EarthCube M4M Pilot Workshop Report

Report Issued: October 28, 2022
Workshop Dates: September 8 + 9, 2022

Nancy Hoebelheinrich, John Graybeal, Melissa Cragin

Introduction:

The EarthCube Office (ECO) at the San Diego Supercomputer Center developed a FAIR Initiative to increase engagement and awareness of the FAIR Principles. Information and training materials were produced to build awareness and support implementation across several aspects of FAIR. These are available on the ECO website as well as in the EarthCube Organization Materials collection held at the UCSD Library.[1]

As a component of the ECO FAIR initiative, we wrote a Supplement proposal to pilot a Metadata for Machines (M4M) workshop based on the workshops developed by GO FAIR (now the GO FAIR Foundation) in the Netherlands. Their workshops and recent derivatives are now running in the Netherlands and other European locations, with over 24 workshops presented to date. These workshops are in increasing demand as participants - including domain researchers, research funders, and data stewards - see positive outcomes.

M4M FAIRification activities focus on machine actionability for metadata, which will support the application of computational approaches to access and use. While the uptake of the FAIR principles is intended to improve data consumption by both humans and machines, the emphasis on machines is a departure from traditional systems that situate humans as the primary consumers of data and metadata. Instead, FAIR approaches seek to create data/metadata that is AI-ready, improve data accessibility for people, and broaden public access to research products.

Initiatives like M4M also provide a process for a community[2] to mobilize around practical solutions for making needed metadata FAIR. M4M workshops bring together a “community” of constituents at either an organizational, subdisciplinary, or project level. Our GO FAIR collaborators note that M4M workshops “are usually intended to kick-start FAIRification efforts” via selection and agreement on “minimally viable metadata components that are modular,

[1] EarthCube Organization Materials Collection
[2] We note that the GO FAIR Foundation M4M team uses the term “community” to mean a specific group of researchers who come together to create a “FAIR vocabulary” and related automated tooling for application in a particular research context.

reusable"[3] and extensible. Design and delivery approaches are still "hardening," as sessions are organized to meet the needs and aims of different constituent groups; planning and iterations of delivered sessions have resulted in different formats and formulations of content delivery through several workshop models. By centering the workshop on concrete recommendations and openly available tools and resources, these workshops offer a rapid and straightforward process for attendees, who can begin making their metadata more FAIR immediately upon completion of the workshop. Piloting the Metadata for Machines workshop approach through the EarthCubeOffice was an opportunity to build on the ECO FAIR initiative and offer another kind of activity to promote the uptake of FAIR actions in the Earth and geosciences.

Pre-Workshop Planning

Participant Development

The M4M organizing team[4] gave careful consideration to EarthCube-related communities that might have been interested in this kind of workshop, given the primary goal for the EarthCube Metadata for Machines workshop to improve the computational approaches to data access and use. We sought Earth Science domain and subdomain communities that were in the best position to take advantage of the M4M workshop, particularly those with the ability to improve vocabulary specification and provide new automation to generate FAIR metadata. Criteria for community participation included those communities that 1) were amenable to working collaboratively to create, maintain and share their data/metadata across domains; 2) understood the importance of standard, AI-ready metadata schemes and community devised and maintained vocabularies for data sharing; 3) were in the process of coming to community agreement on the metadata schemes and vocabularies that they planned to use; and 4) had a mix of experts who could attend the workshop, participate in the activities and contribute to the products planned as an outcome of the workshop.

We gave further consideration to a community's state of preparing their data for sharing, in order to determine the appropriate approach for the format of the workshop and the focus of workshop materials. For example, if the community was fairly early in the process, we could place more emphasis upon describing and illustrating the importance of metadata and vocabularies, especially for making their data more findable and reusable, and then helping the community construct new or adapt existing vocabularies. If the community was at a later stage in the process, workshop materials need only confirm the importance of metadata and vocabularies, and place more emphasis upon showing how metadata and vocabularies can be refined and integrated to increase the interoperability and reusability of their data. In the latter case, we thought that we could also demonstrate the utility and feasibility of automating the workflow associated with metadata creation and vocabulary maintenance. For communities with funded

[3] https://www.go-fair.org/how-to-go-fair/metadata-for-machines/
[4] Melissa Cragin (SDSC & ECO); Nancy Hoebelheinrich (KnowledgeMotifs LLC); and John Graybeal (Stanford University Center for Biomedical Informatics Research)

projects or facilities with more mature metadata management and sharing services, there was less interest in engaging with new processes or tools.

In the process of identifying and inviting community participation we approached the following groups:

- EarthCube Leadership Council
- EarthCube Council for Data Facilities
- ESIP Soils Ontology Cluster
- Tephra community
- BCO-DMO
- An Astromaterials and Geochemistry project

In the end, members of the Astromaterials community working on a specific project attended the workshop along with the participation of Geochemistry and related repository representatives as observers. We discuss the responses of the other communities that we approached in the Lessons Learned section below.

The Geochemistry and Astromaterials communities proved to be in a good position to take advantage of the M4M workshop as they had already invested considerable time in developing a number of metadata and vocabulary standards, but had not finalized the content and representation of their own products. The Geochemists have a metadata standard for cross-domain physical samples - the International Generic Sample Identifier (IGSN) - which will also be used by the Astromaterials for the physical samples of artifacts from NASA's MARS explorers. The Astromaterials community, specifically the domain scientists and research support specialists from the Origins Spectral Interpretation Resource Identification Security-Regolith Explorer (OSIRIS-REx) Project who attended or observed the workshop, are in the process of developing data standards for various data product types that include both basic and auxiliary image data, documents, structured text, 2D and 3D arrays; sample analysis bundles; and software interface specifications. This community is also working toward the archiving and transfer of sample analysis within other geochemistry archives. For the workshop, the demonstration artifacts we used included properties and vocabularies for several specific product types, including polished sections and non-polished sections of data samples, that came from the draft metadata schemes and vocabularies already under discussion by the community.

Workshop Materials Development

Once the OSIRIS-REx Astromaterials community had committed to attending the workshop, we engaged key representatives in discussions about how the workshop could best benefit the community and produce outcomes that would help them move forward in their efforts to document and share their data. We created a survey which both participants and observers were asked to complete to provide more information about participant backgrounds, areas of expertise, and availability for the workshop. We set up several meetings with OSIRIS-REx and related project program managers to discuss the status of the draft metadata standards, and

any contextual issues that could possibly be addressed within the workshop. This background information helped us tailor the workshop materials. For example, we learned that workshop participants had some knowledge of the FAIR principles and how to use them during the research and data lifecycles, especially of the F (findable) and A (accessible) principles, but less knowledge of the I (interoperability) and R (reusability) principles and how to apply them. We learned more about the interrelationships among the draft standards, and which ones would be most productive to use in explaining FAIR concepts, demonstrating tools, and providing opportunities for group discussion. With this background research done, we were able to create (and adapt) productive workshop slides and homework assignments, so as to focus our attention in the workshop on the areas of greatest value to the attendees.

In addition, the organizers/presenters of many other M4M workshops from the international GO FAIR organization were very generous in providing content and information about the delivery of previous M4M workshops. While those workshops were targeted to different communities than the ones we were working with, the background materials (invitations, pre-workshop participation guidelines, homework assignments) and presentation slides and recordings proved very helpful for us. We acknowledged their contributions throughout the workshop.

