Quadcopters or Linguistic Corpora – Establishing RDM services for small-scale data producers at big Universities

Göran Hamrin
KTH Royal Institute of Technology (Sweden)

Viola Voß
University and Regional Library Münster (Germany)

Quadcopters or Linguistic Corpora – Establishing RDM Services for Small-Scale Data Producers at Big Universities

Göran Hamrin
KTH Royal Institute of Technology
Sweden
ghamrin@kth.se

Viola Voß
University and Regional Library Münster
Germany
voss.viola@uni-muenster.de

Abstract

During the IATUL Conference 2017, the authors had many productive exchanges about similarities and differences in Swedish and German higher-education libraries. Since research data management (RDM) is an emerging topic on both sides of the Baltic Sea, we find it valuable to compare strategies, services, and workflows to learn from each other's practices.

Aim: In this talk, we aim to compare the practices and needs of small-scale data producers in engineering and the humanities. In particular, we try to answer the following research questions: What kind of data do the small-scale data producers produce? What do these producers need in terms of RDM support? What then can we librarians help them with?

Hypothesis: Our research hypothesis is that small-scale data producers have similar needs in engineering and the humanities. This hypothesis is based on the many similarities in demands from funding agencies on open data and on the assumption that research in different subjects often creates empirical results which are different in content but similar in structure.

Method: We study the current strategies, practices, and services of our respective universities (KTH Royal Institute of Technology Stockholm and Westfälische Wilhelms-Universität Münster). We also study the work and initiatives done on a more advanced level by universities, libraries, and other organisations in Sweden and Germany (e.g. Stockholm University, Swedish National Data Service (SND), Cologne Center for eHumanities at the University of Cologne).

Results: The talk will give an overview of how we did the groundwork for the initial services provided by our libraries. We focus on what we are doing and in particular why we are doing it. We find that we are following in the leading footsteps of other university libraries. The experiences shared by colleagues help us to adapt their best practices to our local demands, making them better practices for KTH and WWU researchers.

Keywords

research data management; academic library; KTH Royal Institute of Technology; Westfälische Wilhelms-Universität Münster; University and Regional Library Münster

This document contains the speakers' notes for the talk given at the 39th Annual IATUL Conference 2018. The full paper will be published later this year.

For the slides to the talk please see the IATUL repository, http://docs.lib.purdue.edu/iatul/.

Outline

This talk has its origin in the first day of IATUL 2017 and my secret algorithm which optimises your seating whenever you do not know anybody invited to a dinner.

This led to many interesting and productive discussions concerning the similarities and differences between German and Swedish university systems and between library-research support services in the two countries. The following talk is the result of one of these discussion topics.

We will walk you through the research data management situation in our respective subjects, countries, and institutions, trying to highlight important differences and similarities. We will also present some RDM initiatives that have been relevant for developing services for our libraries.

This walk-through – or rather: gallop –, which is a kind of "two talks in one", can of course not paint a complete picture (for which we refer you to our paper) but at least it serves as an entry point for further discussions during this conference.

Two cultures, two countries, two universities

When you look at the questions that researchers ask us librarians about RDM, you often cannot tell from which discipline they are: they all need advice on how to fulfil funding requirements, on where to store data safely, or on how to set up a data management plan.

So could it be that disciplines for example from engineering or the humanities are not as different as they may think – at least regarding RDM?

Back in 1959, Charles Percy Snow gave a lecture called "The Two Cultures" at the University of Cambridge. He painted a grim picture of the division between "scientists" (including applied scientists or engineers) and "intellectuals" (by which he meant humanists and some social scientists).

We will take Snow with us on our walk to see whether the polarisation between the scientists and the intellectuals he postulated can still be found today.

KTH & KTHB

KTH is Sweden's oldest and largest university as well as (according to some rankings) the best technical one. The activities of the KTH Library (KTHB) are as old as KTH, with the first chief librarian assigned at the conception of KTH in 1827.

