done in interaction with the outside of the area of accountability of the Sample Code Exchanger—thus Safety Zone 2. This fact implicates the existence of interfaces between Safety Zone 2 and Safety Zone 3 and accordingly between Safety Zone 2 and the SIM Center. In figure 1 the existence of these interfaces are visible as edges crossing the border of Safety Zone 2.

As the security-sensitive sample tubes are stored within Safety Zone 3, every activity within this zone must be recorded. In the diagram, this requirement is represented by the UMLsec-stereotype ≪provable≫ attached to Safety Zone 3. This indicates that removing and storing of the sample tubes by the Sample Code Exchanger must be recorded.

Beyond that, the access to the freezers is restricted by PET-enforced role-based access control. In the diagram, this is represented by the UMLsec-stereotype ≪rbac≫ attached to Freezer 1 and to Freezer 2 accordingly.

Even more sensitive are the pseudonym links that are stored at the *SIM Center Database*. Analogous to Safety Zone 3, every activity within the Sim Center must be recorded—indicated by the UMLsec-stereotype ≪provable≫—and requires authentication and authorization by the PET-enforced role-based access control—indicated by the UMLsec-stereotype ≪rbac≫. But, unlike the freezers, the SIM Center Database is not part of the biobank. For this reason, data exchange must be done via PET-secured links. In figure 1, the usage of secured links is indicated by the UMLsec-stereotype ≪secure links≫ attached to the SIM Center Database. This means, every communication link to actors from outside the SIM Center Database must have a certain state of security. The most sensitive activity in this use case is saving the pseudonym link. For this, activity there must be a highly secured link that matches requirements concerning confidentiality and integrity. In the diagram, this is represented by the UMLsec-stereotype ≪high≫ at the edge that links the Sample Code Exchanger to the use case associated with this activity. In contrast, of generation of a new pseudonym, a confidential link—indicated by the UMLsec-stereotype ≪secrecy≫—meets the security requirements.

5 Case Study 2: The Popgen Biobank

In a next step of evaluation the popgen biobank of the University Medical Center Schleswig-Holstein was modeled. For this, the report on data management at popgen by Eller-Eberstein et. al. [16] was taken as a basis. Eller-Eberstein describes in this report the flow of data and biospecimens within processes regarding to collection, sampling,

[1] Privacy Enhancing Technology

storing and research in the popgen architecture. In this paper, the process steps regarding to anonymization and research are presented as an exemplar of this case study.

Figure 2 presents an UMLsec-enriched use case diagram of merging and statistically analyzing at popgen. There are four participating facilities in this process:

- the *Pseudonymization Center* (providing a *Pseudonymization Service* for forwarding and pseudonymizing data and biospecimens, and storing the pseudonym-links; it acts as an intermediate for every communication between the other facilities),
- the *Study Center* (responsible for recruitment, data and biospecimen collection, and providing the phenotypes),
- the *Analysis Labor* (extracting DNA from the biospecimens, genotyping, and providing the genotypes), and
- the *Statistical Research Center* (merging pheno- and genotypes, anonymizing and providing statistical analysis).

As the Pseudonymization Service at the Pseudonymization Center, the *Genotype Database* at the Study Center, and the *Phenotype Database* at the Analysis Labor are fully automatic and PET-protected, the only actors in this diagram are the *Data Custodian* and the *Statistical Analyst* at the Statistical Research Center.

The Statistical Analyst is responsible for statistically analyzing the pheno- and genotypes on correlations. This analysis provides a basis for research at popgen.

Prior to statistical analysis, the necessary pheno- and genotypes must be provided and in conformance to the data protection policy at popgen the provided data must be anonymized. These activities lie within the accountability of the Data Custodian. He requests the necessary pheno- and genotypes, merges them by PsID identifier—thus in terms of the donor—, removes all pseudonyms—in case of popgen anonymizes them—, and forwards the merged and anonymous data to the Statistical Analyst.

Requesting the necessary pheno- and genotypes in popgen is supported by the fully automatic and PET-protected pseudonymization Service. This service acts as an intermediate between the Data Custodian and the Phenotype Database at the Study Center and the Genotype Database at the Analysis Labor. Further, the Pseudonymization Service exchanges the pseudonyms on pheno- and genotypes to enable merging in terms of donor. As this service provides access to linked—in terms of donor—pheno- and genotypes, every activity in the Pseudonymization Center must be protocoled—indicated by the UMLsec-stereotype ≪`protected`≫. In addition, usage of the Pseudonymization Service is protected by role-based access control—indicated by the UMLsec-stereotype ≪`rbac`≫—and limited to the Data Custodian. For this, communication is limited to highly secured link—indicated by the UMLsec-stereotype ≪`secured links`≫ within the Pseudonymization Service and the UMLsec-stereotype ≪`high`≫ at the edge linking internal use cases with the outside.

Analogous activities within the Analysis Labor and the Study Center must be protocoled and access to the Pheno- and Genotype Database is done by PET-enforced role-based access controll and via PET-secured links.

