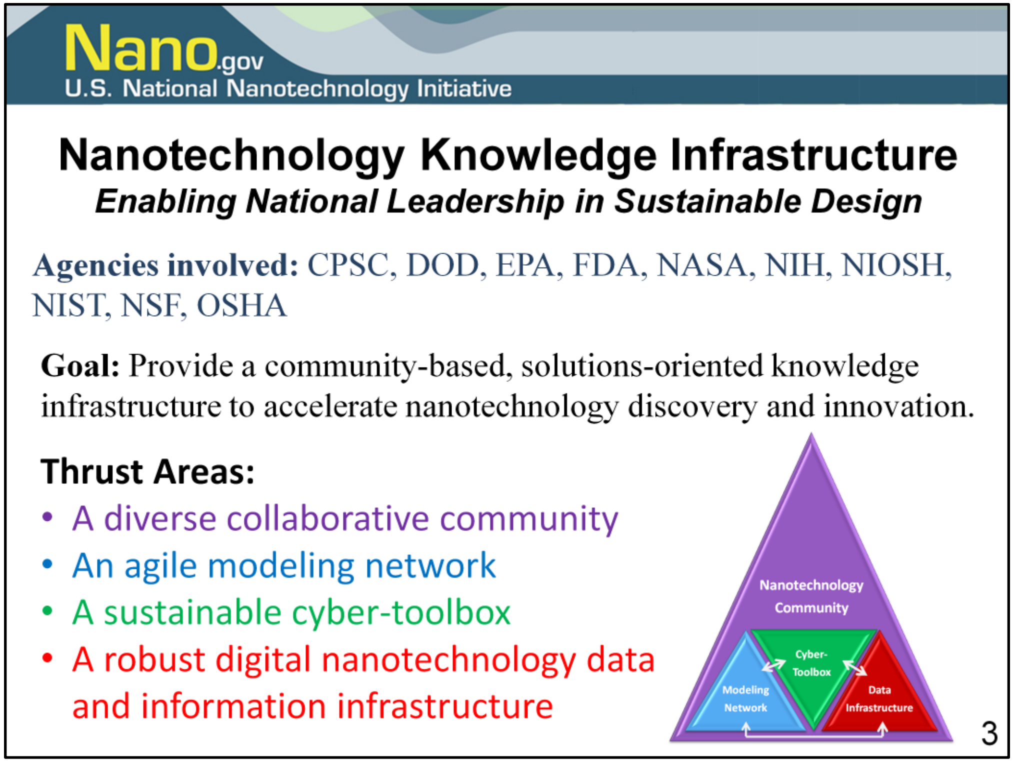
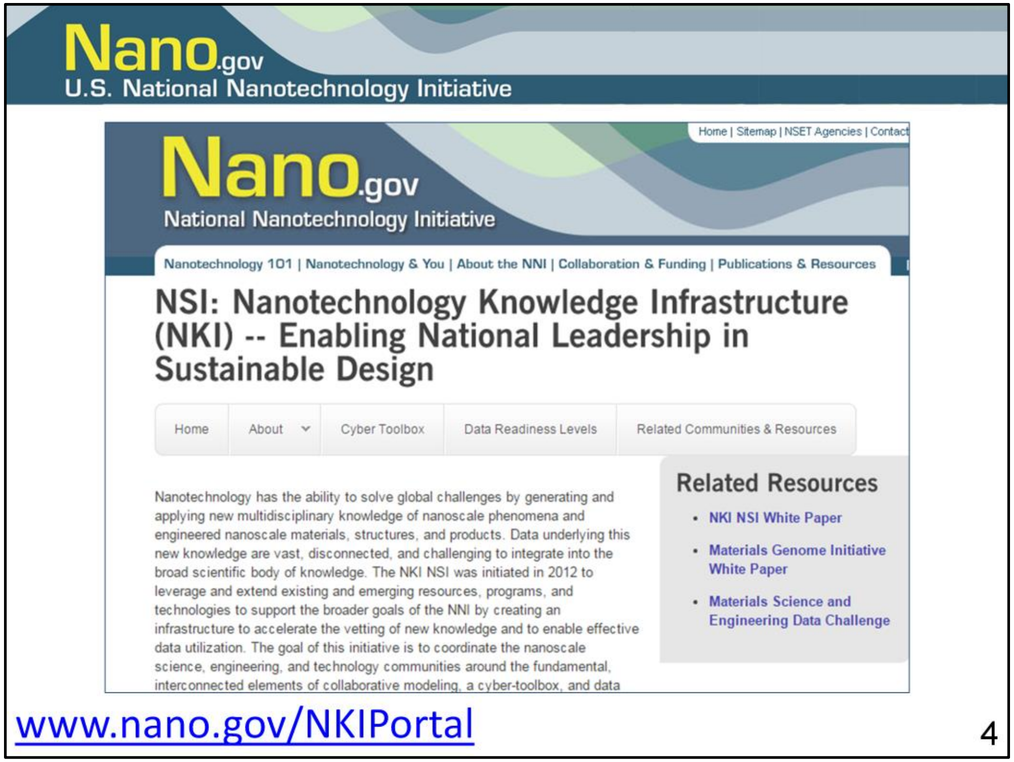
>> Stephen Lehrman: Greetings and welcome to today's Introduction to Nanoinformatics Webinar. My name is Stephen Lehrman, and I'm with the National Nanotechnology Coordination Office. I will be today's moderator. We have great panel of speakers today, and I invite you to read their bios online. By way of brief introduction, Dr. Mark Hoover is Senior Research Scientist with the National Institute for Occupational Safety and Health or NIOSH; Dr. Christina Liu is Program Director at the National Cancer Institute, Office of Cancer Nanotechnology Research; Dr. Robert Hanisch is the Director of National Institute of Standards and Technology's Office of Data and Nanoinformatics, and Dr. Daryl Hess is a Program Director in the Division of Materials Research at the National Science Foundation. Please note that the views expressed by our panel are their own and do not necessarily represent the positions of their respective agencies.

>> Stephen Lehrman: The National Nanotechnology Initiative, also known as the NNI, is a partnership of 20 Federal agencies and departments with activities in nanotechnology research, development, policy, and regulation. More information on the NNI, including reports, nanotechnology news, upcoming events, and solicitations, can be found on our website, nano.gov. These agencies have established five Nanotechnology Signature Initiatives in order to enhance interagency coordination and collaboration. The Signature Initiatives leverage resources and capabilities of the NNI agencies to maximize progress and provide a forum for ongoing communications. The Signature Initiatives are not intended to be a purely Federal activity, but to catalyze communities of interest that extend into academia and industry. In addition to the NNI activities, the Materials Genome Initiative is a multi-Federal agency initiative to discover, manufacture, and deploy advanced materials twice as fast and at a fraction of the cost of traditional methods. Dr. Hanisch will talk more about this initiative during his remarks.

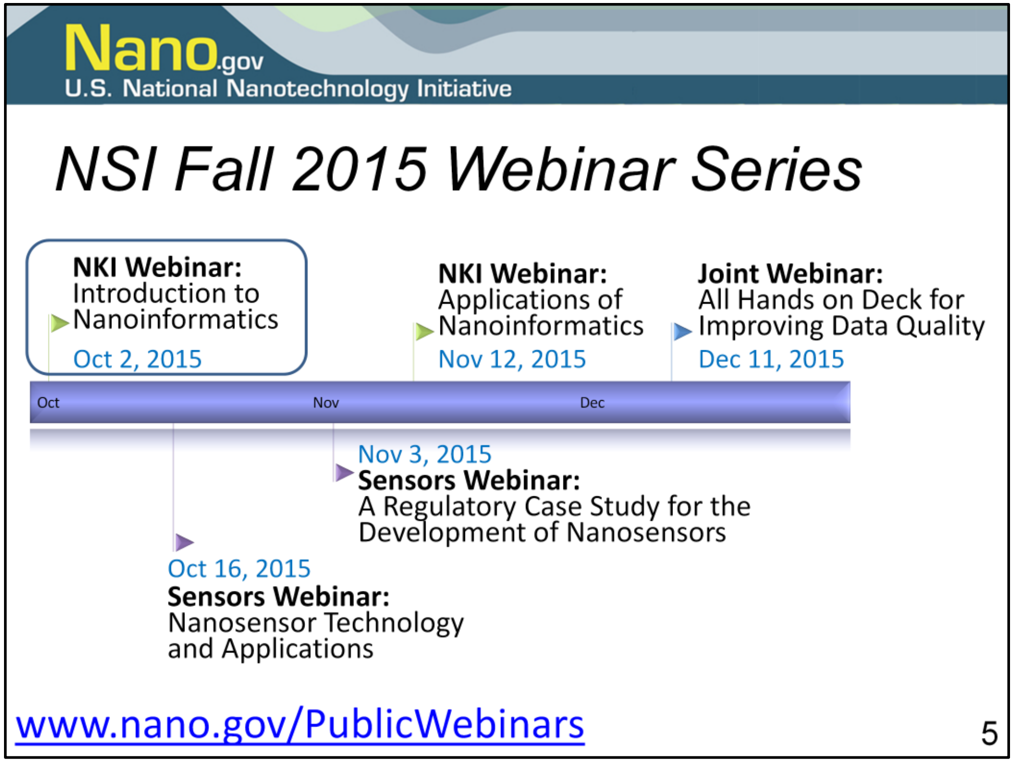


>> Stephen Lehrman: Today's nanoinformatics webinar is sponsored by the Nanotechnology Knowledge Infrastructure Signature Initiative, which involves ten Federal agencies. Also known as the NKI, the goal of this Signature Initiative is to provide a community-based, solutions-oriented knowledge infrastructure to accelerate nanotechnology discovery and innovation. There are four thrust areas:

1. Develop a diverse, collaborative community;
2. Create an agile modeling network;
3. Develop a sustainable cyber toolbox; and
4. Establish a robust digital nanotechnology data and information infrastructure.

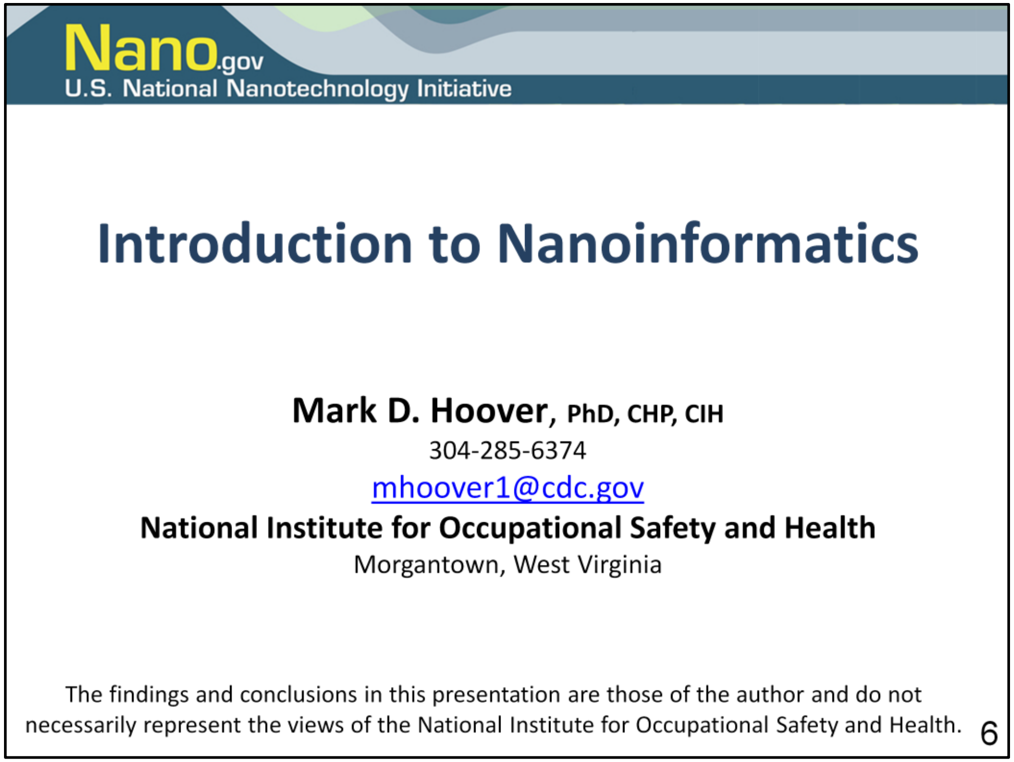


>> Stephen Lehrman: An NKI web portal was recently created at www.nano.gov/NKIportal to share information about NKI agencies and activities, related communities and resources, and a cyber toolbox that includes relevant models, tools, and databases.



>> Stephen Lehrman: This webinar is the first of five Signature Initiative webinars we have planned for this fall. In two weeks, the Nanotechnology for Sensors and Sensors for Nanotechnology Signature Initiative will host a webinar on Nanosensor Technologies and Applications, and registration for that event opens today. In early November, the Sensors group will go through a regulatory case study for the development of a biomedical nanosensor. On November 12th, the NKI group returns with a webinar explaining how to apply nanoinformatics tools and databases to real-world problems. And finally, on December 11th, the NKI and Sensors groups are jointly hosting a webinar on the important issue of improving data quality. You can find out more information about these and other NNCO-hosted webinars at www.nano.gov/publicwebinars.

You are welcome to submit questions via email at webinar@nnco.nano.gov or in the "submit your questions here" window in the webinar interface.



>> Stephen Lehrman: And now I would like to hand the floor over to Dr. Mark Hoover.

>> Mark Hoover: Thank you, Stephen. It is a pleasure to provide this short introduction to nanoinformatics. And my thanks to all of you who are participating in today's webinar. A number of you have been involved for more than a decade now in our community-based efforts to define and apply nanoinformatics. Thank you for sharing the ideas, insights, and experiences that have contributed to the slides I am presenting here today. Others of you may be learning about nanoinformatics for the first time. I hope that our presentations will provide you with actionable information, as well as with a realistic path forward. Because we are a community-based effort, your investment of time in the past, today, and hopefully in the future is essential to our success. We hope that your personal reward for that investment will be an increased ability to adapt and apply nanoinformatics principles for success in both your organizational missions as well as in your everyday activities.

>> Mark Hoover: Our premise is that we can accelerate discovery, revolutionize design, and sustain innovation through what we are calling “a knowledge infrastructure.” We believe that the return on investment for our collective efforts will be significant. And we believe that success in our efforts requires substantial grass roots involvement. We need the experience and involvement of real people doing real things to achieve real outcomes with real materials. Today’s nanotechnology is what it is, although most of us might be hard pressed to clearly state what that is, but tomorrow’s nanotechnology can be what we can make of it. Thank you for engaging with us to define and apply nanoinformatics in a manner that will have measurable and positive impact for you in particular and for the community in general.

Nano.gov
U.S. National Nanotechnology Initiative

Our Approach

Define *Nanoinformatics: Principles and Practices*

***Who** we are ...*

***What** we are trying to accomplish ...*

***When** we are taking action ...*

***Where** we are involved ...*

***Why** we taking specific actions ...*

***How** we are defining and meeting our goals ...*

This is a work in progress ... 8

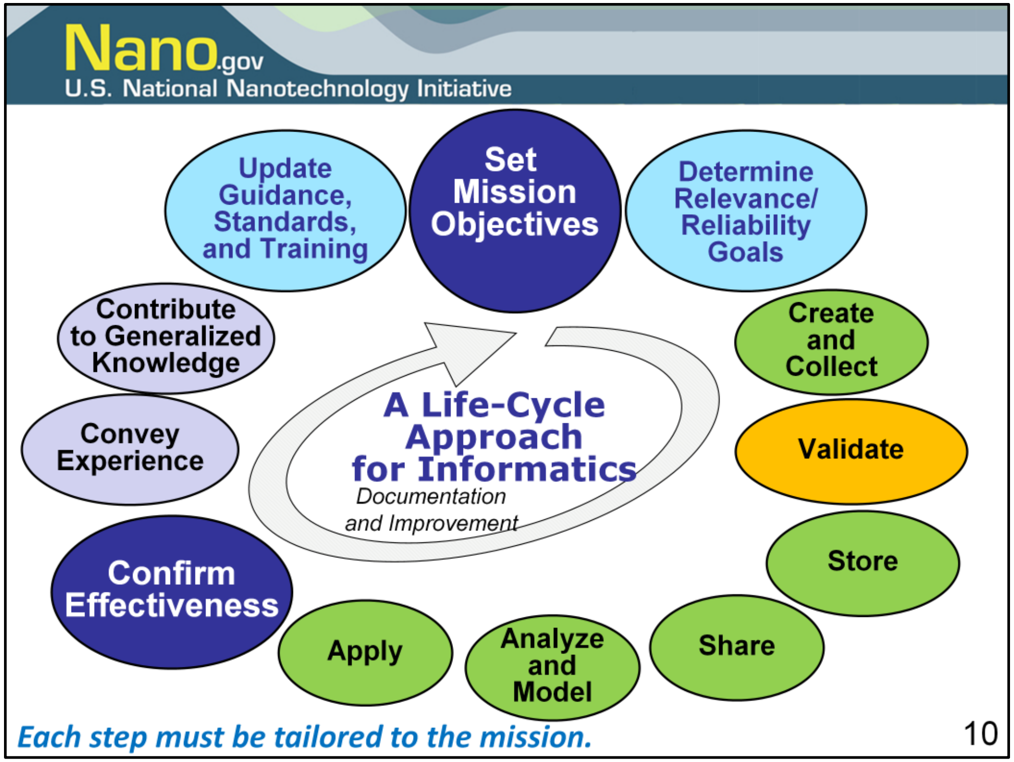
>> Mark Hoover: Our approach to building a robust and sustainable nanotechnology knowledge infrastructure has involved a core effort to define what we hope will prove to be a credible set of nanoinformatics principles and practices. As many of you may already know, we have a community-based collaboration to create a reference text on Nanoinformatics: Principles and Practices, and we invite you to be involved. I hope that many of you will give me a call to talk more about how you can help. As shown in this slide, the process of defining relevant and reliable principles and practices for nanoinformatics has required us to address the classic “Who, What, When, Where, Why, and How” questions. This is truly a work in progress, and we need the benefit of your experience and your specific programmatic and mission needs.

A Working Definition of Nanoinformatics

- The ***science and practice*** of determining ***which information is relevant*** to meeting objectives of the nanoscale science and engineering community,
- and then developing and implementing effective mechanisms
- to *collect, validate, store, share, analyze, model, and apply the information, and then to confirm achievement of the intended outcome* from use of that information,
- and then *conveying experience to the broader community, contributing to generalized knowledge, and updating standards and training*.

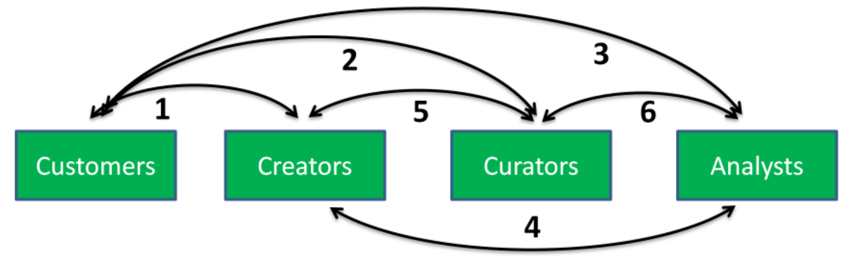
Adapted from http://www.internano.org/nanoinformatics/ 9

>> Mark Hoover: This slide presents our working definition of nanoinformatics. We have created it to be comprehensive in detail because we consider it to be an essential part of our roadmap to get things done. I invite you to read the details of the definition in the webinar slide set that will be available online after today's presentation, and I also invite you to examine the Nanoinformatics 2020 Roadmap document, which is accessible at the Internano weblink listed at the bottom of this slide. In the next slide, let's look at a graphical version of our definition.



>> Mark Hoover: In this depiction of our nanoinformatics life cycle, two elements appear with particular prominence. Those are the key initial step to set mission objectives and then the key assessment step to confirm the effectiveness of our efforts. The details of those steps can be tailored to each of our unique objectives. Along the way, credible actions have to be taken in each of the life cycle steps to determine relevance and reliability goals for the data, and then to create and collect, validate, store, share, analyze and model, and apply that data. After confirming whether we have achieved our objectives and deciding how to improve our performance, we need to convey our experience to the community, contribute to generalized knowledge, and ultimately update standards and training to support the next generation of our work. It truly is a life cycle.

