IT as a Utility Network+ workshop: Libraries of the Future 4

5 December 2013, Oxford e-Research Centre, Oxford

1. Introduction and goals of the day
2. Presentation: Leah McEwan
3. Presentation: David de Roure
4. Topics discussed
    - Data: granularity and preservation
    - Libraries: physical and digital
    - Librarians: changing roles
5. Next steps

1. Introduction and goals of the day

This workshop was the fourth in a series on this subject. The first event took place at the Bodleian Library in Oxford on 17 April 2013 (read the report), a smaller, scoping group met on 18 September 2013 at the University of Southampton followed by a further meeting held at the British Library on 13 November 2013 (read the report – link to follow).

This ITaaU workshop was organised to tie in with the Annual Workshop of the International Association of Scientific and Technological Libraries (IATUL) event Research Data Management: Finding our Role and so the context was research libraries and the science perspective. It aimed to focus on the evolving relationship between research university libraries, professional bodies and consultants by looking at how the emergence of new technological solutions and the design and creation process around these affects their roles and interests. As a result, the goals of the day were to:

- Explore the role of IT utilities in the context of the evolving role of libraries
- Capture key issues in the context of research universities and professional bodies
- Articulate a route forward for investigating and developing findings
- Broaden ITaaU's network of people interested and knowledgeable in this area

2. Presentation: Leah McEwan - Activating the collaboration potential among libraries and scholarly societies in the future of scholarly communication: chemistry as a case study

Background

Leah is the chemistry librarian at Cornell University but is currently on sabbatical, taking a research leave of absence, some of which has been spent with the Royal Society of Chemistry (RSC) working on series of cheminformatics projects. The sabbatical was prompted by Cornell

library's decision, in 2009, to close the chemistry library's hard copy stacks and move 80% of the stock into a high density search facility. She discovered from that transitionary time that access to data matters. If data resources (such as the 5000 volumes of tables of data that did not go into storage) are not available online and are not being kept up then they are effectively dead. There is no point in transitioning to a digital environment and generating data from labs if it's just "out there", in a mess. Making data accessible requires concerted time and effort from individuals or "we're data poor". How can it be made easier, how to have that dialogue with chemists? In addition, to navigate publication mandates, chemists require advisory services. Currently, no one seems to be talking to them about all the messages they are getting, including libraries and e-science movements. It's not a dialogue. These were Leah's concerns and she is focusing on organic chemists as these seem to be the ones struggling most.

Leah also highlighted the Ithaka S+R 2013 report Changing the Research Practices of Chemists. It identified problem areas for academic chemists – they require better knowledge management, infrastructure systems and training – and it contrasted academia-based chemists with those who work in an industrial setting.

RSC projects

With the RSC, Leah is considering the scenario of a younger organic chemist who is just getting tenure and who knows he needs to deal with his data but is not trained to do so. He is organised but only in the sense that his data is stored in workflows and spreadsheets on a local server. He is already used to publishing this data, which is already defined in his head as meaningful data (this is important, it's a development). He needs help to use it in a more digital way and help others to use it. What would the conversation be like for this chemist – what are the pain points? Could the RSC's Chem Spider project help him?

Leah is working on how to help chemists manage their data in this way, by suggesting familiar tools at first, such as working with a master data sheet. Why? You can put a lot of things in there, it's straightforward, a scientist can put it together and it helps their management (e-lab books can come later). The RSC can then take the file and do the conversion and it's not a burden for the chemist. She's also talking to the American Chemistry Society and working with them on workflow issues.

"It sounds like baby steps but that's where we are. That's the concept. Our favourite solution might be technically brilliant but if it's not reachable by the scientists and publishers then it won't take off. We need momentum to get things going and the master data sheet is a nice compromise."

Another project Leah is working on involves looking at data mining or data compiling from published literature. The RSC has digitised its publications from the late 1870s onwards and would like to be able mine the information that's in there. Starting with an exploratory phase covering the last 10 years of digital files, they are considering: what can we pull out and how feasible is it? Could we digitise those data compilations that are in hard copy and how valuable would that be? What makes meaningful data? How can it be organised? What's the context around it and how much of that context do you need?

Leah is also talking to chemists about how they construct their figures and charts as that has an impact on how the data can come out. She is working with publishers and authors on developing better style guides and templates for manuscripts that can help them with their tables and figures.

