RESEARCH STRATEGY

SIGNIFICANCE: As the Human Genome Project (HGP) [1, 2] approached its successful conclusion, NIH was in the process of formulating the Roadmap, which evolved into the Common Fund. At the leading edge of these initiatives was the Biomedical Engineering Research Partnership (BRP), which promoted technology development. For the BRP, Sklar led the development of high throughput flow cytometry, a screening technology which transitioned into the Molecular Libraries Program (MLP) [3] a logical successor to the HGP, to develop small molecule probes for newly discovered genomic targets. For the Common Fund, Illuminating the Druggable Genome (IDG) became the next logical successor to prioritize genomic targets for further investigation. The technology advanced via the BRP by Sklar became a foundation for the collaboration with Oprea resulting in 10 years of participation in the MLP through both pilot and production phases involving GPCR, kinase and transporter targets among others, with Sklar leading both administrative and laboratory efforts and Oprea leading the data science efforts. In the latter phases of the MLP, Oprea, Sklar, and Schürer, who had also been part the MLP since the pilot phase, joined forces for the BARD initiative (2012-2014) [4, 5].

In IDG Phase I, Oprea, Sklar, and Schürer again joined forces for the IDG Knowledge Management Center (KMC), with Oprea as overall PI and leader of the "data organizing core", Sklar leading administrative efforts, and Schürer leading the drug-target ontology development (2014-2017) [6]. All three contributed to platforms and outreach to familiar communities from which they draw and contribute expertise. Given 5-years of partnership, it is natural for Schürer, Oprea, and Sklar to join forces once again for RDOC, which will benefit from this group's many years of PI/PD experience with NIH programs involving technology integration, data science, software development, and program as well as project management. Given our current leadership roles in the IDG Phase I, in addition to our previous work in MLPCN [7] and BARD [4], and Schürer's work in the LINCS (Library of integrated Network-based Cellular Signatures) program [8], another Common Fund project, we are experienced in engagement and facilitation of monthly meetings with NIH program managers as well as external scientific review panels regarding the identification and resolution of programmatic issues. The decade of experience with the MLPCN has strengthened our ability to successfully collaborate with multiple investigators, both within and outside the United States. Our problem-solving skills, the ability to coordinate, and tele-collaborate are supported by the number of chemical probes, investigator-initiated clinical trials, as well as patented technologies and publications from our team. We have managed and participated in consortia (subcontracts) through regular communications such as periodic meetings and conference calls.

This grant partners the University of Miami and the University of New Mexico for compliance with Federal regulations, policies, and guidelines for human subject research, evaluation of risks and protections, ethical oversight, and data and safety monitoring as appropriate. Taken together, our assembled team has already played an important role in Roadmap/Common Fund initiatives through their evolution from discovery technology, discovery informatics, and big data.

INNOVATION: As will be reflected in the research plan that follows, innovation is added into each Aim. This innovation is a reflection of the integration of experienced management, an administrative infrastructure for planning, management, coordination, execution and assessment of the IDG-RDOC activities, and the organizational structures of our Center. For example a collaboration with Genetic Engineering and Biotechnology News [9] will provide an unprecedented level of publicity for an NIH program. Outreach includes modular training, online courses and lectures, open innovation, and social media. The Center balances technological risk and innovation by building on proven technologies, but also pushing for advanced data science capabilities, for example via the proposed hybrid relational and graph storage in the RMS and interfaces to the IDG Portal. The resource and data management approach emphasizes the widely endorsed FAIR principle (slightly expanded for IDG: Findable, Accessible, Attributable, Interoperable, Reusable, Reproducible) based on rigorous data standards, cloud-based metadata management (via modular templates and forms), clear resource sharing policies, and a resource management software system to reduce operational and technological barriers of resource sharing. The relationships of the tasks in the RFA to this proposal are compiled in Table 2 in the Management Plan.

SPECIFIC AIM 1. Coordinate and support the IDG consortium via a highly experienced management and administrative core (MAC).

1.1: IDG Communication and administration: The RDOC management and administrative core coordinates and supports operational and scientific interactions among the IDG Data and Resource Generating Centers (DRGCs), the KMC, and NIH to enable timely and effective dissemination of IDG generated resources,

including datasets, biological and chemical reagents, assays, and biological model systems. The RDOC provides general collaborative and communications infrastructure and organizes regular consortium steering committee and working group tele/video conferences annual consortium meetings. These functions are analogous to those played by Sklar and Oprea for IDG Phase I in their capacity as Chair of the IDG Steering Committee (Oprea), PI of the Admin Core for the KMC (Sklar), and project manager for the Admin Core (Waller). The admin core will also be involved in tracking resources generated by the consortium and facilitating outreach. The goal of these activities is for RDOC to provide overall support for the IDG consortium activities by identifying project requirements, addressing needs, concerns and expectations of all the stakeholders and assisting in balancing the competing project constraints, which may include scope, quality, schedule, resources and risks. Thus, the MAC will maximize the impact of the IDG program as the KMC Admin Core has done during IDG Phase I, linking activities at U Miami and UNM.