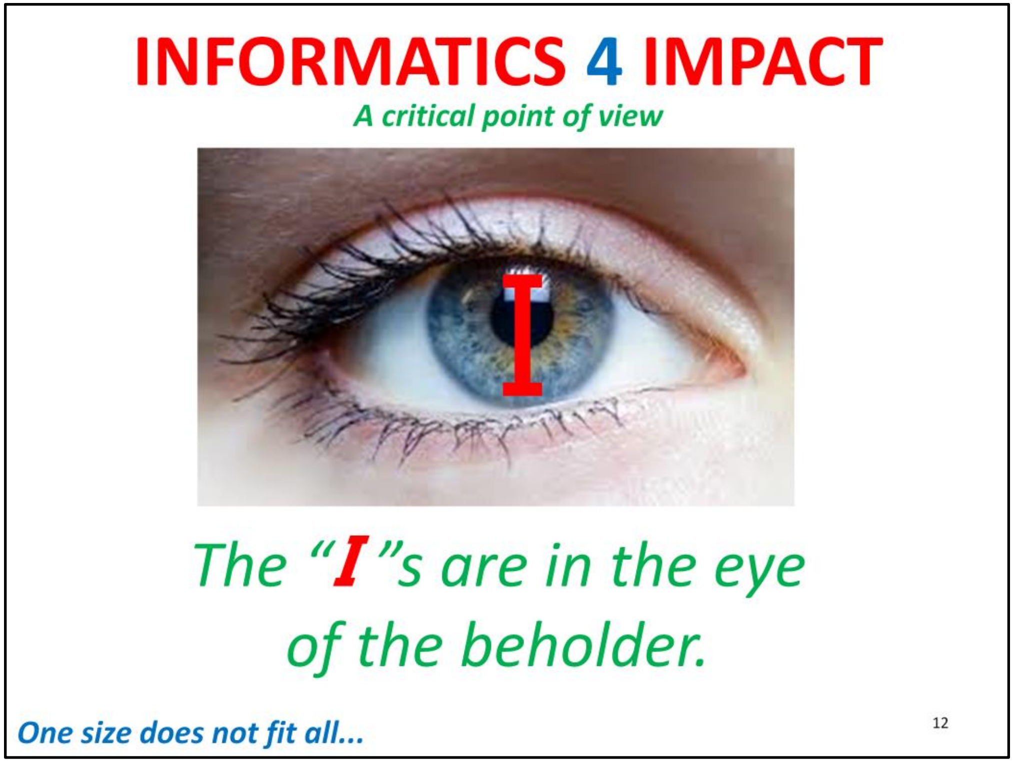
Informatics Roles and Responsibilities



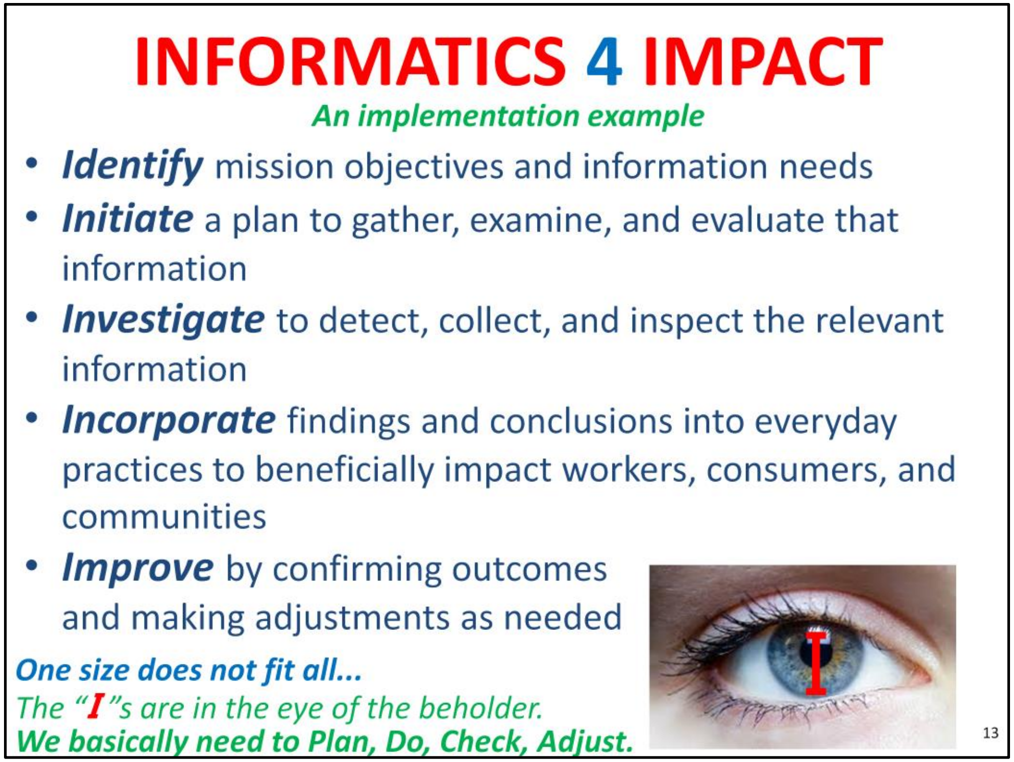
	Set Mission Objectives	Determine Relevance	Collect	Validate	Store	Share	Analyze and Model	Apply	Confirm Effectiveness	Convey Experience	Generalize	Update Guidance
Customers	X	X						X	X	X	X	X
Creators		X	X	X					X			X
Curators		X		X	X	X			X			X
Analysts		X		X			X		X		X	X

Communication and understanding are essential at all steps. 11

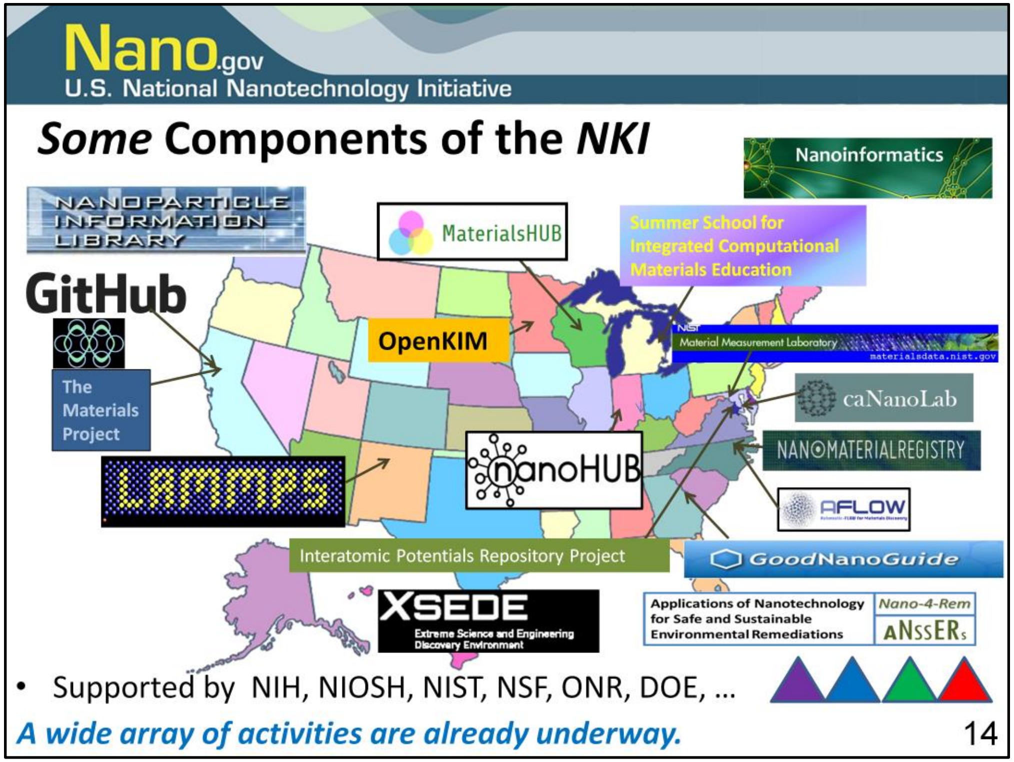
>> Mark Hoover: As shown in this interactive network of nodes and interfaces, we believe that informatics can be applied for impact in our own missions by understanding and carrying out our roles and responsibilities as either a data customer, or a data creator, or a data curator, or a data analyst. Sometimes we wear all four hats. Sometimes we change back and forth among hats. But the pitfalls are clear. Credible actions are required within each of the four nodes and across each of the six interfaces. And actually, there are 12 interfaces if we take into account the fact that each interface must work in both directions. For example, just because the analyst has clearly articulated the data needs to run an applicable model doesn't mean that the data creators or the data curators are going to be able to appropriately translate and respond to those data needs. No wonder so many real-life efforts in the business of science and the business of business go so poorly. Good informatics is really not something for dabblers. There are a lot of moving parts.



>> Mark Hoover: We believe that our ability to personalize informatics for impact in our own missions is key to our likelihood of success. As shown here in what I think is a delightful pun, we believe that the “I’s” are in the eye of the beholder. One size does not fit all. There are a lot of ways that we can view the process.



>> Mark Hoover: In the example shown here, we expand on our use of the "I" words as an informatics-for-impact process in which we Identify our own mission objectives and our own information needs; then Initiate a plan to gather, examine, and evaluate that information; then we Investigate to detect, collect, and inspect the relevant information; then we Incorporate our findings and conclusions into our everyday practices so that we can beneficially impact workers, consumers, and communities; and finally we Improve by confirming the outcomes we achieved and by making adjustments as needed. We believe that informatics-for-impact is a generalizable framework, because we recognize that the details of the process can and must be personalized. One size does not fit all. The "I"s are in the eye of the beholder. We basically need to Plan, Do, Check, and Adjust, which many of you will recognize as the well-established business management mantra of the Deming Cycle.



>> Mark Hoover: In this next slide it is a pleasure to note how many valuable activities are already underway. And this is just here in the U.S. Across the country you can access materials data from the NIST Material Measurement Laboratory, simulation tools from nanoHUB, characterization protocols and results from caNanoLab, examples of good industrial hygiene practice from the GoodNanoGuide, and curated information about specific nanomaterials from the Nanomaterial Registry. You will hear more about some of these resources in today's webinar, and you are invited to find links to them on our nano.gov webpage.

Nano.gov
U.S. National Nanotechnology Initiative

Example initiative for Data Readiness Levels

Concept for Expressing Data Maturity

DRL:	Data Readiness Level
0	Invalid data
1	Raw or unscaled data
2	Scaled data
3	Scaled data with defined precision or noise level
4	Scaled data with defined precision and noise levels, but not related to the larger body of scientific knowledge
5	DRL 4 data related to the larger body of scientific knowledge, but with measurement uncertainty too large for data standards
6(X)	Standards-quality data of X % measurement uncertainty

Readiness level is application-dependent. 15

>> Mark Hoover: In these next two slides I want to share a particular example of just one of the activities we are collectively pursuing in our Signature Initiative on the Nanotechnology Knowledge Infrastructure. As described in our white paper on the NKI section of nano.gov, we want to define what we are calling “Data Readiness Levels” in a manner that conveys the maturity of the data and the suitability of the data for actually doing something useful. Our approach has similarities to the method used widely in the government to assign Technology Readiness Levels for equipment and devices. Our Data Readiness Levels range from invalid data at the DRL 0 level to standards-quality data at the DRL 6 level. The obvious challenge is that the assignment of Data Readiness Levels is intrinsically application-dependent. In other words, materials science data on the durability at various pH conditions of carbon nanotubes might be of standards quality for structural applications but of a lower or even indeterminate quality for biological applications.

Nano.gov
U.S. National Nanotechnology Initiative

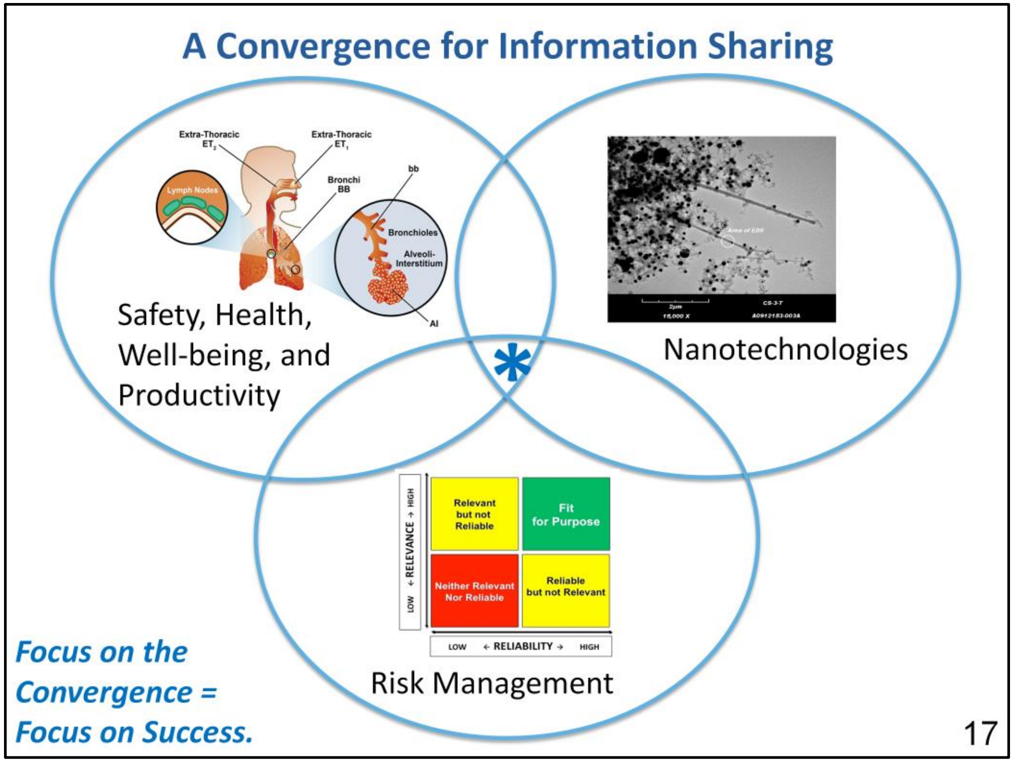
Data Readiness Levels

Summary of DRLs Versus Data Attributes

Attribute	DRL 0	DRL 1	DRL 2	DRL 3	DRL 4	DRL 5	DRL 6
Units		maybe	yes	yes	yes	yes	yes
Precision and Noise				either	both	both	both
Independent Confirmation				possibly	yes	yes	yes
Related to Larger Body of Scientific Knowledge					no	yes	yes
Measurement Uncertainty					specula-tive	high	low
Example or use	little to none	unscaled sensor data	scaled sensor data	scaled data; noise levels defined	major scientific advances	coarse validation of theory	theory refinement and methods validation

Data attribute details are application-dependent. 16

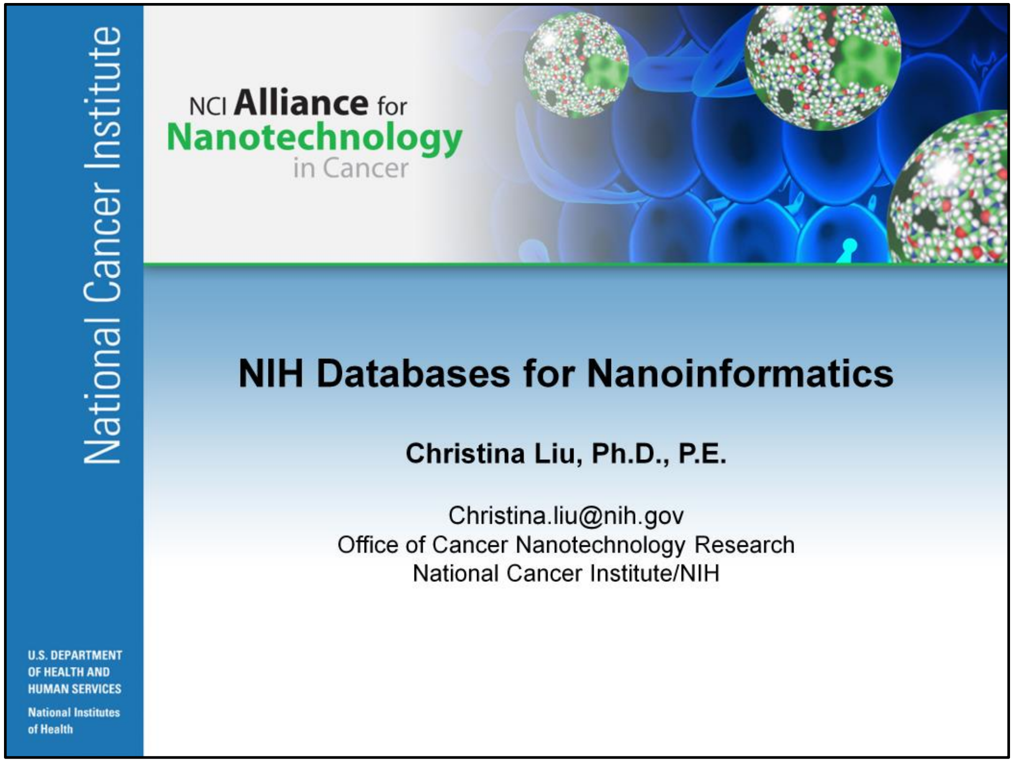
>> Mark Hoover: And in this expanded presentation of the DRLs, we provide some of the data attributes that we consider relevant to their assignment, along with some examples and uses of the various data levels. For example, DRL 0 data are, of course, of little or no value, and it is useful for us to have such data flagged so we can avoid it. DRL 6 data have a quality that merits there use for the refinement of theories and the validation of methods. We think our activity to define the DRLs has value and we invite your comments and contributions though our NKI webpage on nano.gov.



>> Mark Hoover: And now in my penultimate slide of this introduction to nanoinformatics, this Venn diagram conveys our work as a focus at the convergence of three areas. The first is the pursuit of safety, health, well-being, and productivity, which is common to everything we do. Essentially no one wants to do anything in a manner that is unsafe or unproductive, but the conscious convergence of that area with risk management enables us to critically assess the extent to which anything that we do actually affects the bottom line. We need good yardsticks. We need to ask the hard questions. Is what we want to measure, or what we want to study, or what we want to change really important to meeting our mission goals? We can't afford to fritter away our efforts on anything that doesn't matter. And finally, the NKI places the streetlight at the intersection of nanotechnology and commerce. A focus on that convergence equals a focus on success. In addition, our approach to addressing the challenges of nanotechnology is likely to be very relevant to our ability to meet the challenges that may be posed by other areas of emerging technologies.

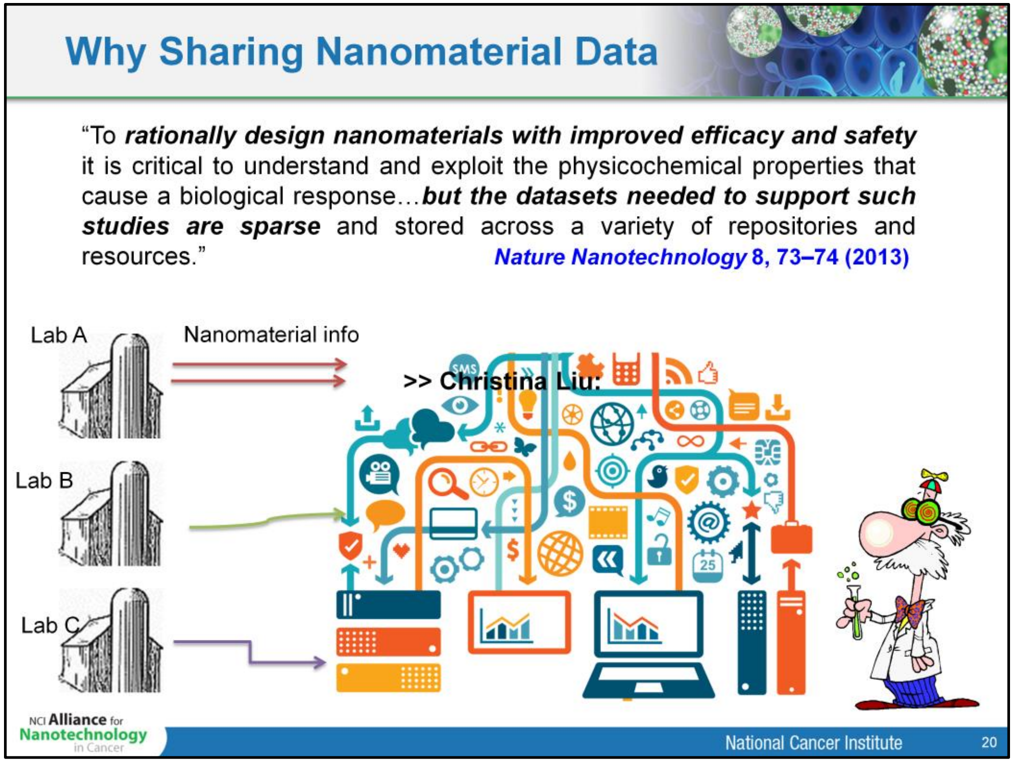
>> Mark Hoover: And in this final slide, we highlight the fact that we are engaged in a four-step process of community action to build and sustain leaders, cultures, and systems for safety, health, well-being, and productivity. We are engaging the community, informing the interested, and hopefully rewarding the responsive -- that is all of you. And we ultimately need to understand and incentivize the reluctant. That old saw of “Lead, follow, or get out of the way” applies to what we are doing. Thank you for engaging with us. We hope that your involvement and the application of what you have learned will prove to be rewarding to you. That concludes my introduction, and I look forward to some interesting discussions with you at the end of our webinar. And please do feel free to contact me at any time to follow up on anything that you feel that I might be able to help with.

>> Stephen Lehrman: Thank you, Mark. That was a great introduction.