Finally, Leah is tackling the case for data re-use relating to chemical safety: can Chem Spider be used as a data safety source and can those connections be leveraged to pull information sources directly into a tool that chemists can have in the lab? The information around different levels of hazard and safety needs to be put right in front of them.

"It's a great opportunity for the chemical safety community to understand how to curate their own sources – to find their pain points and get them to look at that. To get the people who are involved engaged in what it means for their data and information to be moving around in a digital environment."

Disruptive statements

Leah concluded with "disruptive statements" based on her work with chemists. She has discovered that some of the topics and ways the discussions around them are framed are distractions – for the chemists and for their librarians. These are: open access, social networking, data-driven science and big data and undergraduate education. These are "big picture issues" but, on the ground, that's not the conversation point for chemists. It overwhelms them.

Instead, the three starting points need to be:

- Online workflow and technical literacy (they *know* they are not up to speed and are overwhelmed)
- Discoverability and broader impact (could potentially be about depositing data in an open way and this might introduce the topic to them)
- Validation, authority of data

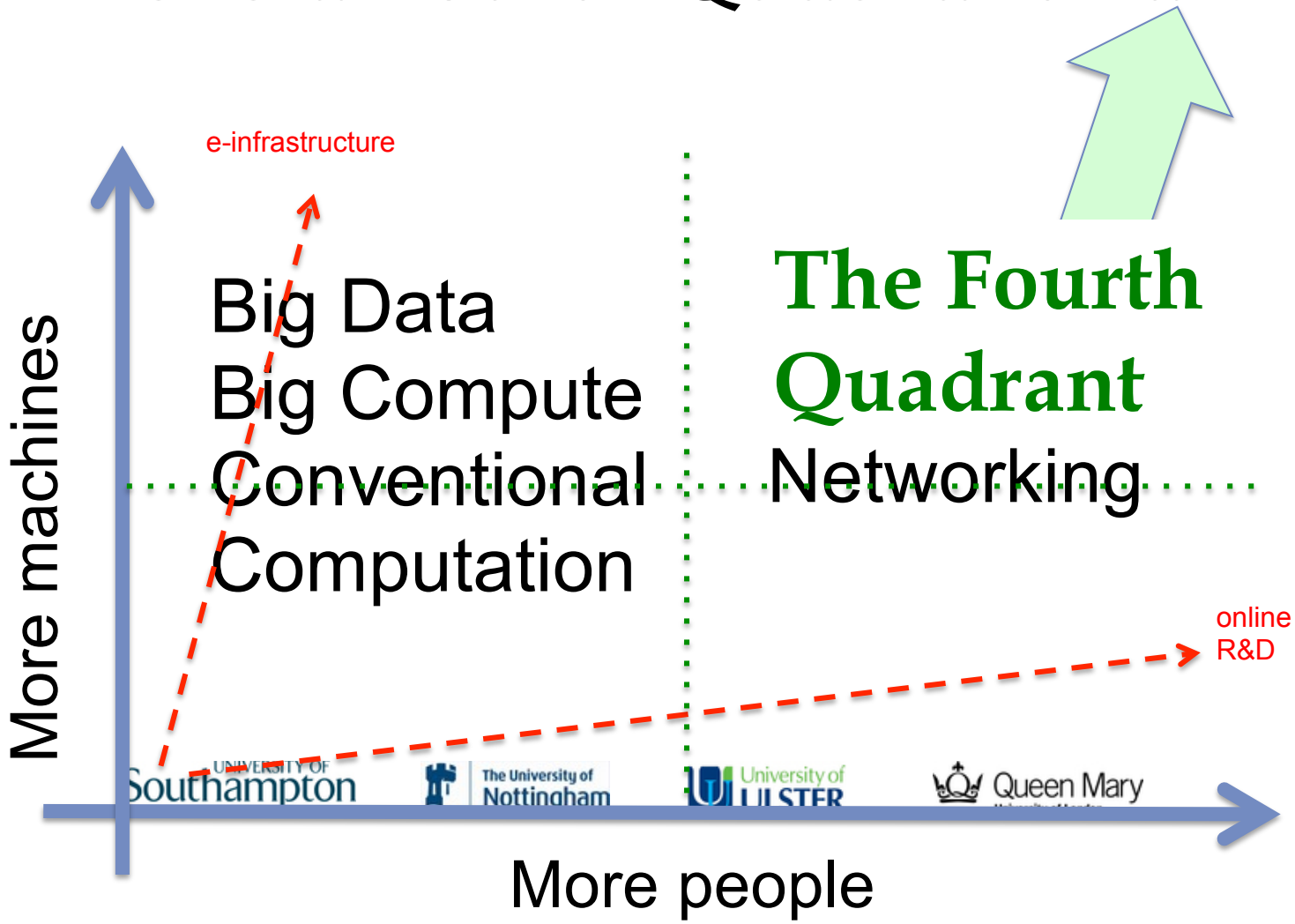
Above all, libraries are about the conversations - whether it's people talking to authors of the papers or the librarians. It's not about a place or about the "stuff" but about the interactions.