Seamless communication and collaboration are critical for a successful outcome for IDG Phase II. In addition to standard tele-/video-conferencing RDOC will set up and administer a **cloud-based communications, collaboration, and task and project management infrastructure**. This infrastructure includes GotoMeeting for video meetings with screen sharing / switching, recording, etc., Slack [10] for team-based communication, Basecamp [11], Teamwork, [12] for team collaboration, document sharing, project management, Trello [13] for task management, Google Docs [14], and Google Sites [15] for collaborative content creation, and Github [16], Confluence [17] for collaborative code sharing, documentation and issue tracking. These components will be set up for sharing non-public information among only IDG members, including posting of meeting minutes, providing cross consortium contact information, and displaying internal presentations, and any files or digital content. Individual IDG members will also be able to share information on their own. This infrastructure also facilitates **management of general meeting logistic** (email calendar requests, distributing agendas, recording the minutes and action items) and supports **coordination of interactions with NIH and non-NIH partners** and **customer interactions** in conjunction with Outreach activities. These functions have been successfully accomplished by Waller and Sklar in the current roles for the KMC Admin Core for IDG Phase I.

1.2: Customer support email will be set up by MAC. At least two individuals (one each at U Miami and UNM) will monitor the account in addition to IDG social media channels to respond in a timely manner to requests for IDG resources and general inquiries by the community and to relay the information to the appropriate group (via the above infrastructure), log the requests and follow-up for customer satisfaction.

1.3: Meeting organization. Coordination with NIH will be extended by MAC for the initial kick-off meeting of the consortium, for which MAC will assist NIH with preparations and management. Subsequent annual IDG face-to-face meetings will be coordinated by MAC together with the IDG Steering Committee enabling all the IDG members to share their developments and progress with the entire consortium, in addition to the 6 External Scientific Consultants (ESC). MAC will be working with and pay for the attendance of the ESC to these annual IDG meetings. These functions have been successfully accomplished by Waller and Sklar in their current roles for the KMC Admin Core for IDG Phase I. The 2016 IDG meeting was acclaimed a success by NIH staff for showcasing the IDG Phase I and rationale for Phase II to outside NIH members and invited pharmaceutical representatives.

Additional biennial meetings will be coordinated, planned, and executed in conjunction with the Outreach group of RDOC. These biennial meetings are international open innovation meetings, allowing for publicizing the IDG resources and establishment of new collaborations with other consortiums such as Structural Genomics Consortium (SGC) [18, 19], Innovative Medicines Initiative (IMI) [20], Accelerating Medicine's Partnership [21], and Eli Lilly's Open Innovation Drug Discovery programs [22]. For these biennial meetings MAC will locate the venue, manage logistics, organize sessions, and aid in partnering with non-IDG consortium. Part of this work will may include developing and launching meeting website and advertising meetings via non-IDG Consortium or via other Outreach partners. Further expansion on these Outreach activities will be explained in Specific Aim 2. Waller and Sklar have managed logistics for meetings at remote sites for IDG Phase I.

An important initial milestone for the IDG consortium is the development of IDG publication policy and internal data sharing. The creation and organization of subcommittees to work on writing these policies, strategic planning, or decision making will be facilitated by the MAC, in conjunction with the other IDG members and the NIH team. In building consensus for these policies, the MAC will work on disseminating questionnaires for feedback and collaborative contributions as we currently do for IDG Phase I. Implementation of this policy will be part of the work described further in Specific Aim 3

1.4: IDG Progress status. RDOC MAC will collect and collate statistics to quantify and evaluate IDG resources and their impact, such as number of reagents, datasets, data points generated, data access via the IDG Portal or public repositories, user requests for data and resource (e.g. reagents), web analytics of the IDG Website and the KMC IDG Portal, publications, conference presentations, and dissemination of reagents and data beyond the IDG (e.g. reagents becoming commercially available, or data being available in public repositories). The MAC will report to NIH on the status of the IDG Consortium as requested.

Outreach led by RDOC on social media and internet for IDG will commence with the establishment of an introductory website for RDOC that could also service and display content from the other IDG consortia members. In addition, social media presence will be created by setting up accounts in Twitter, LinkedIn, and YouTube per acceptance of IDG SC member's accounts for IDG. RDOC will then lend support for posting information and content generated by all IDG members to these social media or Internet sites.

SPECIFIC AIM 2. Facilitate and coordinate external partnerships, outreach and training of the IDG program. RDOC will develop an integrated community outreach, complete with education and training opportunities focused on specific IDG products, such as IDG data, websites and technologies. This will enhance the impact of the IDG program in target-centric drug discovery communities such as molecular, structural and chemical biology, genomics, and medicinal chemistry. Aim 2 will address the tasks for outreach, education on accessing data, developing distribution strategies for IDG resources and, facilitate interactions with other NIH and non-NIH partners, including via biennial international open innovation meeting.

2.1: Massive community outreach. An extensive proposal for generating outreach and media coverage of IDG has been developed in co-ordination between Sklar and Genetic Engineering & Biotechnology News (GEN, [9], see GEN Proposal in the Letters of Support). The proposal includes appropriate means of sharing information about the IDG consortium by generating media content such as podcasts, webinars, overview articles, and expert opinion pieces. Table 1 lists proposed mechanisms for publicizing IDG Phase II and disseminating IDG generated resources, with this potentially enabling partnership with different trade organizations for disseminating tools and reagents.