KTHB currently serves KTH faculty and students on the different campuses with help of around 50 employees. The library provides a selection of research support services. For example, it curates the KTH part of the Swedish publication database DiVA, which enables open access publishing to its archives. KTH-DiVA is also used for evaluation and bibliometrics.

WWU & ULB MS

With about 43 000 students, 675 professors, and an academic staff of 5 050 the Westfälische Wilhelms-Universität Münster – I will use the German abbreviation "WWU" – is one of the biggest universities in Germany. Its 15 faculties cover the main scientific disciplines apart from engineering and veterinary medicine.

A team of 248 colleagues is in charge of the library system. Apart from the "usual" services of a big university library, the ULB has a long history of services for open access publishing. Expanding the research support services to the management of research data formally started in 2017 with the publication of a Research Data Policy. More on this and the WWU eScience-Center later.

RD in Engineering

Research data in engineering are often simple to define; RDM in engineering is not that simple. For example, given the emphasis of quantitative research, it is no big surprise that much data is of the ordinal type. That means that it is easily stored as vectors in a large database and subsequently easy to process, compute, and visualise.

That does not mean that the same data set is easy to manage in the long run. If the data set is meant to be available for further computations in the future, then it must be archived both according to laws and rules and to the demands made by research funders.

It must also be marked with metadata and stored so as to allow for easy access.

Three common examples of data are the following: fluid mechanics data, computing data, and geopositioning data, as a result of an empirical-inductive process.

In our example, the construction of flying vehicles such as quadcopters, all the above kind of research data can be collected during the construction process. The aerodynamic properties of the aircraft are fluid mechanics data, the ability of the steering-system software to quickly adjust to input from sensors is computing data, and this sensor input includes geopositioning data collected from gyros and GPS on the aircraft.

There is research data in engineering that is sensitive. Biomedical data is one obvious example, as well as the aggregated traffic data collected when monitoring commuter systems via GPS or CCTV coverage.

RD(M) in the Humanities

For the sake of simplicity I define “humanities” as ‘everything that is not natural sciences or engineering’.

Research data in the humanities often differs in several aspects regarding the type of data and its usage:

· Research in natural or quantitative sciences is mostly based on measurements or surveys, while the humanities work on representations of cultural artefacts like texts, images, audio recordings, or physical objects;

· while measurements and surveys lead to structured data, data in the humanities is often only modelled during the research itself, through describing, sorting, annotating, etc.

· It can then be saved in different formats and aggregations.

These and some more aspects leave us with a complex situation: diverse types of data in different layers, linked to other data, and “corralled” in specific technical settings that have to be kept as “living systems” to make “useful” reusage or reproduction possible.

Two projects may offer a glimpse into German DH research and their data.

· A project in Leipzig wants to analyse handwritten music scores of folk tunes. The digitised scores have to be transcribed to machine-readable music. As this is not possible via optical recognition software, the project decided on a crowdsourcing approach – for which they had to develop their own platform.

· Meanwhile historians in Münster are working on medieval heraldry. For this, different sources like images, artefacts, architectural information, and texts have to be made available. One of the outcomes will be an ontology of coats of arms to enable the description, documentation, retrieval, and processing of relevant data.

RDM in Sweden

Sweden is a small country. Hence the practical abilities to mimic German institutions are severely limited by size and economy.

But: if the individual institutions in Sweden are too small to handle their research data individually, then they need to cooperate.

They can do so via international initiatives. The canonical example here is “The Human Protein Atlas” project: It was originally initiated at KTH, but is now a global initiative with the aim to map all the human proteins in cells, tissues, and organs, and provide that data open access.

Or they cooperate via national initiatives such as the Swedish National Data Service (SND).

The SND has currently no large data storage capacity and is used mainly for humanities and health sciences. The future of SND includes an expansion of the services to the natural sciences, which could give a possible solution for RDM in engineering.

But to realise the expansion to “SND 2.0”, a large distributed storage solution has to be implemented, as proposed by SUNET, an organisation providing IT services for research and higher education.

There are other solutions like the ad hoc use of the national publication database DiVA, which currently has no database architecture or proper user interface for storing and archiving datasets on a larger scale.

