PhUSE 2014

Paper TT01

A Day in the Life of an RDF Curator

Josephine Anne Gough, F Hoffmann-La Roche, Basel, Switzerland

ABSTRACT

Semantic Technology and RDF have become additional tools for use in Clinical Trials and Clinical Data Standards in particular. At Roche we have one of the first production status RDF based Clinical Data Standards Metadata Registries and have been operating it for 2 years now. We have real end users, a browser, search functionality, web services, and big plans for study workflow automation. But what does it really mean to set up and run a system like this? What resources and skills do you need? What problems do you encounter day by day? This talk is about real life as an RDF curator

INTRODUCTION

The purpose of the paper is to describe the normal working life of an Information Architect in a department that curates an RDF repository containing Clinical Data Standards Metadata. It describes their skills, tasks and experiences in their new role. At Roche we have a full time team of five people and one consultant to work on our RDF repository. The paper is written in conversational first person and is not intended to be an introduction to RDF itself.

WHO AM I?

My name is Amy, Didier, Robin, Ivan. I am an Information Architect (IA) working in the Roche Data Standards Office (DSO). This is a relatively new job for me, in my old job I was in Clinical Data Management, SAS programming. It was a bit of a leap of faith to apply and accept this new position it's a totally new field in Pharma. I found out about the job through a Roche intranet site called the GDSR (Global Data Standards Repository) which is run by the DSO.

So, this Information Architecture job is really about RDF, but what is RDF? It's technology that's been around for about 15 years. RDF stands for Resource Description Framework. It means we can describe all of our data standards about CRFs, SDTM and ADaM in abstract models and then fill those models with content. I'll tell you – it was a bit of a mind shift from SAS, relational databases, Excel or documents! RDF means we can really describe our standards in a machine computable format; it means we can uniquely identify things; it means we can really link things together. RDF is about describing information, formalizing assumptions and linking knowledge together via a network of explicit models. It's really cool - when you get to know it, you get hooked.

The GDSR is Roche's mega RDF Dataset where all Roche and CDISC Clinical Data Standards are defined. We've got everything in there, CRF designs for each Therapeutic area, SDTM, Questionnaires, Labs, Code lists, and more content is constantly being added. The GDSR has a browser for our end users, search functionality, and web services, but I'll explain that later. The GDSR is accessed by anyone in Roche who needs information or to download products about standards. It's available 24/7/365 to the business. The GDSR sort of makes the Data Standards department hip and special. Are we a data standards department that lives in back room offices maintaining our standards in spreadsheets? No – we have a web site, we have download products, and we have visibility…

So what is a typical day? My boss keeps me really busy, we call her the Cheffe - sounds tough but she's nice – Cheffe is a term used for the boss on German building sites, that's what it's like, really a building site! Organised construction. Cheffe makes sure that we are properly trained and that training is part of the job and boy is there a lot to learn!

The learning list for this job comes in several buckets. First the technical framework, we need to learn about RDF of course but in addition there is RDFS, some OWL, SKOS, SPARQL, XML and XSLT. Then we need to know about web services and HTTP. I also need to know the "domain", that means knowing at least the CDISC standards that I'm working on (I'm so glad I knew some of these already!), then Roche also have their own CRF standards and EDC system. I need to know ISO 11179 since it forms the base of our metadata registry.

……..And it doesn't stop there! We need to learn about tools such as TopBraid to edit our RDF models, oXygen to develop XSLT code and the GDSR itself has administration features, validation code, search configuration, and browser/screen design and development. We do product development so I need to know how to do user requirements, product development and testing. Sometimes I'm a bit overwhelmed by the amount there is to learn, but each week I begin to understand more and more, it all starts to fit together, and it's so interesting that it almost carries you with it.

Oh…and I almost forgot the crème de la crème of this job is learning about the RDF models themselves and the actual task of modeling. The GDSR consists currently of some 40 odd linked models, some simple, some complex but each one carefully crafted. When you first look at some of these models it's like Chinese spaghetti you think you'll never understand what or why. But when you learn the domain and work with models making changes to the content (I'll come onto that later) it's a journey of discovery – hey the world is no longer restricted to 2 dimensional SAS datasets for me I can walk the models linking CRFs with SDTM, SDTM with ADaM. One day I want to be able to design my own models….I think that's what we (The IAs) are all aiming for.

So back to my day job, well every day I can split my activities into three main areas:

- Day to Day changes to the standards in the GDSR
- Product development
- Modeling

DAY TO DAY CHANGES TO THE STANDARDS

People think that standards are fixed but as we all know in the business they are constantly being refined and added to. There are two other departments in the Roche DSO group who look after the Collection (CRF) and Tabulation (SDTM and ADaM) standards and who run Governance committees for that. When they've had their discussions and they want to change a standard then they send our IA group a service ticket via a system called JIRA telling us that we need to change the GDSR. Today I have a bunch of requests to do: they want to change one of the CRF labels on one of the Medical History forms, the SDTM group has new annotations for some new CRF fields, and there is a brand new Questionnaire to be uploaded and a new Lab conversion unit for the standards.
What this means is that I have to edit the RDF models themselves with the changes using our tool TopBraid, QC them and then make them publically available in the browser. This was a bit scary at first but you kind of feel proud when that change appears in the GDSR browser. As for new standards my mate is currently working on new CRFs for Hematology, Oncology and Ophthalmology.