Workshop Logistics

Although previous M4M workshops ranged from one week to 4 hours in duration, we decided that our best options, given the needs of the OSIRIS-REx community, were to schedule two four-hour days in sequence. This allowed us to provide background information on the GO FAIR Framework, demonstrate use of the CEDAR and BioPortal tools in generating FAIR and AI-ready metadata, and allow time for workshop participants to make progress on a workshop product—the refinement of one of the key vocabularies that was part of the draft metadata schema. See the agendas for Days One and Two of the workshop attached as Appendix A. Also, see the presentation slides for Days One and Two.[5]

We followed the format of previous M4M workshops by making a distinction between active participants in the workshop and passive observers. Using that distinction, we had 6 participants and 5 observers in the workshop although not every person was able to attend both days. We recorded the sessions and made them available to both observers and participants for a short period of time after the workshop.

Workshop Outcomes

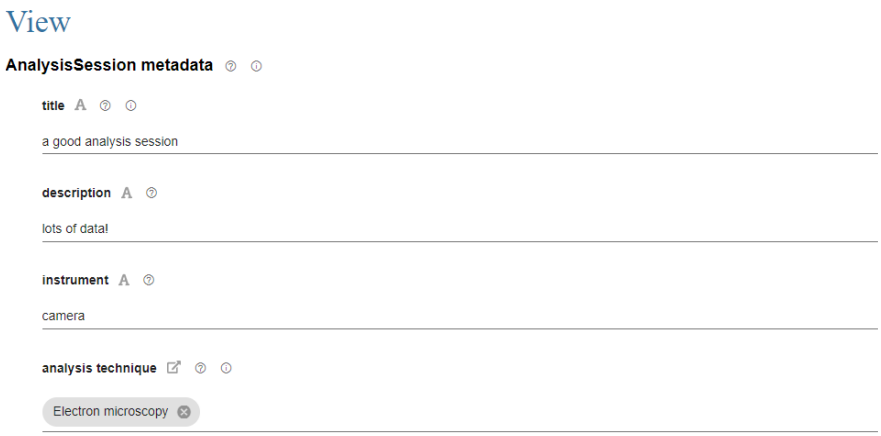
Outcomes for the workshop included products created by the workshop team, and the modifications and products created by the attendees.

[5] EarthCube Materials Collection

The workshop team developed an alpha version of a collection of EarthCube M4M vocabularies for sample metadata definitions in BioPortal. (See: https://bioportal.bioontology.org/ontologies/ECM4M-MD-VOCABS). The vocabularies were created for the digital images of various types of physical samples including Non-Polished Section Container Type, Non-Polished Section Sample Type, Polished Section Container Type and Polished Section Sample Type. They then used these vocabularies to create example templates, elements, and fields for the metadata specifications used in the workshop, and to fill out one of those templates as an example during the workshop.