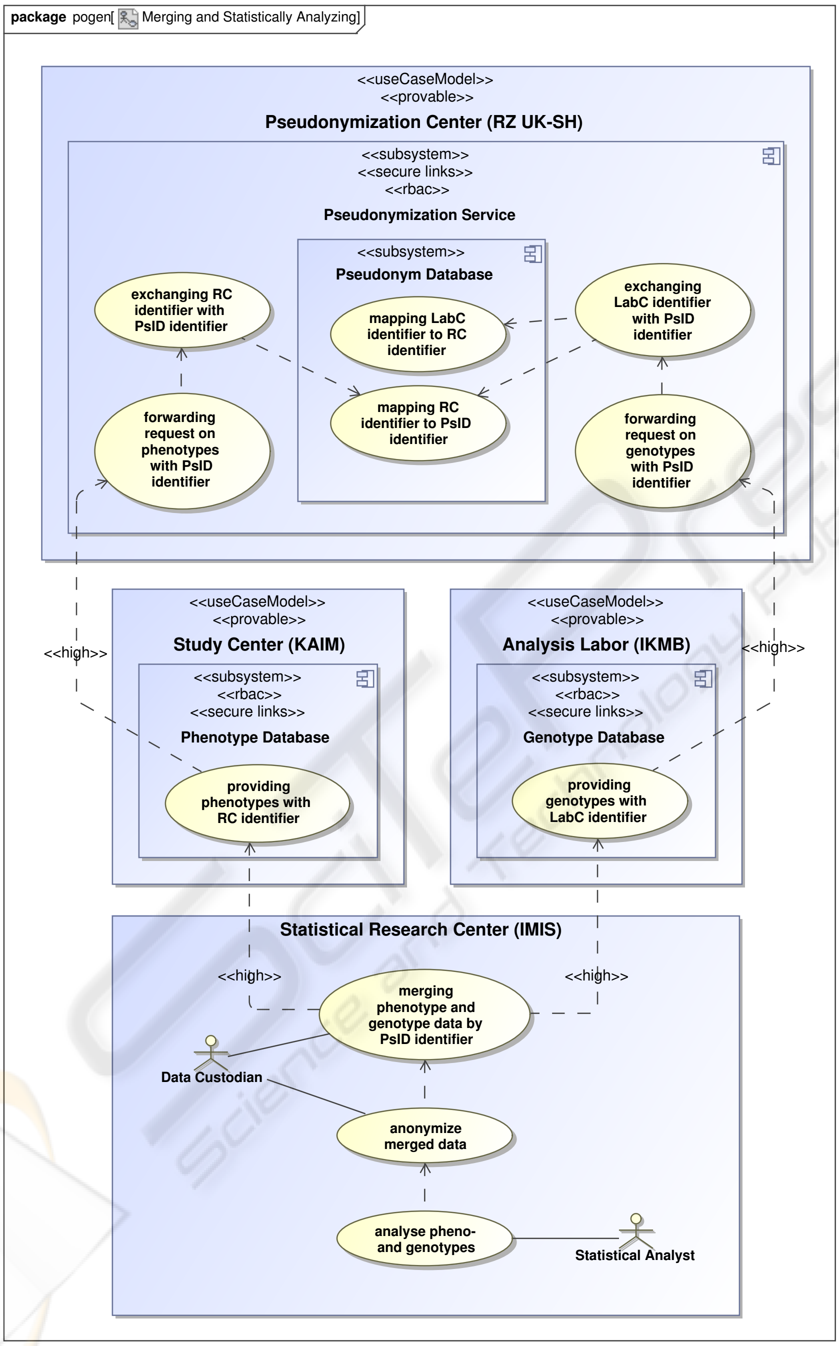
package pogen[Merging and Statistically Analyzing]
<<useCaseModel>>
<<provable>>
Pseudonymization Center (RZ UK-SH)
<<subsystem>>
<<secure links>>
<<rbac>>
Pseudonymization Service
<<subsystem>>
Pseudonym Database
exchanging RC identifier with PsID identifier
mapping LabC identifier to RC identifier
exchanging LabC identifier with PsID identifier
mapping RC identifier to PsID identifier
forwarding request on phenotypes with PsID identifier
forwarding request on genotypes with PsID identifier
<<high>>
<<high>>
<<useCaseModel>>
<<provable>>
Study Center (KAIM)
<<subsystem>>
<<rbac>>
<<secure links>>
Phenotype Database
providing phenotypes with RC identifier
<<useCaseModel>>
<<provable>>
Analysis Labor (IKMB)
<<subsystem>>
<<rbac>>
<<secure links>>
Genotype Database
providing genotypes with LabC identifier
Statistical Research Center (IMIS)
<<high>>
<<high>>
merging phenotype and genotype data by PsID identifier
Data Custodian
anonymize merged data
analyse pheno- and genotypes
Statistical Analyst

Fig. 2. UMLsec-enriched use case diagram of merging and statistically analyzing at popgen.

