The Cathedral and the Bazaar

Matthew Addis, Arkivum
RDMF14, York, Nov 2015
http://dx.doi.org/10.6084/m9.figshare.1597674

I’m going to spend 5 minutes talking about why I think RDM is like the Cathedral and the Bazaar - and how this might help us with RDM from a systems integration perspective.

http://dx.doi.org/10.6084/m9.figshare.1597674
Arkivum works with over 30 UK Higher Education Institutions and has been involved in a wide range of discussions and projects involving RDM systems integration.

We’ve produced lots of pretty pictures like this of how various systems for RDM can fit together.
These are all in a report we published recently as part of a JISC Research Data Spring project. It’s licensed under creative commons and you’re very welcome to download and use it.

http://dx.doi.org/10.6084/m9.figshare.1476832
Looking back on the report and what we continue to see, it strikes me that there are interesting parallels with the Cathedral and the Bazaar.

Almost 20 years ago, Eric Raymond published a seminal work in the software engineering community that compares the ‘Cathedral’ model of software with that of the ‘Bazaar’.

http://www.unterstein.net/su/docs/CathBaz.pdf

Eric said that large corporations such as Microsoft build their products like Cathedrals. Hundreds of software masons working away for years at a time in splendid isolation. The product is released only when they deem it worthy.
Maybe institutions can be viewed as creating RDM Cathedrals? Years of chiseling the mighty stones of CRIS and IR. A compliance altar at which to worship the 10 commandments of RCUK.

The doors finally get flung wide, accompanied by an education and awareness raising campaign on the virtues of RDM. But the researchers don’t come. The Cathedral remains empty.
Why? Because researchers like Bazaars not Cathedrals.

Eric Raymond described Linux development as a Bazaar. Somehow a community comes together and as is by magic a wonderfully robust and cohesive system emerges from what appears to be chaos and disorder.
And this is what a research Bazaar looks like. This is part of a brilliant poster from the Utrecht University Library at the Force11 conference earlier this year. A diverse and colourful explosion of tools that researchers actually use for doing research. It’s all about collaboration, publication, reputation and promotion. This is where all the action is. This is where RDM actually happens.

http://dx.doi.org/10.6084/m9.figshare.1286826
Eric stated 19 key features of the Bazaar model that leads to its success. Maybe these can help guide us in the approach to RDM systems integration? I don’t have time to look at all of them, so here’s a few that caught my eye.

First, and Eric said, ‘scratch an itch’.

The more an RDM solution solves an immediate problem for researchers the far more likely they are to not only engage, but crucially help in its construction. Itches for researchers mostly occur when data is active, so that’s the place to start.
Second, Eric said release early and release often. Get something simple out there and do it fast. Give researchers something to play with, to build upon. The Linux model is one of lots of simple tools. Each tool has a job to do, each one does it well, and each one scratches an itch for someone.

In RDM systems integration, this means we need to help researchers to solve their problems. In turn, they will help develop and integrate RDM tools. Get them more involved from the outset.
And finally, one more point that Eric made that I really like is that, ‘all problems become shallow’ given a community of developers.

What he means is that if there are enough people working on a solution then there will always be someone who quickly sees how to get round any problems. Researchers aren’t short of bright ideas.

In RDM systems integration, a problem that looks difficult from a library or central IT perspective might get solved or fixed much more rapidly if a community of researchers are actively involved in creating the solution.
And to make all this a bit more concrete, in the last 60 seconds of this short talk, here’s one example. This screenshot is how computational chemists at Imperial use their HPC facility. They run a simulation and when it’s finished the only step needed to make sure the data is both managed and published is to click a single ‘publish’ button.

This button causes the complete data set of inputs, outputs and settings to get put into a local chemical database repository run by the group. It is simultaneously uploaded to Zenodo or Figshare. Metadata is automatic based on what the researcher has already used to describe the simulation, a DOI and handle gets minted, and the researcher can immediately reference that data in publications they might be writing. No faffing about downloading/uploading datasets from the HPC facility, no waiting for an institution to QA the data and provide a DOI. It’s all immediate and it scratches the itch. Data management and data publication using one button.

I think systems integration to support effective RDM is going to need a lot more of this sort of thing. Eric presented the Cathedral and the Bazaar as being one or the other. In RDM systems integration, I think it’s a case of both. Systems integration can connect the research Bazaar with the institutional Cathedral. And Eric’s observations can help us do this a lot more effectively.

http://pubs.acs.org/doi/abs/10.1021/ci500302p