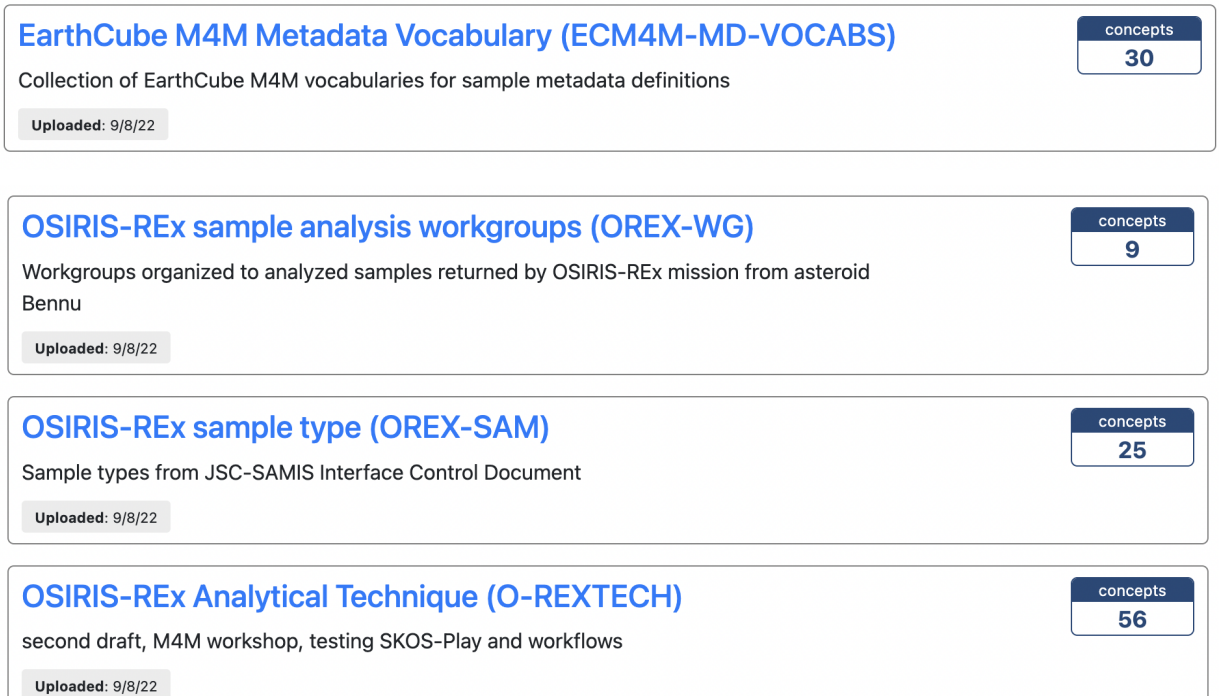
During the workshop, attendees created a draft metadata template for "analysis sessions" conducted by domain scientists in which the digital images of the physical samples are analyzed that were taken during the OSIRIS REx mission and brought back from near-Earth asteroid 101955 Bennu. The metadata template documents key contextual properties associated with the sessions. Workshop participants thought it would be useful to build a template for these sessions because there were a number of vocabularies that had already been drafted for the project which could be easily built in future using the CEDAR tools. The metadata template and associated vocabularies (e.g., for instrument types) could then be easily used or integrated with other data management tools that a project team was building to facilitate the automated collection of information and reduce the amount of time the domain scientists would need to spend on creating metadata for each session. See Figure 1.

Figure 1. Partial view of draft metadata template in CEDAR OpenView.



The group also made improvements to the original vocabulary, and created and submitted 3 vocabularies of their own, which they submitted to BioPortal. See Figure 2.

Figure 2. Summaries of draft vocabularies in BioPortal created by workshop participants.



The group developed and refined these vocabularies[6] during the workshop sessions, and in the homework period between the two sessions. Issues identified while doing the homework were addressed in the subsequent sessions, giving everyone confidence that the technical approaches were viable.

Lessons Learned

Our assessment is that this pilot was successful. We met the aims of the workshop proposal, learned about the intricacies of community engagement and delivery, and received positive feedback from the participants regarding outcomes. This assessment is based on feedback received during and after the pilot workshop by both participants and observers who found the level of explanation about the FAIR principles to be appropriate, and the process and tools demonstrated to have significant potential for continued or new adoption in their own environments. We found that the format of the M4M workshop is flexible enough to adapt to the specific needs of a community while also following the more generic explanations of the FAIR principles that emphasize the importance of community-sanctioned metadata standards and vocabularies. For example, based on specific feedback, the M4M workshop could also help the

[6] These URLs link to the vocabulary created by organizers, followed by the vocabularies created by workshop participants:
https://bioportal.bioontology.org/ontologies/ECM4M-MD-VOCABS
https://bioportal.bioontology.org/ontologies/OREX-WG
https://bioportal.bioontology.org/ontologies/OREX-SAM
https://bioportal.bioontology.org/ontologies/O-REXTECH

Geochemistry community come to agreement on metadata schemes and vocabularies for data transfer into internationally-shared community data archives.

Other positive aspects for this workshop included:

- Productive pre-interviews with key technical attendees which resulted in good coordination, well-articulated examples, and better interactions with the highly technical participants during the workshop. The back-and-forth between trainers and participants illuminated the utility of many technical aspects of the M4M approach.
- Participants who were fully engaged and made significant effort (as part of their homework) to develop new vocabulary content and integrate recommendations from the workshop. Many also showed a ready grasp of the technical tools and materials (as would be expected given their advanced technical level).

