PART 2

Scott Renton, Digital Development
So, Gavin and myself are both very involved in tackling this issue, but from slightly different angles; even after an attempt at choreographing our halves of this particular tango, we're inevitably going to walk over the same boards a few times (hopefully without treading on any toes). It'll drum a few of the points home, I'm sure, though!

I'd like to talk about this from the perspective of digital development, so effectively why, as a techie, I had an interest in tackling this issue, so I'll mention the problem, why we did what we did, the design decisions involved, how it fit into our workflow, and where, with the benefit of two years' additional experience and interest in the community, we can see this going.

Apologies, I tend to explain this through the medium of skeleton.
So, from my perspective, I'm very interested in workflows, and despite 7 years of hacking away at this, there is still something of an atavistic perspective to how we get image data in the department. A lot of digitisation is based on readers' orders being passed to the photographers, and this actually still involves paper forms. Therefore, you're only going to get a small amount of identification metadata for a digitised page, because it needs to be copied out. Most of this stuff isn't really catalogued, and those that hold the purse strings will be far likelier to pay for a new camera than a cataloguer.
This generally means there's a lot more that could be captured for an item than we've got our skeleton records, but we need meat on the bones. This particularly macabre scene is from one of our most popular collections—the Incunabula—our earliest printed books, effectively, and very popular (mainly because there was no political correctness around the gruesome images therein). That's generally what, if we're lucky, will go straight onto a record, along with the EXIF data that comes directly from the camera but where's the rich, contextual information, and where are the tags that tell us "skeleton", "horse", etc.
The systems that we had prior to this didn't have really any capability for enrichment, beyond hiring interns to come in and properly describe our images through the LUNA backend. LUNA does have an annotation module - annotations are stored in an database, so we can look at and garner this information but nobody knows about it, as it's pretty well hidden, so we've never been able to get very much. We needed something extra.
So, we decided to build the games. Gavin's shown you what they look like - very deliberately lo-fi, with the space invaders feel, almost to emphasise their simplicity. They're really tenuously described as games - there's nothing particularly clever going on, and they've only worked because they're surprisingly addictive! We could do so much more if we had the time.
How we built it

• Emphasis on lo-fi
• LAMP (Linux, Apache, MySQL, PHP) - not cool!
• Front-end design decisions - retro feel
• EASE authentication
• Part of the Library Labs site

We built them using straightforward lamp: it’s really just form data writing directly to a database, although that can be surprisingly powerful. There is authentication, using the University’s single sign-on module, and we put them directly onto our Library Labs site, which is where all of our investigative and innovative work goes.
And again, as Gavin said, we have had to take a number of decisions as to how to control the flow of what we get, and to try to incentivise the users. Random images seemed to work best, based on collections that were safe to surface and needed tags most. We went initially for a moderation module to try and exert a human eye over some of the data before anyone else got to see it - we are using stopwords downstream for this kind of thing - but the problem has been a massive backlog, because the crowd aren't involved at this point and approved moderators don't have time. We probably need to bite the bullet and ditch this. The voting module is a crucial part of the game, an easy way to score a lot of points and the thing that determines whether data gets into the core system. The scoreboard is hugely effective in incentivising users - I've watched people get very worked up trying to win during a session. We've introduced a "star" branding for high scoring players, and we've also tried to make games more complex, introducing clocks and set numbers of images (see the Art game), as well as easter eggs (the hidden Dolly the Sheep).
Authenticity

- Moderation
- Voting module
- Thresholds
- Demarcation
- Could we do this better?