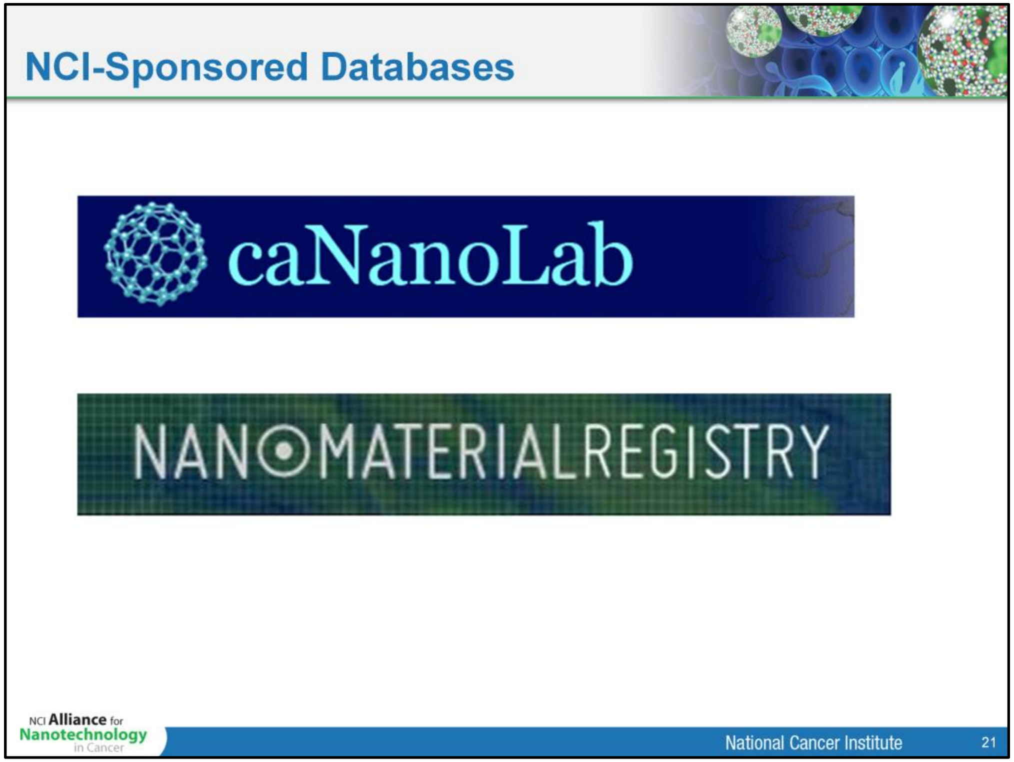


>> Stephen Lehrman: Our next speaker is Dr. Christina Liu of the National Cancer Institute.

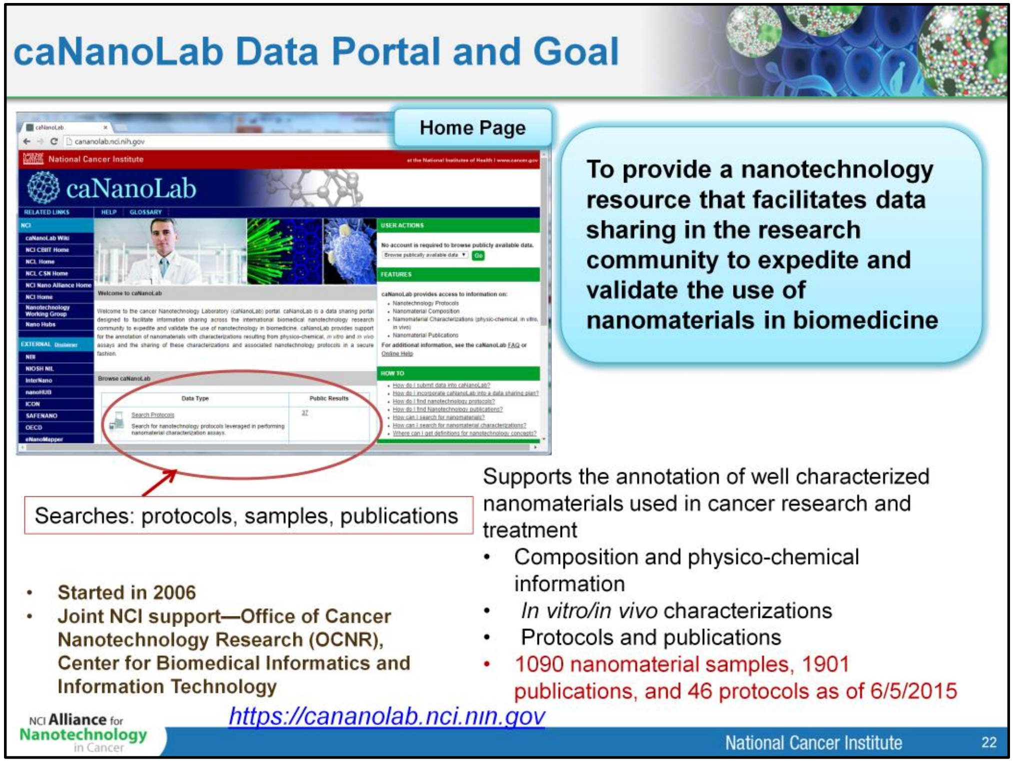
>> Christina Liu: Thanks very much, Mark, for the introduction to nanoinformatics. As Mark indicated, the basis of nanoinformatics, in fact, any informatics is data, data, data. NIH has awarded many biomedical, bioengineering research grants on nanomaterials development and their applications to disease detection and therapeutics. These awardees have generated a great wealth of nanomaterial data. So today I will show you two NIH supported data bases for nanoinformatics.



>> Christina Liu: The potential benefits of sharing nanomaterial data are significant, as indicated in this nature nanotechnology paper in 2013, to rationally design nanomaterials with improved efficacy and safety. Although the situation is getting better, the reality today is that different NIH-funded labs generate a variety of nanoparticles based on their understanding and expertise of the nanomaterials. These labs published their experimental findings and patented their nanoparticle inventions by themselves. It is very possible that there are lessons learned from Lab A and Lab C that can potentially help Lab B's nanoparticle designs and studies. Without knowing this information, Lab B could have spent additional years just to stumbling into similar pitfalls while designing and refining their nanoparticles, which is a waste of time and research money.



>> Christina Liu: Here are two NCI-sponsored databases to help the nanomaterial and nanoinformatics communities: Cancer Nanotechnology Laboratory (caNanoLab) and Nanomaterial Registry.



>> Christina Liu: caNanoLab is a data sharing portal designed to facilitate information sharing across the international biomedical nanotechnology research community to expedite and validate the use of nanotechnology in biomedicine.

A brief history, caNanoLab was established in 2006 by NCI to support the annotation of nanomaterials with characterizations resulting from physicochemical, in vitro and in vivo assays and the sharing of these characterizations and associated nanotechnology protocols in a secure fashion. From the caNanoLab home page, you will be able to search protocols, samples, and publications. As of June this year, there are more than 1000 nanomaterial samples, about 2000 publications, and 45 protocols in caNanoLab.

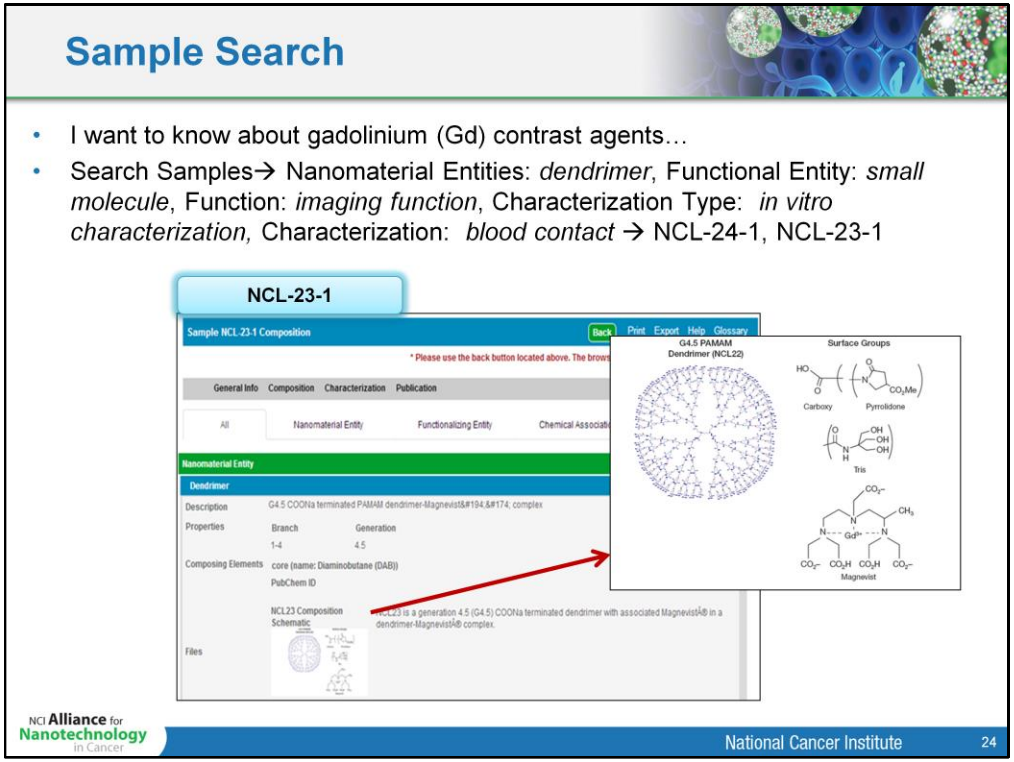
Summary of Data Available in caNanoLab

Particle Type (# of entities)	Metadata
Biopolymer (43)	Name, Type (e.g. DNA, Peptide, RNA), Sequence
Carbon Nanotube (50)	Average Length, Chirality, Diameter, Wall Type (e.g. DWNT, MWNT, SWNT)
Dendrimer (74)	Branch, Generation
Emulsion (88)	Is Polymerized, Polymer Name
Fullerene (16)	Average Diameter, Number of Carbons
Liposome (34)	Is Polymerized, Polymer Name
Polymer (188)	Initiator, Is Cross Linked, Cross Link Degree
…	…

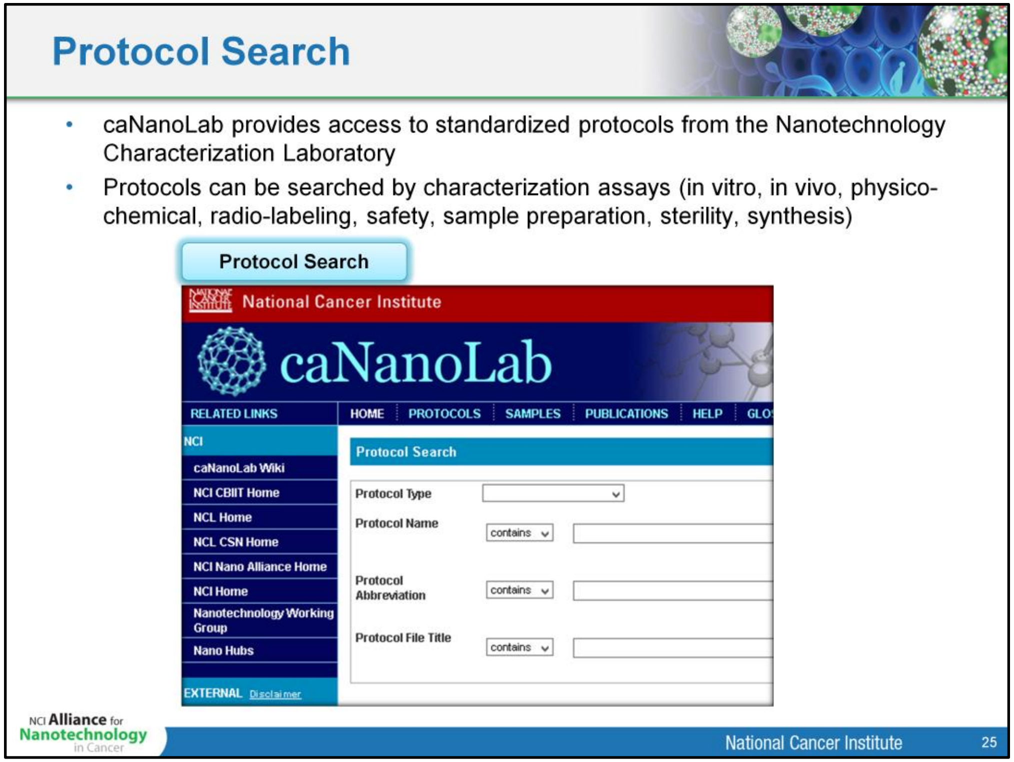
NCI Alliance for Nanotechnology in Cancer

National Cancer Institute 23

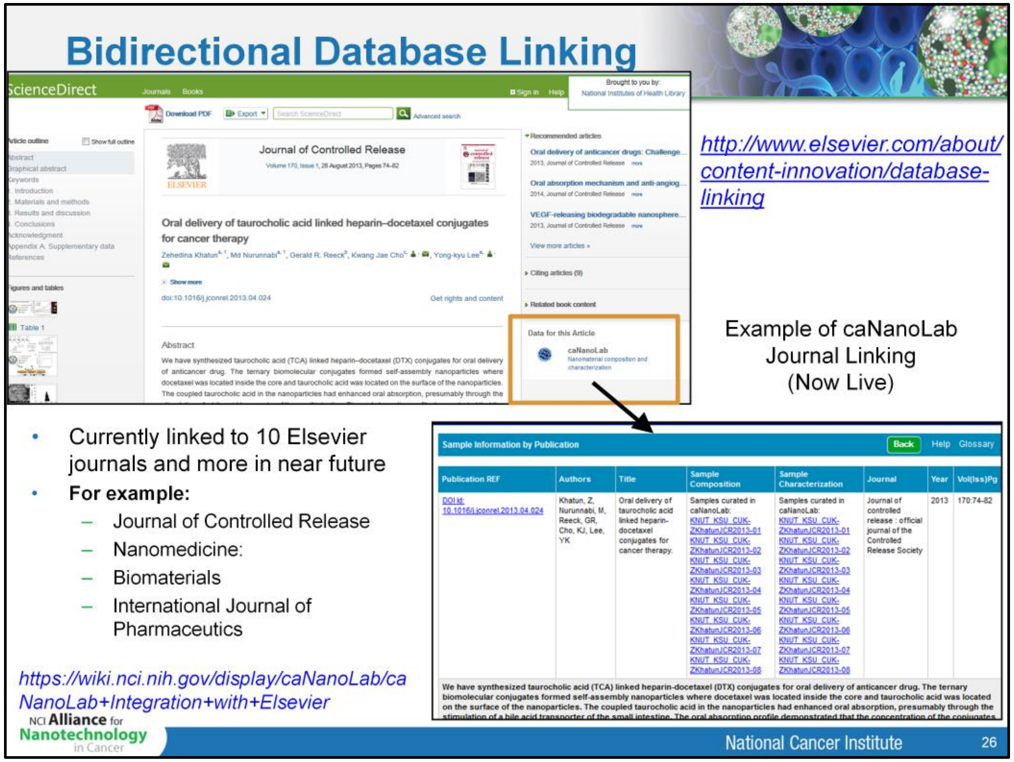
>> Christina Liu: Here is the summary of nanomaterial data available in caNanoLab, which include their types and numbers and what information each type of nanoparticles has.



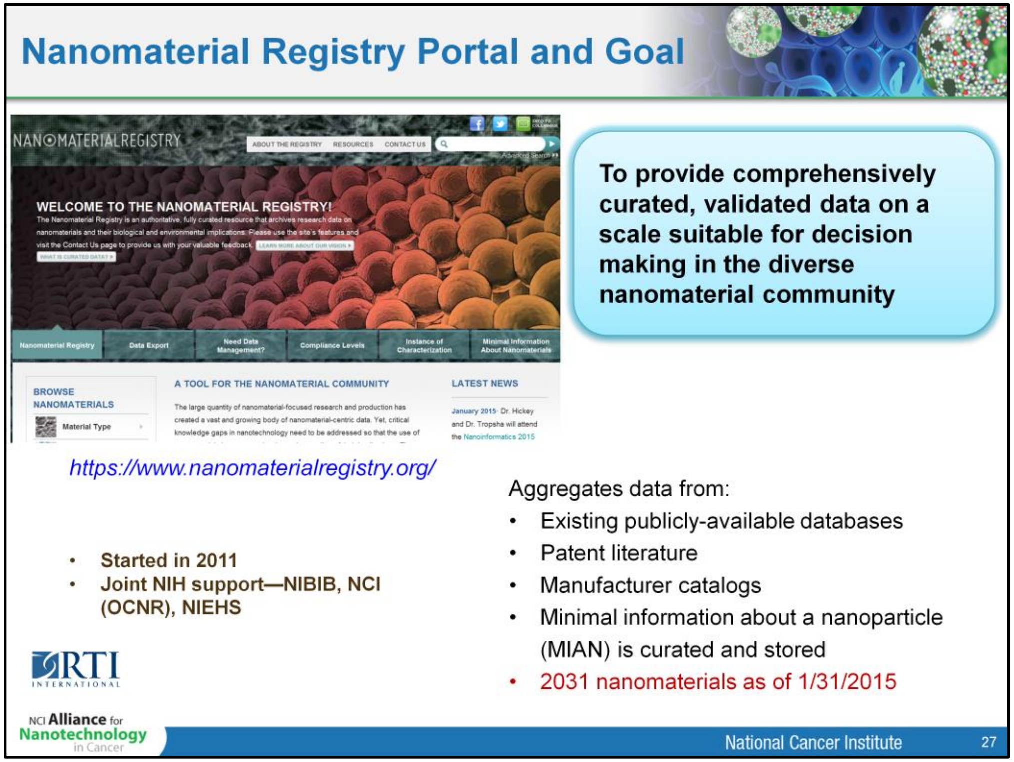
>> Christina Liu: From the portal, you can select sample search to look for a particular nanoparticle and its related properties. Based on the search criteria shown here, two samples showed up about gadolinium MR contrast agents. In the output screen, you will find nanomaterial information, including general composition properties inherent to particles, such as chemical names and formulas, as well as composition metadata.



>> Christina Liu: You can also do a protocol search form the portal. caNanoLab has a collection of standardized protocols from the Nanotechnology Characterization Laboratory. For your information, the Nanotechnology Characterization Laboratory is an intramural arm of the NCI's Nano Alliance and provides a characterization support to evaluate clinically promising nanomaterials and establish their physical, pharmacological, and toxicological characteristics. Protocols can be searched by the characterization assays listed on the slide. The predominant protocol types currently available are for physicochemical and in vitro characterizations.

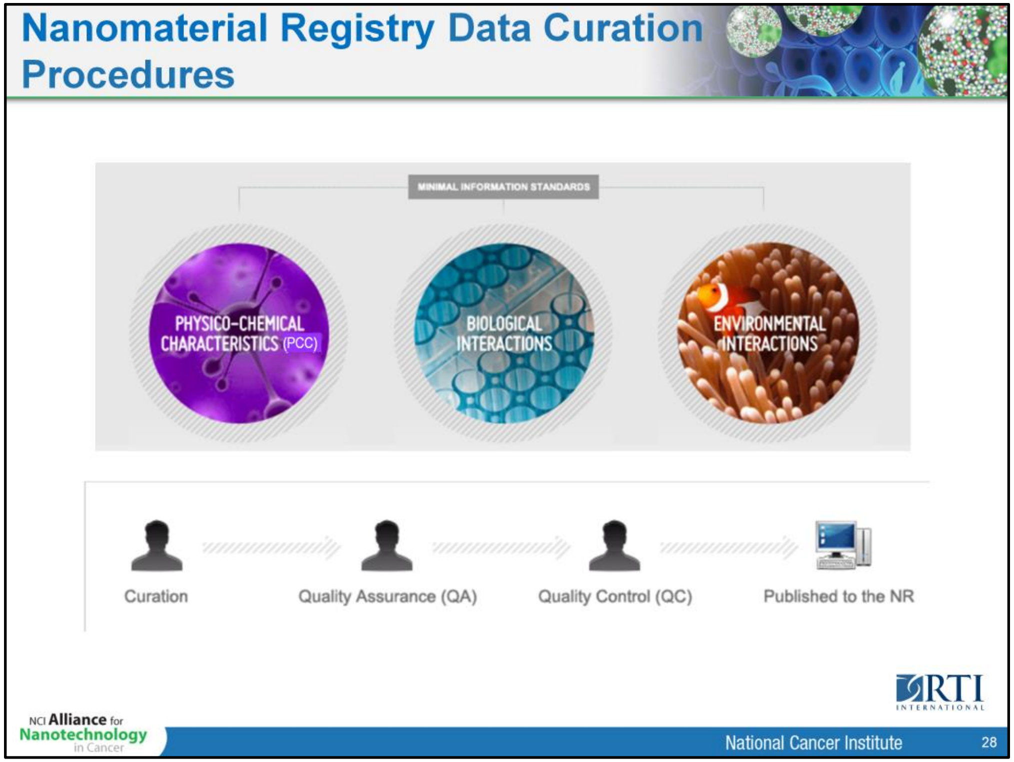


>> Christina Liu: To encourage standardized input of nanomaterial information through publication, caNanoLab worked with Elsevier for bidirectional linking between publication and data curated into caNanoLab from this publication. Currently there are ten Elsevier journals that have a link from their sites to caNanoLab.