6 Conclusions

To summarize, the case studies presented in this paper show that it is of use to describe processes, roles, and their relations within biobanks by using UMLsec-enriched use case diagrams. Furthermore, they demonstrate how the modeling supports the description of biobanks for a data protection audit:

1. Interfaces between different areas of accountability are visible as edges crossing their borders. This enhances the detection of hidden data flow.
2. Usage of protocoling, role based access control, and secured communication are modeled as UMLsec-stereotypes. This enriches the model by important information necessary for the evaluation of data flow.

Thus, modeling processes in biobanks by using UMLsec-enriched use case diagrams significantly improves the description of a biobank with the objective of a data protection audit.

7 Future Work

Even if use case diagrams meet the requirements on describing biobanks in terms of data protection audits rather well, the following questions arise:

1. What about activity diagrams? As activity diagrams are used for modeling data and control flows, they might be a powerful completion of describing biobanks in terms of data protection with respect to items of responsibility.
2. Is it possible to describe other characteristics concerning data protection (e.g. characteristics of pseudonymization and anonymization and responsibilities on data and control flow)? And if yes, is it necessary and possible to significantly extend UMLsec for this purpose? UMLsec meaningfully allows such extensions. Experiences during the case studies lead to the assumption that modeling the responsibilities on data and control flow may be possible by a new defined stereotype ≪`responsible`≫.
3. As UMLsec provides a basis for proving the achievement of security, could UMLsec be a basis for a formal argumentation within the data protection audit meeting the standard of proving security in safety-critical systems presented by Jürjens?

References

1. Unified Modeling Language: Superstructure. Version 2.1.1 (formal/2007-02-03). Object Management Group. `http://www.omg.org/docs/formal/07-02-03.pdf`
2. Kreische, D.: Geschäftsprozessmodellierung mit der ”Unified Modeling Language (UML)” (in German). Dissertation at the University Erlangen-Nürnberg (2004). `http://deposit.ddb.de/cgi-bin/dokserv?idn=972544232`
3. Jürjens, J.: Secure systems development with UML. Springer-Verlag, Berlin Heidelberg New York (2005)

4. Business Process Modeling Notation Specification. Final Adopted Specification (dtc/2006-02-01). Object Management Group.
   `http://www.bpmn.org/Documents/OMG\%20Final\%20Adopted\%20BPMN\%201-0\%20Spec\%2006-02-01.pdf`
5. White, S. A.: Process Modeling Notations and Workflow Patterns. Object Management Group, Business Process Management Initiative (2004). `http://www.bpmn.org/Documents/Notations\%20and\%20Workflow\%20Patterns.pdf`
6. Keller, G., Nüttgens, M., Scheer, A.-W.: Semantische Prozeßmodellierung auf der Grundlage ”Ereignisgesteuerter Prozeßketten (EPK)” (in German). Scheer, A.-W. (Hrsg.): Veröffentlichungen des Instituts für Wirtschaftsinformatik, Nr. 89. Saarbrücken (1992).
7. Best Practices for Biospecimen Resources. National Cancer Institute (2007).
   `http://biospecimens.cancer.gov/practices/`
8. Data Protection - Complete Audit Guide. The Information Commissioner’s Office, UK.
   `http://www.ico.gov.uk/upload/documents/library/data_protection/detailed_specialist_guides/data_protection_complete_audit_guide.pdf`
9. Reng, C.-M., Dembold, P., Specker, Ch., Pommerening, K.: Generische Lösungen zum Datenschutz für die Forschung in der Medizin (in German). Medizinisch Wissenschaftliche Verlagsgesellschaft, Berlin (2006)
10. Pommerening, K., Schröder, M., Petrov, D., Schlösser-Faßbender, M., Semler, S.C., Drepper, J.: Pseudonymization Service and Data Custodians in Medical Research Networks and Biobanks. GI Jahrestagung (1) 2006: 715-721
11. Biobanks: Obtainment, preservation and utilisation of human biological material. Swiss Academy of medical science (SAMS),Basel , Swiss (2006).
    `http://www.samw.ch/docs/Richtlinien/e_RL_Biobanken.pdf`
12. Luttenberger, N., Reischl, J., Schröder, M., Stürzebecher, C.S.: Datenschutz in der pharmakogenetischen Forschung - eine Fallstudie (in German). DuD Datenschutz und Datensicherheit 28(6) (2004).
13. Luttenberger, N., Stürzebecher, C.S., Reischl, J., Schröder, M.: Der elektronische Datentreuhänder (in German). DIGMA Zeitschrift fur Datenrecht und Informationssicherheit 5, 1, pages 2429, 3 2005.
14. Brief Report on the Data Protection Audit. Independent State Centre for Privacy Protection Schleswig-Holstein (2003). `https://www.datenschutzzentrum.de/audit/kurzgutachten/a0303/a0303_engl.htm`
15. Pfitzmann, A., Hansen, M.: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management A Consolidated Proposal for Terminology.
    `http://dud.inf.tu-dresden.de/Anon_Terminology.shtml`
16. v. Eller-Eberstein, H., Gundermann, L., Krawczak, M., Schreiber, S., Wolf, A.: Datenmanagement bei popgen (in German). GI Jahrestagung (1) 2006: 729-735