When the updates to the models have been finished then I will have to publish the download products – what does that mean?? Well it's all very well having a fancy RDF database but people need the standards in other forms for their daily work. So we have various download products for them. For example we create CRFs in PDF format automatically so that Medical History form change I made was available for our users in a PDF CRF document within minutes of me making the change. The SAS programmers also get Excel spreadsheets about all of our SDTM set up; we also make CSV files for lab unit conversions which are used again by the SAS programmers. Basically we make products for the business based on the information contained in the standards – that's what it's all about really making the standards, automating with the standards, driving efficiencies and quality in the business through standards products. So that brings me onto the next group of tasks I work on.

PRODUCT DEVELOPMENT

How do we get the data out of RDF to create products? We call that Web Services and what we get out of the GDSR is the standards data in XML documents. I could do a whole presentation on just this one subject but we haven't got time, so I'm just going to say that we can configure our XML output via RDF which makes it really powerful. What do I mean by that? Well for example our SAS programming team wants the structure of SDTM delivered in XML for their conformance checks. We ran off our basic web service for that and they came back to us asking if we could change the way the tree in XML presents the data—sure we can! All down to our configurable web services.

Today I have to make a change to our most complex product to date; the EDC study builder. Basically we take the information about all of our TA data collection standards (or Therapeutic Area CRFs) and create a file ready to upload into our EDC (Electronic Data Capture) system as templates for the online study data entry. The EDC system accepts an Excel/XML upload sheet so we convert the XML from the RDF repository into the particular format required by our EDC system using a language called XSLT… Now I have a programming background but still this XSLT language is weird. Cheffe sent me on an XSLT course and I was still a bit confused but one of the other IAs, who has been here longer than me, helped me to practice on small examples and now I understand it.. This piece of XSLT code is 2000 lines long I'm glad I only have to edit it and not write in from scratch! OK so to give you an idea about the whole subject of product development, the total list of GDSR products in production to date is as follows:

Operational CRF in PDF
SDTM and Code lists in Excel and XML
Lab Analytes in Excel and CSV
Lab conversion factors in CSV
Lab synonyms in CSV
Questionnaires in Excel and CSV
EDC study build for each TA (Therapeutic Area)
PK test codes and units in CSV

We are working on some new products for example, creating submission CRFs for esub, and SAS code to transform our EDC SAS datasets into our SDTM format.

The most visible product to our end users is the GDSR browser itself. RDF in motion. The RDF triple store (name for an RDF database) at the heart of the browser runs on a server with 24/7 availability, backup and service desk support. The front end is a model driven user interface. What that means is that the IA team has control over the screens via yet another set of RDF models. So when we have to add a new set of standards say for example Biomarker into the browser, then we have to design the UI, configure it in RDF, test and then release.

MODELING

The third part of my job is modelling or should I say at the moment I'm just a learner. How do you learn to model in RDF? It really is a skill that has to be acquired, Cheffe calls it an apprenticeship and I think she's right. You need to first understand RDF. This is done by reading and training and changing existing model content as in those JIRA requests I mentioned before. It helps to look at models that experienced modelers have already built based on a domain that you are familiar with. So for example I know the SDTM standard and then if I walk through the SDTM RDF model I learn about the power of RDF when it represents SDTM. Understanding RDF is a layering process.

The next step is to experiment on your own so Cheffe has set me on a project. She gave me the PhUSE/CDISC ADaM model from the GitHub. Then she says read everything about ADaM and then see how it is represented in RDF. Then she says examine how our Roche ADaM submission datasets are designed and see how we need to extend the PhUSE/CDISC ADaM model to represent the Roche Data Analysis datasets. I started out drawing loads of balloon pictures wondering if I'm getting anywhere. When I reached a certain point I discussed it with Cheffe and the team, I should say that our team includes a very experienced RDF modeler, and gradually piece by piece I begin to understand the task and build the options. We will have a group IA discussion to decide on the final model that way everyone learns. I have to say our team really works as one. We all started in this job over the past 12 months, we all learn together and we discuss openly. It's a great environment to be in.

My colleagues have also started to learn about modeling; Amy has a big job to redesign our Data Collection schemas to handle a Schedule of Activities, Didier he has to figure out how we will represent Biomarker data.

CONCLUSION

My days are very busy and often very long but I am learning loads and I can see my work has a direct impact on our end-users, be it changing the way the browser looks, giving the programmers the new CDISC code lists from last quarter, or discussing the User Requirements to automate our EDC study build. All in all it's very satisfying and I really enjoy my job. I was already an advocate of data standards, but what I see and practice every day is that by making data standards machine computable they can drive the business. …and to be honest from what I have learnt, you could not achieve all that we do without having at the heart of it an RDF triplestore.

ACKNOWLEDGMENTS

I'd like to thank my team for their willingness to jump into all sorts of deep ends and come up swimming time after time. Amy Klopman, Robin Köger, Ivan Robinson and Didier Clement. Especially thank you to Frederik Malfait the creator of the Roche GDSR for giving us such interesting jobs.

RECOMMENDED READING

Dean Allemang and Jim Hendler, *Semantic Web for the Working Ontologist* (Morgan Kaufmann July 2011)

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Josephine Anne Gough (Cheffe)
F. Hoffmann-La Roche Ltd.

B.670 R.313, Malzgasse 30
CH-4070 Basel, Switzerland
Work Phone: +41 61 688 76 05
Fax:
Email: Josephine_anne.gough@roche.com
Web: