CIDOC 2015 - New Delhi, India
September 8, 2015

CIDOC Session Report

How Can We Achieve GLAM? Understanding and Overcoming the Challenges to Integrating Metadata Across Museums, Archives, and Libraries.

On Tuesday afternoon, 8 September 2015, a panel of eight colleagues representing museums, libraries, and archives gathered at the CIDOC conference to discuss ways to overcome the challenges to integrating metadata across our three sectors. For years, researchers, teachers, and the general public have called for integrated access to collections and collection information. The most tangible results thus far have been large-scale aggregations like Europeana, as well as institution-specific projects to implement federating searching across various in-house collections. The panel explored the environmental challenges to integration and suggested future directions.

Moderators:
David Farneth, Assistant Director, Getty Research Institute, USA
Lorraine A. Stuart, Chief of Archives, Museum of Fine Arts, Houston, USA

Invited Panelists:
Gabriel Moore Forell Bevilacqua, Professor, Archival Science, Universidade Federal Fluminense, Brazil
Emmanuelle Delmas-Glass, Collections Data Manager, Yale Center for British Art, USA
Monika Hagedorn-Saupe, Deputy Director, Institut für Museumsforschung SMB - PK, Germany
Jennifer Riley, Associate Dean, Digital Initiatives, McGill University, Canada
Regine Stein, Head of Information Technology of the German Documentation Center for Art HistoryBildarchiv Foto Marburg, Germany
Reem Weda, Information Specialist, RKD - Netherlands Institute for Art History, The Netherlands

Introduction

Co-moderator Lorraine Stuart and each panelist discussed the challenges to metadata integration. In the interest of time, co-moderator David Farneth forsook his turn, but his planned comments appear in this paper. The panel then presented solutions to overcome the identified challenges.

In the limited time allotted, panelists touched upon topics that ranged from economic to technological to – most challengingly – the practice and culture of the different professional sectors. It was felt that the effects of limited resources could be mitigated to some degree by large institutions and governments continuing to provide leadership and tools to smaller cultural organizations. A novel idea was to tailor resources toward

documented user interests. Technology was presented as a means to augment limited resources with tools that aided the descriptive process, including a seismic shift toward controlled vocabularies as a function of systems, rather than the process of cataloguing. While presenting its own challenges with large-scale search results, Linked Open Data (LOD) was perceived as a means to effect this seismic shift as well as to deliver more context-based information. Among the challenges presented by professional/sector culture, the need to rethink descriptive practices in order to harmonize conceptual models was addressed. The relaxing of sector-specific descriptive standards was presented as essential in the repurposing of metadata for semantic-based technologies. While cross-training was hailed as a means to overcome unawareness among practitioners, the larger cultural challenge of indifference among leadership was largely unaddressed. In conclusion, it was felt that only when institutions and professional organizations recognized these efforts as central to their missions would they succeed on a wide-scale level.

Challenges

LAS
Last year I served as chair of the Museum Archives Section (MAS) of the Society of American Archivists. Serving in that capacity – as well as working for 20 years in a museum setting - has a lot to do with why I find this discussion of library, archival and museum practice compelling. Early in my chairmanship it became obvious to me that there is a deep divide – and it extends beyond the digital divide that we will discuss later and indeed contributes significantly to it. Moreover it is a divide not only between large, well-funded programs and small ones, but also often between the disciplines of libraries and museums. My observations were confirmed by a survey of the MAS conducted earlier this summer. Although small and non-scientific, the results offered insight into the state of the museum archives profession in the U.S.

The results indicated that nearly 60% had never participated in any shared metadata project. Of those that had, very few use METS; a large portion rely on spreadsheets or direct data entry. The most common reason given for the latter is that the material had not been previously catalogued at the level required by the project. Which brings me to the point – in fact 2 points – that I think are relevant to today's discussion. The first is that half of the respondents indicated that more than a third of their collections lacked sufficient description. I believe this reflects the reality that 70% of the museum archives lack even a single position dedicated exclusively to arrangement and description. Bear in mind that the majority of collections that lack a single processing archivist contain literally millions of documents. As we entertain the ideal of shared descriptive metadata, we need to recognize the simple fact that archivists cannot share what has not been created. This is a fundamental impediment to the success of the GLAM projects.

As to the second point, the survey found that while half the collections are under described, 40-50 percent of archivists are attempting single item cataloguing for visual images and another quarter are attempting to create single item cataloguing for born-digital documents, which is a massive undertaking.