At Stockholm University a data repository is available via Figshare services, and also a Big Data service available at an extra cost.

An example for a domain-specific solution is the “Tilda” system of the University of Agricultural Sciences that will collect climate data.

RDM in Germany

Germany is a federalist country with 16 states, 17 Ministries of Education and Research, and over 425 universities and scientific organisations – it comes as no surprise that RDM is dealt with in many a place, constellation, or context.

So this is only a very rough overview.

It starts with a lot of paper – or PDF files: different institutions have published statements, principles, or recommendations. They all define RDM as an important strategic task for science and politics, and they all ask for a coordinated approach.

One important outcome was a proposal of a national research data infrastructure called NFDI for all scientific disciplines.

In two states so far there are groups working on the NFDI from the states’ institutions’ perspectives. This is a typical German example of how national ideas are broken down on state level, resulting in more initiatives and more papers. But hopefully they can also help raise awareness and act as a kind of mediator between the federal government, the state governments, and the institutions.

There are some groups working on a nationwide level like the German branches of the Research Data Alliance or of the European networks DARIAH and CLARIN.

Several initiatives have collaborated on two platforms “Forschungsdaten.info” and “.org” for general information about RDM, and on “Forschungslizenzen.de” about copyright and licensing. I expect more sites like these to come up, as “only three” is quite few for German standards…

On a local level, several universities can be seen as “role models” for the development of services in Münster. For example the Cologne Center for eHumanities with its Data Center is a

good place for “espionage”.

There are many papers to read, abbreviations to learn, and initiatives to follow in Germany.

While “many people working on a problem” can lead to many good ideas, it can also lead to duplicate structures and developments. So hopefully the different initiatives will keep contact and more central ideas will be introduced.

In Münster, we will try to reuse many solutions developed elsewhere and to cooperate with others to create synergetic effects.

RDM at KTH(B)

At KTHB, we have staff with relevant subject qualifications. Besides expertise in library and information science, there is proficiency in chemistry, biochemistry, ecology, mathematics, and computer science.

But without support from highest management, RDM services can scarcely be productively implemented. So, although we must work by responding to the few questions that we are asked, we need a formal mandate and funding from the KTH board.

After receiving that, the next step will be to enrich the publication policy with RDM statements. This also has to be decided on by KTH management for further implementation on KTH School level.

I see KTHB as the primary developer of that RDM policy. For that, we have to actively participate in RDM networks and monitor RDM policies already in effect.

We started our RDM support by forming an informal working group with people from Archive, IT, Research Office, and the National Infrastructure for Computing.

We documented the current state and the future plans in a report for our Chief Librarian, and we started to attend selected networking or information meetings in order to meet and engage with researchers and other parties at KTH vital for RDM.

We have built, but not yet published, a support web site with Q&A. We have also started to improve staff knowledge on GDPR and RDM.

We are currently awaiting a formal mandate from the KTH president. After receiving that, we may continue our work, probably with the recruitment of special competencies necessary to expand and scale up our support services.

RDM at WWU & ULB MS

The WWU started RDM activities in about 2015.

By that time, it could look back on 15 years of cooperation between the library, the central IT services, and the administration. This alliance, called “IKM”, coordinates the planning and maintenance of digital infrastructures and services.

The WWU rectorate asked the IKM group to develop an RDM strategy for the university. This put the topic on the “official agenda” – which was an important step for the establishment of the matter. After a survey and intense discussions, a Research Data Policy passed the senate in 2017.

Of course RDM questions had been raised before – but only occasionally, mainly in the humanities.

Since 2016 the WWU had developed a “digitalisation strategy” regarding every aspect of teaching, research, and administration, and RDM and DH are important factors in the new development plan.

The nucleus of the eScience strategy is an eScience-Center with a Service Point RDM, a Service Point Digital Humanities and other Service Points to follow.

The eScience-Center is affiliated with the ULB, while the responsibilities and competencies for RDM are shared among the IKM partners.

On our to-do list there are e.g.