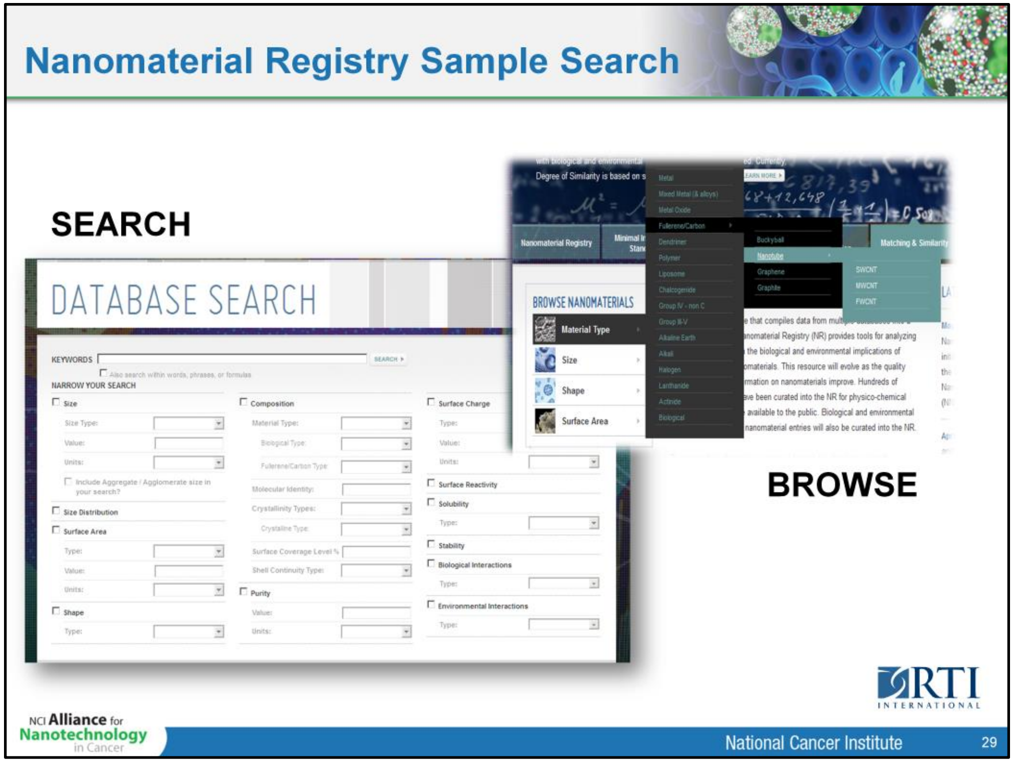


>> Christina Liu: In contrast to caNanoLab which collects data from publications, the Nanomaterial Registry collects validated data from a broad field of accessible nanomaterials sources relevant not only to medical applications, but also for the environmental implications of nanomaterials and their impact on human health and safety. One goal of the Registry is to provide comprehensively curated, validated data on a scale suitable for decision making in the diverse nanomaterial community.

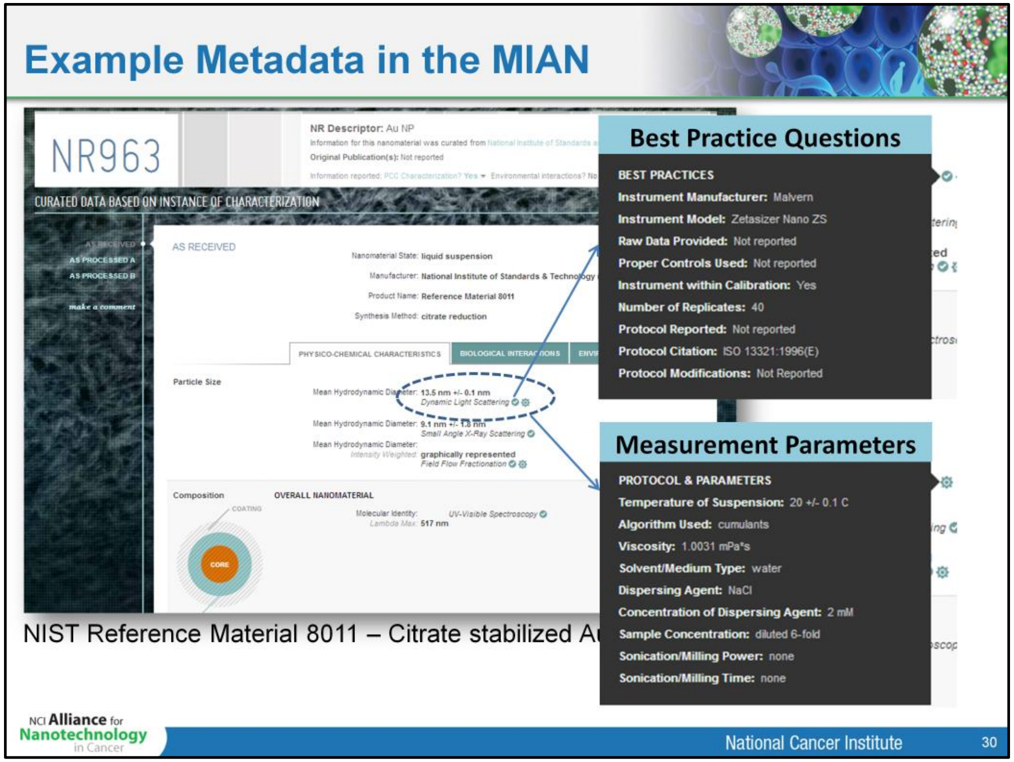
Brief history, nanomaterial registry started in 2011 by a joint effort from NIH Institutes. It aggregates data from existing publicly available databases, including caNanoLab, patent literature, and manufacture catalogs, based on Minimal Information About Nanomaterials, also called MIAN, to simplify the complexity of handling the variety of nanomaterial data. As of January of this year, there are more than 2000 nanomaterials that have been curated by the Registry.



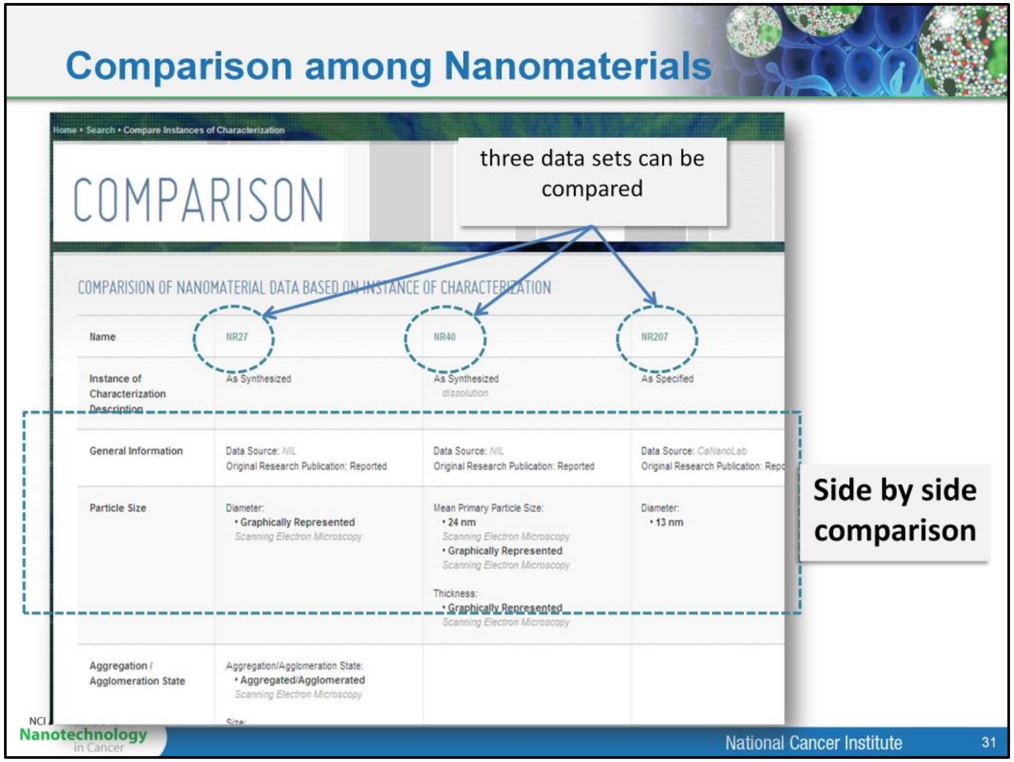
>> Christina Liu: Currently, Nanomaterial Registry provides curated data in three categories: Physico-Chemical Characteristics, Biological Interactions, and Environmental Interactions. Based on the MIAN, the physicochemical characterization (PCC) data and metadata captures not only characteristic data, but also the protocols and parameters associated with each measurement of those characteristics and general information about the nanomaterials, such as production or synthesis technique, product name, lot number, and manufacturer. The curation process is rather standard where the curators identify, collect information, and through QA and QC processes, the data is then published.



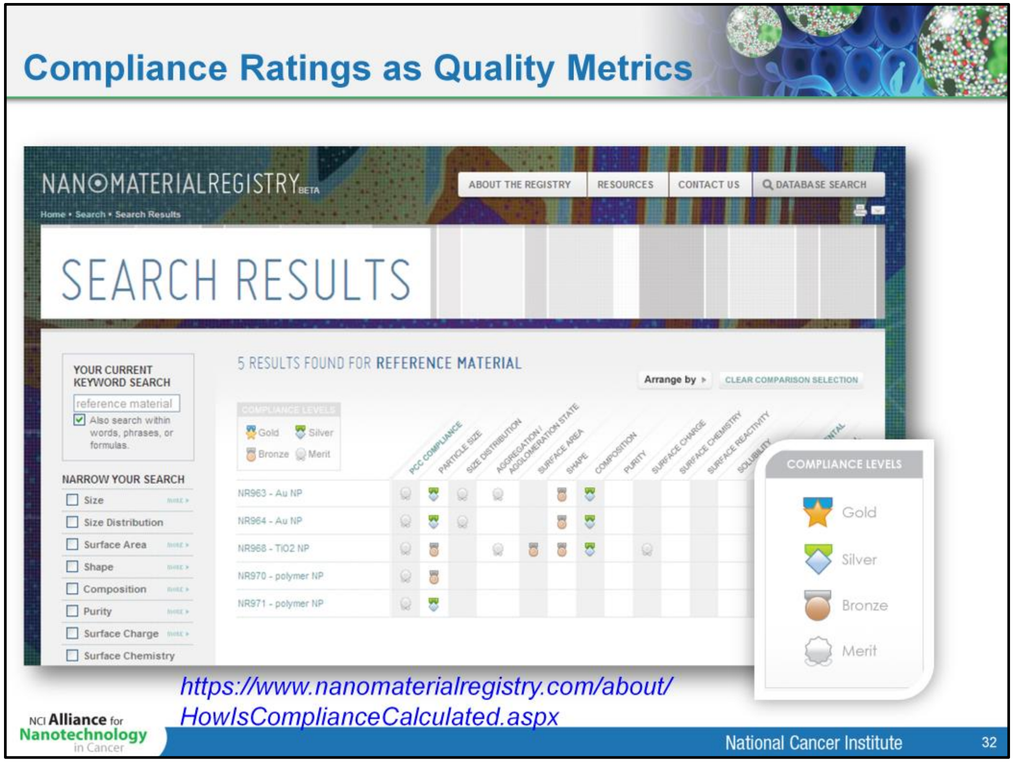
>> Christina Liu: Here is a screen shot of the search and browse utilities of Registry. You can type in as much information as you know into the search window or browse from the pull-down menu.



>> Christina Liu: This slide shows the metadata you will get from the nanoparticle you found: particle size, diagram, composition, etc. The information presented is based on MIAN and you will be able to find a list of information on best practice questions and measurement parameters.

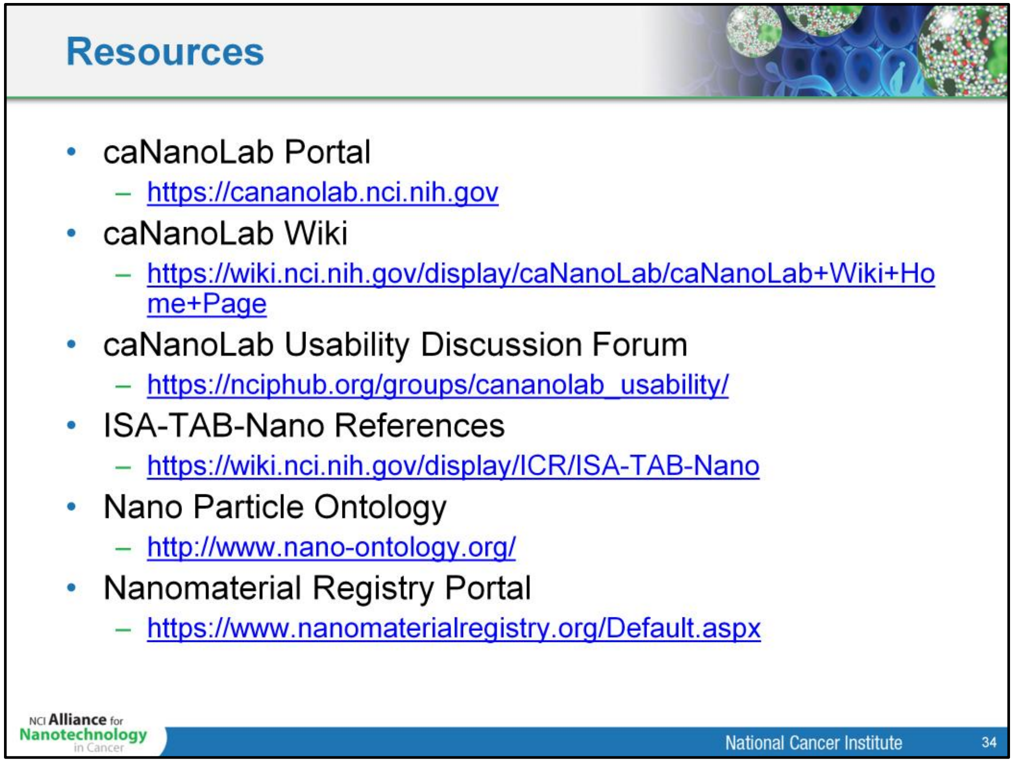


>> Christina Liu: You can also perform comparison of up to three nanomaterials. A simple rule-based matching system was established to query all nanomaterial entries that are at least 10% similar to the original nanomaterial entry. You can also choose what information to be included for side-by-side comparisons.



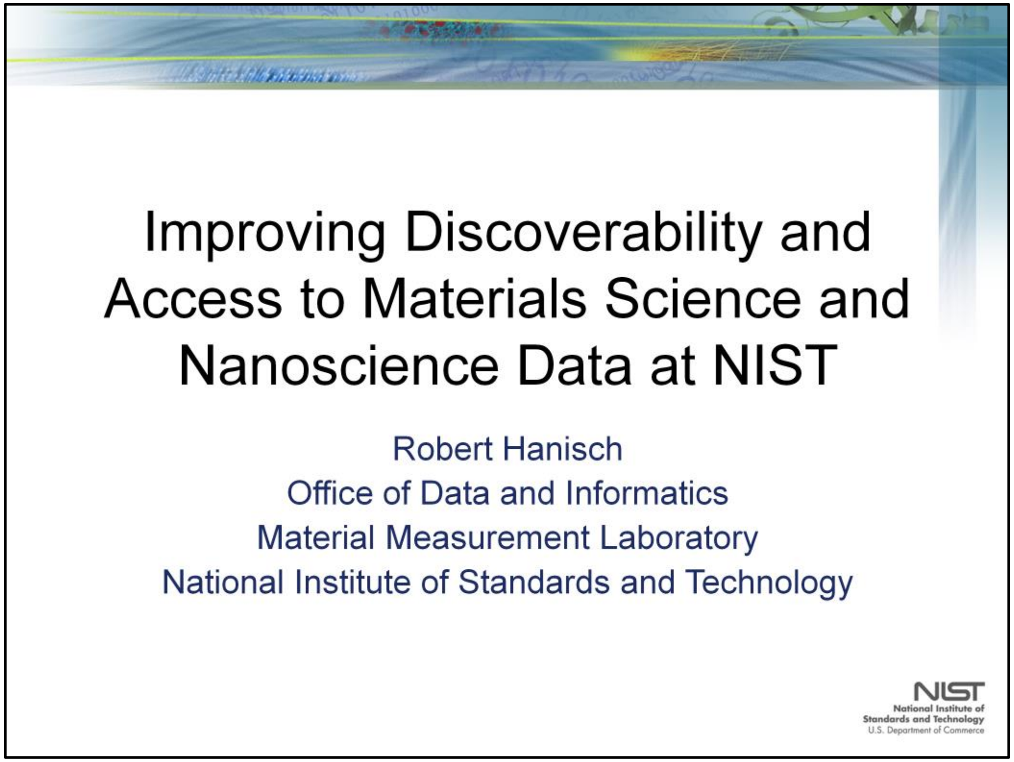
>> Christina Liu: In order to improve the overall quality of information about nanomaterials in diverse studies, the compliance level feature of the Registry provides a metric on the quality and quantity of characterization for each nanomaterial entry, based on MIAN and certain equations. Registry compliance medals are given in decreasing order of compliance score range: gold, silver, bronze, and merit.

>> Christina Liu: Finally, this is very relevant to the NIH grant applicants, there is a section in the NIH application that requires the inclusion of a data sharing plan. Nanomaterial Registry can provide an appropriate data sharing plan for the nanomaterials proposed in an application. It even has a pull-down window listing all NIH grant mechanisms so the user can find the right language to use for their NIH applications.



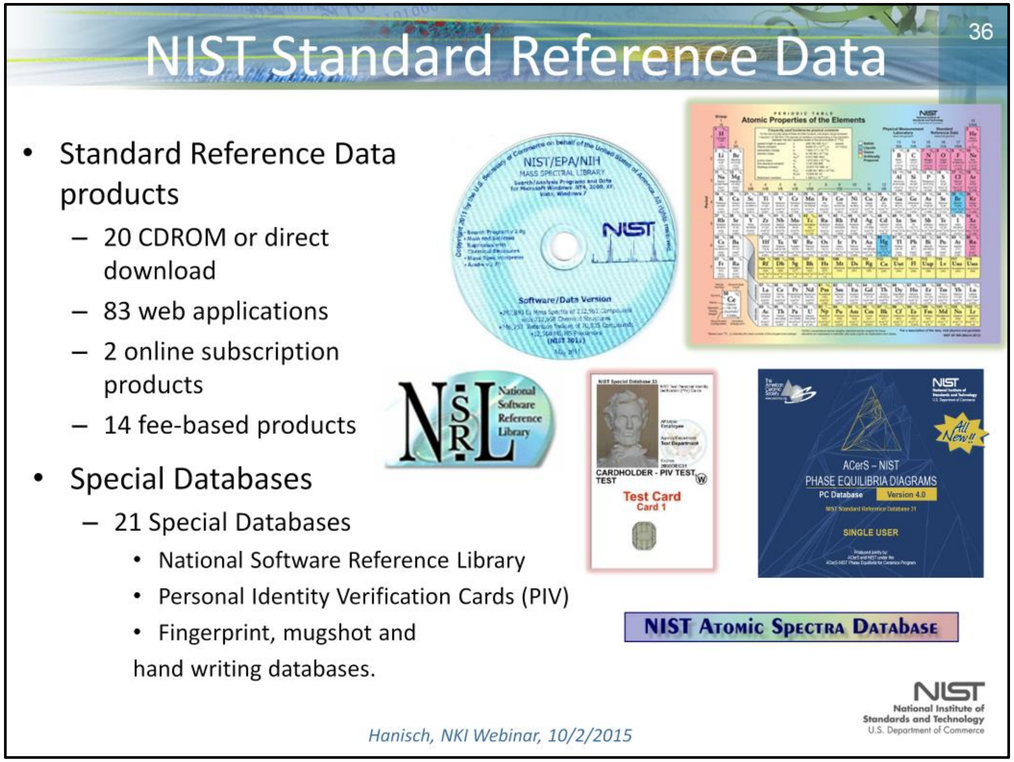
>> Christina Liu: My presentation was intended to be an overview of the utilities of these databases for you as “customers” to give you a sense of what they can do for you. The links here will give you additional information on other databases’ capabilities. This concludes my presentation and I appreciate your attention.

>> Stephen Lehrman: Thank you, Christina. Clearly, there is already a wealth of information and data available on nanomaterials that users can access.