3. Presentation: David de Roure, professor of e-research at University of Oxford, director of the Oxford e-Research Centre - Social Machines in the Library … OR Libraries in the Social Machine

To understand the future of libraries we need to understand the future of scholarship; social objects; and social machines.

ITaaU Network

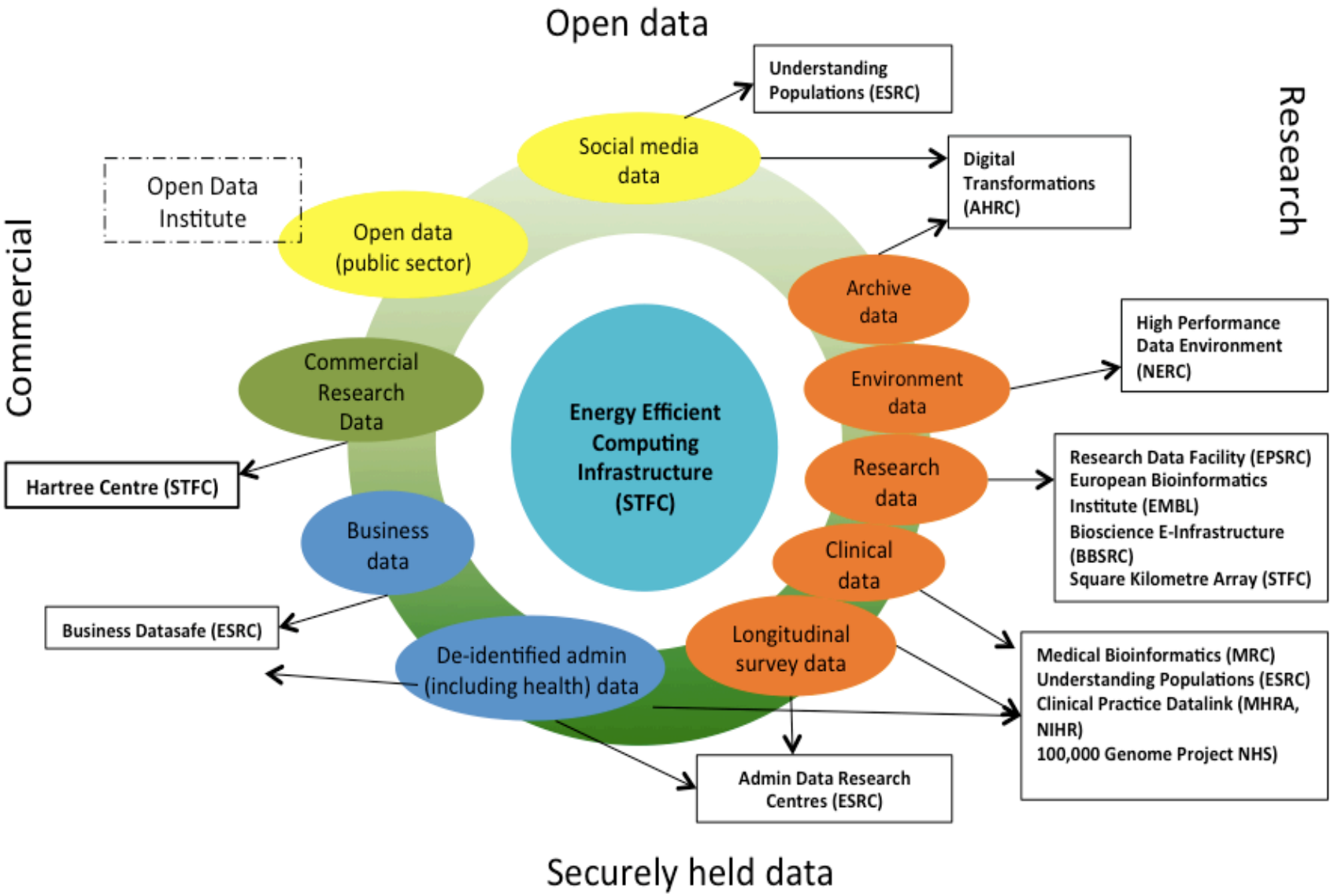
This is a Fourth Quadrant Talk



As time goes on, the future becomes the fourth quadrant. In the future of scholarship we are seeing an increase in citizens and scholars alike and a data deluge affecting every discipline, from social scientists to architecture.

This is leading to methodological changes - if you start with the data rather than the hypothesis then you can find out different things. Does it make obsolete other ways of doing things? Or is it just another tool? We are having to reconsider the power of correlation over causation – we can now do things with prediction without understanding what's going on. Whereas traditionally social scientists have scaled the problem they are trying to solve to the power of the computer on their desk, they are now trying to work with much more data and in real time and having to build big computers to do the analysis. This is a sudden complication in social science with big data causing changes in methods. New forms of data show that old methods do not work and we have to do something different.

ITaaU Network



ITaaU Network

Experiments are increasingly automated – just look at overnight results, notifications and automatic re-runs. Self-repair is happening. Systems are sitting there doing research. The machines are users too. Take Zooniverse, of which Galaxy Zoo is part. Through crowdsourcing, humans are being used to classify images, becoming part of the machine, and scientists are using that data. But citizen scientists are also getting good at some of the science and they are communicating their discoveries via forums. Humans are spotting stuff the machines may not. It's a social machine but what's social and what's machine?

Social objects

