Development and an evolution of the data model of a CRIS system – Example from FRIDA to CRIStin

Grete Christina Lingjærde, Geir Magne Vangen
Centre for Information Technology (USIT), University of Oslo, Norway

Summary

The main focus of this article is a systematic development of computer systems with good data models. Example of a data model being changed through its life cycle, the national CRIS-system CRIStin now used by 160 institutions in Norway, is presented. Mapping to CERIF is briefly discussed and international standards touched upon. Also the future of international systems and interoperability is mentioned.

1 The lifecycle and development of a data model

A central and important part of a CRIS system is the data model. A solid and generalized model will be able to meet future changes. Even though a good data model can endure a lot of changes, a good viable system must change its data model as time goes by and new needs, new requirements, and new possibilities appear. New technology, new areas of the application, more knowledge of the domain, rationalizations gains and new possibilities increase the need for changes and are the main drivers of change in many cases.

Some changes might be just minor or additions to the data model, while others might cause comprehensive changes in the data model. Still we will claim that a good data model as a base will be a strength and that some changes will not necessarily cause a large amount of work in order to get the application to satisfy the wanted and needed new functionality and effects that might be a substantial change of the systems functionality.

Why do we focus on this issue? Some will say it is elementary knowledge in any systems of a data centric nature. Our experience is that a lot of systems are not modelled properly and that this causes a lot of problems in the different systems in the Higher Education sector. Lack of skills in data modelling and use of methodology, when structures are complicated, will result in poorly functional systems. Since the topic is not as hot as web services and other new technologies, it is easy to forget to keep proper focus of the heart of the system, the data model both in the beginning and in the lifecycle of the system.

In this article we will examine how the data model of the FRIDA system, which was developed to cover the need for registration of research data to the individual institution, was changed to meet the national needs as well as the institution’s needs. The new data model simplifies integration, research, administration, reporting of data to the institutions and the government. We will discuss the background, prerequisites, and the processes in the development of the new data model.
Research is performed in an international arena with co-authors from different countries. Next level of development could be to make a model that was not so tied up to a certain country’s specifications and at the same time not being a lean system. We still need to have a system that allows the specifications of different national regulations and rules, and at the same time satisfies needs across national borders. Some issues and challenges must then be solved. One example is the identification of a researcher across different countries and within a country. Another example is the support of different national calculation rules for publication points.

1.1 Standards

In the area of study information Europe has developed different standards. Based on these standards different initiatives for developing systems that exchange this information both on national level and across countries have been initiated. Are there any incitements for making international standards for the CRIS area? Different challenges and issues in connection to the European CERIF-model are discussed. We will also look at the possibility to translate our national model of CRISTin into the CERIF model and challenges in this process will be discussed.

1.2 Development method

We will also have a look at how data modelling and analysis fit with the different development methods that we use. There will also be a discussion about how to make sure that the data models develop in a sound way during the systems life cycle. The time line in the development will also be discussed.

We will have look closer at our development methods. With development methods such as Scrum and Extreme Programming, we must make sure to keep our focus on the data model in order to ensure against fragmentation and improper generalization of the model.

Standard tools and methods are used in system development in our department. Focus is on development involving extensive interaction with end users. Considerable time and effort is spent on specification of system requirements and on subsequent analysis of the data model. All systems are based on a well-founded data model.

2 System development at USIT (University Centre for Information Technology, University of Oslo)

System development consists of four major activities:

1 Specification of requirements. In close interaction with the users we work out a specification of requirements. We do this in order to understand the user's needs and to understand the area of work and problems. Relevant questions are: how they work today, working tasks, routines, new solutions that are wanted, which people are involved etc. We are splitting activities to a level of detail according to what is useful to comprehend. We arrange meetings etc. with end users. This phase is time consuming both to the development team and to the users involved. In this phase it is important to find general
solutions for the same task at all institutions. We have experienced that the users appreciate the method. They are now eager to learn from each other to find the best solution. We have also found that different institution have very similar needs.