The lack of resources is not the sole reason that I have some reservations about digital curation and the single item treatment that often accompanies it. Archivists place a huge value on context and the concept of original order. Any treatment of materials that isolates a letter, for example, from the ongoing correspondence of which it is associated, loses a critical part of its evidentiary value.

I am grateful to CIDOC for inviting archivists here to get our perspective. Before concluding, I wanted to mention that the International Council of Archivists will be releasing a draft of a Conceptual Model for archival description this winter. It will map to the CIDOC CRM. I hope that it will address some of the differences in our descriptive practices.

GB

For me, the basic challenge is one of description, in that archives do not describe at the same level of description as museums and libraries. This is the basic obstacle. Another challenge relates to the image representation of the collections. It is easier for libraries and museums, where you usually have one-to-one representation of metadata to object. It is very difficult to do this with only archival series and collection-level description, and it often leads to difficult challenges for the researcher who is looking for specific images.

In Brazil, almost all of the initiatives trying to tackle this problem are approaching it from the library perspective, which has very well established standards for description. These do not adapt well for museums and archives, which are both still working on developing descriptive standards. In the archival community we have ISAD(g), which may need to be rethought conceptually. For example, when you look at what a title means in ISAD(g), it can mean almost anything: series, subject, type of document, etc. This needs to be more precise.

There is also the issue of scale. There are very large institutions that have the technological resources for addressing this problem, but also very small ones that may not even have one IT professional on staff.

So, in summary, how can we achieve this without structured metadata and a general lack of technological expertise and support?

RW

In my daily work I am confronted with several issues related to the integration of metadata, aside from the inevitable editorial questions relating to thesaurus concepts.

These are mostly related to promoting the use of standard vocabularies in the heritage sector. For example, one must decide which of the standard authority vocabularies are best suited for your specific situations and topics. I also find myself constantly

reminding people of the importance of using unique identifiers or persistent web addresses. Then one must convey the guidelines for construction to users, because most people will divert from these guidelines if allowed, and thus undermine the efficiency of the thesaurus. Then there are the challenges associated with getting updated information to the users (such as changes in descriptors and preferred terms) as well as encouraging the heritage sector to contribute knowledge and new concepts to shared controlled vocabularies that function as standards.

EDG

I would echo much of what my colleagues have already said. It starts with the training of the staff in our cultural institutions. This is something that CIDOC has been working on for a long time and will keep working on.

It is clear that when we are talking about data integration we are not talking only about IT problems – systems problems – we are also talking about issues related to the data. In my experience working at various institutions, it is hard to find the person in each institution who will be the data champion; someone who knows the data really well, has a good command of data standards, and is able to explain the narratives and research developed by the curators. I am thinking about someone who will also be in charge of the long-term management of the data, and who is concentrating on its meaning, context, function, integrity, use, and reuse no matter what system the data is in. This is a first challenge that we as a community should try to overcome.

It is difficult to do collaborative projects across museums, libraries, and archives, and of course Gabriel has mentioned the problem of the differing metadata standards and levels of description. The way these collaborations have been handled in the past is by making metadata crosswalks from one schema to another. And what usually happens? In order for all the partners to share their metadata they end up "dumbing down" their metadata to some degree, which is probably exactly the opposite of what we came to do. What we really want are rich datasets that better enable researchers to do their work. So the traditional model of implementing crosswalks has not really worked very well. It is clear that we need better tools to support good semantic integration and representation of our data.

On the topic of researchers: it seems that there is general consensus that only a small fraction of the public using our collections – i.e. the serious researcher -- is looking for rich and expert data. But I would like to challenge this assumption. I don't think that only researchers are looking for rich metadata. I think the mainline, educated public is also looking for complex information, and we should be considering them as well. We talked about context earlier. This is also a challenge to the integration work that we are trying to do, because we usually focus on the objects and forget about the context but of course it is the context that enables cultural heritage objects to come alive in a wonderfully complex way.

MHS

One point specific to the museum area is that until just a few years ago, museums have been using mainly stand-alone systems, because their mission of research and exhibition has focused on their own collections. When they were collaborating with other museums it was mostly for temporary exhibitions or special research projects. So within the museum community we have standards, but they are often in-house standards. This means that while we are thinking about how we can collaborate with libraries and archives, we also have to be thinking about how we can connect to other museum collections as well. There we have various issues, some of them shared with libraries and archives. We have a huge amount of content and large backlogs, especially in ethnographical and classical cultural heritage museums, which may have thousands of objects not fully catalogued. In addition, we are just starting to get some uniformity among all of the in-house standards and systems that have been in place for years. This takes time.