Big data elephant v sense-making network of people and computers. We still need some notion of the social object. The paper is an identifiable chunk of knowledge. Artefacts are the social objects. We are re-interpreting and re-imaging that. The challenge is to foster the co-constituted socio-technical system ie a computationally-enabled sense-making network of humans and machines sharing *social objects*… not just papers but data, models, software, narratives – new digital artefacts we call *research objects*.

There is a yin and yang of data and method. Data needs to be matched with method (script, protocol etc), of doing with the object. Software tends to get forgotten but it lasts longer than hardware. We need to get people to take it seriously.

The Myexperiment platform was created six years ago as a way to to find, use and share scientific workflows and other research objects, and to build communities. It is niche but proving provocative and it is changing scientific workflows. You find out what people want to share in scientific workflow – slides, pdfs etc. So the social object becomes a pack. Analysing the packs has produced the R Dimensions – the 12 Rs of the e-research record: reusable, repurposable, repeatable, reproducible, replayable, referenceable, revealable, respectful.

Social machines

ITaaU Network

"Real life is and must be full of all kinds of social constraint – the very processes from which society arises. Computers can help if we use them to create abstract social machines on the web: processes in which the people do
the creative work and the machine does the administration… The stage is set for an evolutionary growth of new social engines."

Tim Berners-Lee, Weaving the Web, 1999

We've been thinking social machines for a long time but not really used the name. There are lots of different social machines from Wikipedia (not programmed on day one but the behaviour has emerged socially) to Amazon mechanical turk. Is Twitter a social machine? Hashtags are social machines –they emerged, humans defined around it. Example of CAPTCHA – getting you to transcribe digitise texts as well as how you use the machine.