Table 1	Type of Content
1	Overview article written by Druggable Genome team (1500 words).
2	A resource list with links to relevant papers, databases, websites, etc.
3	Q&A focused audio webinar with two druggable genome experts with powerpoint slides.
4	Podcast (audio only). GEN will interview druggable genome expert(s).
5	Infographics and/or posters.
6	Suggest Thought leaders/collaborators GEN can approach for additional content like point of view articles, case studies, or lab profiles.
7	Give list of experts that GEN interviews for a review article.
8	Provide links to videos we will embed on our site with a short summary
9	Put GEN on your press release distribution list and that of our collaborators such that GEN are aware of any "news" from these organizations and can write news briefs.
10	Provide a list of relevant databases, websites, or scientific apps that GEN can review and feature.
11	Provide a list of recent and relevant papers that GEN can write short reviews of.
12	Feature Druggable Genome digital properties (like NewDrugTargets.org) as part of GEN's highly popular "Best of the Web" and "Best Science Apps"

NIH has already approved policies regarding this outreach effort following consent of current members (PIs) of the IDG Phase I Steering Committee. In short, these GEN outreach activities should offer a balanced representation of IDG's achievements (inclusive of all members), and allow for the information provided to GEN not to put at risk the IDG PIs possible future publication and intellectual property. For any representation concerning the IDG Consortium as a whole, the PI must gain prior approval from the IDG Consortium. NIH's involvement will follow the normal NIH policy/communication guidelines and will only speak with the consultation of the NIH Working Group and IDG Consortium.

The relationship with GEN is predicated upon the following considerations. Since 1981, GEN's news and technology focus spans the entire bioproduct life cycle, including drug discovery, early-stage R&D, applied research (e.g., omics, biomarkers, and diagnostics), bioprocessing, and commercialization. GEN's team of scientific & technical writers produce articles that inform research and process scientists via: Print magazine 21x annually; GENengnews.com website updated daily; GEN Highlights daily newsletter; GEN Facebook,

Twitter, You Tube, LinkedIn accounts; and GEN Touch mobile app. GEN's content development has 36 years experience during which GEN has been researching and packaging highly useful content in ways that resonate with the needs of researchers working in discovery and development. GEN uses different formats, from traditional print to video and interactive data to infographics. GEN has a total audience reach of 414,212 life science professionals for information of the key applications of new and popular life science technologies used in drug discovery and diagnostics. GEN is known particularly for its circulation among pharma and biotech entities. GEN will provide arm's length neutral coverage of the IDG Phase II regardless of selected entities. For IDG, staged coverage throughout 2017 starts in April/May (IDG Phase I) and continues releasing content elements for at least one year (bridging) to Phase II to garner maximum targeted readership and impact. GEN is responsible for finding commercial sponsorship by identifying and securing one or more life science tool and technology vendors as sponsors of this content.

The current Admin Core of KMC Phase I (Sklar, Waller, and Oprea) is committed to transitioning GEN coverage of IDG Phase II regardless of the entity selected to lead RDOC. We do not receive compensation or other special consideration from GEN for our efforts.

2.2: IDG Scientific Webinar series. We will organize a monthly Webinar lectures given by invited experts in areas relevant to the IDG project including drug discovery, target validation, computational biology, data science, systems biology, chemical biology, and medicinal chemistry. We have already recruited several experts to participate in the Webinar series (as examples, see letters of support: Mark Musen, Stanford / BD2K / NCBO, Avi Ma'ayan, Mt Sinai NY / LINCS, Melissa Haendel, OHSU / Monarch / UBERON, Susanna Sansone, Oxford / Biosharing, Henning Hermjakob, EMBL-EBI / Reactome, Yanli Wang, NLM / PubChem, Mario Medvedovic, U Cincinnati / LINCS, Rajarshi Guha, NCATS / IDG, Shawn Gomez, UNC / IDG, Lucila Ohno-Machado, UCSD / bioCADDIE). Lectures will be posted to the IDG YouTube channel with approval of the presenters.

2.3: Open Innovation meetings and workshops to include Eli Lilly (letter of support, Michal Vieth, Eli Lilly), SGC, IMI, International Mouse Phenotyping Consortium (IMPC), ChEMBL (letter of support, Andrew Leach, EMBl-EBI), We plan to hold two of these meetings in Miami, in coordination with the Miami Winter symposium. Miami is an attractive location and we have secured additional support for such meetings (letter of support, Charles Luetje, Chairman, Department Pharmacology, University of Miami).

2.4: Advertise IDG resources and engage the scientific community via social media to include industry and academic researcher. LinkedIn group, Twitter, YouTube channel, MOOC at Coursera (see below).

2.5: IDG Training and Education Program (ITEP): ITEP will offer introductory lectures and videos based on IDG-specific queries and demonstrations, aimed at preparing experts, occasional users, physicians as well as citizen scientists for the in-depth exploration of the understudied proteins. ITEP will provide an overview of several data mining technologies, as well as advanced analytics to evaluate "dark" proteins, with emphasis on the proteins of interest for the IDG consortium, making extensive use of the *NIH Commons* where possible. *ITEP will be comprised of three modules:* 1) an e-learning module will consist of online lectures, initially in a web-based system (end of Y1) but supplemented by videos, (end of Y2); 2) theoretical challenges such, DREAM open competition and other crowdsourcing activities, specifically derived from DRGC case-studies (proteins of interest to the IDG), comprising exercises, literature reviews and a wiki/chat module to facilitate on-line discussions; 3) specific half-day face-to-face (F2F) symposia, organized in conjunction with major scientific conferences, for learners that have completed modules 1 and 2. These modules, including digital outcomes from Module 3, will be shared on the NIH Commons virtual space. For side dissemination, we will also adopt this course as a **Massive Open Online Course (MOOC)**, for example in the Coursera platform [23], and we will have the opportunity to share these modules at UNM (see letter of support, Mara Steinkamp, Director of the STMC Education Core).

ITEP Module 1: Built on existing and future (IDG Phase II) developments generated by the IDG KMC, the IDG Consortium and other IDG partners, this module will provide an introduction to data science and knowledge management, primarily focused on methods and tools for data integration, archival, management and articulation of IDG-specific resources, based on human proteome-wide efforts. *Outcome:* Learners are expected to become competent in data acquisition and interpretation.