Authenticity is an important point- we have a threshold of net 2 votes allowing something to get into the core system, because we don't have enough activity to make it higher. That said, the fact that we clearly demarcate the tags to keep the public/scholarly separate probably means we have the leeway to do very little. It is difficult though, and is a regularly bone of contention. I have spoken to people whose metadata librarians would be horrified at the very thought of what we're doing.
The proliferation of the games has been an interesting point. We started off with one, endless game, and let that fly, and then we added the Art game, which had a bit more structure and sense of competition to it (we fudged it so that the items physically on show always came up first for that particular event, so they’re very well-described). The Where's Dolly was made to look like a Wellcome/Roslin glass plate slide and, as I say, would cause the game to haemorrhage excitement if Dolly The Sheep came up. It was pretty straightforward to skin a new game and give it distinct data, so we could easily do more of the same (but at roughly the same level of complexity).
Workflow ‘in’
Then, just a little bit about how we get the data back into core systems- mysql data exported into csv, and through the LUNA API, it gets onto the main record.
For collections sites (not all our collections see LUNA as their golden copy), it goes into the DSpace database via a Java curation task, and is pretty heavily grilled for stopwords at this point.
So, here's how we did. It's not bad off approximately 20 hours of play - we get, as Gavin's said, no action when we don't have an event on, and it is something we need to address.
So, that's what we've done, but there's a lot that can still happen. We had some informatics students looking at metadata games using this as basis last year, and it gave us some ideas. We have an intern signed up for the summer to give them an overhaul - that's to make them responsive initially, but possibly to do the things they currently do more elegantly, and certainly engagingly. Gamification (if that's a word) should be addressed: I initially got involved in this from the selfish perspective of trying to meet our users' needs of making the collections searchable, and not to any great end making it an altruistic "bring the community together". It was all about free labour doing the things we couldn't do. We need to find ways to bring the users back, beyond competitiveness and incentives, and it is one of the challenges of this internship. Can we do things with computer-generated metadata, recognition of shapes etc, or OCRing? Should we be using OAuth as a means of getting in, and hooking the sites up to Twitter and Facebook to allow greater security, engagement and ease of access?
And then, there’s this new thing that we’ve been getting into in the Library Community called IIIF. This is a framework for interoperability across images, with the intention of getting away from silos. The idea is the image is only ever hosted once, and can then be reused by bespoke viewers and in customised presentations by “photoshopping" on the fly using a URL. That's one side of what it can do, so it allows us to beautify our websites by bringing in Deep Zoom functionality embedded into the page, means we can get rid of jpegs, and it means we can often hide collections that are revealed in two places, so we don't need to keep metadata in sync! The other side is that it presents a standard for interoperable PRESENTATION, where you may have multiple images relating to one item, to determine the sequence, direction etc of how these should be shown, and crucially ANNOTATION can come in at this point.
Just to give you a couple of examples of this: the existing game has a link to the LUNA system for high-res zooming, that opens a separate window- this will get away from that. Using the OpenSeadragon viewer, we will be able to zoom directly on the page, which is very useful for checking what is there to be tagged.
And then, for crowdsourced material, IIIF is a much more standardised way of bringing in annotations for translations and transcriptions. I think it is something that we will set our intern off to look at, getting away from the mysql inserts, html form data approach. Last year, we got funding for an intern to look at a transcription and translation tool for researchers - as opposed to "the layman", and she came up with something way beyond our expectations. I'm not sure these screenshots quite do it justice, but she used IIIF (image API) to bring in content, and IIIF (presentation API) to manage and administer translations. She uses Leaflet IIIF and Draw to allow vectors to be drawn round the area we're interested in, and stores the annotations that people submit, and, yes annotate again, as well as votes and discussion on the annotations. She's also introduced an array of keyboards so you can transcribe in more or less any language, although she struggles a bit with languages which are no longer with us in that form (and of course a lot of this manuscript data is in such a form!). It's not live yet, so I can only show screenshots, but I will let you all know when it is available!
Questions and more info

• Email
  • gavin.willshaw@ed.ac.uk
  • scott.renton@ed.ac.uk
  • lddt@mlist.is.ed.ac.uk

• Sites
  • http://libraryblogs.is.ed.ac.uk/librarylabs/ (blog)
  • http://images.is.ed.ac.uk (image repository)
  • http://collections.ed.ac.uk (collections repository)
  • http://iiif.io (IIIF website)