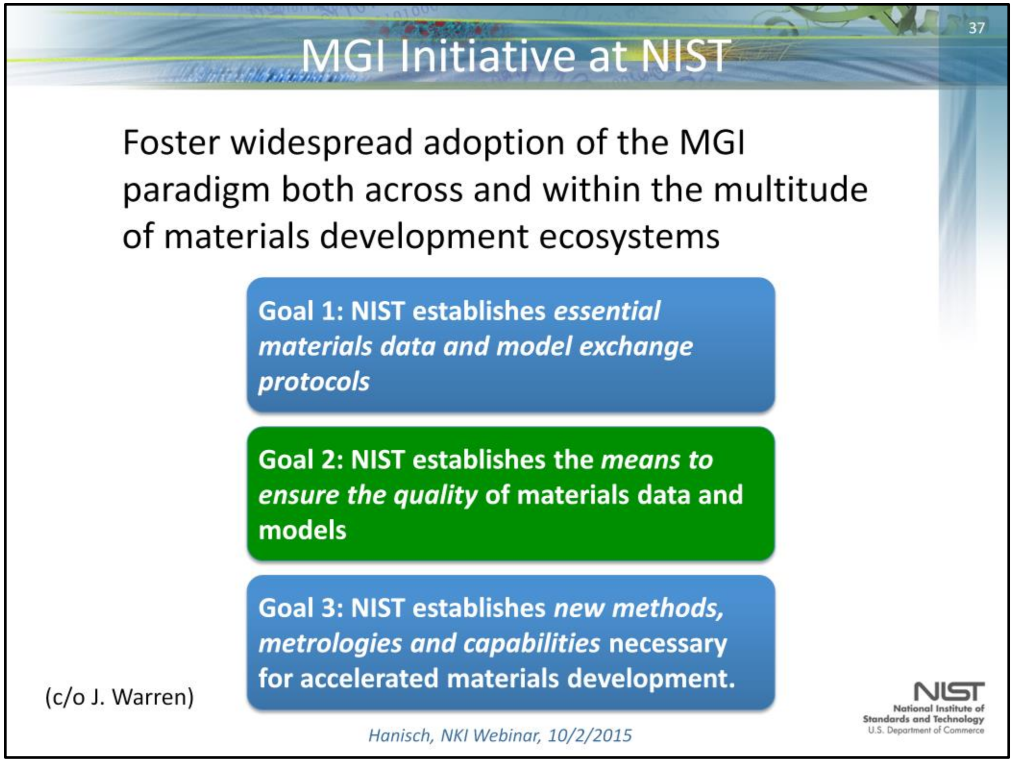


>> Stephen Lehrman: Our third speaker today is Dr. Robert Hanisch from the National Institute of Standards and Technology. And he's going to talk about the Materials Genome Initiative, as well as some other things.

>> Robert Hanisch: So I would like to give you a quick overview of some of the activities we are engaging in at NIST, particularly the Material Measurement Laboratory in my Office of Data and Informatics, around materials data and nanoscience data as well.



>> Robert Hanisch: I first need to point out and remind people that NIST is the host of standard reference data products. We have over a hundred databases, many of them online and freely available. Several of them are fee-based according to the Standard Reference Data Act. But these data are already available, many of them pertinent to areas of interest of this community. And we are right now working on a project to modernize access to these databases through newly developed application programming interfaces and updated websites.



>> Robert Hanisch: First, to talk about the Materials Genome Initiative a bit, my colleague Jim Warren at NIST. The MGI activities at NIST focus on establishing the essential materials data and models that are needed to support the MGI program, also establishing the means to ensure the quality of the materials data and models. And then we're also working on new methods, metrologies, and capabilities to help accelerate materials development.

MGI: Data Discovery and Interoperability

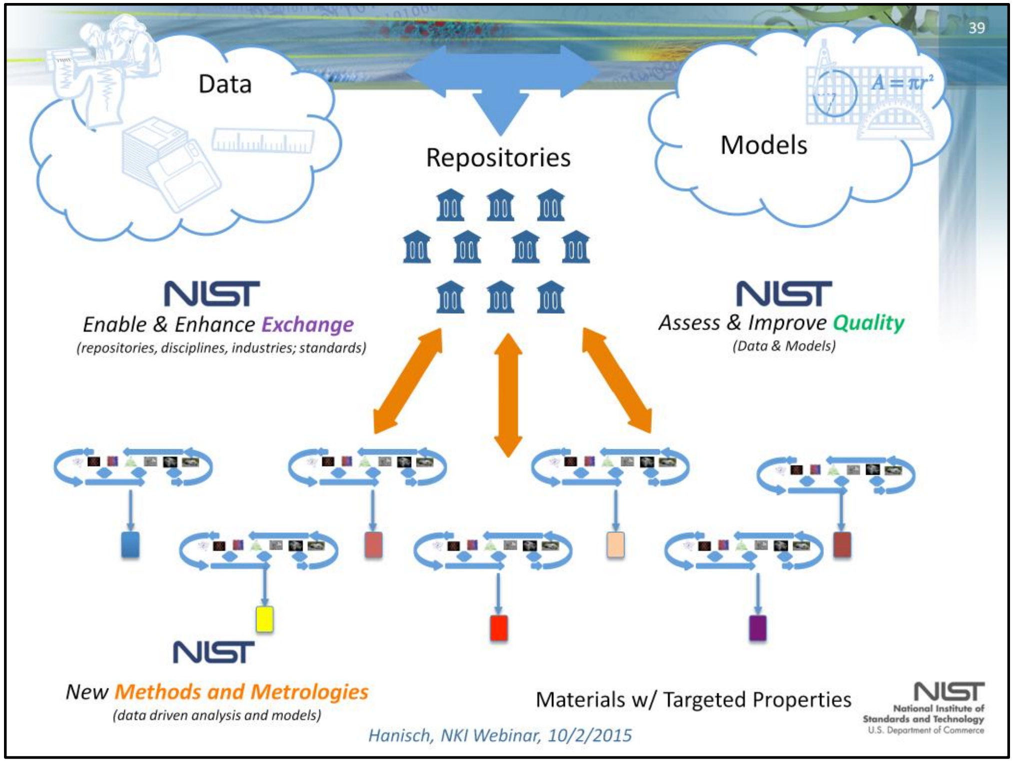
38

- Develop and deploy repositories
- Develop and disseminate materials informatics infrastructure
  - Enable data discovery through tools and standards
  - Capture data from scientific workflows and archival sources
  - Engage with stakeholders to determine needs and disseminate best practices
- Integrate across length and time scale
- Build and test infrastructure through pilots

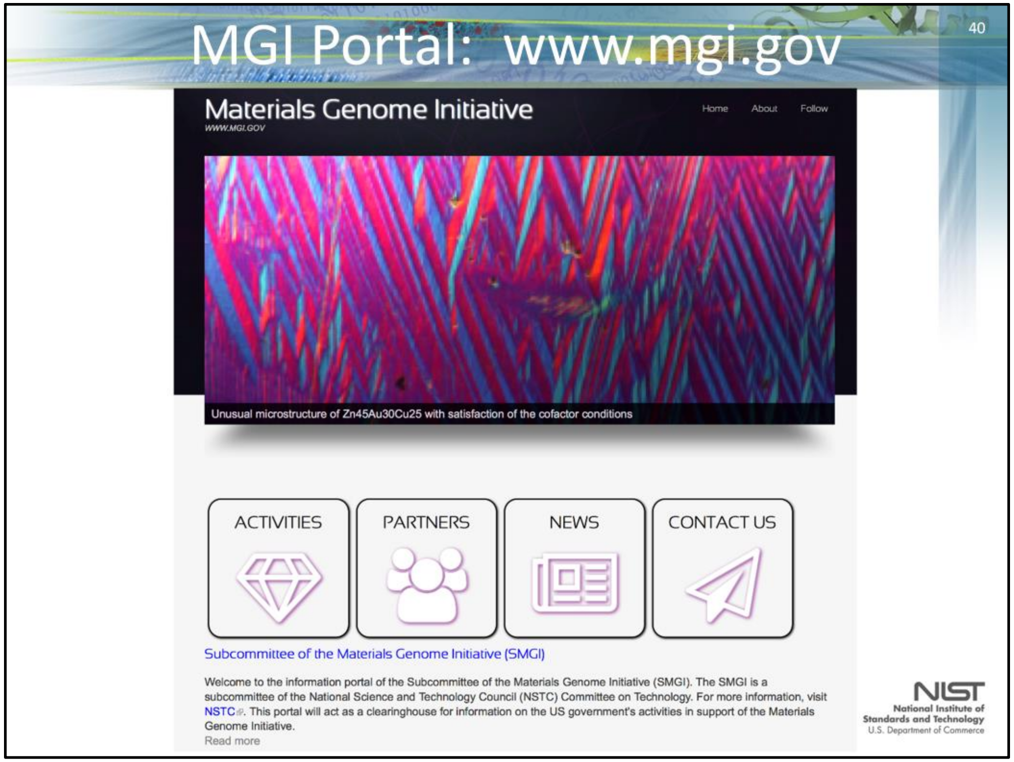
NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Hanisch, NKI Webinar, 10/2/2015

>> Robert Hanisch: We are also focusing on data discovery and interoperability by developing and deploying repositories -- I will show you an example of that in a moment; developing and disseminating materials informatics infrastructure that makes it much easier to annotate data, associated workflows, and to work with the stakeholders to determine their needs and to disseminate best practices.



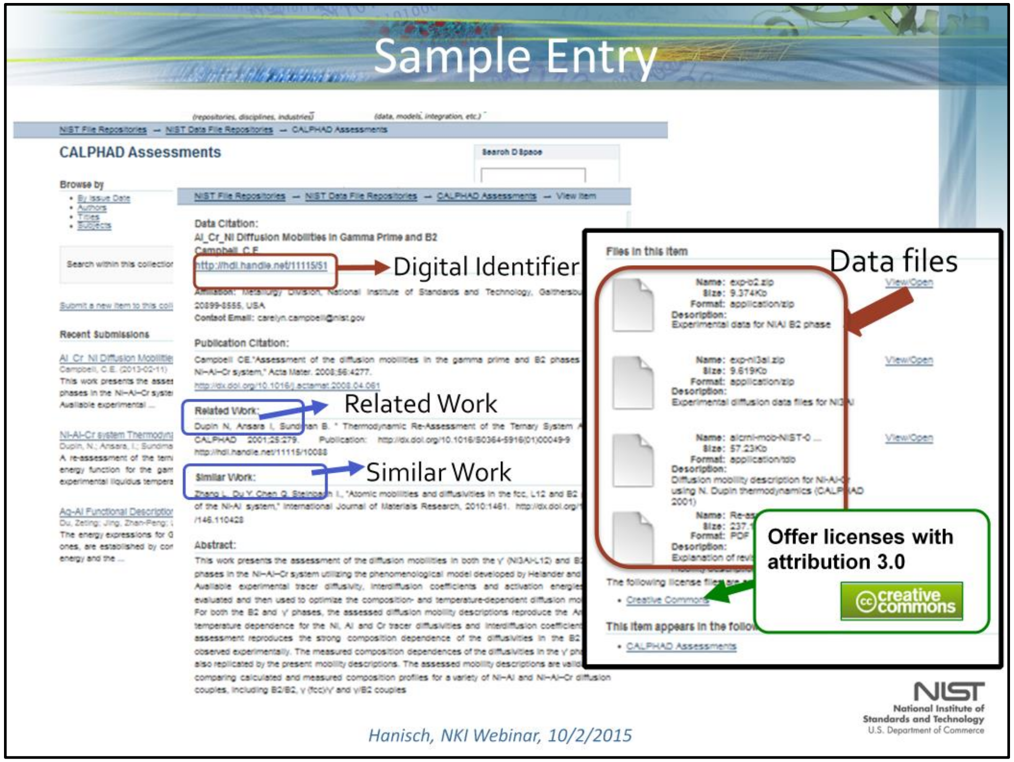
>> Robert Hanisch: Essentially, we're envisioning a distributed but federated system of data repositories, models, and data flows, processing flows that are centered on particular materials with targeted properties. These can work all together seamlessly in a federated system.



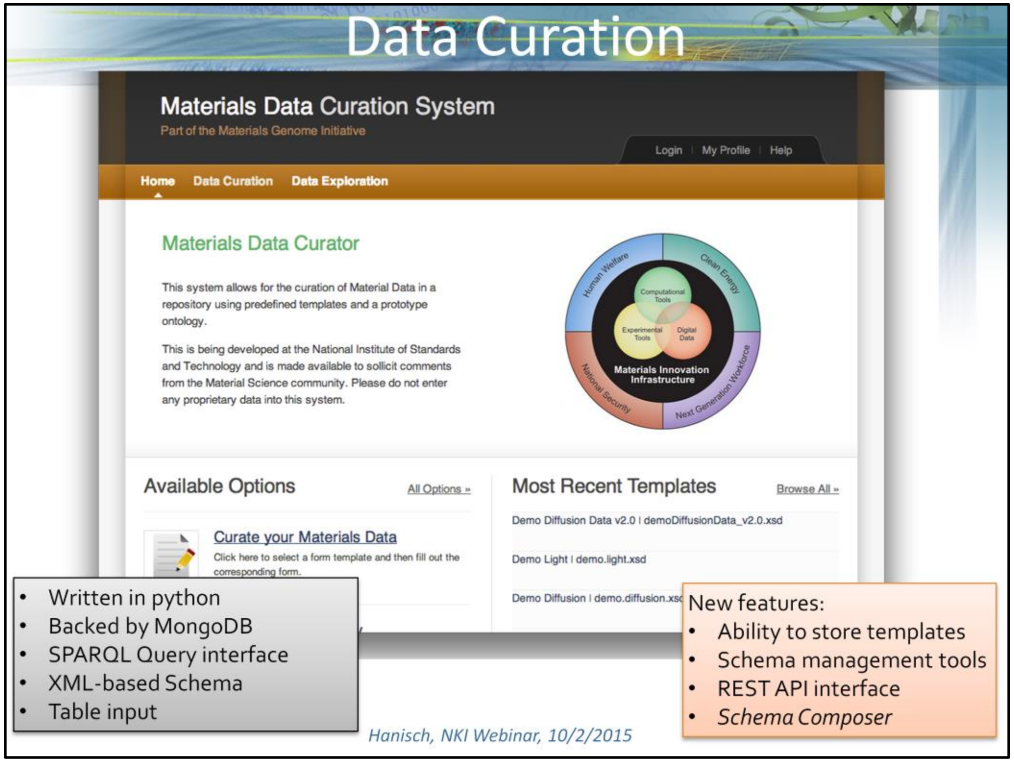
>> Robert Hanisch: We've developed the national portal for the Materials Genome Initiative at mgi.gov. This site is maintained and populated by staff at NIST in my group and in close collaboration with Jim Warren and the MGI program overall.

>> Robert Hanisch: We are hosting a repository for materials data at NIST using the DSpace technology. This repository is there to facilitate collaboration in various research communities in materials science and nanoscience.

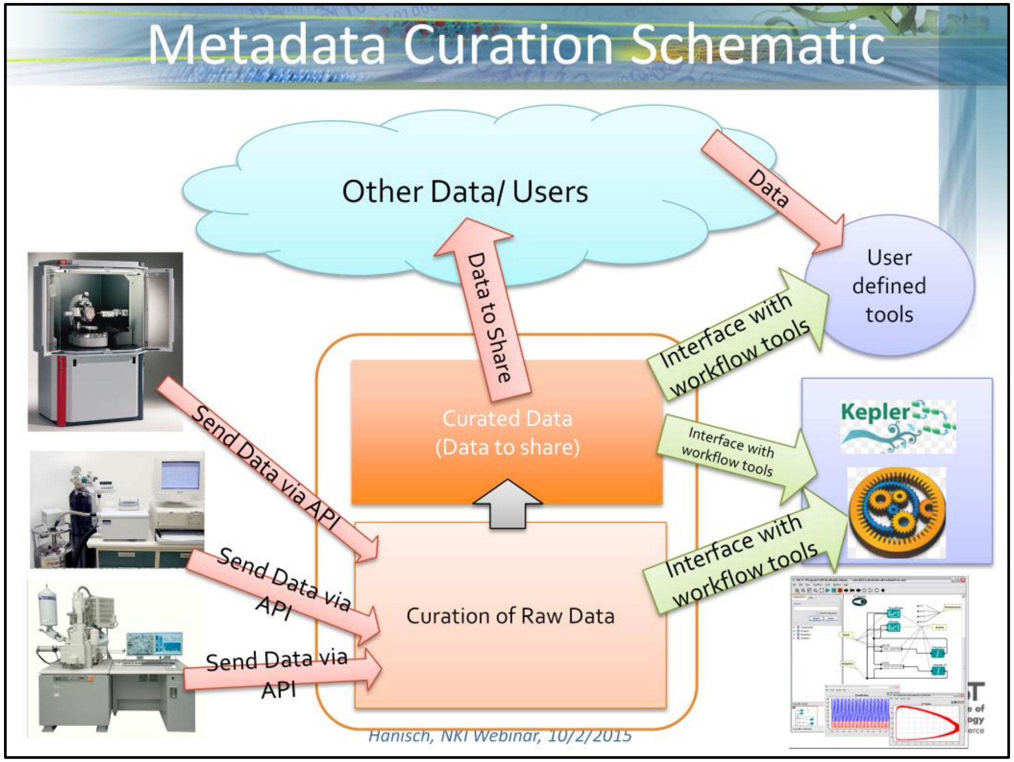
>> Robert Hanisch: There are now 16 so-called communities, collaborations that have data in the materials data repository at NIST. The communities can control what level of public access they wish to have for their data. So things that are very early in their development, close collaboration, can be shared just amongst that research team. When data products have been developed that are ready for dissemination, they can be recategorized and be made fully open to the public.



>> Robert Hanisch: A typical entry here is for these phase diagrams where you can have a description of the data. You get automatically a persistent identifier to the data collection. You get links to related work and similar work, and here is the actual list of data files that are in this collection. You can also provide licenses that describe what type of access and reuse is permitted.

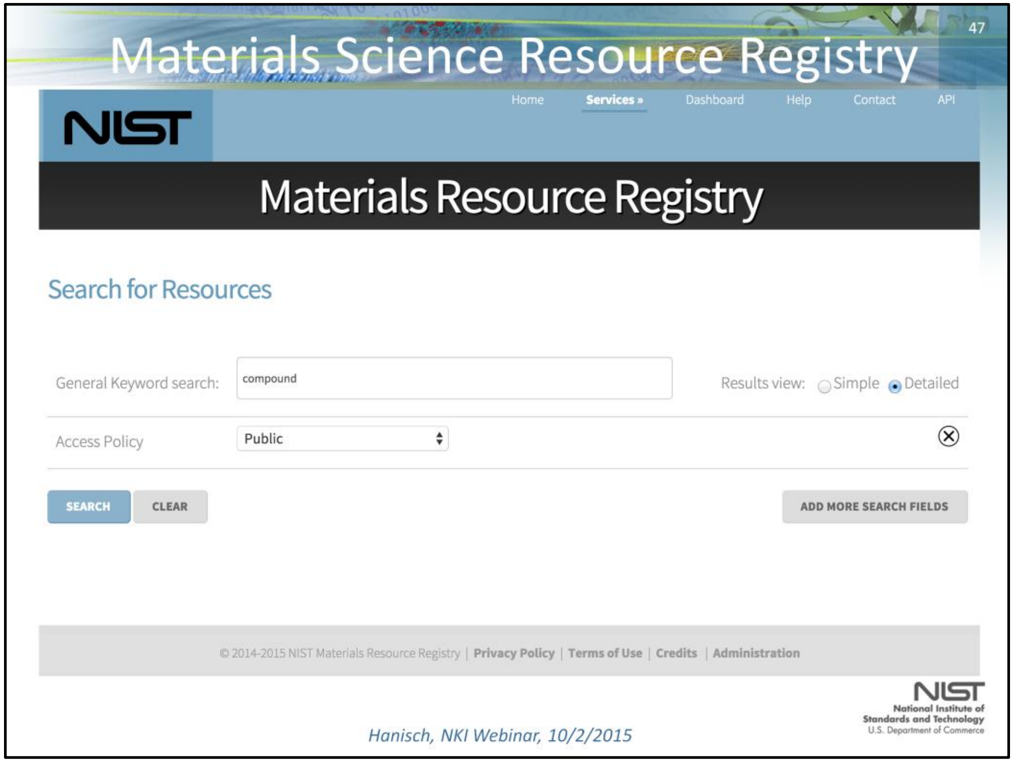


>> Robert Hanisch: At a much lower level, we are developing a materials data curation system that allows you to develop the metadata and schema to describe particular types of data and datasets. Whereas the materials repository is much more broad-reaching, the materials data curation system focuses on describing particular types of data and relationships. This is written in Python, backed by Mongo DB. It has a SPARQL Query interface, and the schema are stored in XML. So you can develop templates that can be reused by other people. There are schema management tools and a REST application program interface.