· Developing a repository for research data that has to be interlinked with the document repository, the research information system, and ORCID

· Developing different tools and an “eScience Cloud”

· Coordinating WWU projects and activities.

Of course there is a lot to discuss for each of these steps.

The same applies to the lessons we have learned so far. Most are consistent with the reports of other libraries. The most important is perhaps that we have to be “prepared for everything and everyone”: scientists with a first draft idea for a project or with an elaborated data plan, scientists in the middle of a project or with a finished project – and all of them with a colourful variety of data types and formats.

“Same same but different”? – Conclusions

So: what have we learned from our short run-through?

One prejudice one might have when thinking about “people with screwdrivers” vs “people with dictionaries / Old French dramas / or oil paintings from the Baroque” could be true: while engineers are fully aware of the fact that they are handling data, this is not necessarily true at least for “traditional” humanists. But the closer we come to digital humanities the smaller this gap becomes.

Both data from engineering and the humanities are mostly based on artefacts. But while humanist data can come in many different “flavours”, engineering data is mostly of the ordinal type, making it relatively easy to handle.

This entrains the biggest difference: the demands on the technical infrastructure. If it were only for “keep it safe”, relatively simple storage solutions would be enough for both sides. But if it’s about “keep it alive and running”, things are more complex on the humanist side.

Things are the nearly the same again when it comes to funding requirements, and regardless of the discipline the willingness to share data depends more on the personality of a researcher than on his or her discipline.

Both Sweden and Germany started working on RDM at nearly the same time, following discussions about good scientific work and open access.

In a small country like Sweden with central players some things are easier than in a big country with a decentralized structure. But small countries don’t have the same financial and staffing capacities, which may be an opportunity for close cooperation between institutions, while in Germany the risk of duplicate structures and developments is immanent also in RDM.

An example of a similarity is that in both countries many RDM policies have little or no connection to the actual needs of the researchers. They have to be complemented with more specific recommendations which should take into account the respective disciplines.

Here we could easily pass country borders: Swedish engineers could find inspiration at German universities, while German humanists could have a look as Swedish recommendations – or the other way round.

Apart from the differences in size and number of staff, KTHB can concentrate on those subjects

taught at KTH, while the ULB has to be “prepared for everything”, being the library of a big university where nearly every subject can come along.

Nevertheless WWU and ULB can learn from KTH(B) e.g. in the area of engineering data, which might also be of interest for physics or other technical subjects at the WWU, and of course there are overlaps in biotechnology, chemistry, or computer science.

Regarding engineering and also architecture, the other academic library in Münster, the one from the University of Applied Sciences, could be another interesting partner for joint considerations on RDM.

A problem all libraries have to tackle is keeping track of RDM developments. For this and for deciding which services to implement it’s important to train the staff continuously and to establish a close cooperation between library and faculty – and between other libraries.

As the discussions between us have shown also “cross discipline / country / library type” exchanges can be fruitful for discussing RDM.

We are now coming full circle in our talk. Let us return to our starting point, Snow’s lecture.

Already in 1959 he understood the importance of open data:

In the postscriptum of “The Two Cultures”, Snow mentions that he originally meant to title his lecture “The Rich and the Poor”. He thought that higher education was going astray and not able to fulfil the important task of improving public health and the general well-being by means of the scientific revolution. Snow believed that supporting scientific progress in all countries was the only reasonable way to a better life for this planet.

This view is unfortunately not outdated. In the current framework for scientific publishing, most data and publications are behind a paywall or locked-in in closed repositories or on local computers or USB sticks. This reduces the possibility for less-developed countries to access the current state of research. In this instance, the approach to open RDM is vital.

Does anyone remember the outbreaks of Ebola in Africa or Zika in South America?

In the first case, no structured collection of data from the spread of the disease or patient history was made; instead this data is located on individual hospital teams’ computers. In the second case, data was collected both in more detail and in a more structured manner, helping to handle this outbreak more efficiently.

While reading Snow today might prove useful, discussions with colleagues from all over the world are for sure very useful and interesting.

So we are looking forward to hear about your RDM ideas and experiences!

Thank you.