Collaboration to Clarify the Cost of Curation

4C
Collaboration to Clarify the Costs of Curation

COOPERATION

Report of Focus Group Meeting on Nov 3rd 2014

In London at

The Wesley, 81-103 Euston Street, London, NW1 2EZ, http://www.thewesley.co.uk/

Project funded by the European Commission within the Seventh Framework Programme		
Dissemination Level		
PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Version History

Version	Date	Changed pages / reason	Modified by
0.01	8 Jan 2015	First draft	KH
0.02	16 Jan 2015	Edited	DS
1.00	22 Jan 2015	Final version	KH

Attendees

4C:

Joy Davidson, DCC

Magdalena Getler, DCC

Sarah Middleton, DPC

Diana Sisu, DCC

Participants:

Catherine Hardman, University of York

Grant Denkinson, University of Leicester

Tim Clark, Harvard University

Gareth Knight, London School of Hygiene and Tropical Medicine

Stephen Grace, University of East London

Event schedule:

- 13:00 Lunch
- 13:30 Welcome
- 13:40 Introduction to CCEx
- 14:00 Time for individual familiarisation with the tool
- 14:30 Focus group discussion on CCEx
- 15:15 Coffee break
- 15:30 Overview of Roadmap, http://4cproject.eu/roadmap
- 15:40 Focus group discussion on roadmap http://4cproject.eu/rmfeedbackand
- 16:00 END

Focus group analysis

1.0 Methodology

Focus group interviews are popular research method for exploring what individuals believe or feel as well as to understand why they behave the way they do.

The main aim was to understand, explain, and

- Find out what users think of CCEx
- To identify what users want / expect from the tool
- To find out what are the common problems when using it
- To collect recommendations for improvement
- Find out if they would have concerns about sharing their cost data and if so, why. What could be done to mitigate these concerns?

Data Analysis: Process

1. Familiarisation with the data (listening to the recording, reading the transcript – 43 pages, reading the observational notes and summary notes after the interview)
2. Developing categories. Look for patterns: repeating ideas, larger themes. Also
    a. Consider the actual words used
    b. Consider the frequency and extensiveness of comments (how often a comment or view is made)
    c. Intensity of the comments
    d. Internal consistency
    e. Specificity of responses (greater attention is placed referring to personal experience)
3. Did we answer our research questions?
4. What theories develop?
5. Implications
    a. What does it mean?
    b. What major themes emerge?
    c. Is the knowledge acquired something we already know, or is it new?
    d. Does the knowledge confirm a hunch?
    e. How does the knowledge change our perspective?
    f. How are participants' environments or past experiences related to their behaviour and attitudes?
    g. **What else do we need to know?** (Additional topics for the next focus group)

2.0 Focus Group Interview

The focus group interview took place on 03.11.14, The Wesley Hotel, London.

FACILITATOR**:** Joy Davidson and Sarah Middleton

OBSERVERS/ NOTE TAKERS**:** Diana Sisu and Magdalena Getler

SIZE**:** 7 participants (respondents). Participants were selected on the criteria that they a) would have something to say on the topic and b) would be comfortable talking to the interviewer.

OVERALL IMPRESSION: The atmosphere was friendly and people felt comfortable in expressing their opinions.

LANGUAGE**:** Language used (selected quotes in yellow, below), special vocabulary – See Conclusions on p. 9.

NON-VERBAL INTERACTIONS: impact of the group dynamic. Three respondents dominated the discussion; more reserved participants could perhaps be probed more.

2.1 Questions asked at the focus group

- What are your genuine thoughts on the CCEx tool?
- Do you find it easy and intuitive to use?
- What do you like best/least about it?
- What are the barriers in using it?
- Are there redundant features?
- Is the language understandable or is there too much jargon?
- Would your organisation submit cost data? Who in your organization would be responsible for submitting the data?
- What kind of support are you currently using for costing digital curation?

2.2 Developing Categories

Feedback was extracted and categorised into Scope of the tool, Problems/uncertainties encountered by users, Suggestions for extra features/improvements, and Comments on the interface/workflow.

2.2.1 The Scope of the tool

- It was not immediately clear to participants what the tool does, how does it work, who it is aimed at:

  R6: *I wasn't too sure what costs we were talking about. I'm used to doing planning ahead for services and working out the kind of business models that we're going to be talking about, how many staff we're going to need the kind of hardware, infrastructure kits and that kind of thing. It wasn't at all obvious to me that what we're talking about here really seems to be depositing data for long-term preservation…*

  Because for me curation needs more than that and I think that needs to be clear up front what this is actually…what you're costing. I was confused at first because I didn't know what it was I was costing. I was trying to cost a service and none of it made any sense…

  just what the scope is because it's not necessarily obvious on the front page...

