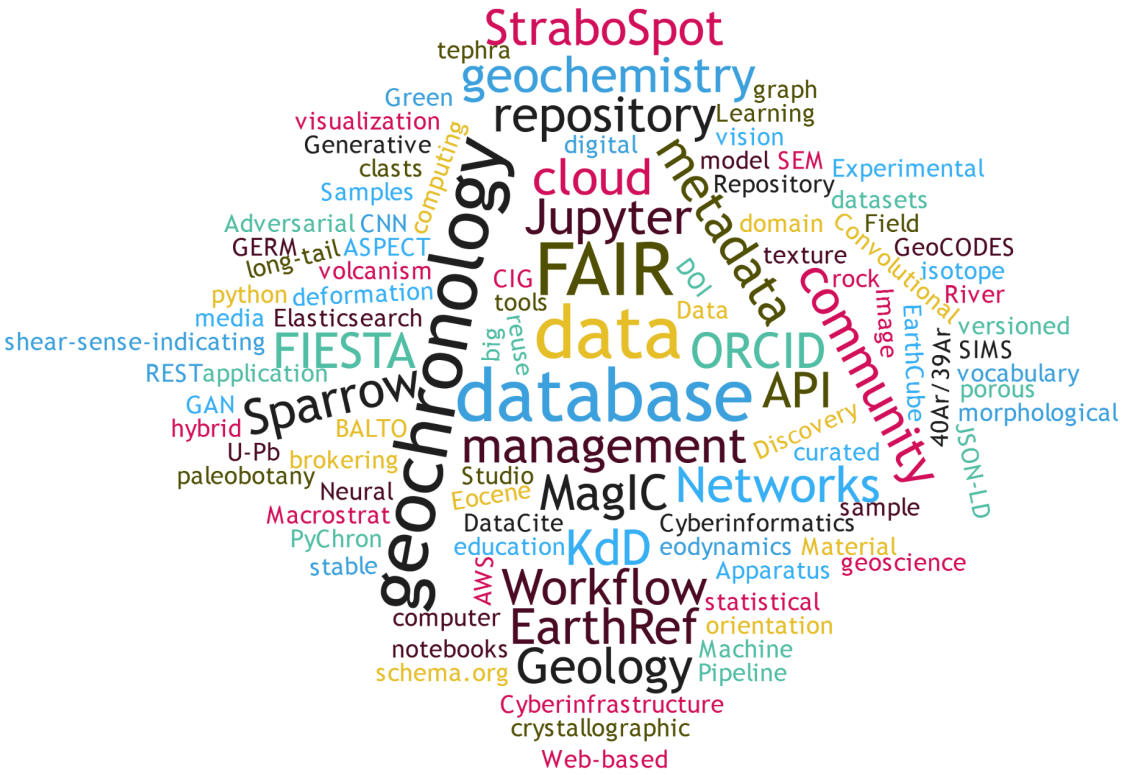
Wednesday Poster session
2021 EarthCube Annual Meeting
Wednesday, June 16, 12-130 PDT



ID	First	Last	Title
1	Stephen	Kuehn	Making tephra data FAIR and connected through community-driven best practices for digital data collection and documentation
2	Nicolas	Roberts	StraboSpot2: A new, intuitive iPad application for the collection of geologic field data
3	Basil	Tikoff	StraboSpot Workshops and Community Engagement During Covid
4	D. Sarah	Stamps	Seamless Access to Long-Tail and Big Data in Earth and Space Science via the EarthCube Brokering Cyberinfrastructure BALTO
5	John	Clyne	Project Pythia: A Community Learning Resource for Geoscientists
6	Stephen	Iota	Automatic Detection and Classification of Rock Microstructures through Machine Learning
7	Joshua	Davis	Quantifying Crystallographic Preferred Orientation Textures in Deformed Rocks