ITEP Module 2: Data analytics and workflows developed for the IDG knowledge graph (e.g., content from DGCR and IDG KMC, as structured archived in Pharos, including existing and future workflows) will be used to explore understudied proteins from the three families of interest for the IDG. These theoretical challenges will

use the NIH Commons Credits business model to support execution of IDG-specific pipelines and to host any public facing software and services derived as outcome of Module 2 challenges. Challenges include, but are not limited to computational predictions of protein structure/function/pathway/role in disease for IDG Phase II proteins, prediction of drug use and repurposing for these proteins, as well as development of novel data analytics and pipelines specifically geared to support IDG (for example, auxiliary proteins for ion channels, which are outside the IDG Phase II gene list). Challenges will be tailored to the educational background of each participant. For example, patient advocates and citizen scientists will be asked to participate in data synthesis, validation or exhaustive reviews, in collaboration with the IDG Consortium, we will establish a process to enable experimental verification of these challenges. Those that are deemed successful will be forwarded to the IDG KMC, for incorporation in Pharos and IDG KMC pipelines. *Outcome:* Learners are expected to become competent in data and prediction interpretation, as well as experiment planning and feedback (hypothesis confirmation or falsification).

ITEP Module 3: Participants who successfully complete the first 2 modules will be invited to face-to-face (F2F) meetings, where they will be asked to give seminars based on their own IDG specific research. F2F participation is not mandatory, and travel funding will not be available to support learners. Optionally, learners will attend such meetings virtually, via webinar presentations. Digital outcomes of ITEP Module 3, such as F2F presentation videos, webinars, slides, reports, papers, will be shared on the NIH Commons and the IDG portal (as appropriate). *Outcome:* Learners are expected to gain competence in scientific communication skills.

Benefit for the community: By completing the ITEP Modules, learners will learn how to specifically interrogate the understudied proteome, with emphasis on critical thinking and practical applications (including close cooperation with experimental groups, such as the Type A, B and C DGRCs). One such example is the New Mexico IDeA Network for Biomedical Research Excellence [24]. NM INBRE champions biomedical and community based research excellence in the state of New Mexico through the development of innovative, supportive and sustainable research environments for faculty and students, community engagement health initiatives, while building a network of lead scientists and educators at the state, regional and national level. As per letter of support from its Director, Dr. Shelley Lusetti, we anticipate that IDG-RDOC will benefit the community, first in New Mexico and, if successful, throughout the entire IDeA network.

Benefit for the IDG Consortium: ITEP outcome is clearly designed to benefit the IDG in 3 ways: outreach and community awareness; development of problem solving tools for IDG specific needs; and, for successfully responding to Module 2 challenges, increase in IDG specific knowledge.

Education expertise: Oprea has taught a wide variety of topics, from human physiology and biochemistry (1991-2006) to lead and drug discovery (2001-2010), as well as informatics in drug discovery since 2005. He has organized 3-5 day informatics summer schools and workshops on 3 continents since 2006, typically attended by 30-40 participants. In addition to UNM, Oprea has held a guest professorship with (and taught at) University of Perugia (2009), Technical University of Denmark (2010-2012), University of Gothenburg (2012-2014), and is currently a guest professor at the Universities of Copenhagen and Gothenburg (2017-2019). For the development of ITEP modules, Oprea will rely on the combined expertise of the IDG KMC, the IDG-RDOC and the IDG DGRCs, as each center will be invited to contribute source materials (for Module 1), and particularly challenges (for Module 2). With approval from NIH staff and the data owners, Oprea intends to mask parts of the newly developed DGRC data, so that the optimal models are selected.

SPECIFIC AIM 3. Steward the development, adoption and implementation of consistent metadata standards and consortium-wide processes and policies for IDG resource sharing and dissemination. We aim to maximize the scientific impact of the program, IDG-generated resources, such as datasets and collections, biomaterials, reagents, experimental models, assay kits, protocols / SOPs, training materials / tutorials, and more must become widely disseminated and integrated into the wider biomedical ecosystem. To achieve that goal, RDOC will lead the development, implementation, adoption and publication of guidelines, technical specifications and formal policies that guide all aspects of resource dissemination and sharing including internal and public data releases, embargo periods, data analysis and QC, authentication and quality criteria of physical reagents, error and discrepancy resolution, data descriptors, metadata, provenance, data repositories, and resource attribution / citation. In addition to coordinating technical specification and consortium policies, it is the responsibility of the RDOC to assemble and make available detailed digital descriptions of each resource including how they can be accessed. While the RDOC will receive and transmit requests, and assist in fulfilling requests, the actual transactions of physical samples will be handled directly between the producer of the resource and the requester or via a specialized intermediary, for example a

commercial supplier.

Guiding principles to assess the utility and impact of science resources include the degree to which they are Findable, Accessible, Attributable, Interoperable (digital only), Reusable, and Reproducible. The FAIR principles [25] have been developed for digital resources (e.g. scholarly data). Here we propose to adopt them (with appropriate extension) for all IDG resources. These guidelines have been designed and jointly endorsed by academia, industry, funding agencies, and publishers. For (physical) chemical and biological resources, these guidelines also align with authentication and quality assurance best practices and requirements that are in place at NIH, publishers, research laboratories and commercial suppliers.