Two documents are produced:

a. Deconstruction of activities. A document where activities are deconstructed to a suitable level. The different activities are represented graphically as boxes in a standard and unified manner. Also called WBS (Work Breakdown Structure)

b. Description of activities. A description of all the activities in the activity deconstruction. Who and what is involved in the activities. All documents in this phase are made with the same notification and standard layout etc. (similar to UML).

2 Data-modelling. In this phase the data model is constructed based on the activity description. The data model is developed using NIAM and describes the structural and functional aspects of the system. NIAM (Nijssen’s Information Analysis Methodology) is related to ORM (Object Role Modelling). Both are linguistic-based methods with a graphical representation. A few senior analysts who all have been involved in the specification phase construct the data model. The work ends up with a proposal of a data model, which then is presented to the end users. It is the responsibility of the analyst to explain the data model to the users in such a way that they are able to understand the consequences of the model.

3 Implementation. When we developed the student information system, we used a prototype the users must to test at the start of the project. The system is made of a kernel that can easily be integrated with new components (modules). These components are developed in a manner that makes them easily modifiable. Users test the different components as soon as they are implemented. Examples of components are the exam module and the registration module.

4 Testing. In this phase users test the system. Testing is done in all the phases (unit testing, system testing, requirement testing, module testing). In all the above phases the development team works in close interaction with the users. It is our experience that a system based on a good data model has relatively few errors.

Iteration of stages 1-4 above. Depending on the situation, further iterations may start with phase 2, 3 or 4.
Figure 1. Developing phases

Prior to the development of FRIDA, CRISatin and the student information system FS, NIAM and Activity deconstruction had been used successfully for several years at USIT; hence there was no need to change methods.

NIAM was originally an acronym for "Nijssen's Information Analysis Methodology", but more recently, since G. M. Nijssen was only one of many people involved in the development of the method, it has been generalized to "Natural language Information Analysis Method". In 1989, Terry Halpin formalised NIAM in his PhD thesis. The result is called, "Object-role Modelling", or ORM. There are now several dialects known by the collective name “Fact Based Modelling” (FBM)

Object Role Modelling (ORM) is a conceptual modelling method that allows the semantics of a universe of discourse to be modelled using natural language and diagrams. Fact based modelling, has been used commercially for more than 30 years as a data modelling methodology, and has recently become popular not only for engineering but also as a graphical notation for the modelling of business rules, XML-Schemas, data warehouses, requirements engineering and web forms. [1] Ref: http://en.wikipedia.org/wiki/Object-Role_Modeling

There is a tradition at USIT for focusing on the development method as well as good design practices. Firmly grounded, general and flexible data models and heavy interaction with the users are important ingredients. We adopt new tools, which support our design practice, which focus on the requirement and analysis phase and methodology. The FS system has been developed in that spirit, in close collaboration with other that make use of the methods, both at the Department of Informatics at the University of Oslo and elsewhere.

A new development cycle is initiated when:

- New features are required
- Existing requirements change over time
- New modules are integrated into the system (like the "exchange student" module)
- There is a demand for new types of integration to other external systems
3 System development for systems in production

Most large systems must meet requirements in an ever-changing reality. “The moving target” is an expression in system development highly relevant to our systems in the sector of higher education and research administration. Rules are being changed and new proposals for a system to cover greater areas and enhance functionality are requested. Changes in IT-technology will produce new possibilities and set new demands, and so on.

In the development process we have not faced any crisis. The fact that we have developed a system for institutions with quite different properties from those of the University of Oslo has led to requirements that at first glance appear to diverge. We don't consider this a problem, but rather as an opportunity to make more general solutions that are more robust and flexible to changes and needs of the future. We simply get more background information and more examples.

3.1 The data model and its importance

Common structures and understanding increase the possibilities of reuse both of content and metadata and will benefit the developing of different systems related to these structures. To get a common understanding of the universe of discourse it is important to use a structured model. It is simpler and easier to exchange data between systems when you can move data that are highly structured to a less structured data system than the other way around. Therefore it is essential those authority systems are highly structured and that they are the basis for developing standards.