Another challenge is that museums use different terminology, or the terminology they use may have different meanings. Museums want to provide access, but if you discuss this with museum directors it means more exhibitions, make your collection visible, maybe even have an online collection, etc. In libraries these days, "access" generally refers to online access, which is not the case for museums. This is one more area in which we differ.

JR

As far as these challenges go, I may have a slightly different take than some of the other panelists. Coming from the library world, I see the value in the amount of predictability, control, and authority we have given to our metadata, however, I think this reliance and trust, and the value we have placed upon control over our metadata, will be our greatest challenge to integration because we have three different approaches, and we are all convinced that ours is the best. Each sector has built their standards over many years, and the professionals in each sector feel equally protective of their standards. LOD is about bringing together data that comes from different traditions. If we enter into collaboration with the idea that everything has to be mapped into our model, we are going to fail. The value that we place on control will be the thing that causes us not to move forward. Our challenge is to get out of that mindset. We need to take the value of what we have done and make it OK to use it in a new environment and in a different way that may not have the same level of authority control. We need to start to welcome metadata from other sources that may not be as trusted or may not conform to our traditional models.

As Gabriel said, there are technical challenges, but I think these can be overcome with the right amount of resources. We need targeted work in the technical realm to make progress. The technical problems are very real and we need to fund them. We need to find a way to get the right people working on these problems and that, together with the cultural shift I am proposing to accept metadata done in different ways, will enable us to move forward.

RS

I would like to focus on two aspects. In the area of different descriptive standards, we have already worked a lot on different crosswalks and achieved some results. I would slightly disagree with Emmanuelle in that I think that some of the crosswalks allow for in-depth integration and allow for broad access across different sectors – although it is true that in practice we can see quite poor implementations only. However, this is in my opinion rather due to 'laziness', and lack of ambitions to really cope with the complexity behind, than to the method of working with crosswalks. I also think that controlled vocabulary and authority control are equally important for any kind of federated search across collections and institutions. This problem also relates to access. We not only want to link things together just for the sake of linking, we have to be able to find them and browse them. If large-scale aggregation of metadata just means bringing it all together in one place, which has been the technically preferred method so far, it is of very limited value. But providing access across different collections means something different and this is the real challenge. This can be seen in the large-scale aggregations like Europeana. Of what use is a search result with more than 1.000 hits, and you cannot further refine it? So a big challenge is to harmonize collection-specific indexing rules with the requirements arising from a large-scale aggregation.

Secondly, we are typically talking about projects running at rather short term, most often two or three years. They are usually put on top and not really integrated with the primary tasks and workflows of an institution. They often use temporary staff who leave after the project, and they often implement project-focused workflows alongside the daily work. And, in the end, it turns out that it is really difficult to integrate the results into the existing infrastructure and to perpetuate the workflow. This kind of sustainability obviously requires resources from permanent staff, and these resources are rarely included in the project.

* * *

Solutions

For the second part of the session, each panelist had the opportunity to respond to what others had to say as well as propose solutions for moving forward more effectively.

LAS

I won't take too much time for my comments because I am interested in what the panelists have to say. About backlogs, I think that our generation of archivists is like Janus, as David said the other day; we are looking both forward and backwards. We are dealing with huge backlogs of analog materials while at the same time trying to get a handle on our born-digital materials. Once we do the latter, I think digital tools may start augmenting our current arrangement and descriptive practices. For instance, I know that

the Frick Museum [in New York City] is using image recognition software for their photo cataloguing. These advances will be helpful.

I am interested in hearing from the panel, do you perceive controlled vocabulary becoming more or less important as things develop?

GB

I think that controlled vocabularies are one of the solutions to this problem and we should work more on them. As for some other possible solutions, one of the most important is the way that we train professionals. I am teaching future archivists in Brazil, and I know that these kinds of issues are not talked about as a key problem. This would be a very important change. I think the ICOM CRM could provide a lot of help in dealing with this gap. Of course using integrated vocabularies such as the Art and Architecture Thesaurus, the Union List of Artist Names, and several other initiatives, as Reem has told us: there are a lot of different standards and authorities, so we have plenty of options.