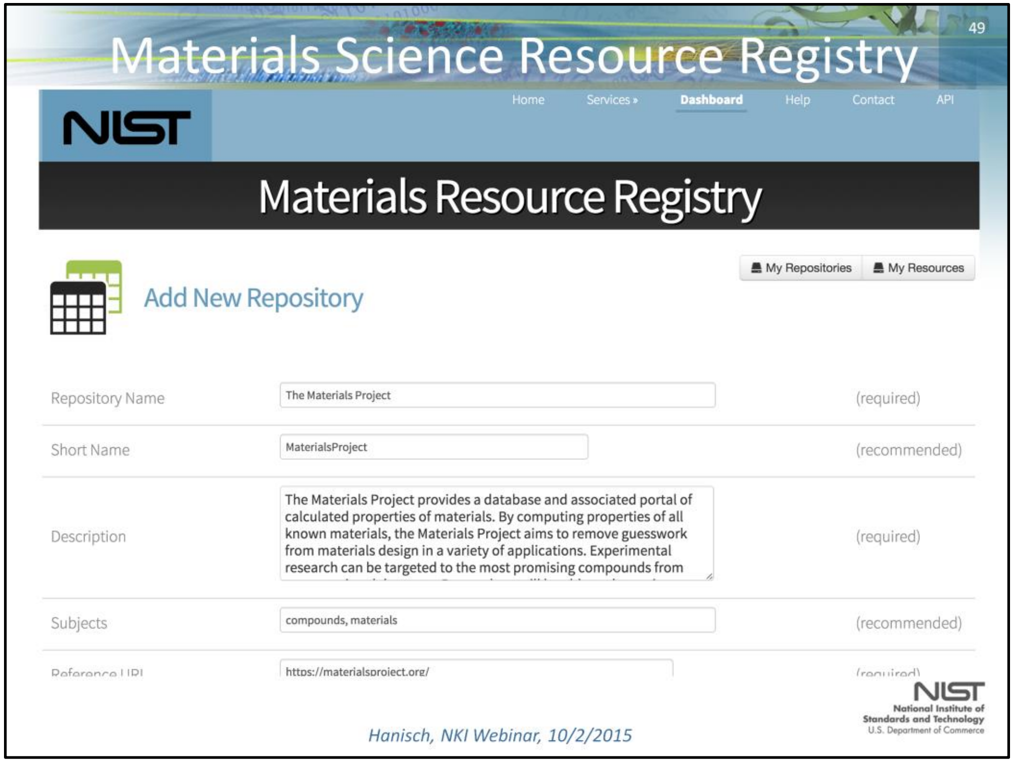
>> Robert Hanisch: Basically, this allows you to aggregate or collect data from instrumentation or from simulations. You can decide what to share, and you can interface with various workflow tools.

>> Robert Hanisch: We're also working in the context of the Research Data Alliance where we have a RDA/CODATA Materials Data, Infrastructure, and Interoperability Interest Group. If you are not familiar with RDA, I suggest you have a look at the rd-alliance.org website. There are now over 3,000 people participating in the Research Data Alliance from, I think, about 60 different countries. And the RDA held its sixth plenary meeting just last week in Paris bringing together over 600 scientists and investigators in all aspects of research from physics and materials science to anthropology and linguistics. It covers the entire gamut.

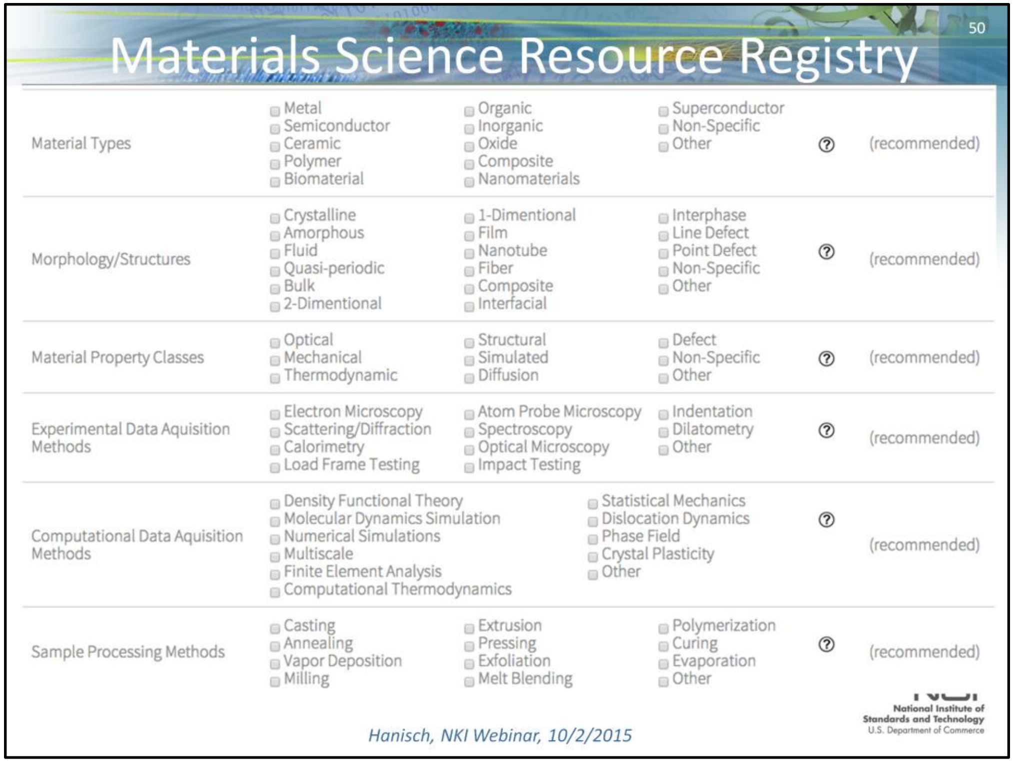


>> Robert Hanisch: But within this context, we are developing a Materials Science Resource Registry where we will collect information, metadata about data collections that span the entire spectrum of materials science and nanoscience. This is right now a pilot project because we feel it's important to work in the international context to agree on the types of metadata that will be tracked in this system.

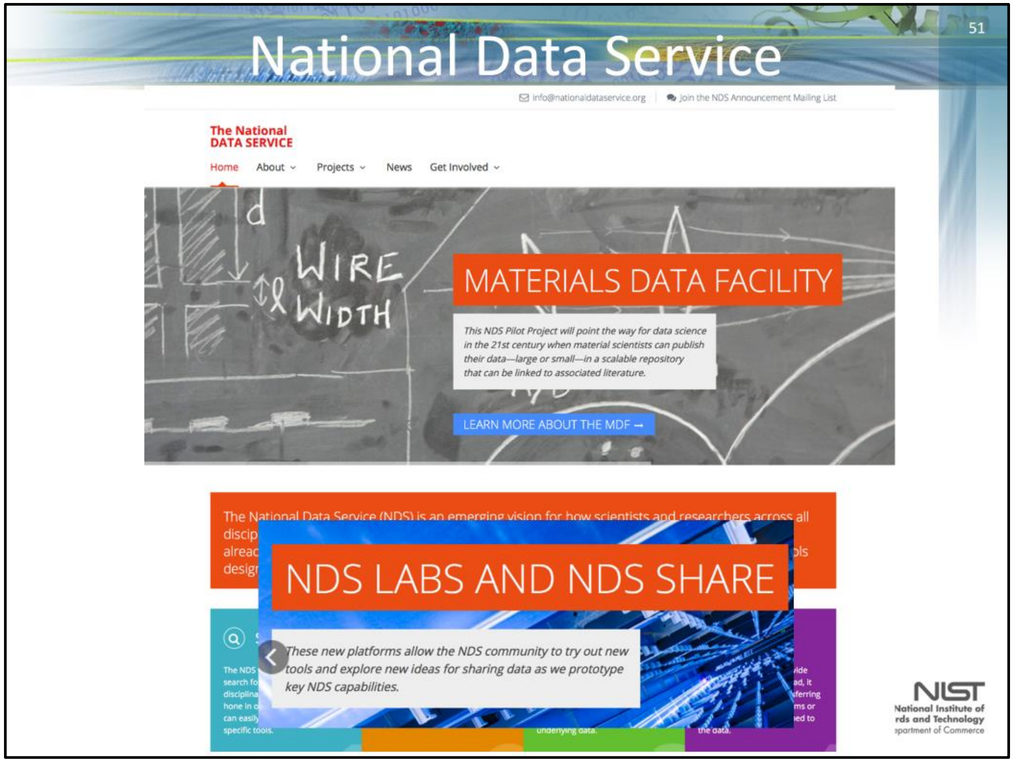
>> Robert Hanisch: We will be tracking organizations, data collections, datasets, data services, and even software related to materials and nanoscience research.



>> Robert Hanisch: You can dynamically add new collections to this system so that as you have a new data service or data collection, you can add it to the facility so others can easily discover it.

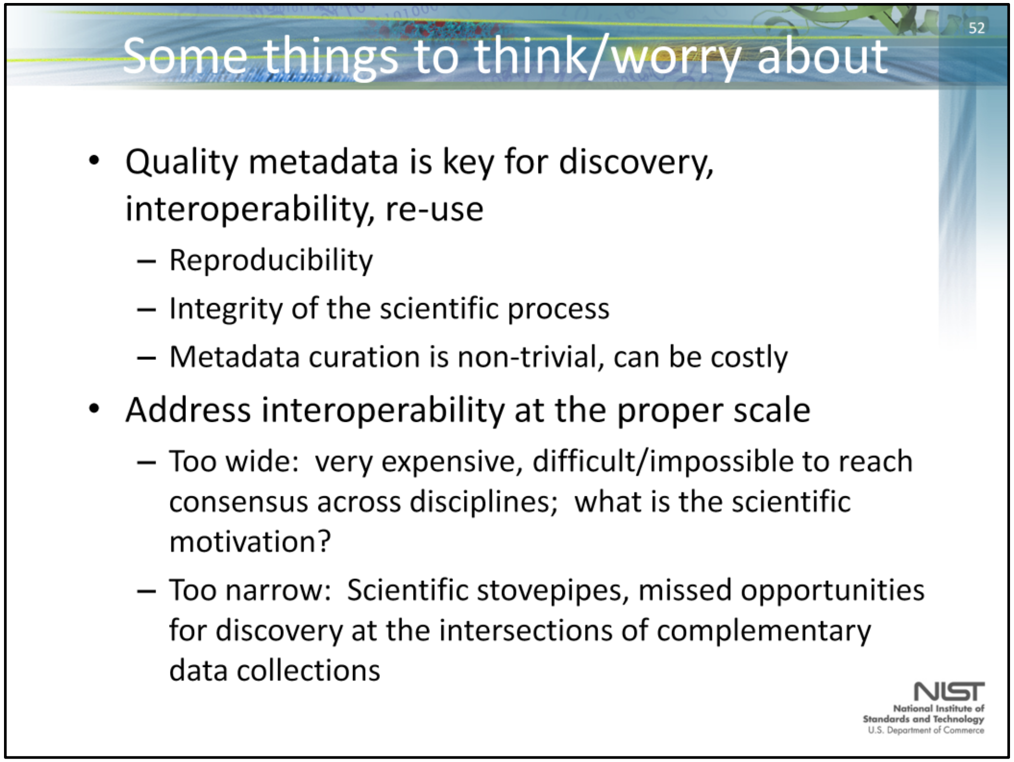


>> Robert Hanisch: This slide shows the type of metadata that we are considering. Again, this is a pilot. We will be working through a new RDA working group to reach global consensus on the set of metadata terms that we need in order to enable global discovery.



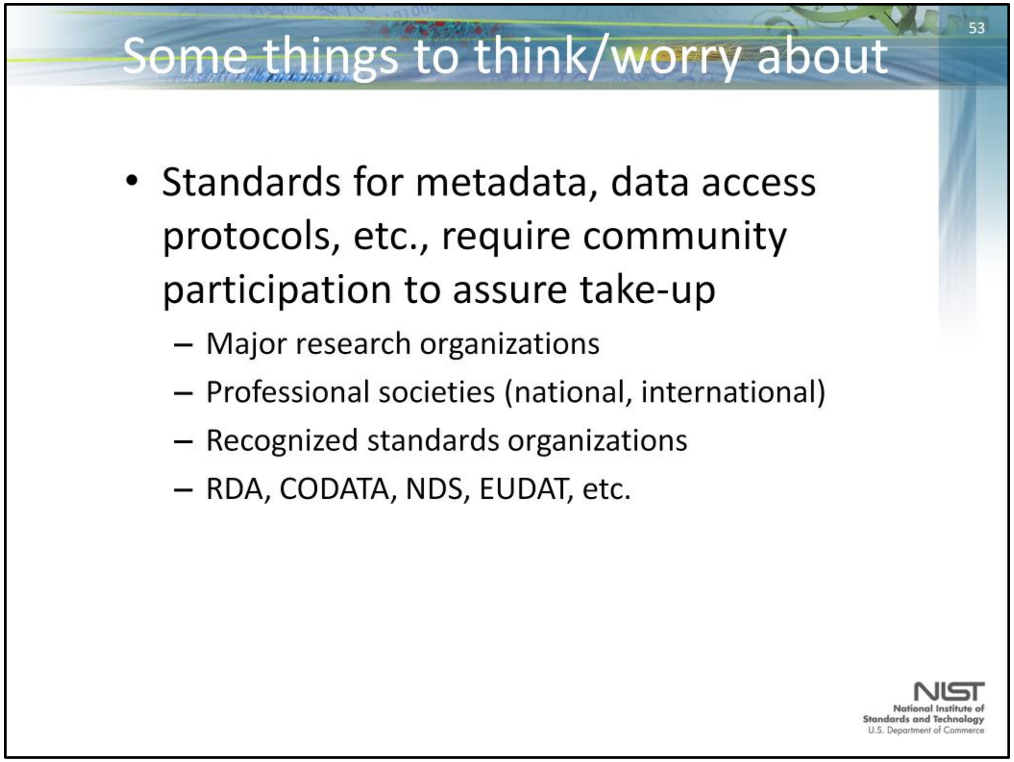
>> Robert Hanisch: Within the U.S., we are also working with the National Data Service, with funding from NIST. We will be setting up a materials data facility that will be layered on top of hardware services hosted at the National Center for Supercomputing Applications in Champaign-Urbana. There is also a generic facility within the National Data Service called NDS Labs and NDS Share, which are platforms for deploying applications and for sharing data using the computational infrastructure available at some of the national supercomputing facilities.

I should say also that there is a plenary meeting of the National Data Service Consortium coming up later in October at the supercomputer center in San Diego. Registration is still open. If you go to the National Data Service website, you can find instructions there on how to register.

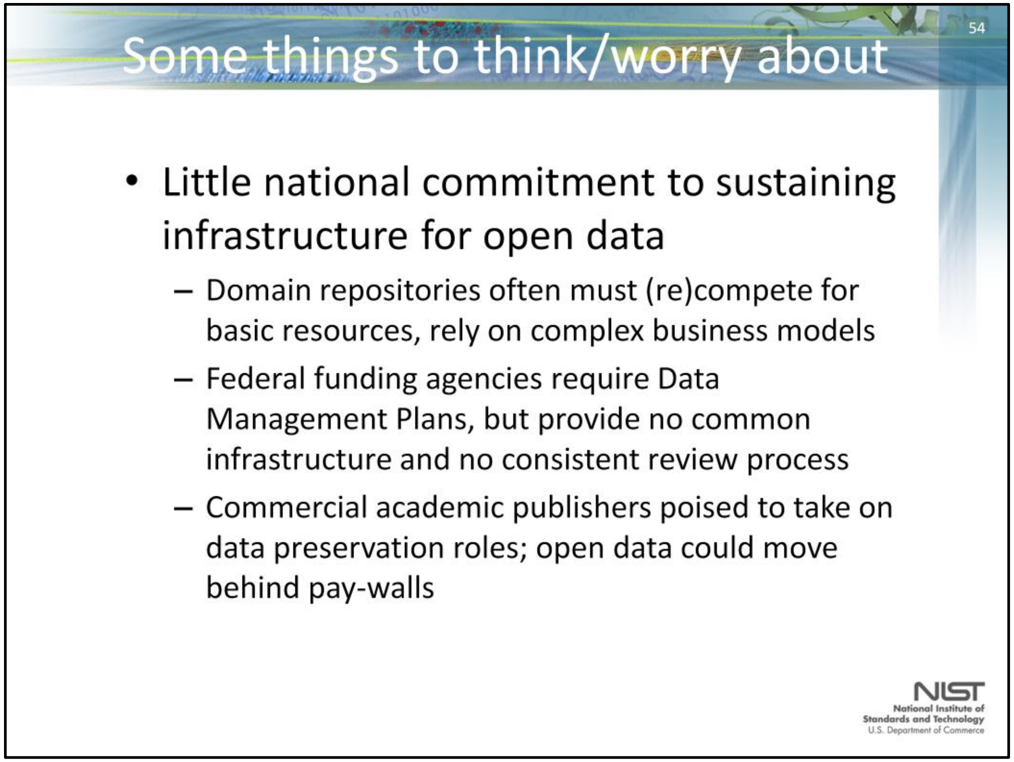


>> **Robert Hanisch:** I just want to close with some things that I think about and I worry about in terms of data dissemination and discovery. Quality metadata is key for discovery, interoperability, and reuse. This is key for reproducibility; it is part of ensuring the integrity of the scientific process. But metadata curation is non-trivial and can be costly. A colleague of mine, Carl Lagoze now at the University of Michigan, told me many years ago, you can never overestimate the amount of effort it takes to do quality curation. It is important, but it does cost human resources.

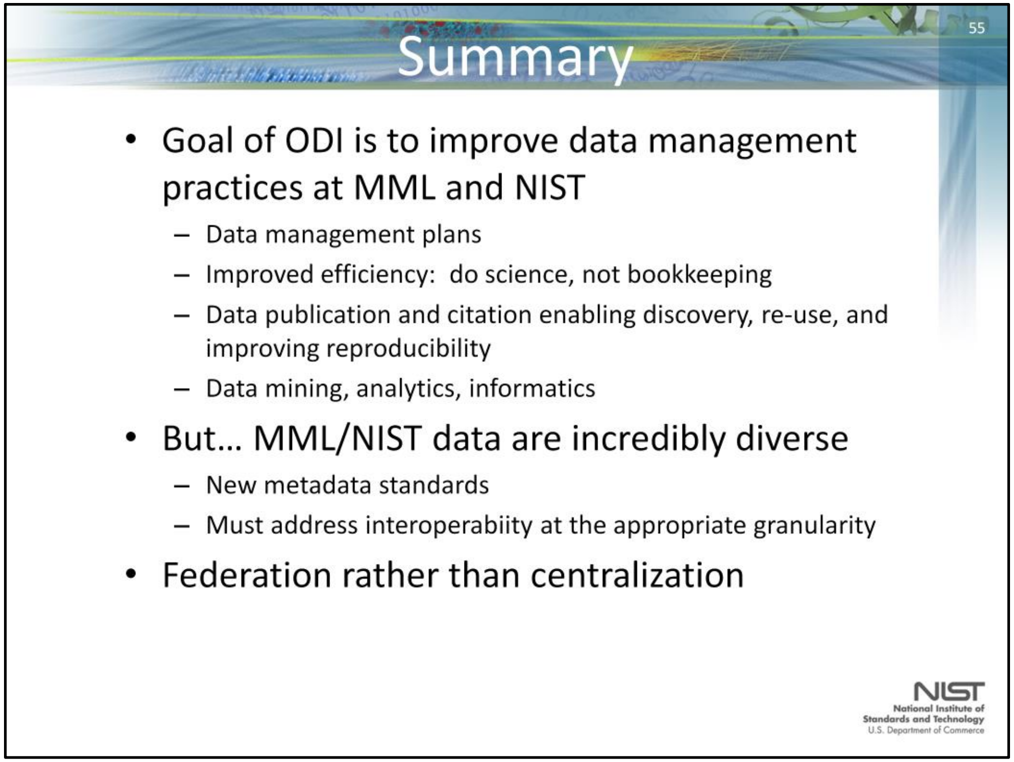
A corollary to that is I think we have to address interoperability at the proper scale. If you try to cast a net too wide, it will be expensive; difficult, if not impossible, to reach consensus across disciplines; and it's not clear what the scientific motivation would be. On the other hand, if you are too narrow, you can have scientific stovepipes, missed opportunities for discovery at the intersections of complementary data collections. They always say the most productive part of an ecosystem is at the boundary between zones, between forest and field. And this is exactly the sort of thing that we want to foster.



>> Robert Hanisch: Standards for metadata and data access protocols require community participation to assure take up. There's a “not invented here” syndrome that pervades. So you really have to work across communities, including major research organizations, professional societies, and recognized standards organizations.