8	Julie	Newman	StraboMicro: Sharing and contextualizing microscope information of rocks
9	Noah	Phillips	Squishing and twisting experimental rock deformation data into the Strabo data system: Progress in developing the StraboExperimental application
10	Sarah	Ramdeen	iSamples (Internet of Samples): Cyberinfrastructure to support transdisciplinary use of material samples
11	Ellen	Currano	Introducing PBot, the Integrative Paleobotany Portal
12	Masa	Prodanovic	Visualization and Reuse Competition for Digital Rocks Portal Datasets
13	Ari'El	Encarnacion	A Pipeline for Automatically Classifying Shear-Sense Indicating Clasts in Photomicrographs
14	Rupert	Minnett	Integrating FIESTA Domain Data Repositories with EarthCube Tools
15	Nicholas	Jarboe	Quick Startup and Long-term Sustainability of Domain Data Repositories Using FIESTA
16	Daven	Quinn	An extensible interface for geochemical data sharing atop the Sparrow laboratory management system
17	Benjamin	Bruck	A Cyber Pipeline for Collection, Management, and Exploration of 40Ar/39Ar and U-Pb Geochronology Data
18	Benjamin	Linzmeier	Connecting a SIMS laboratory to the Sparrow data system

A Cyber Pipeline for Collection, Management, and Exploration of 40Ar/39Ar and U-Pb Geochronology Data

Submitting Author: **Benjamin Bruck, University of Wisconsin-Madison**
Authors: Benjamin Bruck, Daven Quinn, Brad S. Singer, Brian R. Jicha, Jake Ross, Mark Schmitz

Needs for intercalibration and standardization, and the geoscience-wide goal of building integrated, data-driven 4-D digital Earth models, have prompted geochronologists to implement cyberinformatics systems to collect, manage, and share data. The WiscAr Lab is building an integrated pipeline for exposing 40Ar/39Ar data to end-users, including synthetic databases such as Macrostrat, Neotoma, and the Paleobiology Database (PBDB). This workflow integrates the PyChron data-collection and analysis software and the Sparrow data management system.

PyChron links spatial location and stratigraphic data with samples upon receipt; these data are automatically retained throughout the process of sample preparation, irradiation, analysis, and data reduction. PyChron archives analytical and spatial metadata on GitHub in JSON format, which Sparrow automatically ingests through a customizable schema-based importing tool. Sample information is also incorporated into Sparrow through input forms and a data-sheet style bulk editor, which pull descriptors directly from the Macrostrat API for ease of use. The WiscAr Lab Sparrow portal (http://wiscar-sparrow.geoscience.wisc.edu/) provides data interfaces for

lab-level metadata to be accessed by synthetic databases; individual end-users can search, navigate, and visualize radioisotopic age data and metadata.

Most of the 20 year WiscAr data archive has been ingested into Sparrow, providing a rich data set to share with the community and address geologic questions. We present a prototype integration between the WiscAr lab workflow and Macrostrat to develop an age model for the Early Eocene Greater Green River Basin, WY. This model employs recently obtained 40Ar/39Ar dates, U-Pb dates from the parallel Sparrow pipeline developed at the Boise State Isotope Geology Laboratory (IGL), and data surfaced from published literature using GeoDeepDive machine reading tools. This effort provides a template for other regional model comparisons atop the Sparrow and Macrostrat public APIs, and for any dataset tracked by a Sparrow-enabled lab.

Project Pythia: A Community Learning Resource for Geoscientists

Submitting Author: **John Clyne, National Center for Atmospheric Research**
Authors: Anderson Banihirwe, Drew Camron, John Clyne, Orhan Eroglu, Max Grover, Julia Kent, Austin Koorz, Matt Long, Ryan May, Kevin Paul, Brian Rose, Michaela Sizemore, Kevin Tyle and Anissa Zacharias

Scientists working in many geoscience disciplines rely heavily on computing technologies for their research. Numerical simulations, run on supercomputers, are used in the study of climate, weather, atmospheric chemistry, wildfires, space weather, and more. Similarly, a tremendous volume of digital data produced both by simulations, and through observations made with instruments, are analyzed with the help of powerful computers and software. Thus, today's scientists require not only disciplinary expertise, but also high-level technical skills to effectively analyze, manipulate, and make sense of potentially vast volumes of data.