To maximize a resource in each of these FAIR dimensions, high-level requirements can be defined. *Findable*: In addition to the IDG Portal (at the Knowledge Management Center, KMC) resources should be registered and indexed in a searchable catalogs or repository and also findable via a Web Search. That requires unique persistent resource identifiers, standardized deep metadata annotations that conform to industry / community best practices and also detailed textual descriptions. *Accessible*: Datasets should be retrievable by their identifier in an open and free protocol preferably from a public persistent repository and initially from the IDG Portal; material resources must be available from one or more suppliers (originally the DRGC); request and access to resources will be mediated via the RDOC MAC (compare Aim 1; e.g. customer support email, phone, social media channels, consortium policies, etc.). *Attribution*: Resources should be clearly attributed to who generated them and provenance of the production and validation of a resource must be kept in the resource descriptions. *Interoperable*: Digital resources need to be represented in standard formats, have standardized formal metadata annotations using controlled vocabularies with qualified references. *Reusable*: Digital resources require a clear data usage license (all IDG data should be open source), and biological and chemical reagents or model systems require formal terms-of-use for either unrestricted use or a material transfer agreement (MTA). *Reproducibility*: Sufficient detail how the resource (digital and material) was generated, typically following domain-specific reporting guidelines, many of which are available via the Minimum Information for Biological and Biomedical Investigations (MIBBI) Portal [26]. Here we describe our approach to implement the most important requirements.

3.1: Resource metadata standards specifications for IDG resources. In the LINCS Data Coordination and Integration Center (DCIC) we have created metadata standards specifications to obtain key details to describe and annotate datasets, assays, reagents and experimental protocols, all of which are vital to reproduce biomedical studies [27]. Our group has developed and published metadata standards for small molecules, cell lines, primary cells, embryonic stem cells, induced pluripotent cells, differentiated cells, proteins, antibodies, and a general class of other reagents [28]. Many of these data fields are linked to previously established standards such as, Minimum information about a bioactive entity (MIABE) [29], Minimum Information About a Cellular Assay (MIACA, [30]), Minimum Information About a Cellular Assay for Regenerative Medicine (MIACARM) [31], the eagle-I resource discovery system [32], and standards put forth by the International Human Epigenome Consortium [33]. Additionally, field names and controlled vocabulary for content of many fields are linked to ontologies contained within Bioportal [34] such as BioAssay Ontology (BAO) [35, 36], The Ontology for Biomedical Investigations (OBI) [37], Cell Line Ontology (CLO) [38], Medical Subject Headings (MeSH) [39], and many others. Based on our previous experience we, in close coordination with the IDG DRGCs and KMC, will develop IDG metadata standards specifications for all resource categories. We aim to have the first version of these standards defined in Q2 of the project and release a subsequent update annually. These specifications will then be implemented in templates to collect the information in the Resource Management System (RMS, Aim 4). IDG data standards will be integrated into the wider biomedical ecosystem via Biosharing [40]; LINCS data standards are also hosted there (see letter of support from Susanna Sansone, Oxford, [41]).

3.2: Data and resource repositories and -indexing. In addition to facilitating the hosting of IDG resources via the IDG Portal, the RDOC will assist in the deposition of IDG datasets including detailed metadata annotations into appropriate public long-term repositories, for example PubChem [42, 43] for small molecule drug screening data (see Letter of support from Yanli Wang, NIH / NLM, PubChem), GEO for gene expression data, or PRIDE, MassIVE, PeptideAtlas via the ProteomeXchange [44] for proteomics mass spectrometry data. As we have done with LINCS datasets, we will index IDG datasets in the bioCADDIE Data Discovery Index DataMed [45] (see letter of support from Lucila Ohno-Machado, UCSD, bioCADDIE) and in the Omics Discover Index [46] (see letter of support Henning Hermjakob, EMBL-EBI, omicsDI). These catalogs index a variety of metadata, such that datasets can be identified by protein target, cell type or model system, small

molecule, etc. The RDOC can also facilitate IDG dataset publications, for example in Nature Scientific Data (editorial board member Schürer). According to the requirements and priorities of the consortium, we will consider other resources, for example eagle-i [32] for various reagent types, or Protein Model Portal [47] or PMDB [48] for protein structure models.

3.3: Authentication and QC of reagents and experimental models. The RDOC resource descriptions will include any documentation of authenticity and quality of biological and chemical reagents or experimental models that are provided by a DRGC according to consortium resource specifications and policies. The RDOC will put in place measures that such documentation is part of the resource descriptors that are made publicly available via the Data Portal.

3.4: Resource identification system and persistent identifiers. All resources will be registered via a Resource Management System (RMS, Aim 4), where the submitted reagents, samples, assays, datasets, etc. will be tracked and cross-referenced. RMS tracks the submitted Center-specific (and/or batch-specific) IDs and assigns a unique IDG resource-specific ID, which will be mapped to external reference ontologies and global identifiers (e.g. PubChem [42], Cellosaurus [49], NIF antibody registry [50], etc.). All resources will be registered in the MIRIAM registry to provide unique, perennial and location-independent identifiers [51]. The Identifiers.org service [52] which is built upon the information stored in MIRIAM, provides directly resolvable identifiers in the form of Uniform Resource Locators (URLs). This system provides a globally unique identification scheme to which any external resource can point and a resolving system that gives the owner / creator of the resource collection flexibility to update the resolving URL without changing the global identifiers. Together with mappings of IDG resources to reference ontologies and registration of datasets into public data repositories, this approach also addresses persistence of digital resources created in the IDG program beyond the funding duration of the project. In the LINCS program we have taken a similar approach; LINCS DataSets have unique (LDS) IDs and are mapped to, e.g., LINCS Small Molecule (LSM), LINCS Cell Line (LCL) IDs. All of these entities can be uniquely globally referenced based on their global and unique persistent identifiers and they currently resolve to the corresponding LINCS Data Portal landing pages [53].