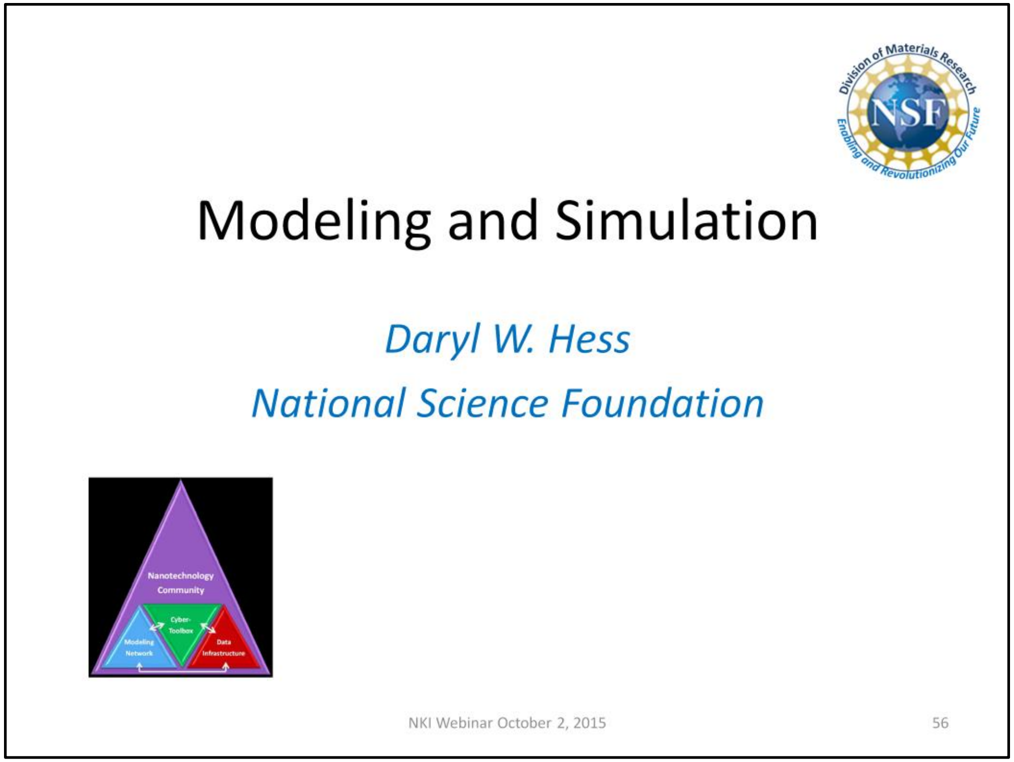
>> Robert Hanisch: I worry, too, that our national commitment to sustaining infrastructure for open data is lacking. Domain repositories often have to compete or recompete for basic resources. Federal funding agencies require data management plans. But there's not a common or consistent infrastructure to support the data that are produced. And commercial academic publishers are poised to take on data preservation roles. This could be a positive thing, but it also could result in open data moving behind paywalls. We have to approach this interaction with the publishing community with some caution.



>> Robert Hanisch: So to close, my Office of Data and Informatics is working to improve data management practices within our laboratory and generally at NIST, but our data are incredibly diverse. Just as we see the global community of data and materials and nanoscience is incredibly diverse. We need metadata standards, and we need to address interoperability at the appropriate level of granularity. I always talk about federation rather than centralization. It's impossible to hold all of the resources in one place. And as data grows, more and more federation is the way to go.

Thank you very much.

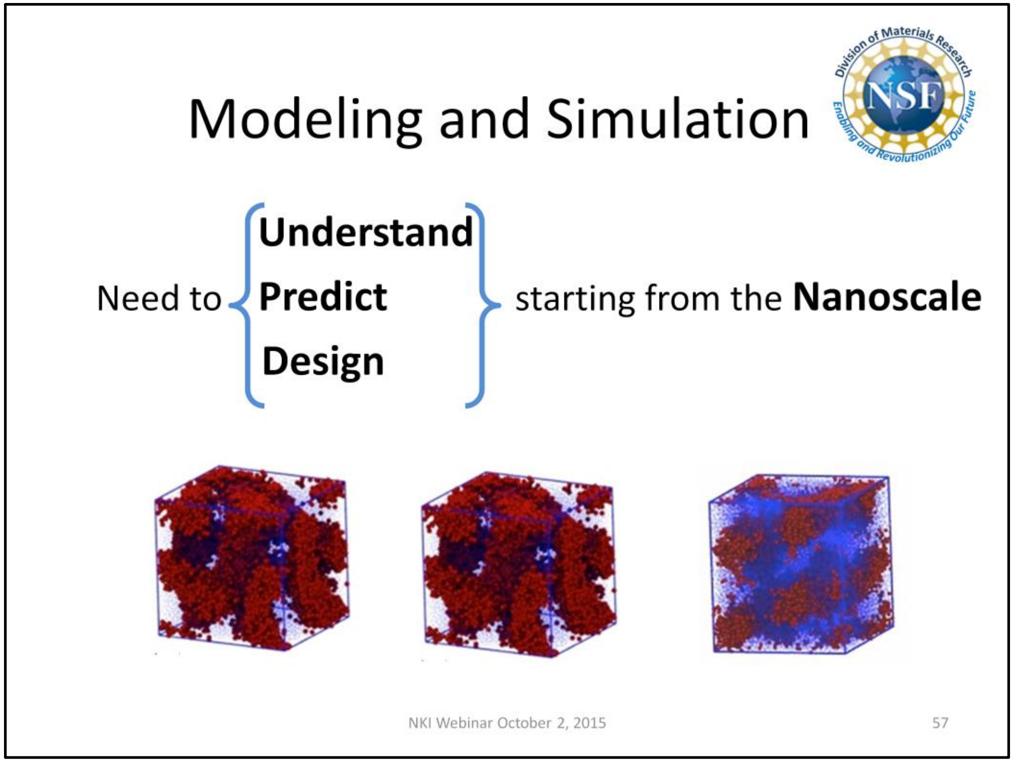
>> Stephen Lehrman: Thank you, Robert. Certainly a lot of activity going on at NIST and in the international community.



>> Stephen Lehrman: Our final speaker today is Dr. Daryl Hess from the National Science Foundation. And he's going to talk about some of the modeling activities that are going on.

>> Daryl Hess: Thank you. I will say a few words about modeling and simulation and the supporting infrastructure -- or more precisely cyber infrastructure -- element of the NKI called the cyber toolbox. The cyber toolbox contains parts of other elements, including data and education on how to use the existing NKI cyber infrastructure, as well as training those who will build future cyber infrastructure.

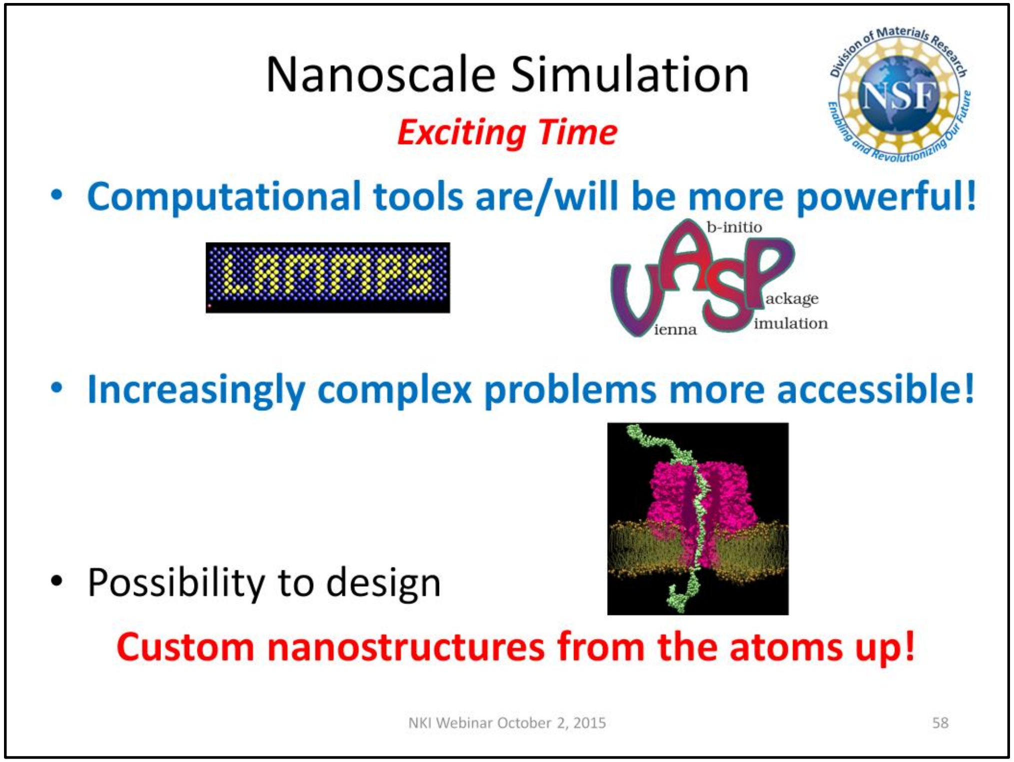
The perspective is perhaps more physical than the preceding presentations, but hopefully it will be apparent that modeling and simulation crosscut the inherently interdisciplinary field of nanoscience and nanotechnology.



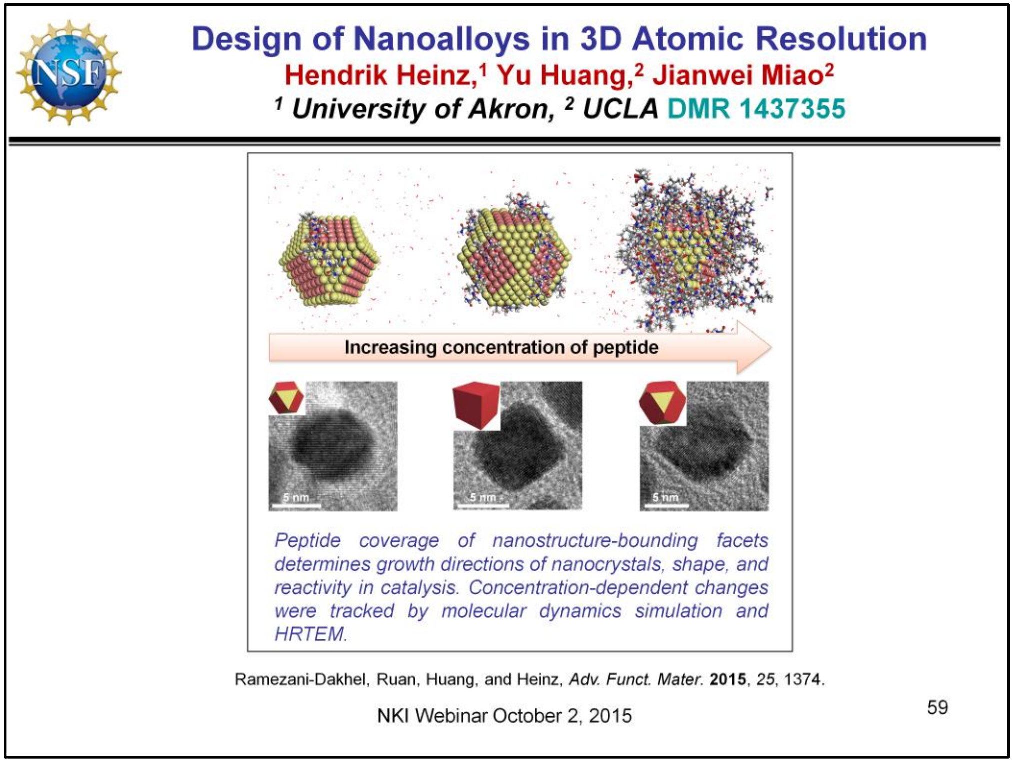
>> Daryl Hess: Modeling and simulation play an important role in advancing science and technology. Through utilization of computation, our conception of how the world works down to the scale of atoms can be brought to bear to advance our understanding and to discover and exploit new phenomena at the nanoscale. Modeling can be used to make quantitative predictions, to stimulate and advance technologies that begin at the nanoscale, such as new electronic devices or predictions that relate to the potential toxicity of nanoparticles. Reliable modeling techniques enable the design of new nanostructures that form the building blocks of new materials with desired properties or new electronic devices, as well as the design of experiments and protocols.

Models themselves play an important role. Among these roles is to bridge gaps in fundamental theories, so that progress can be made. Shown here are course-grained molecular dynamics simulations that explore the properties of tapered block copolymers in which a polymer that is soft and facilitates ion transport is bonded to another polymer that provides mechanical strength to form the di-block copolymer by a smoothly varying composition-changing region.

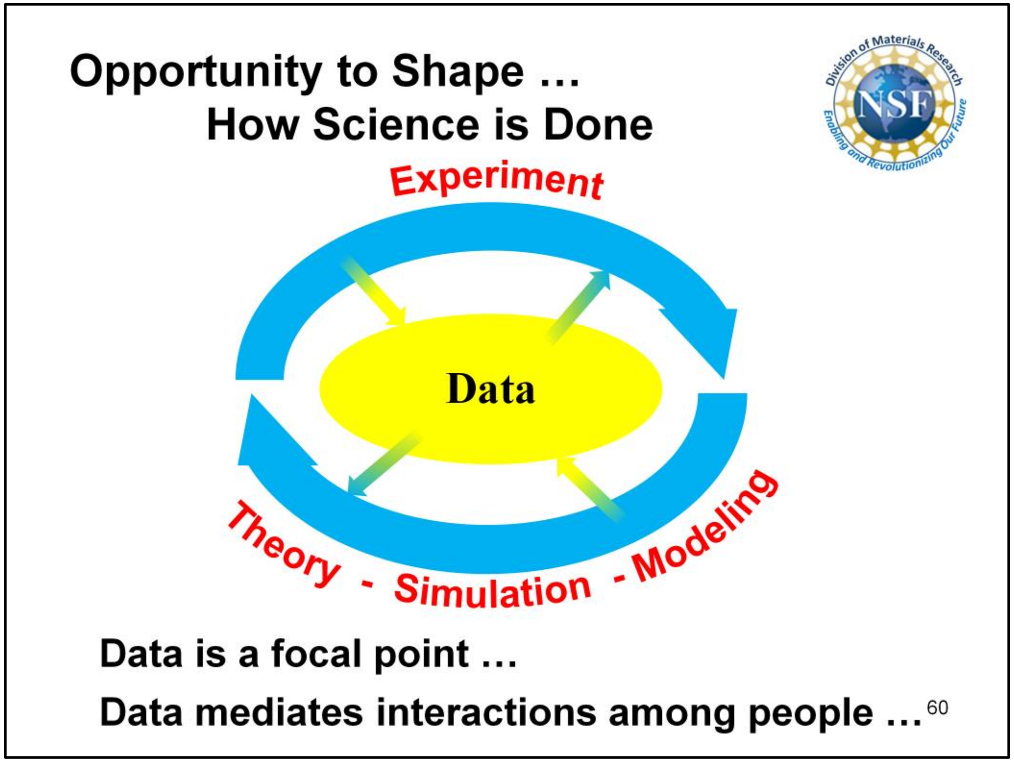
Adjusting the taper allows one to control the structure and dynamical properties of the resulting polymeric material.



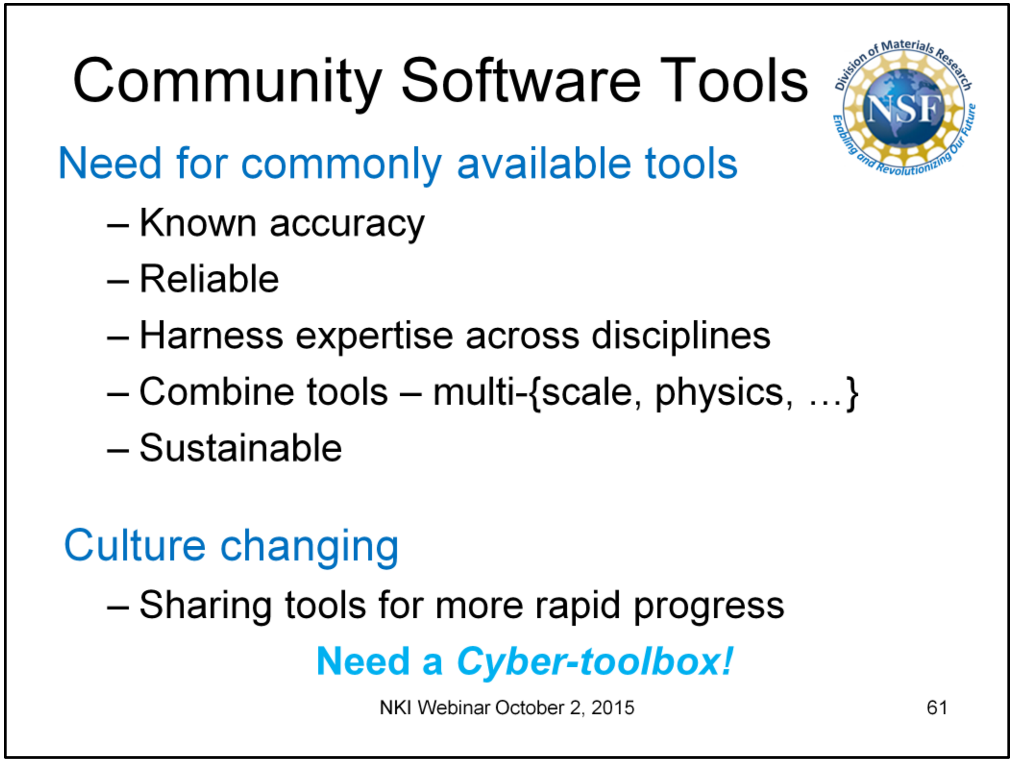
>> Daryl Hess: This is an exciting time. Good software is required in order to perform simulations that yield meaningful results. In the current time, powerful computational tools, such as the molecular dynamics program LAMMPS -- developed at Sandia National Laboratory -- and the density functional theory-based code VASP, enable discovery and innovation starting at the nanoscale. But this is only the beginning. Tackling real-world problems with inherently high complexity requires even more sophisticated computational tools and the ability to combine them in flexible ways. While the translocation of DNA in a nanopore, with complexities introduced by the species of the counterions and the chemical structure of the pore, enables the simulation of a possible high-speed DNA sequencer, this is but a glimpse of the possibilities and the frontiers. While the investigator of the previous slide thought to explore properties of polymeric materials from monomer sequences on up, there remains the challenge to build nanostructures and materials from atoms up, in the spirit of the MGI.



>> Daryl Hess: The synergistic interaction of theory, simulation, and experiment, as supported by NSF's Designing Materials to Revolutionize and Engineer our Future, takes a bold step in this direction. This highlight shows an example from recent work. Molecular dynamics simulations working together with experiment provide insight into the mechanism of selective peptide recognition of platinum nanocrystals and peptide-directed growth into specific shapes. The adsorption strength of the peptides depends on the spatial location of the surface and on the peptide concentration in addition to the actual peptide sequence. Quantitative correlations between preferences in facet coverage and binding energies with nanocrystal shape, size, and yield have been elucidated and can be helpful for the rational design of more complex structures.



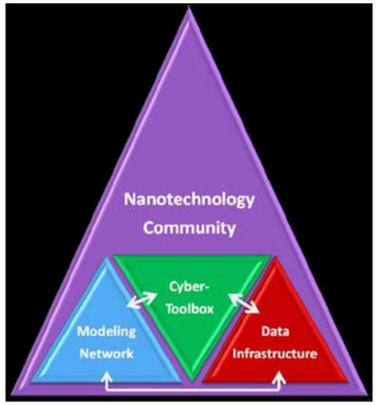
>> Daryl Hess: The previous example highlights how theory, modeling, and simulation can interact in a synergistic way. In the course of doing science, data plays a key role. Experiments create data. Data may stimulate new experiments. Experimental data may stimulate simulation which in turn creates data that can stimulate experiment and further simulation. The simulations themselves may act on the data to help interpret the results of experiments or delve more deeply into the correlations hidden but underlying a group of simulations, as in the modern high throughput simulation strategies. It is important to keep in mind that data is a focal point to scientific progress. It also mediates interactions among people.



>> Daryl Hess: Community software enables simulation of known accuracy and the uniform application of the concepts that form our understanding of the nanoscape and the world around us. Transformation ensues when theoretical and algorithmic innovation can be translated rapidly into software and data that other researchers can access, inspect, verify, and use as a tool for guidance and discovery. This is, in fact, a goal of the software infrastructure for sustained innovation at NSF.

Community software enables researchers to harness the fruits of research in other disciplines and bring them to bear on the inherently interdisciplinary problems that characterize the challenges of our time. Community software must be a sustainable and achievable goal through the partnership of domain science and cyber infrastructure specialists, and propagation to the next generation that arises with unique skills to advance science and responsively develop new community software to engage future challenges.

The Cyber-toolbox



NKI Webinar October 2, 2015 62

>> Daryl Hess: Through the organized sharing of software tools, progress can be made more rapidly. A community cyber toolbox is needed. But a cyber toolbox is not your grandparent's tool box.

The Cyber-toolbox

Vast and "Organic" !

NKI Webinar October 2, 2015 63

>> Daryl Hess: Rather, it is vast and organic like this meadow, and the flowers that are blooming in it.