Two technological trends have emerged over the past decade that are having a sizable impact on scientific productivity: the use of Python and the Scientific Python Ecosystem for analysis, and the migration of workflows to cloud computing resources. Making the most effective use of these and other related technologies requires a degree of specialized technical literacy that is rarely found in a geosciences education curriculum. The goal of Project Pythia is to provide a public, web-accessible training resource to help current and aspiring earth scientists learn how to use the Scientific Python Ecosystem and Cloud Computing. Pythia covers a range of topics from beginning Python programming to advanced subjects such as developing scalable workflows. Steered by classroom needs and community demand, Pythia follows an open development model. This poster will provide an update on the current state of Project Pythia, and describe how to get involved with this community-owned effort.

Introducing PBot, the Integrative Paleobotany Portal

Submitting Author: **Ellen Currano, University of Wyoming**
Authors: Ellen Currano, Claire Cleveland, Dori Contreras, Rebecca Koll, Douglas Meredith, Shanan Peters, Mark Uhen, Andrew Zaffos

Paleobotanical data is severely underrepresented in major publicly accessible databases, even though fossil plants represent the best record of ancient terrestrial environments. A major impediment to the inclusion of paleobotanical data in databases at meaningful levels of taxonomy is that plant parts are most often preserved separately, with varying potential for taxonomic resolution. Paleobotanists therefore commonly use morphologically-based, informal taxonomies (morphotypes) rather than traditional Linnaean classifications. The names given to a particular morphotype are inconsistent among research groups, and currently there is no data-management infrastructure that allows comparison and synonymizing of morphotypes among regions or time periods or with published formal taxonomies. As a result, a large proportion of the millions of fossil plant specimens housed in museums worldwide are, together with their spatio-temporal occurrence data, inaccessible for inclusion in studies of paleobiology, paleoclimatology, Earth system modeling, macroevolution, and macroecology.

In 2020, our group received EarthCube funding to address these problems by creating PBot, The Integrative Paleobotany Portal. PBot will consist of an online workbench and graph database that allow the paleobotanical community to (1) create novel, dynamic, community-sourced character schemas for describing plant fossils; (2) enter and browse informal and formal taxonomies via morphological characters; and (3) maintain a community forum for commentary on PBOT contents. It will work with and complement existing databases (iDigBio and the Paleobiology Database) to enhance utility and accessibility of paleobotanical data, allowing paleobotanists to easily fulfill NSF data management plans. Further, the online workbench will provide a standardized resource for fossil plant description and data entry that will benefit students and professionals, as well as fossil enthusiasts. To create a system that is "of the community, by the community, and for the community," we have organized two large virtual workshops to introduce the PBot design and solicit input as we design and refine the system.

A Pipeline for Automatically Classifying Shear-Sense Indicating Clasts in Photomicrographs

Submitting Author: **Ari'El Encarnacion, Sonoma State University**
Authors: Ari'El Encarnacion, Cinthya Rosales, Kevin Drake, Gurman Gill, Matty Mookerjee

We are constructing a machine learning (ML) powered, automated pipeline for classifications and detections of shear-sense-indicating clasts in photomicrographs. Classifications include sinistral (CCW) or dextral (CW) shearing, and detections refer to location of clasts in photomicrographs. Multiple ML tools and techniques are employed in this pipeline. Most tools rely on Convolutional Neural Networks (CNN). CNNs are a type of ML model that imitates the neuron connections in the human brain, allowing for complex shape recognition by extracting image features at different scales and complexity. Our pipeline has 4 stages: Preprocessing, Clast detection, Shear-sense classification, and a user-facing iOS app that displays classification and detection results. Preprocessing includes image denoising using a total variation L1 filter, and image resizing to a standardized width using bilinear interpolation. Clast detections are done using YOLOv5, a CNN based model that outputs a bounding box around a

detected object of interest. Shear-sense classifications are handled via a CNN model supported by transfer learning, a technique leveraging an existing, trained CNN in order to perform classifications with a smaller, secondary custom CNN. CNNs are typically trained with millions of images. However, our base dataset is about 100 images. To supplement data deficiency, we are using NVIDIA's Style-GAN2-ada, a Generative Adversarial Network (GAN). A GAN uses a generator and a discriminator to pass created and critiqued data back and forth, eventually allowing for a vector of noise to generate an image indistinguishable from the original object category. Currently, over 6,000 clast images have been generated. A subset of these images has been used to train InceptionV3, a CNN, for classification of shear-sense clasts: 252 CCW and 437 CW clasts. InceptionV3 was tested on 53 CCW and 117 CW clasts The average accuracy of predictions for CW were 86% and 88% for CCW. YOLOv5 was trained on our base dataset of 92 photomicrographs with bounding boxes of 303 CW and 130 CCW clast tails. YOLOv5 has detected CW & CCW clasts in 10 photomicrographs with 32% and 15% accuracy respectively. Average confidence score of correct detections was 25%. Our iOS app will employ our pipeline in order to provide automatic classification & detections to the user. This will provide users with vital data and feedback provided on app-generated results will benefit our pipeline. Future work includes improving classification and detection accuracy, assembling the pipeline, and opening the iOS app to users.