  R6: *I think in general it helps to be able to explore all this stuff to know what you're doing but you kind of need to know that before you actually start it which comes back to this idea of having to sign up and representing a university before you can even get an idea of what the processes actually involve. I think you need to be able to see what you're going to do and what kind of tool this is much more clearly before you have to sign up to anything. I didn't understand it at first, even though it looks lovely. I mean, it's a beautiful looking website, but I didn't understand it… at first when I came to this, I thought, oh, good, I've got a service I'm planning. I can do this and then I'll go to the CEO and say, can I have £2 million and I'll give you a curation service. Of course that's not the tool for this and it's not at all obvious that that's the case when you first look at the main website...*

2.2.2 Problems/uncertainties encountered by users

2.2.2.1 Conceptual:

- Participants agreed that they tend to add costs first and calculate the total later. It seemed *the other way round* to provide a cost upfront, as it is in the tool:

  R7: *I can see doing this two different ways round. The way it's done is we're going to spend this much money, 50 per cent of which is ingest. The other way, the way I'd probably get that is well, how much time are we going to spend time ingesting, how much hardware we need, how much software we need, what is the total? So it's sort of the other way round because I wouldn't know a percentage...*

  R7 continues: *But I can see that when I actually do this I'd probably do it the other way round. We'd say, well, for this much stuff we're thinking we're going to need about this much hardware and we're going to need this much people, the total being this much…*

  Respondents would prefer to asses resources first (look at purchases and staff mapping), then map to activities (decide how they would spend it per activity – ingest, storage, access) and get the total.

  R1: *When you're adding a cost unit it's confusing to be presented with the activity mapping first of all before you've actually worked out what resources you're allocating?*

- The issue of comparing costs with other organisations. Some participants expressed concerns about comparing their organisation's costs with other organisations. R4 could see the tool used to compare costs internally.

  I didn't really know what I was putting my data in for. I suspect that what I'd probably end up finding most useful for this tool is for me putting in lots of data on different levels about our own costs for me to compare against itself rather than compare with someone else. I do think that we're so different actually in what we're trying to achieve that to try and…that comparison thing, it sounds awful because I know that's the whole point of your project...

  R1 agrees it only make sense if comparisons are made between similar organisations:

  But there must have been an impulse to look at that benchmarking because that's clearly part of what we're obviously setting out to do. I think there is a value in benchmarking as long as you're certain that you're comparing…

  R4 adds:

  That's right, like with like…

  R6 would like to see more questions about curation levels in order to make comparisons more meaningful:

  I'm just to trying to think of this from the point of view of my librarians. They're going to be a bit terrified because in the costs of their curation during this kind of thing, senior management are going to look at it and say, that's bloody expensive, because what you've got on here, you've got the start of it and you can see how many copies people are making of it. There are other ways of measuring the kind of curation level in particular our librarians would be very nervous that there isn't a way of indicating that all the metadata surrounding the deposited data is actually reviewed *by a librarian, so it has that personal review which obviously puts the price up sky high. They're never going to get away with it but they want to. It's a very different level of curation they're hoping to offer. It's going to make the prices look totally incomparable to the other averages you've got unless you can put it in a tick box or some slightly more complex option…*

and continues

Yes, but there's three other parameters involved. Firstly whether all of the metadata has been reviewed on ingest. Secondly whether these copies you're capturing are spinning disk or tape, because that will make a difference to the costs. Thirdly whether there's geographical separation of where the different copies are being held because that will also increase, less of an issue than the other two factors, but they all allow people to make more accurate comparisons…

- Sharing costs: R3 didn't think the BBC could ever share cost information with anyone, even anonymously, because of the risk that tabloids could got hold of it and see how much public money had been spent on x, y or z then however well-intentioned it was it could be held against them.

- Reliability of cost data entered. This knowledge is shared between IT, Library, Research Support Office and reasearchers, not one single unit has the whole picture for costs:

R2: *How do you evaluate the reliability [inaudible 03:12] sadly they'll be less qualified to do that. Also I think you'll probably all admit it's incredibly time-consuming. Even working out the 75 per cent of your time [inaudible 03:30] when costs are [inaudible 03:33] funding may come from different sources as well. I'm not sure that, say nobody [03:40] in the BBC could actually sit down and do it and if so, it would be quite time-consuming. So people will input all sorts of numbers…*

Related to the reliability issue is the issue of staff time: who has time to enter cost data. Most institutions do not employ a research data manager who has the time to chase up cost information so whose duty is it?