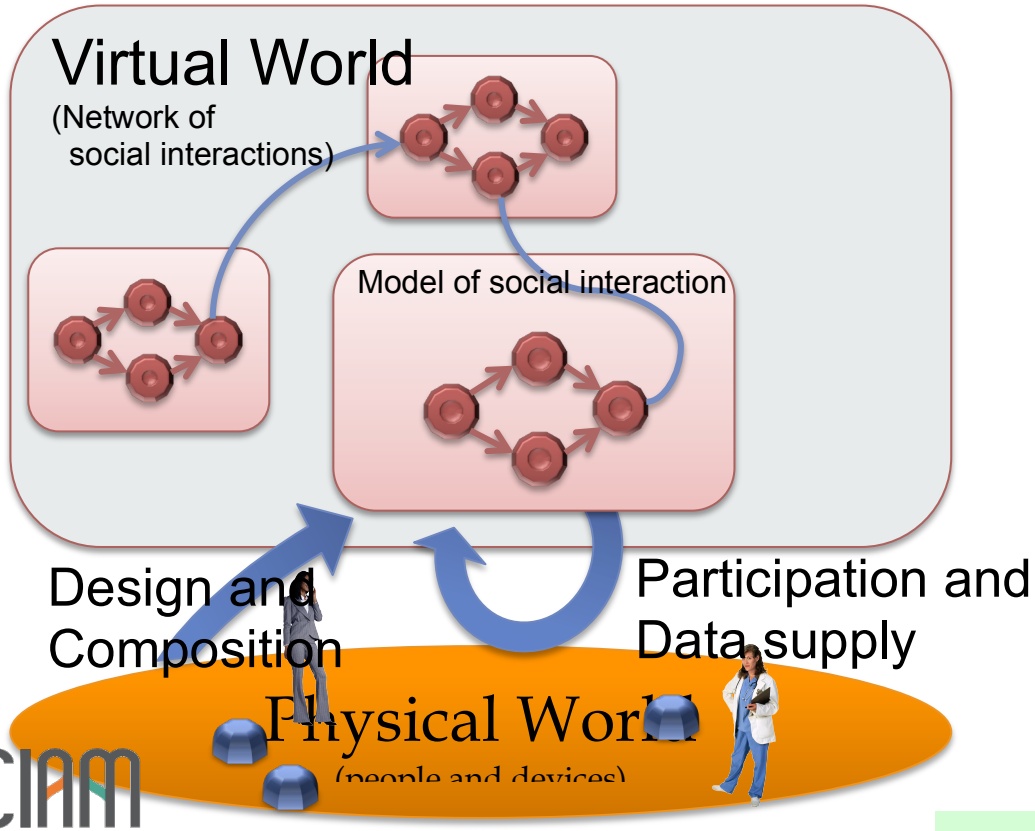
"The myExperiment social machine protected by the reCAPTCHA social machine was attacked by the spam social machine so we built a temporary social machine to delete accounts using people, scripts and a blacklisting social machine then evolved the myExp social machine into a new social machine…"

In scholarly communications there is an evolving scholarly social machines ecosystem. Lots of things coming out that are social machines and are part of the process such as Zooniverse. But we're also seeing different models for peer review and data publishing etc. There is a set of machines in the scholarly communications space – it's not a programme you run and it finishes but an evolving experiment. What's the next thing? Work out what behaviour you want to affect within the ecosystem and then what intervention will cause that behaviour change. We're trying to understand that through case studies of existing social machines and building new ones.

University of Southampton | The University of Nottingham | University of Ulster | Queen Mary University of London

ITaaU Network

Building a Social



SOCIAM

Dave Robertson

University of Southampton · The University of Nottingham · Ulster University · Queen Mary University of London

ITaaU Network

Discussion points

Are libraries:

- Knowledge acquisition machines?
- Knowledge publishing machines?
- Recommendation social machines?
- Annotation social machines? (Data – we should be able to find it, use it and add value to it)
- Provenance social machines?

- Can we build a library out of Zooniverse Citizen Science Projects?
- Can we build a physical library out of devices and new modes of interaction? If we walk into this physical space with our smartphones, have we created a library?
- Can libraries be transient?
- What have we left when we walk out?

University of Southampton | The University of Nottingham | University of Ulster | Queen Mary University of London

Recognition: we recognise social machines but we also don't. We use them rather than recognise them eg Wikipedia. How does a social machine gain recognition by the people I want to consume it? Should we be able to cite a social machine? Citation is powerful. It's how these things become **proxies for reputations** and how reward structures are built around them.

Training and teaching: people can be reluctant to delve into this. They may have core skills but they have to engage with social machines. Transparency and transferability of process is a fundamental aspect of training.

4. Topics discussed

i. Data and library granularity – the paper is all?