Automatic Detection and Classification of Rock Microstructures through Machine Learning

Submitting Author: Stephen Iota (University of Southern California)
Authors: Stephen Iota, Junyi Liu, Ming Lyu, Bolong Pan, Xiaoyu Wang, Yolanda Gil, Wael AbdAlmageed (University of Southern California); Gurman Gill, Matty Mookerjee (Sonoma State University)

Geologists need help classifying microscope rock images of sigma clasts; a type of mantled porphyroclasts widely used as kinematic indicators in rocks. Knowledge about the sheer sense of sigma clast during formation (either CCW or CW sheering) gives insights into rock formation history. This work reports on early investigation of machine learning techniques for automatic detection and classification of sigma clasts and their rotation from photomicrographs. Convolutional Neural Networks (CNNs) are used to extract and leverage defining features of sigma clasts, such as shape, color, texture, and tail direction to improve accuracy. We leverage existing models that are pre-trained on very large collections of images, and use transfer learning techniques to apply them to microstructure images. We used YOLOv3 to identify different sigma clasts in a given image. We also experimented with other large pre-trained models such as ResNet50, VGG19, Inceptionv3 with two additional layers trained specifically on our dataset. In order to facilitate exploration of different models with different settings, we are developing a computational experimentation environment to visualize different CNN network layers, classification heatmaps, and comparative metrics. Finally, since models perform better when more data are available, we are developing a web application to collect additional data from geoscientists and incentivize their participation in open science. The website allows researchers to upload images of rock microstructures, showing them the classification of the

images based on the best models available, and allows them to correct any errors which can be used to improve the models.

Quick Startup and Long-term Sustainability of Domain Data Repositories Using FIESTA

Submitting Author: **Nicholas Jarboe, Scripps Institution of Oceanography, UCSD**
Authors: Nicholas Jarboe, Rupert Minnett, Anthony Koppers, Cathy Constable, Lisa Tauxe

As scientists look for FAIR-compliant data repositories for their datasets, they should prefer domain-specific repositories that have a data model compatible with their datasets. Placing data into a singular data model allows for easy reuse and for all datasets to be easily compared and combined in a homogeneous repository. This ease of use is one of the main goals of the FAIR initiative and will facilitate quicker "time to science" and allow new scientific questions to be investigated. Unfortunately, the high cost in time and money of creating domain-specific repositories and their ongoing upkeep has prevented the creation of data repositories for many, if not most, scientific domains. We have created a software stack called FIESTA (Framework for Integrated Earth Science and Technology Applications) that requires an order of magnitude less effort and less cost for implementation and maintenance of a FAIR-compliant repository than creating one from scratch (https://earthref.org/FIESTA). FIESTA features include minting of data DOIs, ORCID iDs for authentication and author identification, a sophisticated data search/filtering system, data versioning, EarthCube compatible schema.org/JSON-LD metadata contribution headers, web and API data uploading and downloading, and many other features. The main work to be done is deciding on a data model that will work for the community when the data repository is being created. We have two data repositories implemented that are using FIESTA: MagIC (https://earthref.org/MagIC), a rock and paleomagnetic data repository; and KdD (https://earthref.org/KdD) for partition coefficients. We also are actively pursuing the development of FIESTA repositories for 40Ar/39Ar dating, geologic cores, and geomagnetic models. Please see our companion poster "Integrating FIESTA Domain Data Repositories with EarthCube Tools" for information on the technical details of the FIESTA system.