2.2.2.2 Technical:

- Terminology used in the tool:

R6: *Is there a list of cost units, because I was a bit baffled at first as to what the cost unit really is…*

R5: *I had the help page open to remind myself what the resources are and services…*

R7:*We know that they're both going to be generic but getting a feel of the hierarchy between a data set and a unit would be helpful because a unit is a part of a thing in a thing in a data set. That makes sense when I look at it now, but because it's generic the data, it's not obvious what that means.*

- Asset types:

R7: *Actually we probably won't split them all up individually for this sort of costing because even in one department I might have 50 different types of things. They'll just come under, kind of, instrument data. As I say, we'll start to think later human readable or not or that sort of thing but initially saying, here's some data we'll need someone else to describe what it is…*

- Data volume vs number of files. In some organisations costs are associated with number of files rather than their volume.

R1: *Can I just make a very small comment on the data volume option? It would be useful to have a megabytes option because not all collections that I deal with are gigabyte size. I wasn't entirely sure whether that was the data volume of the SIP or whether it was the VIP or all of them all together...*

R4: *I had a similar or the same sort of set up issues that Gareth's got is that I was unsure because we work on…we had a bit of a discussion of this on the table, we at ADS work on file types and numbers of files. It's not really about the gigabyteage because some of our more complex files might actually be quite small but we could spend more staff time on them and the gigabyte stuff is associated with our archival storage costs. The numbers of files and file complexity is associated with staff cost…*

- Graphs. Some respondents struggled to understand the chart:

  R4 comments

  you have your nice little pie chart that works all your percentages out and then your file types, your data types, we did have a specific question about whether graphics was the same as images, but it gives you a skewed assessment of the amount of work that's been going into an archive if…so most of our archives might be word process files or spreadsheets; in gigabyte terms they're quite small. We might have 100,000 images; in gigabyte terms they're massive but in terms of the percentage of time we spend on them, very small because we'll batch process them and they're easy. So that didn't reflect properly in my case how we would divvy up the work associated with it, all the costs associated with it...

  R7 says

  I've noticed the graph doesn't show the percentage that we put in…

  And later adds

  I haven't tried to download just when I see activities, if I hover it gives me the amount, which is great, but the thing I put in was the percentage…

2.2.3 Suggestions for extra features/improvements

- An option to upload an excel spreadsheet with data sets (or reintroduce a submission template)

  R 7: *Well, if I had a hundred data sets, having to type them all in to get a graph, I can just take it straight from Excel and do that. I couldn't do the comparison, but I could do a lot of the other things without the tool…*

  R1 suggests…

  Does it sound like there needs to be some kind of standard Excel template that people can use to import stuff into and then we can do the comparisons?

- An option to share costs internally before publishing

  R1 inquiries

  Can you share your in development stuff that you are putting into here with other people before it's published?

- Inserting a 'department' field: participants were concerned about appearing to represent a whole institution.

- Being shown an example of what the tool does (also see Scope of the tool)

  R6 says

  I think in general it helps to be able to explore all this stuff to know what you're doing but you kind of need to know that before you actually start it which comes back to this idea of having to sign up and representing a university before you can even get an idea of what the processes actually involve. I think you need to be able to see what you're going to do and what kind of tool this is much more clearly before you have to sign up to anything. I didn't understand it at first, even though it looks lovely. I mean, it's a beautiful looking website, but I didn't understand it…

  Facilitator1 adds the project is in the process of developing a wizard, a video tutorial.

  R6: *Yes, that would be really handy, just so you know what they're doing and can check that it is going to be useful, because at first when I came to this, I thought, oh, good, I've got a service I'm planning. I can do this and then I'll go to the CEO and say, can I have £2 million and I'll give you a curation service. Of course that's not the tool for this and it's not at all obvious that that's the case when you first look at the main website…*

- Be able to run a report to help with a business case

  R4 makes a suggestion

  There is a thing about who is this for and how you would perhaps use it best. I think a lot of it is about making the case to someone or something and that comes back to the import and export functions. It would be very, very nice to be able to run a report because what I want to be able to do is either send a PDF or slap two sides of A4 on a table and say, this is where we need investment. This is where we're overspending. This is why we're doing this. This is why we're doing that. That might be actually quite a nice way of doing things.