As for the problem of scale, there are some simple and very basic solutions for smaller institutions. For smaller museums, 80% of the user inquiries are about information about exhibitions, artists, or a specific work of art. So these would be the three most important data elements to start with. This might be part of a solution for smaller institutions.

RW

I completely concur with Gabriel that training of museum professionals in the use of standard vocabularies is a very important aspect for moving forward. It is very different between the smaller institutions and the larger – which often have the "not invented here" syndrome.

Network level aggregation projects like Europeana can help, because they can work on innovative solutions that are normally out of reach for most heritage institutions. This is especially true if the network aggregator can put forward clear requirements and, if necessary, help with preparing the data. The problem of levels of description remains. The aggregator can act as a data manager and can try to redirect to standards and disambiguation. Large-scale projects like Europeana do need to put a lot of effort into managing expectations and creating trust in the project. Like Jenn said, we do sometimes have to allow that our metadata is used in a slightly different way than it was created for.

I think improvements from the ground up are possible when institutions make a strategic plan that defines what they want to achieve with their digital documentation and how they plan to achieve it. Smaller projects can be started in-house to explore the situation and can be directed towards the goals that are set in the strategic plan. I can imagine this will lead to quite a bit of data cleaning.

Training museum professionals in the use of standard vocabularies is good practice. But the knowledge should fit within a larger development within museum policies and in information management systems. The use of standards should be specified in policy and the usage should be made easier by the organizations that maintain standards/ authority files and by the software vendors that specialize in museum systems.

Software vendors can make annotation tools, help create and use web services, and improve search ability within the incorporated standard vocabularies. Everyone is talking about mapping metadata, but no one is doing it. Mapping is also a good way to improve the interoperability of metadata. But it would be great if a knowledge system would be able to address this issue.

In the Netherlands the ministry of education and culture, together with some large network organizations, like the Royal Library and the National Archive have started a program called National Strategy for Digital Heritage to improve the situation.

EDG

I agree with most of Reem's comments. I think that LOD holds fantastic promise for the cultural heritage sector, and we are in good hands especially with the CIDOC CRM ontology. There is good documentation on it and there are institutions that have already implemented it. It's a really good semantic framework that keeps the granularity and specificity of each dataset even though it integrates them with other datasets. The CIDOC CRM addresses this problem of dumbing down metadata and rather encourages each institution to keep its own level of specificity and vocabularies.

As for controlled vocabularies, they will be very useful in the future. They are not going away. Quite the contrary: they will serve as the common ground between GLAM, regardless of differences in the metadata schemas and cataloguing practices.

I think we can start with some basic steps in terms of data management. We know that in this digital world, we are not mapping spelling and terms anymore, but rather unique identifiers for concepts and entities, which is what Reem just talked about. And so adding this workflow to the cataloguing process would be a very important first step that could be leveraged later on.

MHS

In these days of the internet, more and more material data are available online worldwide. This means that research becomes easier, but, as Regine pointed out, you might retrieve millions of records; how do you find what is relevant for you? For that I think we need vocabularies, but we also need multi-lingual vocabularies or linked vocabularies so that you can really understand if your retrievals are relevant to your research question. So this topic of how to be better able to find the right information you are seeking, in other

words, context-based retrieval, is something that we have a chance of providing using a combination of linked data and structured multi-lingual vocabularies. This is something we should work on in the future.

JR

I will talk about LOD as well, but I want to say a few words about a different topic at the end. I think our solution here is to see the power of LOD technologies for what they actually do: bring together the different vocabularies of the different communities. One data hub can make a couple connections with some "same as" relationships to another data hub, and that then links those two data sets completely and totally together, and any system can track this activity fairly seamlessly because they know where the connections between the concepts are. We can look at some big LOD implementations such as OCLC's WorldCat, which is a big bibliographic database for libraries. When they made their database into LOD it contained 15 billion triples. That's a lot of data. We cannot continue to use our current processes of having someone copy catalogue or view every record every time it comes to us. That amount of data is just far too big. It's the question of scale that Gabriel was talking about. If we look at Europeana, which Regine mentioned, it has four billion triples, aggregating metadata from all over the European Union. We have an opportunity to change our practices and stop exerting so much control over our metadata and to accept data that is only "kind-of" fitting together.

We have heard a lot of people on the panel so far talk about the utility of controlled vocabularies and I agree, except that I think the utility of controlled vocabularies is not to create metadata but to create the Rosetta Stone that brings the different data sets together, which might be structured or unstructured. The controlled vocabulary becomes part of the system rather than part of the cataloguing process. And I think that is a big shift that we need to make in order to move forward.