3.5: Resource Landing Pages and attribution and citation of resources and datasets. Globally persistent uniform resource locators (PURLs) also facilitate citation and attribution of resources. In LINCS, the citation PURL for a collection or a dataset resolves to the repository of the LINCS Data Portal Dataset Landing Page [53] composed of the dataset collection description, entity metadata, and underlying data files. The dataset description credits the creator of the dataset; acknowledges the funding source; and links relevant information like associated assay(s), source repository, protocol(s), processing pipeline(s) and related publications. The metadata entities and their relationships to the dataset are further annotated by linking to external sources and databases. This approach to dataset citation provides flexibility to adopt to variations in data the processing and publication pipelines, including versioning and provenance. Our approach to dataset and resource attribution and citation was guided by the Joint Declaration of Data Citation Principles (JDDCP) [54]. Here we propose a similar scheme to implement, in collaboration with the KMC, Resource Landing Pages that include all relevant information, metadata, and annotations to describe and attribute the resources to their creators and allow citation via global identifiers (PURLs). These Landing Pages will be made accessible via the IDG Data Portal (compare Aim 4, see letter of support from Rajarshi Guha, lead developer of the Pharos Data Portal at NCATS).

3.6: IDG resource submission, registration and publication. In the LINCS project, we have been developing the end-to-end processing pipelines and supporting informatics infrastructure to receive, standardize and harmonize, register, validate, integrity check, and publish LINCS data along with detailed and standardized metadata descriptions for reagents, experimental conditions, assays, processing pipelines and general information (author, center, related publications, etc). In six Data and Signature Generating Centers (DSGCs) the project generated a wide variety of cellular perturbation response data including transcriptomics, proteomics, epigenomics, biochemical and imaging [8]. Metadata standards provide the foundation for harmonizing and integrating these resources [28]. Our already developed infrastructure includes user authentication, graphical user interfaces, APIs, data storage, transfer protocols, a relational database schema with several integrated data stores, business rules, and several processing pipelines to allow users to submit dataset files and metadata, reagent, experimental and assay descriptions (with controlled metadata) and to validate, standardize, register (assign IDs) all entities and track their relationships. Several reagent types are further annotated and cross-referenced to external resources such as PubChem [42], ChEBI [55], ChEMBL

[56], PDB [57], and DrugBank [58] for small molecules, CLO [38], disease (DO) [59] and organ / tissue (Uberon) [60] annotations for cells. The entire data packages after processing are made available via the Dataset Landing Pages (and currently Small Molecule and Cell Landing Pages) in the LINCS Data Portal and programmatically via RESTful APIs.

We propose to adopt this successful operational approach, substantially re-using already developed infrastructure and software tools to manage IDG resources (Figure 1). The timely development of resource standards specifications requires an experienced team and close collaboration among the different groups and stakeholders in a research consortium. We will start that process immediately and stage it based on the data and resource generation timelines of the individual DRGCs. However, even if standards are defined and agreed, it can be challenging to communicate and obtain all details of generated data and resources in the required format. There are technological, operational, and management barriers. To overcome technological and operational challenges we will deploy the RMS (Aim 4). Without such a system, it has been our experience to be very time-consuming to obtain comprehensive and standardized resource descriptions. To overcome differences in priorities among consortium members, we will work with each group individually, provide training and have regular meetings to reach consensus and communicate timelines and specific actions. The LINCS Data Working Group (chair Schürer), which meets monthly, played an important role in developing and then implementing standards in the LINCS Consortium.

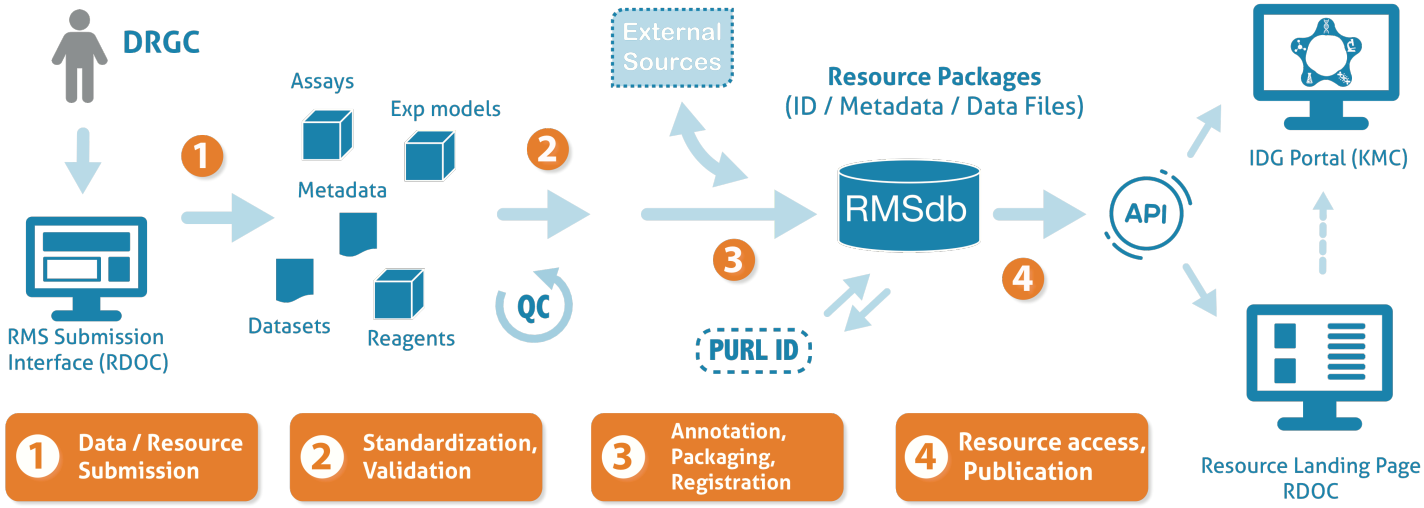

Figure 1. End-to-end process of submission, standardization, annotation, packaging, registration of IDG resource descriptions and their publication in Resource Landing Pages and at the IDG Portal at the KMC (also see Figure 2).