- Being able to calculate and compare salaries (FTEs).

  Both R4 and R1 mentioned this and others approved. R1 mentioned a tool called Sirius which could perhaps be plugged into CCEx. The discussion on the salaries issue arose both during Sarah's presentation and later on.

2.2.4 Comments on the interface, workflow

- We think it is fair to say that users in the given time were not able to create a mental model of how the tool worked. For example, for returning users it was still not clear how to add more costs (having to go to 'Compare costs' in order to add data didn't make sense either).

  R7: *I think one of the things if I was using this, one of the things I quite often do is add more data. So from the page at the beginning I have to go to compare costs to do that, which is not what I'm doing. I'm putting in data. I'm going to compare costs in a minute. Then I have to go to manage cost data sets to add another one. There isn't an add button here…*

  Facilitator2 suggests:

  So in terms of coming back to what Catherine was saying and the value of having your own organisational data, would you prefer to have adding your own data as a separate field altogether and then maybe once you've got that then go into compare costs? Would that be a more…?

[Participants agree].

R7 adds: *having said that there's not many buttons and I'm used to computers and things, I would just press all the buttons until I find it so I'm not too worried. It's not for completely naïve users. I'd work it out very quickly...*

R4: *because I had the same problems as X, I wasn't actually sure what I was adding when I was adding a cost unit and why I was adding it or how I added it. So I just basically stabbed around randomly*

[People laugh].

- Sliders vs manual entry: during the hands-on session, R1 stated the sliders were useful but they would have also liked to be able to enter figures manually.

- It was not clear also how you could delete data, where was the delete button:

  R7: *The other thing I'd want to be able to do is delete... It's not under management, which is where I'd expect to be able to delete…*

  R4 says: *Well, I haven't even found that. Oh, there's delete…*

  [Everyone laughs]

- Registration: R6 failed to receive registration confirmation. Fortunately, Facilitator1 was able to speed it up manually. This issue had arisen before during one of the previous usability testing sessions, when one user had to wait at least 9 minutes for registration and was tempted to create another account.

3.0 Themes

- The scope of the tool – unclear
  - What the tool does, how it works, who it is aimed at
- Uploading and comparing costs a major issue
  - Worries about reliability of cost data entered (by multiple staff)
  - Staff time – who has time to enter cost data?
  - Also doubts about meaningfulness of such comparisons (especially global comparisons)
  - Suggestion for more questions on curation levels to make comparisons more significant
- Terminology not clear
- Graphs not clear
- Suggestions for improvements:
  - Being able to upload excel spreadsheet with organisation's data
  - Option to share costs internally before publishing
  - Add 'department' field
  - Being shown an example of how the tool works
  - Being able to run a report to help make a business case
  - Being able to compare and calculate salaries

4.0 Conclusion

The language employed by users demonstrates that it is not immediately obvious how CCEx works, what costs are being added and why and how to employ CCEx to obtain accurate cost figures. R7 says *It's not for completely naïve users. I'd work it out very quickly...* while R4 says *I wasn't actually sure what I was adding when I was adding a cost unit and why I was adding it or how I added it. So I just basically stabbed around randomly*.

Others described it as time consuming and this fact may affect the quality of data users will enter. R3 said: *Also I think you'll probably all admit it's incredibly time-consuming... I'm not sure that, say nobody in the [organisation] could actually sit down and do it and if so, it would be quite time-consuming. So people will input all sorts of numbers.*

R4 and R7 asked to have examples, case studies of how CCEx can help make a difference and R4 even mentioned a prize. They seemed sceptical that people will use the tool; they think serious incentives are needed.

The Focus Group was a great method for assessing CCEx. Visualising people in action highlighted difficulties people encountered both technologically and conceptually.

The interface in itself seemed easy to use. The two major obstacles encountered were the jargon and the fact that people failed to see adding costs was a two-step process i.e. adding cost data sets then cost units.

The main problem however, is at conceptual level. People find it difficult to build a mental picture of the process they have to go through. First of all, users do not think of budgets in terms of data sets and various units associated with it. Secondly, the categories used to breakdown costs do not make sense or, better said, they are presented in the wrong order. It is not the way users normally calculate costs and, after trying out the tool, they did not appear inclined to follow the CCEx way either.

Please note, the group understood the importance of sharing and comparing costs so they are not short on motivation. There is a certain reluctance to share data due to worries about confidentiality and reliability of data, both of which can be easily alleviated with careful communication.