In the library community there's a cataloguing theorist from the 1960s named Jesse Shera, who has what he calls the two laws of cataloguing. The first is that no cataloguer will accept any other cataloger's record, because it's not good enough. And the second is that no cataloguer will accept his or her own record six months after it was created. It's so funny because it is true. The value that we are adding by this obsessive level of control is so low. We have to get over it. The LOD movement is showing the need for systems to show conflicting information, to show the provenance of information so that people can decide for themselves if they want to believe it, and it is driving systems that bring in human intellect only at critical points, such as making the decision to bring together one data set with another data set by connecting two controlled vocabularies and thus making the connections that everyone is talking about.

I will just briefly mention a couple other ideas that I have. I think from the library point-of-view we should look at the archival community's practice of multi-level description. One should not assume that everything needs to be described at the item level. We need to look to the museum community to understand the value of interpretation of objects. Library metadata pretends to be neutral and museum metadata does not make that

assumption. People want interpretation, and it adds value to the system. I think we can also look at things that both public and academic libraries are excited about such as 3-D scanning and printing. What a huge opportunity to work more closely with museums. There is an opportunity with the app culture of museums, to have users participate in metadata creation. Then the metadata from those apps can be brought into a LOD file. If we step back and see what our commonalities are, there many opportunities to be explored. There is a lot of room for leadership at every level, from the directors down to the people actually doing the work.

RS

LOD is the way to go, but we have to realize it is just a framework. We can do almost anything within the framework, so it is really a matter of quality.

At this point, the most important activity across the three sectors is harmonizing their conceptual models. You may know there is official cooperation between the IFLA FRBR (Functional Requirements for Bibliographic Records) Review Group and the ICOM-CIDOC CRM-SIG, the Working Group on the Conceptual Reference Model. Interestingly, the actual work on developing an object-oriented formulation of FRBR, the FRBRoo, that is a compatible extension of the CRM has been ongoing since 2002/2003. However, it has not been until August 2014 that the FRBRoo model in its version 2.0 was officially endorsed by the IFLA FRBR Review Group as a valid ontology for semantic relations embedded in descriptions provided by libraries. This is an indicator for how long these things take so we need to take a large breath and be persistent.

The harmonization of the conceptual models of the library and the museum world, and similar activities have been started with the archive sector as well (as mentioned by Lorraine), is eventually giving us the common ground to actually understand the conceptual issues that we encounter in information integration across the sectors. So this is the “high-level” conceptual part.

Given this foundation, in my opinion we should further specific projects that lead the way for a larger scale change, so I would focus on “ground-up” effort. On the network level, I think we need to put much more focus on quality rather than quantity, and on community building which allows people to perpetuate this work. This is a long-term process and we need to promote this fact to our directors.

Another target could be the educational programs of each sector – maybe some kind of cross-sector module which may integrate into the CIDOC training program. We need “cross-border commuters” – people who really understand the specifics of each sector on a practical ground and can facilitate cooperation.

DF

My thinking aligns very closely with Jenn's comments. I would reiterate her point that we can no longer think of controlled vocabularies as cataloguing tools, and their power can only become fully realized in a networked environment of computer talking to computer. Controlled vocabularies are very expensive to build and maintain, and we must find ways of extending this work to the community while maintaining the quality and integrity of the data.

Even though this panel has understandably focused on solutions to the metadata problem, we might also formulate some top-down solutions by trying to understand better the upper-level political and cultural challenges to integrating access to GLAM collections. It seems to me that the leaders of even our well-resourced institutions have yet to endorse this work. The reasons are probably complex, but they might include a combination several factors, such as: 1) we have not effectively articulated the value of doing so; 2) supporting research is not a central institutional mission; or 3) issues related to competition, branding, ownership, and the like, overshadow the benefits. We should also provide stronger advocacy to our professional organizations, which, to my knowledge, have not yet bonded together to articulate and prioritize this work as an important goal. This lack of commitment on the part of professional organizations may be part of the reason why current professionals and training institutions have been slow to embrace it.

As for the value of "top-down" versus "bottom-up" approaches, it would be good if both approaches could happen in tandem. A few convincing, cross-sector "bottom-up" projects as a proof-of-concept would help us to demonstrate the value of this work, and the knowledge gained would help us to convince the leaders of our institutions and build enthusiasm within our professional organizations.