Our approach of developing data standards and identifiers as well as collecting and curating resource annotations has been endorsed by several collaborators in the IDG Pilot program, Bryan Roth (UNC, IDG GPCR), Shawn Gomez (UNC, IDG Kinase), the LINCS program, Avi Ma'ayan (Mt. Sinai, LINCS DCIC), Mario Medvedovic (U. Cincinnati, LINCS DCIC), and also by experts in standards development, ontologies, and data management, Mark Musen (Stanford)), Melissa Haendel (OHSU, Monarch Initiative), Susanna Sansone (Oxford, BioSharing), Lucila Ohno-Macado (UCSD, bioCADDIE), Henning Hermjakob (EMBL-EBI), all of whom has provided letters of support.

SPECIFIC AIM 4. Develop, deploy and support the IDG Resource Management System (RMS). To overcome technical and operational barriers for obtaining resource annotations in sufficient detail and according to standards specifications (Aim 3) we will implement the IDG Resource Management System (RMS). RMS supports the submission, processing and exchange of standardized descriptions of IDG resources, which fall into the general categories of dataset/collection, biological or chemical reagents or materials, experimental model systems, assays/kits, training materials/tutorials, including authentication or documentation of resources. The system will be flexible enough to also support other types of resources. RMS will substantially reuse components we have been developing for a dataset submission system for LINCS along with the backend infrastructure including user authentication, graphical user interface elements, metadata templates, APIs, data (file) storage, transfer protocols, a relational database schema with several integrated data stores, business rules, and various processing pipelines and scripts. This system will shortly go live at the BD2K LINCS DCIC DataPortal [61].

The RMS system and integration with the KMC and the CEDAR (Center for Expanded Data Annotation and Retrieval [62]) metadata management environment are conceptually illustrated in Figure 2 and the key features and characteristics are described in detail below. The system will be hosted and supported at the Center for Computational Science (CCS) at UM (see letter of support Nicholas Tsinoremas, executive director).

4.1: Cloud-stored templates and forms. The RMS submission forms will leverage metadata templates and APIs available from CEDAR (Center for Expanded Data Annotation and Retrieval, BD2K Center of Excellence, Stanford University, see letter of support Mark Musen). Storage and management of modular metadata templates and filled forms in a cloud environment makes it easier to discover, access, mix-and-match, re-use, and analyze metadata for diverse resources. CEDAR integrates directly with Bioportal ontologies, allowing direct linkages of field name identifiers and controlled vocabulary for field content. In addition, CEDAR enables controlled vocabulary through specialized field types such as radio buttons, check boxes, and drop down menus, all of which have a user select from preset options. Template instances containing completed metadata fields are stored on CEDAR's servers, and are accessible through a REST API and exportable in a JSON schema format. These enable access to metadata forms from the RMS servers to move the data to our (local) repositories and, if needed, re-format the hierarchical data easily. Additionally, CEDAR enables tracking of new template versions and provenance of the users who completed individual instances. We have already implemented that functionality with the CEDAR development team for LINCS including a Keycloak-based Single Sign On (SSO) system. CEDAR greatly simplifies the creation of user interfaces to capture controlled vocabulary resource descriptions and facilitates the subsequent processing of that information.