Making tephra data FAIR and connected through community-driven best practices for digital data collection and documentation

Submitting Author: **Stephen Kuehn, Concord University**
Authors: Stephen Kuehn, Marcus Bursik, Simon Goring, Samuel Kodama, Andrei Kurbatov, Kerstin Lehnert, Lucia Profeta, Sarah Ramdeen, Kristi Wallace, Douglas Walker, Daven Quinn

Tephra, or fragmental material ejected from volcanoes, has a unique array of applications in science. Research into tephra production, distribution and characterization supports the use of tephra as a tool for global, interdisciplinary research, and tephra deposits are of critical use to geoscientists as widespread time stratigraphic markers and indicators of past volcanic activity. The interdisciplinary use of tephra as a research object has exacerbated issues related to data access, integration, and cross-disciplinary data use created by incomplete documentation and a lack of standardization within the broader community. Simultaneously, the explosion of

cryptotephra studies has made traditional regional and lab specific data compilations insufficient for many applications because the potential volcanic sources that must now be considered may be spread across an entire hemisphere or further.

Recommendations for best practices in tephra studies, from sample collection through analysis and data reporting (https://zenodo.org/record/4075613) have been developed through EarthCube funded activities. These are being incorporated into digital tools and data repositories to improve the effectiveness of data sharing among the diverse community of tephra researchers and users. Here we report on 1) a new set of templates for samples, methods, and data reporting, 2) a new tephra module in the StraboSpot field app, and 3) new implementations at SESAR and EarthChem, including a tephra community portal. We also report on re-usable analytical method descriptors, now implemented at EarthChem, and the association of sample and quality control reference material data for individual analytical sessions. Data linking is facilitated by extensive use of unique identifiers including ORCIDs for people, IGSNs for field sites and samples; DOIs for publications, data, and methods; and Smithsonian IDs for volcanoes and eruptions. These developments allow users to follow simple workflows to archive data and facilitate faster access to key research by secondary users.

Connecting a SIMS laboratory to the Sparrow data system

Submitting Author: **Benjamin Linzmeier, University of Wisconsin - Madison**
Authors: Benjamin J. Linzmeier1,2, Daven Quinn1, Casey Idzikowski1, Shanan E. Peters1, Kouki Kitajima1,2, Noriko Kita1,2, Chloë Bonamici1,2, and John W. Valley1,2
(1) Department of Geoscience, University of Wisconsin – Madison, Madison, WI 53706
(2) WiscSIMS Laboratory, University of Wisconsin – Madison, Madison, WI 53706

Secondary ion mass spectrometry (SIMS) produces abundant, precise measurements of stable isotope ratios from a suite of solid materials. Spatial precision (1-10 μm spots) and integration with petrographic images make SIMS uniquely suited for investigating the composition of small, rare, or zoned materials. During a typical analytical session for light stable isotopes (δ18O and δ13C), hundreds of points on sample and standard materials are measured. Data-reduction involves correcting for instrumental mass fractionation using the measurements of standards that bracket samples. Currently, data are reduced in Excel workbooks and shared as supplemental datasets with papers.

The Sparrow data system (sparrow-data.org) is a flexible, integrated PostgreSQL database, webserver, and JavaScript user interface platform designed for combining, managing, and distributing analytical data and sample metadata to end-user applications. A critical step in populating the Sparrow platform is cleaning datasets from diverse Excel-based data reduction workflows.

Here we share our procedures for automatically parsing and importing the data archive from the WiscSIMS laboratory. Import procedures using the R scripting language include standardizing column names, classifying analyses as standards or samples using regular expressions, identifying standard-sample-standard brackets, calculating bracket statistics, and creating JSON

consistent with Sparrow's structured representations. Import is implemented in a Docker container that writes data into the database using Sparrow's API. Our importer is flexible for the addition of new analytical methods and standards. Future development of Sparrow for in situ analyses will focus on integrating data from various instruments (e.g. SEM, EPMA, Raman) with SIMS datasets. Current integrations of these data use map-making in QGIS, which is particularly important for aligning and integrating images with point data across instruments built by multiple manufacturers. A particular focus in the future will be creating a data reduction platform for more complex datasets and sharing images that are important for providing petrological context.