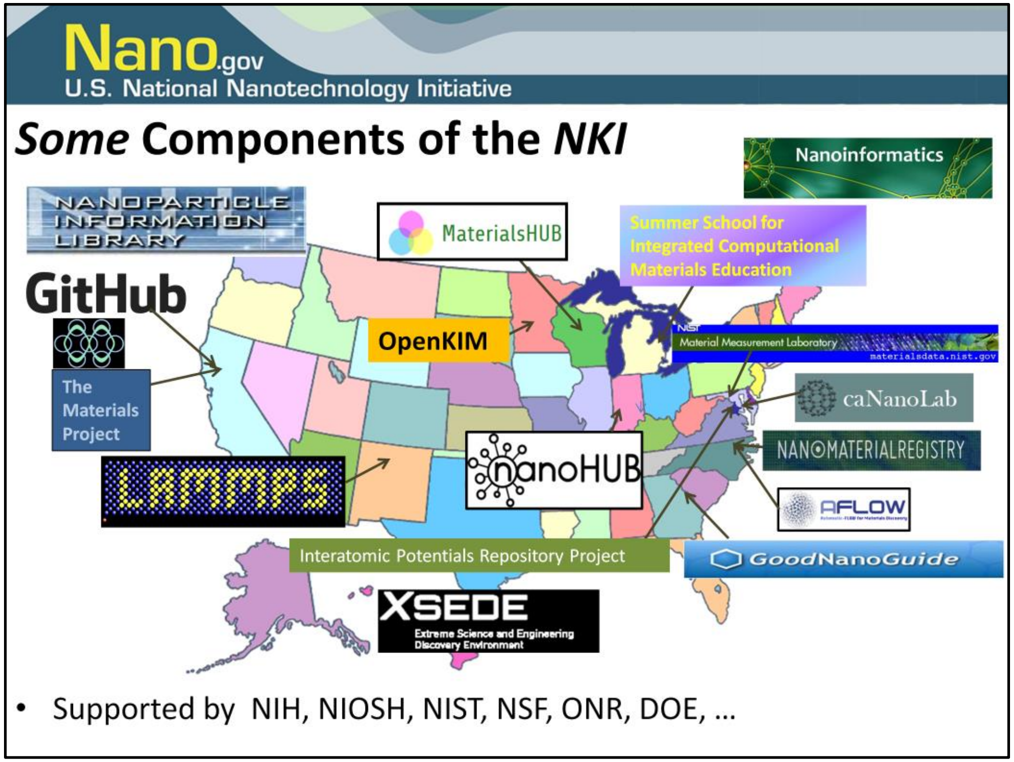
Cyber-toolbox

- Distributed across the internet
  - GitHub
  - SourceForge
  - nanoHUB
  - Others …

- Find tools through community infrastructure
  - Workshops and conferences
  - Expert communities → “Stack Exchange”
  - Schools to the curriculum

    SS for Integrated Computational Materials Education
  - Others …

>> Daryl Hess: The cyber toolbox is vast because it is distributed across the Internet. It lies on the websites of many applied mathematicians, theoretical and computational chemists, physicists, engineers, and more. The network of the cyber toolbox effectively has hubs, including GitHub, SourceForge, and nanoHUB -- the latter providing unique computational and educational services to the community. Human infrastructure plays a key role in understanding how to use the cyber toolbox -- through workshops, conferences, professional society meetings, and many others around the world, and the interaction with expert communities, as in Stack Exchange. Summer schools and nanoHUB play an important role in infusing computation and knowledge of the cyber toolbox into the curriculum.

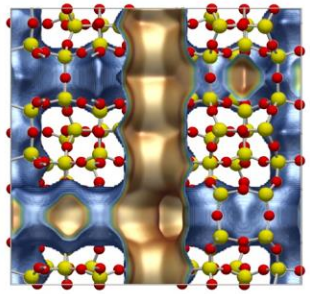


>> Daryl Hess: The components of NKI are distributed across the country and the Web. They provide access to data and computational tools that will advance science and technology starting from the nanoscale. This simple and incomplete map provides some perspective on the investments of agencies over time to develop the foundations of the NKI: software, data, computational resources, and educational resources, with some elements that do not fall neatly into any of these categories.

Cyber-toolbox

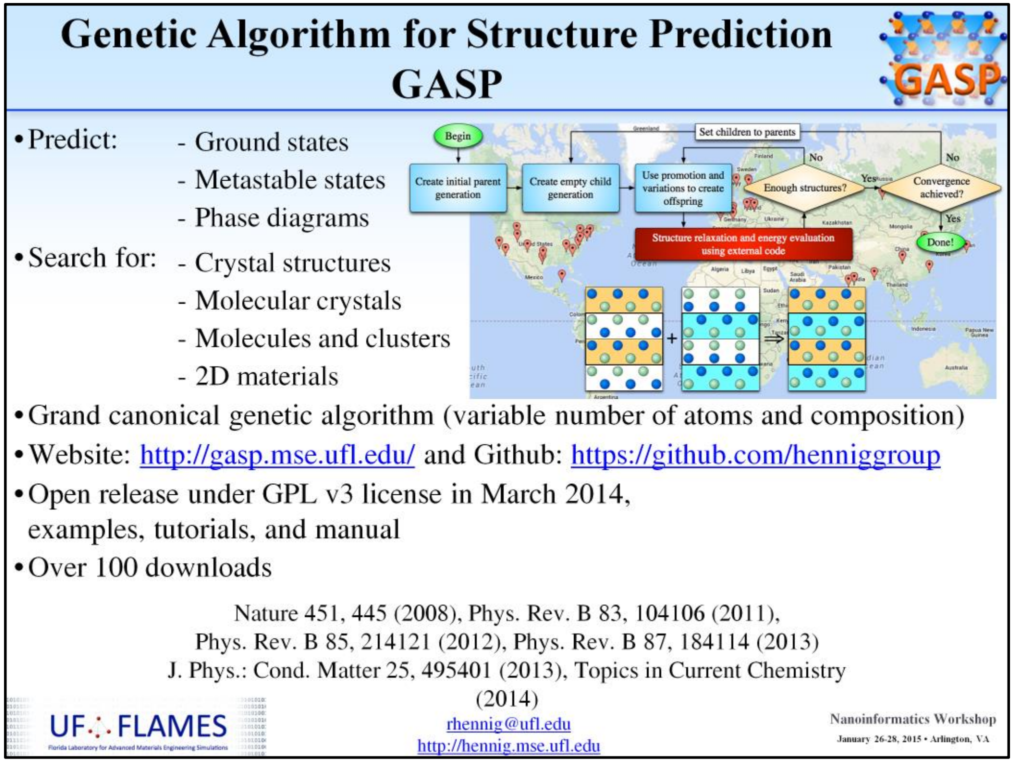
Examples

The RASPA code is available online at GitHUB under the GNU General Public License and is already being used by a start-up company.

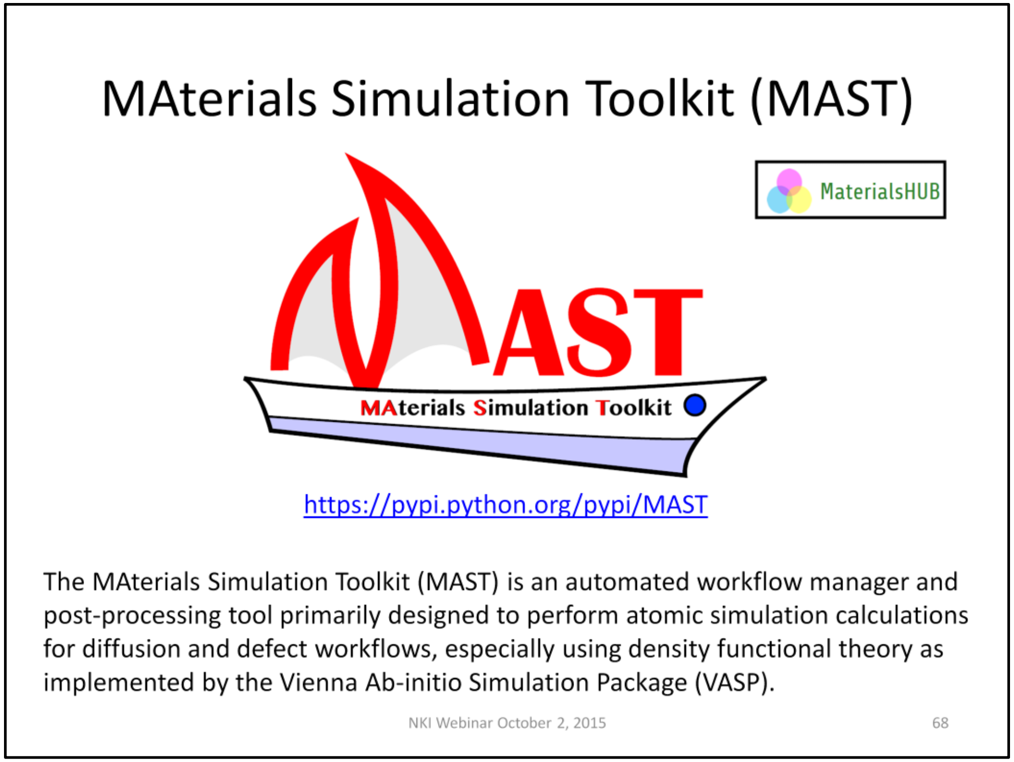


Molecular-level structure of a nanoporous material that can be simulated with the RASPA code.

>> Daryl Hess: Some examples of things in the cyber toolbox funded through different mechanisms. Shown here is the culmination of many years of work. Randall Snurr and collaborators released the new molecular simulation code, RASPA, to the community. This project was started under previous NSF support by David Dubbeldam when he was a post-doc in Snurr's group from 2005 to 2008. He is now a professor in the Netherlands, and the two groups, along with the group of Sofia Calero in Spain, have continued to develop the code. The code implements the latest state-of-the-art algorithms from molecular dynamics and Monte Carlo. The code is particularly well suited for nanoporous materials that are important in solving a variety of energy and environmental problems. This software is released under a GNU general public license and is available at GitHub. The RASPA code can be used by researchers at a variety of disciplines to provide molecular dynamics and Monte Carlo simulations and is already being used by a startup company.



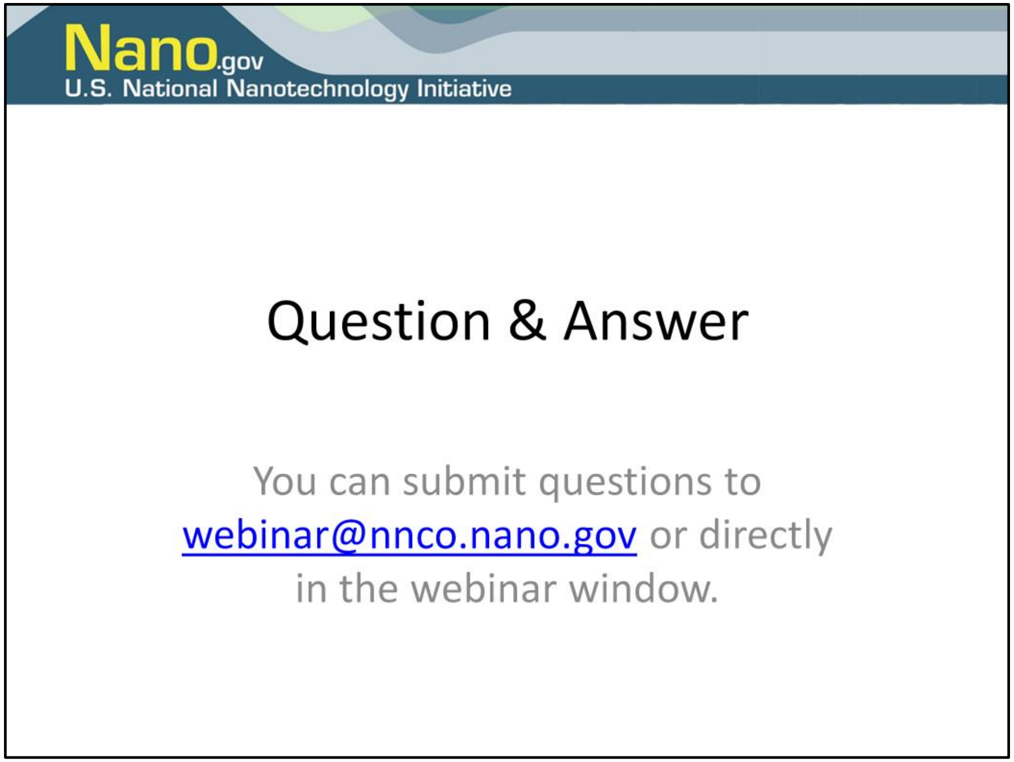
>> Daryl Hess: The development of GASP is supported through the Software Infrastructure for Sustained Innovation program at NSF. It contains a heuristic search algorithm to solve atomic structure prediction problems. It includes most successful techniques described in the literature, as well as a few important new ones. It can search for crystalline structures, as well as molecules and clusters. It is interfaced with a number of codes to calculate total energies, including both VASP and LAMMPS. The investigator here, Richard Hennig, has also contributed to the VASP code itself, highlighting another route for contributing to the software infrastructure of the community.



>> Daryl Hess: The final example here is MAST, which is also funded through the Software Infrastructure for Sustained Innovation. Mast is availble through what is now becoming the Materials Hub. The material simulation toolkit enables users to easily manage the calculation of directed complex workflows for developing large datasets of defect and diffusion properties from first principles calculations, which can be used to enhance understanding and to speed up materials development in multiple application areas, energy being one of them. MAST includes a range of well-developed tools and workflows for popular calculations, for example the determination of charge defect formation energies with proper band gap and finite size corrections and multifrequency model diffusion coefficient prediction. MAST is being used to assess strain effects on oxygen migration in perovskites and to construct the database of impurity metal element diffusion in alloys.

>> Daryl Hess: Thank you very much.

>>Stephen Lehrman: All right. Thank you, Daryl.



>> Stephen Lehrman: We're now at the point of our webinar where we invite our listeners to ask questions of our panelists. So, again, you can submit your questions by going to the window on the website or to webinar@nnco.nano.gov.

Scientific journals are beginning to require authors to deposit their data into repositories when they submit papers for publication. Do you think this is a growing trend? Why is this important to the nanotechnology community?

>>Stephen Lehrman: So the first question that we have is: *A lot of the scientific journals are beginning to require authors to deposit their data into repositories when they submit their papers for publications. Why does the panel think that this is important to the nanotechnology* community? *And what sort of data repositories, some of which have been mentioned already by our panelists, are being used?*

>> Robert Hanisch: Yes, I think it's a growing trend. I think it's very important that researchers should have access to the data that typically sits behind images or plots in a paper, but in the past has generally not been available. This I see as part of this whole move toward reproducibility and integrity in the scientific process. I do, as I mentioned in my presentation, have a little bit of concern here about having proper partnerships with the academic publishing community such that we have assurance that this data remains publicly available and not behind paywalls. So we just need to proceed in this way with some caution. I also very much favor the idea that data repositories, for what's typically called "long-tail data," should exist along with major national facilities and with government agencies so that we can assure long-term and open access.

>> Christina Liu: Also on this point of view, I presented the bidirectional link between caNanoLab and Elsevier journal publisher. At this time it is really helpful because previously authors submitted whatever data they had in their format. There was no requirement of any certain type. But if they are linked to caNanoLab, it would be in a more organized format, and that can be easily extracted for later use. So it's good for both sides: for authors, that they can have a place to present their full data in all its richness and link to their article without worry about taking up space or anything like that, and for readers because they can extract the data easier and minimize the extra work that they are just stumbling around, as I said.

Nano.gov
U.S. National Nanotechnology Initiative

Is there a charge to users for accessing caNanoLab?

>>Stephen Lehrman: So, Christina, sort of a follow-up question: *Is there a charge to users for accessing caNanoLab?*

>> Christina Liu: As far as I know, there is no charge. You just need to be registered to use it.

How does the National Science Foundation award their funding? What are some of the different mechanisms?

>> Stephen Lehrman: Good. Next question we have, which I'm going to address to Dr. Hess, is: *How does the National Science Foundation award their funding? What are some of the different mechanisms?*

>> Daryl Hess: Is this a mechanism question? By mechanism, the standard mechanism, of course, is to submit a proposal, and the proposal is reviewed. According to the merits and competition with other proposals, funding takes place. Other specific things that are relevant to my talk in developing Software Infrastructure for Sustained Innovation, for example, is one of the important solicitations for helping to create the infrastructure that science needs to do the science, to engage the frontiers. The Designing Materials to Revolutionize and Engineer our Future is NSF's answer to the MGI solicitation. As broad as it is, it will engage various facets of nanotechnology, as my highlights demonstrated. And, of course, the NSF core programs are also available by the same mechanism of submitting a proposal through the appropriate deadlines. And, finally, I think I should say this is a mechanism that's important to the community because the community can decide what it needs in order to do the science and to have a decent chance of getting funded.

What is the rate of data deposition in some of the databases that have been mentioned?

>> Stephen Lehrman: Okay. We have another question from our listeners. And the question is: *What is the rate of data deposition in some of the databases that have been mentioned?*

>> Robert Hanisch: I could just comment on the Materials Data Repository that we're operating at NIST which has grown from three or four research communities to 16 in just a matter of a few months. We have capacity there for, I think, around 40 terabytes of data which is -- we're well below that at the moment. But I think we will see the rate of increase go up in the near future as the research community becomes aware that this facility is available to them to use.

>> Christina Liu: With regard to caNanoLab and Nanomaterial Registry, as you know, the majority of these operations right now identify relevant information and curate it and present it in the more understandable format for the community. CaNanoLab was actually created for The Cancer Nano Alliance Centers and platforms that have been funded by the cancer nanotechnology group to deposit. They are required to deposit their nanomaterial information into caNanoLab. So that's a starter, and as the requirement of the deposition of material information into the NIH-funded grants become more rigid, I believe there will be more deposition of these nanomaterials data from the general NIH-funded awards.

>> Stephen Lehrman: Daryl, did you have a question you wanted to ask Robert?

>> Daryl Hess: What kind of data do you have on your 40-terabyte drive there?

>> Robert Hanisch: I showed the list of research communities. It’s phase diagrams. There is a nano group that is using it. We also have some genomics data. It really is quite broad in scope. And we are eager to reach out to others, particularly in the MGI community, to utilize this data sharing platform.

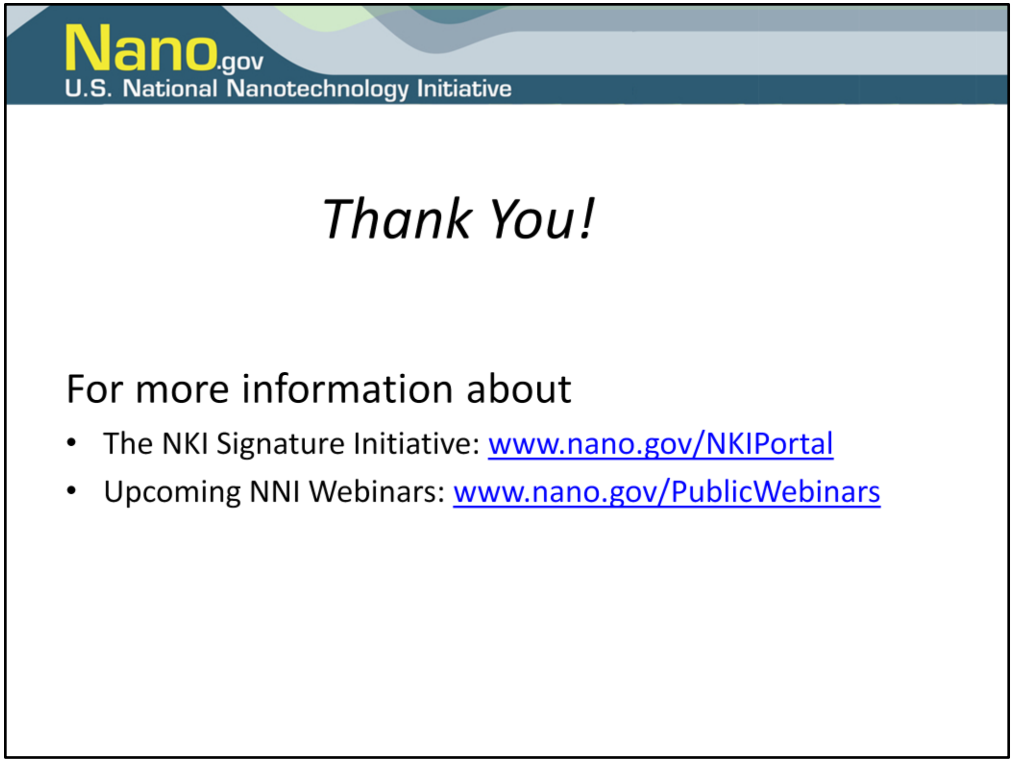
Nano.gov
U.S. National Nanotechnology Initiative

What is your advice for newcomers to learn about databases, how to use them, and also how to contribute?

>>Stephen Lehrman: Okay. We have one final question from our audience to our panel: *What is your advice for newcomers to learn about databases, how to use them, and also how to contribute?*

>> Robert Hanisch: That's why we're building this materials resource registry because it's actually fairly difficult for newcomers to get a grasp on all of the resources that are out there and available. With a resource registry, we hope to allow, basically, one-stop shopping to find the resources of interest to you and then provide links to the actual data downloads and documentation. We're just at the beginning of this. I think some of the things that we heard about before maybe are more directly ready now. But this discovery problem is a big one. And it's going to require, I think, this level of global collaboration that I talked about before in the context of the Research Data Alliance to really solve it.

>> Mark Hoover: I would like to emphasize the capturing of experience as groups come together, and they say, "We have this objective, and here's an example of how we use this database and how we integrated this process and how we talked on the phone or we exchanged protocols." It's got to be kind of -- not exactly a cookbook, but we have to understand if you are going to do this, you need to do certain steps. And the experience of how it worked or didn't work that we have is going to be inform people more quickly of how to proceed.



>> Stephen Lehrman: We have reached the end of our hour. And I'd like to thank all our panelists for taking the time to participate and their great presentations. I would also like to thank our listening audience for calling in to this presentation. The NNCO will post a transcript and the presentation slides for this webinar on the nano.gov website. And I'd like to remind you that our next Webinar in our fall series is on Friday, October 16th at 12:00 noon Eastern. That is on Nanosensor Technologies and Applications, and you can register for that at www.nano.gov/publicwebinars. With that, again, my thanks to our panelists, and that concludes our webinar for today.