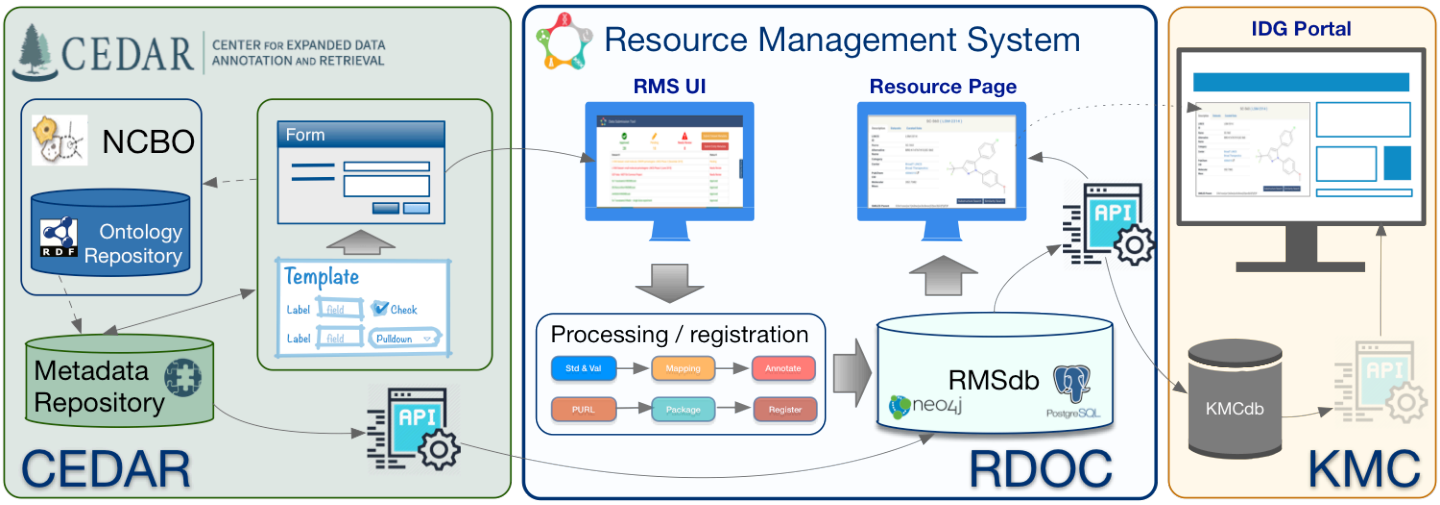

Figure 2. Key components of the RMS (User Interface, processing /registration pipeline, RMSdb, resource landing page and APIs) at RDOC and interfaces (via API) to the CEDAR metadata management environment and the KMC IDG Portal.

4.2: The overall process works as follows. The RDOC will create metadata templates for the specific resource categories (described in Aim 3) and which are linked to controlled vocabularies where possible. RMS guides DRGC scientists through the process of annotating their datasets, regents, and other resources via these (selectable) templates / forms. Metadata are stored in CEDAR and the data are transferred to the RMS using APIs available from CEDAR. We expect validation and standardization of the submitted information will be semi-manual involving RDOC scientist (at least in the initial versions of RMS). RMS will track the status of each submitted resource. As RMS advances, we will add improved functionality to validate and enhance resource annotations. That may not be practical for all types of resources in which case a manual component will remain, but will be manageable with our personnel. RDOC will directly contact the DRGCs if there are issues that need resolution. Once validated, the status of a resource description will be updated to "published" and can then be accessed through the APIs and Landing Pages (see below).

4.3: Data Storage. All data will be stored in a relational database (PostgreSQL [63]); our current system is the LINCS MetaData Repository (MDR). As we advance the system, we will consider a hybrid storage environment containing an additional graph database (Neo4j [64]) to more effectively store and retrieve connected data elements, such as datasets and reagents in different roles (perturbation, model system, analyte) and complex resource annotations for example from ontologies. Proposed hybrid storage may facilitate the integration of IDG-generated resources with the body of scientific resources processed at the KMC; we will of course coordinate our efforts accordingly.

4.4: Services and APIs. A set of services will be written to serve the data to the users (KMC and others). Contextual services will pull information from this hybrid store and create summary documents traversing the data from several starting points that correspond to entities submitted by the DRGCs. The services will be visualized through a set of REST APIs for use by software developers (such as the KMC) who wish to access the information for their own systems. The public set of APIs will be a subset of the full RESTful API suite developed during the project.

4.5: Integration with the KMC. The IDG Portal at the KMC will be able to access the resource descriptions via the RMS services / APIs (Figure 2), making them easily searchable upon indexing. Pharos is the current IDG Portal and could readily interface via the proposed solution (see letter of support Rajarshi Guha, Pharos lead developer at the IDG KMC NCATS site). In addition, via services described above, we propose to create Resource Landing Pages, which would include all content of a resource including the unique global identifier (compare Aim 3). These Landing Pages could be directly incorporated into the IDG Portal. The exact mechanism of integration will be coordinated depending on the KMC technologies and operations.

4.6: Technology. To build RMS we will significantly leverage existing infrastructure and tools we have been developing for a Dataset submission system for LINCS, including user authentication, graphical user interface elements, metadata CEDAR templates specifications, APIs, data (file) storage, transfer protocols, a relational database schema with several integrated data stores, business rules, and various processing pipelines and scripts. Relational database management will be done using PostgreSQL. Neo4j will be used for storing and accessing inferred data. Service components will be written in Java. REST API components will be primarily written in Java and deployed on Apache Tomcat. The web-based user interface will be developed using HTML5 and JavaScript, and will include at least the Angular [65] and Bootstrap [66] JavaScript libraries. Throughout the project period we will remain flexible on our choices for technology and strive to apply the latest best practices and techniques to maximize the longevity and interoperability of the systems developed as part of this project.

OUTREACH AND TRAINING. Outreach for publicizing the work produced by RDOC and enhancing the visibility of the resources developed within the IDG consortium will be carried out by appropriate members of RDOC via conference participation, organization of open symposiums, publications, organizing and participating in webinars, and sharing content on websites and social media. Initial outreach elements will be accomplished in collaboration with GEN, as outlined further in Specific Aim 2. The primary goal for RDOC is to establish a presence for RDOC and IDG consortium on the internet. That mechanism will be utilized for updating the scientific community and public in general on the progress being made within the IDG consortium.

Further outreach will be accomplished by offers of specific training as described in Aim 2 as the ITEP. These educational modules will offer introductory lectures and videos both within the IDG Consortium as well as external to the consortium. Initial workshop and/or symposium will be coordinated with the tasks of developing data sharing policy and establishing metadata annotations. MAC will organize this workshop with NIH collaboration to educate consortium members on current policies on data sharing with follow-up focused on annotation of data being shared. An important element for success is ultimately developing standardization of metadata annotation (FAIR) within the IDG consortium and solicitation of feedback. MAC will lead the development of associated surveys and questionnaires, hosted on our website.

With the development of the Resource Submission System, webinars and workshops will be held to enable IDG consortium members to submit their products. The utility of this system will be extended to non-IDG members in concert with the biennial open symposium to solicit broader acceptance of these data standards.