An example below is an European Learner Mobility (ELM) model that is a European standard for student results/assessments. This is the first model in a series of models to support interoperability between European IT systems that support student information. http://wiki.teria.no/display/EuropeanLearnerMobility/European+Learner+Mobility.

The ELM project was initiated back in 2007 and financed by the European commission, and carried out by the European Committee for Standardization – CEN. This first model was made by an expert team, and express information in the Diploma / Diploma Supplement documents. The model – EuroLMai became a European standard in 2011. http://research.cen.eu/research/Details.aspx?id=5978354

A new project – ELMO - started in 2012 and is a spin off of EuroLMai. The goal for this project is to agree on data representations (eg XML schemas) for EuroLMai to enable the industry to make interfaces for exporting and importing data regarding student results and for supporting the exchange of students.
4 The history and evolution of data model FRIDA to CRIStin

We will give a short description of the original FRIDA model. This original data model was enhanced with some new modules that were developed such as research groups and projects. In this phase a lot of attention was paid to the data model and the data model became more generalized, a more general concept was developed, called a presentation object. This object could be of type project, research group etc. We don’t have any entity called project, but we do have objects called presentation that can be of type project. This makes the system easy and quick to augment with other types in the future; for instance presentation of a research infrastructure.

Several institutions within the sectors Higher Education, Health Centres and Research Institutes have shown interest in a research documentation system. In 2008, the Ministry of Education and Research appointed a committee to consider the possibility of achieving one common system for these three sectors for registration of scientific publications. The committee also considered the need for a research documentation system that covered other research activities in addition to scientific publications. A premise for the solution was that data needed to be registered only once across the institutions, while the control of and responsibility for the data quality remained at the individual institution. In the autumn of 2009 the Ministry decided that a restructured FRIDA would be the research documentation system for 150 institutions in Norway. The new system was renamed CRIStin.

Since it is based on the highly structured data model FRIDA, we were able with relative simple adjustments to model CRIStin in a way that satisfied both the institutional perspective and the national perspective. Still, a lot of the logic and functions had to be changed and the existing application needed considerable reprogramming to be able to adapt to this change.
A prerequisite for the solution was that FRIDA was built on a well-designed data model. This makes it relatively simple to change and adapt the system to meet new demands, altered needs and a changing future. Another prerequisite for the solution was the Virtual Private Database (VPD) features in Oracle that was used to simplify maintenance. One physical database supports all the institutions, but the data are logically separated for each institution. The fact that each person employed by an institution or associated with an institution, is uniquely identified by his or her Norwegian social security number was also integral to this solution. The restructuring of FRIDA increased the sharing of common data and at the same time maintained the institutional perspective or view of the system.

The solution requires that we maintain the institutional specific “databases”. However, various components is shared or partly shared among the institutions. Furthermore, we introduced two concepts and divided the data between these concepts: A global part, that is a common /shared
part and a local part that is the institutionally specific part. The global components/concepts are those that do not need to be registered more than once, and where it is acceptable to share information both with respect to registration and to presentation between the institutions. The local components are those that do not contain duplicate information, but contain institutionally specific information where only a certain institution has the rights to register data or change the data content. Examples of global and local components are: Global person contains information that uniquely identifies person, name, title, etc. Local person contains information about employment at the institution. The institutions themselves decide who is allowed to access this information.

Global Scientific publication contains information about the title, references to the journal etc. When an author is referenced from the Global scientific publication it refers to the Global person. Local Scientific publication contains mainly information concerning the approval of the post/entry. The institutions themselves are responsible for approving data that are reported to the Ministry. Examples of this type of information:

- Whether the post/entry has been approved or not
- Id of the user who approved the post/entry
- Date for approval.

Other local data connected to publications are for example institutionally specific questions and answers concerning a submit service to institutional archives. Figure 3 shows a high-level model of the relations between global and local concepts.

Although the conversion from FRIDA to CRISlin seems to be simple and easy, we admit that we spent a long time finding the solution. Senior analysts had a lot of discussions about the problem. At some time we thought we could never do it. And when we found the solution, it seemed so easy.