Data preservation role: not just books, the role of the crowd

Preservation of process – key part of what we're about

Safety process matters

- No difference between the physical and the digital with respect to function – it's all knowledge. But with digital need tools…Form, format, content mix-up? But does it change when machines come into it? At the end of the whole process is usually a human.
- Safety – tolerance level with safety and machines v humans. Chemists may be less tolerant of error than some other disciplines. Or does every subject discipline think it's special? Methodology is crosscutting between disciplines. Need to know when there is a need to be granular and when it's a general issue.

- Process is more difficult to preserve than pieces of information.
- One way of looking at this is from the discovery angle – how easy it is it to find the things you want? That might be a way to determine if there is the right level of granularity. Physical and digital objects need to be linked in both directions. There is a critical need to have an identifier for your object – then you at least have a chance of finding it. Granularity of key identifiers might be subject-specific eg organic chemist might be a compound whereas for a physical chemist it might be the theory.
- Metadata should come first and object second but in some areas just having the metadata first would be a starting point. But critical point is exposing how that metadata arose. Then you can work out if your question was framed in the right way. Encourage people to provide that metadata. And then discovery becomes easier.
- We need to educate researchers that their responsibility is not only to do the work but to do the work and then organise it in a way that others can find it.

You do it for your own good but if you share it could it be useful for others.

- Proposal: find out how to persuade people to put metadata into things.

ii. [social machine question] Can we build a physical library from Zooniverse?

Where and how do the conversations take place?

Can we build a physical library from devices?

- Why would we want to build a physical library from Zooniverse? Accessibility becomes an issue if you try to make the digital physical. But can we build a physical library akin to (rather than from) Zooniverse? Something more like it in its collaborative approach to providing knowledge and data that can be accessed by a broad number of people whether physically or virtually or via a hybrid version?
- Bringing people together always makes the conversation easier. Idea of the library as a hub – central place for people to come to, to access the information and expertise of librarians and colleagues etc who know what is stored there.
- There is a move away from departmental libraries to central libraries – the mega library rather than local cornershop library. Would it be easier if we could browse the objects virtually as well as physically? The objects can have barcodes and tags and be browsed virtually to find out if your book is there before you go and get it, if the book is the only source of the knowledge.

iii. With changing roles do we need a new name for libraries and librarians?

Professional personal development

- Rebranding not renaming is required. Need to lose mentality about physical space and journals. If there's a case for maintaining the role of librarians then there's a role for it being a profession. With that comes a professional development duty. What do they need to pass on and do? How do they develop themselves?
- Idea: reinventing the departmental librarian role – propose prototyping the embedded librarian. Someone with discipline knowledge, sitting in the faculty with consultant expertise, can advise on the research process and how to find information and how to structure research, record and preserve it, but still have a connection to the mothership, the central library. Guiding people towards the

training they need as well as looking at the bits and bytes that come out of the machines. A triage role and a filter - the first point of contact and an expert in that area.

- They should be a member of the faculty team, undertaking research like other faculty members. What can they do research-wise in that domain as part of the tenure? How could they be an equal among the faculty?
- Note: Jisc funded a study with UCL in 2010 to consider the role of the embedded research data manager.
- Career progression might be an issue, with the risk of being an outpost. However, consider the GP model: they are in the community but they do have strong connections with local hospital.
- In which disciplines could this idea be tested – the most challenging? Computer scientists? Engineering?
- Note: Cornell university has been looking at this model. Purdue university might be a good model to look at where an embedded librarian is faculty.
- What would be the motivation for a library person joining a research group? To see what they could contribute with their set of skills and find their niche? Their motivation might not be getting their name on a paper.
- Possible sales pitch for the librarian – for the library you'll find out more about your users; for the librarian you'll improve your skill set, discover ideas that you could implement eg how to encourage the researchers to submit metadata, conversations about the language (do they even use the word metadata?), a learning outcome.

5. Next steps

- Agreed that the "embedded librarian" idea should be taken forward with a call for proposals for people who want to take on this role as an investigatory trial. The librarian would take a split faculty role for a period of time as a secondment with a view to testing and understanding if this idea would work. Suggested that some sample scenarios are created so that potential secondees could understand what it involves. ITaaU would need a report.
- Possibly, another pilot project on recording process for provenance purposes and reproducibility. Using dissertations for recording process?
- The next ITaaU Network+ workshop on libraries of the future will take place in the spring at Aberdeen library and will consider "library and community". Date of 25/26 March tbc.