Despite the success of the workshop, there were challenges to developing and presenting this workshop including:

- Difficulty in finding an appropriate and sufficiently motivated community to participate
- Lack of familiarity with the M4M approach
- Lack of Geoscience-based exemplars
- Early stages of impact measurements for the GO FAIR framework
- Finding the appropriate timing for the workshop in projects' lifecycle

Seeking an appropriate community: One of the most challenging aspects of developing and presenting the workshop was the difficulty in finding an appropriate and sufficiently motivated community to participate. This was true of the early stages of M4M workshop development in the EU as well; significant support and uptake by a funding agency in the Netherlands has increased uptake. Recently, the M4M workshops that have been conducted in Europe by the international GO FAIR organizations seem to have benefited from post-workshop engagement with participants who provided descriptions of workshop utility both between and among researchers of various subject domains. In addition, governments of European countries such as the Netherlands, Germany, and the United Kingdom have supported and funded work toward implementing the FAIR principles as a way to encourage open and efficient scientific research. Neither of these motivators are present in the United States at this time.

Lack of familiarity: Given that this workshop was funded as a "pilot" project, the lack of familiarity and communication among research communities about the promise of the GO FAIR framework and the M4M workshop is not surprising. As more M4M workshops are developed and planned, we are confident that the utility of the M4M approach will become more visible as more communities take part in them. Note that M4M workshops are targeted for delivery to other communities in the U.S. per direct requests from those projects.

Lack of Geoscience-based exemplars: Another disadvantage that we faced in finding a suitable community for our pilot workshop was the lack of a Geoscience-based exemplar that showed the value of the GO FAIR framework's approach to Geoscience research. Although there are very successful exemplars in the health and bioinformatics sectors, specifically those

that were and are very important for early research on dealing with the COVID-19 pandemic in Africa (See VODAN Africa), the GO FAIR approach is less well-known in the U.S., and less noted by Geoscientists who would, understandably, be more convinced by seeing similar success in their own disciplines.

Lack of impact data: Another factor that proved a challenge for our efforts to engage participants for the workshop was that broad assessment of the GO FAIR framework is still in the early stages of deciding the means for measuring the impacts of using GO FAIR tools and techniques to implement the FAIR principles. This is an area of ongoing effort and concern by the international GO FAIR organization that also sees the need for this kind of data. We found that when the broader assessment data are not available, some groups reacted with some skepticism, particularly when they are moving ahead with their own efforts to create effective automated workflows to improve the collection of metadata for their research data.

Project timing: Factors associated with the natural timing of development within a project or community proved to be the case for one of the most promising communities we approached: the Tephra community, which is focused on research related to a unique volcanic product (tephra) that plays an unparalleled role in understanding past eruptions, long-term behavior of volcanoes, and the effects of volcanism on climate and the environment. This community has already developed some vocabularies and a tool for gathering metadata, and is in the process of distributing information about the vocabularies they have developed. We thought that perhaps the community could benefit from an M4M workshop that provided a process for automating the production of metadata and vocabulary refinement, and thereby improve the interoperability of their data and reusability of their vocabularies. The community did not follow up on invitations to participate in the pilot workshop, but may be interested in future workshops that can be tailored to support and assist them in their efforts to standardize the metadata workflows for their research products.

Another promising Oceanography community was deeply immersed in rebuilding their infrastructure for gathering metadata. So, although there was great interest in the potential for further automation of the metadata creation/maintenance and data curation activities of their research support specialists, the timing for the workshop did not align with an ongoing architecture revision. There does seem to be good potential to involve both the Tephra and the Oceanography communities in future M4M workshops.