The FRIDA system was restructured into the CRISlin system between the summer of 2010 and the autumn 2010, and by October 2011 about 160 institutions in Norway were put into production.

10 institutions in HE sector were using the FRIDA system.

5 Adapting to the CERIF model

When the predecessor to CRISlin, FRIDA was developed (2002 -2003), CERIF did not contain the entity ResultPublication with all the information and structure we needed at that time. The ResultPublication modules were the most important part of FRIDA because it had to be able to deliver data to the performance based budgeting model of the Norwegian Government. Although we use a different data model and syntax, it will be possible to translate the CRISlin model to a CERIF–model for all the relevant and common objects like Person, Project, ResultPublication, organizational unit etc. It is important to be able to exchange data with other CRIS-systems in the future. At the moment this has not been addressed or in focus as much as needed. From a well-structured model it is possible to map the data to another structured format at the same level or lower.

In CRISlin there are concepts that are not in CERIF and vice versa. For example a general questionnaire module is modelled and implemented in CRISlin. This module is used in a submit
service which allows the submission of journal papers and other content to Open Access Systems and questions and answer concerning legal rights. The questionnaire module concerning ethical issues related to both publications and projects has also been implemented in CRIStin. Using this functionality, researchers may choose to confirm that they vow to follow discipline-specific ethical guidelines (such as the Helsinki declaration). They may also choose to declare that a publication adheres to the Vancouver guidelines for co-authorship. In addition, researchers must confirm that they have obtained the required approvals and contracts for a project. The goal is to increase general awareness regarding sound scientific practices and the correct use of and accurate registration of data.

Other concepts in CRIStin that we do not find in the CERIF models are Presentation entities that can be of type; research groups, organizational units, research centre including a brief description of each research units and lists of associated researchers and all unit-related publications and projects. Research visits abroad such as sabbaticals are not easy to find in the CERIF-model. Concepts that are not in CRIStin are Facility, and Equipment. CRIStin is planning to use the general concept presentation concept to present infrastructure.

6 International arena

Research is performed in an international arena with co-authors from different countries. Next level of development could be to make a model that was not so tied up to a certain country’s specifications and at the same time not being a lean system. We still need to have a system that allows the specifications of different national regulations and rules and that also satisfies needs across the national borders. Some issues and challenges must then find a solution.

An example is the identification of researchers across different countries and within a country. The system must have a data model that handles, and supports different identification schemes both institutional and preferably national like in Norway with the candidate key Social security number that uniquely identifies each person in Norway. Our opinion is that when we have order at the national level we can start thinking internationally. (You cannot get rid of national disorder by pushing the disorder out internationally!) In order to get an international identification, the system in each country must allow a researcher to have an ID from different countries and associated to him or her if they have been researcher in more than one country.

We believe that a service at international level could be established. Each national system can then get a unique identifier based on their national identifier. The international identifier of the researcher is only useful when the researcher is linked to other data such as publication and projects. Therefore it is important that we get the suppliers of bibliographic data to use this common identifier in connection with their data. Additionally, other important international players and stakeholders could use this identifier or demand the usage of it for instance in application processes for Research funding. It is also necessary in an international perspective to have a common organizational register at top level.

The data model of an international CRIS system must support different national calculation rules, for example of publication points. The data model must reflect these different calculation rules and different calculation can be programmed based on the rules in the current system.
7 Conclusion

Here is a summary of some of our experiences when developing administrative information systems.

- A well-structured and generalized data model is the secret behind our successful systems
- The data model must evolve over time
- The development process must contain activities so that the data model is audited/focused on as new demands and more knowledge of the domain appear
- The system developing and modelling phase must not be left to the user or expert user of the system
- Our experience is that analyzing and data modelling skills are hard to find
- Standards become more important in communicating with other systems
- Most development processes have too little focus on data modelling

References


http://www.ormfoundation.org/

Contact Information

Grete Christina Lingjærde
USIT (University Center for Information technology)
University of Oslo
Forskningsveien 3B
0313 Oslo
Norway

g.c.lingjarde@usit.uio.no