The ESIP Soils Ontology Cluster was interested in attending the pilot workshop as observers; however, they were not available to attend on the dates we planned. The Cluster was particularly interested in motivating their participants to take part in hands-on activities that would help create agreements within their own vocabularies. Given the flexible nature of the M4M format, the workshop can address these specific phases of the process by adjusting which facets are emphasized. By adjusting the content to some extent, this Cluster might benefit in the future from an M4M workshop format that explores a process by which such agreements can be developed and documented.

Based on these challenges, developing strategies for identifying and recruiting potential participant groups could be very productive for future workshop organizers. Strategies could include how and when to recruit participants, and how to analyze where a project or community is in terms of metadata support and development. As we learned in developing this pilot workshop, after a project has started, potential M4M participants become increasingly wedded to the path where funding and efforts have already been invested for a metadata approach, no matter how good or poor, or FAIR or unFAIR, that approach may be. Documenting these kinds of strategies and developing criteria for when to apply them in order to better identify and recruit participants would not only be more productive for workshop organizers, but also likely to reduce costs for project coordination and content development. For this pilot event, activities for each community that was approached had to be repeated to some extent to learn about their needs, and navigate their interest and potential participation. This process of "customer development" is necessary to assess specific needs and fit, and to develop a plan of delivery that meets the community's expectations.

Next Steps & Future Opportunities

Preparation

As noted, there are several Geoscience communities that could be approached for future M4M workshops in Geochemistry, Tephra, Oceanography, and Soils Ontology. In each of those situations, however, it is important for the communities themselves to identify their participants and available timelines to maximize the upfront planning for workshop engagement.

To prime the communities, future workshop organizers would benefit from an effort to pull together some information about the GO FAIR framework, including the M4M workshops, that would better demonstrate the advantages of the approach and its successes even if they are not directly related to the Geosciences. Having such information available would ease the efforts required to "market" the M4M workshops. Efforts would be better spent responding to the special needs of communities and tailoring workshop format and materials to those needs in a dynamic and timely response.

Timing and Engagement

A key element in engaging communities to participate with the M4M training process is the relative timing of its introduction in the community's activities. As indicated above, the greatest opportunity for receptivity and adoption of any new approach comes at the beginning of a project, before a commitment to participate. We can describe this concretely using the following hypothetical project milestones. Assuming the funding organization supports the goal of hosting M4M workshops and the approach is introduced at the corresponding Milestone, we estimate the percentage likelihood a community project will accept training in an approach, and actually adopts that approach, as follows:

Milestone	**Accepts Training?**	**Adopts Approach?**

Before the proposal is written:	90%	75%
Before the grant is awarded:	70%	50%
Before development starts:	40%	25%
After development starts:	20%	10%

These likelihood estimates are based on our experiences with these and other technology approaches, and they align with the GO FAIR Foundation's reports when working with organizational mandates for metadata, data sharing, and other 'noble objectives' that require effort on the part of community project's staff.

In most projects, the longest part of the project's life cycle occurs after development starts (the fourth line in the chart above), so there is a relatively small window of time to influence most projects. It is easy to appreciate that early buy-in from the funding organization—even before solicitations are issued—strongly impacts the adoption of effective metadata and harmonization technologies over time. (This is true for **any** desired technology or best practices approach, not just the M4M training we discuss here.)

Other things also motivate engagement and adoption, as the discussion on Pre-Planning reflects. Expectations of funding agencies are paramount in driving the interest of communities, and explicit requirements from the funding agency that are expressed in the solicitation (either for addressing general principles like FAIRness and data harmonization, or even more directly for adoption of specific approaches, perhaps with explicit repositories or tools) can drive adoption even more strongly.

FAIR Infrastructure Development

The goals of Open Science necessitate the development of FAIR vocabularies and related information infrastructure. As with the Timing and Engagement chart above, intentionally scheduling programming to aid in the FAIRification of resources would facilitate development of the information infrastructure needed to produce domain-based resources and other community-driven outcomes that enable domain interoperability and reuse. Increased capacity around FAIR vocabularies will also serve to build awareness and implementation across domains, and drive efficiencies resulting from the scaling of training services.

The fact that this workshop was pre-funded made it feasible for this community. That is, we posit that self-organizing at a project or team level to develop proposals and seek expertise for this kind of information development work is challenging and not a priority. Rather, funding via multiple federal agencies could support a variety of approaches, such as:

- Projects requesting a workshop to meet the elements of their Data Management Plan;
- Program Officers or agencies providing support at the Program level for projects to request supplemental funds to hold a workshop, thus moving larger communities toward more interoperability;
- Directions within Solicitations or Requests for Proposals to include plans and budget for FAIR training or FAIRification actions;

- Direct support to program awardees from the funding organization to attend a federally-organized workshop.

Conclusion

The GO FAIR Metadata for Machines (M4M) workshop was successful as a pilot designed to explore an approach to mobilizing Geoscience research communities around practical solutions for making needed metadata FAIR. We found a research community engaged in Astromaterials and Geochemistry research willing to participate actively in an M4M workshop that is in a very good position to take advantage of the opportunity to come together to leverage the work that had already been done to create standardized, community-agreed upon metadata standards with their concomitant vocabularies. During and since the workshop in September, members of the OSIRIS-REx and SAMIS projects have used workshop materials to further document and refine a metadata template and vocabularies that they plan to build upon and then make available to the domain researchers they support.

Judging from the enthusiastic reaction of workshop participants and observers, the tools and processes demonstrated in the M4M Pilot workshop proved useful and promising for their future work. While there were a number of challenges involved in finding an appropriate audience and tailoring the workshop content and format, organizers see very good potential in these kinds of workshops. M4M workshops can help other Geoscience communities develop, document and maintain metadata standards and vocabularies that support the machine-actionability and workflow automation critical to making their data and metadata findable, accessible, interoperable, and reusable (FAIR).

Acknowledgements

We thank the teams from the OSIRIS-REx and EarthChem communities for their effort and feedback as participants in the workshop, as well as the EarthCube community groups who participated in the start-up engagement processes. This work is supported through the NSF award #1928208.

Appendix A: Workshop Agendas

DAY 1 - Build machine-actionable controlled vocabularies

#	**Duration** Total: 4 hrs	**Agenda Item**
1	0:10	Welcome: Self-introductions of presenters and participants Workshop: Conceptual scope, motivations and practicalities Summary: Short introduction about the workshop
2	0:30 GO	Introduction to Automating FAIR using linked, machine-actionable (meta)data
3	0.45 GO	How to build machine-actionable controlled vocabularies
	0.15	Break
4	0:15 GO	How to build domain-specific controlled vocabularies
4A 4B 4C	1:30 0:30+ each 0:30 0:50 0:20	Exercises (your choice!) in building domain-specific controlled vocabularies 4A: Pick one or more vocabularies to improve 4B: Create a new vocabulary (sheet) for your domain 4C: Set up a workflow with GitHub & BioPortal Pose / discuss advanced vocabulary topics
5	0.15 GO	Round-off and what happens next Summary: A summary of the first day of M4M and a prelude to day 2 of M4M.

DAY 2: Build a machine-actionable metadata template

#	**Duration** Total: 4 hrs	**Agenda Item**
0	0:10 GO	Review and issues from Day 1
1	0:45 GO	How to build machine-actionable metadata templates
2	0:20 GO 0:40 GO	Step 1: Make a copy of existing template (e.g., Generic Dataset Metadata Template) Step 2: Create metadata fields and assign controlled vocabularies
	0:15	Break
3	0:30 GO 0:20 GO 0:10 GO	Step 3: Add created or new fields to metadata elements and add RDF properties Step 4: Structure new or copied template by composing and replacing elements Step 5: Make machine-actionable metadata by filling out the form
4	0:20 GO	FAIR Orchestration and Integration with other systems
5	0:15 GO	Questions and Discussion
6	0:15	Round-off and what happens next? Summary: What we did during M4M and discussion of next steps.