Integrating FIESTA Domain Data Repositories with EarthCube Tools

Submitting Author: **Rupert Minnett, Oregon State University**
Authors: Rupert Minnett, Nicholas Jarboe, Anthony Koppers, Cathy Constable, Lisa Tauxe

Successful domain data repositories must find a delicate balance between specificity to a scientific community and sufficient usage within and outside these communities to justify the effort of creating and maintaining the repository. However, reducing this effort makes establishing domain-specific repositories for smaller scientific communities far more viable. EarthRef.org repurposed many of the components developed for the Magnetics Information Consortium (https://earthref.org/MagIC) to create the Framework for Integrated Earth Science and Technology Applications (https://earthref.org/FIESTA) to reduce the barriers in deploying a FAIR-compliant data repository that is tailored to a specific scientific domain yet readily integrated with existing EarthCube tools and other services. FIESTA repositories are deployed to EarthrRef.org's hybrid cloud infrastructure at Oregon State University and Amazon Web Services where scalable instances are managed along with redundant and persistent elastic storage. Each repository is configured with a single hierarchical data model including field validation resulting in support for deep indexing of contributed datasets and subset search results for download. Users are authenticated with EarthRef's ORCID Member subscription and given access to EarthRef's JupyterHub (https://jupyterhub.earthref.org) for online dataset processing and interacting with FIESTA's REST API (https://api.earthref.org). Versioned datasets are constructed in a private workspace with share links for collaborators and reviewers, minted with DataCite DOIs, and annotated using JSON-LD structured metadata with schema.org vocabularies that are crawled by several search engines including EarthCube's GeoCODES, EarthCube's Data Discovery Studio, EarthCube's Macrostrat, EPOS-MSL's Data Catalog, and Google's Dataset Search. As a proof of concept, two repositories for distinct communities are currently deployed with FIESTA: MagIC (https://earthref.org/MagIC) for paleo- and rock-magnetism, and KdD (https://earthref.org/KdD) for partition coefficients. Also, we are pursuing the development of FIESTA repositories for 40Ar/39Ar dating, geologic cores, and geomagnetic models. Please see our companion poster "Quick Startup and Long-term Sustainability of Domain Data Repositories Using FIESTA" for additional information about FIESTA.

Quantifying Crystallographic Preferred Orientation Textures in Deformed Rocks

Submitting Author: **Matty Mookerjee, Sonoma State University**
Authors: Joshua R. Davis & Matty Mookerjee

Geologists frequently want to infer deformation geometry and intensity from patterns of crystallographic preferred orientation (CPO) in deformed rock. We describe computational tools to support this kind of inference, with an emphasis on statistical rigor and handling a wide range of deformation models. We forward-model the development of quartz CPO using the Taylor-Bishop-Hill theory, starting from an initially uniform orientation distribution. We quantify the final CPO by partitioning orientation space into bins and counting the orientations in each bin. Based on the multinomial distribution, we define a likelihood function that compares an observed CPO to a predicted CPO. This function can then be used in maximum likelihood estimation or Bayesian Markov chain Monte Carlo simulation, to fit the model to the data and to assess the uncertainty in that fit. We test this approach on synthetic data sets with realistic uncertainty, and discuss the effects of discretization error and sampling error.

StraboMicro: Sharing and contextualizing microscope information of rocks

Submitting Author: **Julie Newman, Department of Geology and Geophysics, Texas A&M University**
Authors: Julie Newman, Randy Williams, Jason Ash, Nick Roberts, Noah Phillips, Alex Lusk, Basil Tikoff

StraboMicro provides tools for image management and data reporting for geologic data that is observed and analyzed at the microstructural scale (e.g., from thin sections). StraboMicro is part of StraboSpot (StraboSpot.org), a digital datasystem for geologic data from the macro- to the microscale, including field, laboratory, and experimental rock deformation.

The first phase of StraboMicro development, now in beta testing, is an image management system with basic data collection capabilities. The application includes the ability to add images from any source (e.g., optical microscope, scanning or transmission electron microscope) to a project, complete with metadata concerning instrument type and operating conditions. For each image, the user may define the scale and the orientation (i.e. field-oriented thin sections). Any image added to the system can be used as a basemap, on which other images may be overlain, scaled, and rotated. The basemap has a coordinate system that relates all microscale data to geographic coordinates (for field samples) or instrument coordinates (for experimental samples). Scale is tracked between images arranged in this hierarchical manner, and it is possible to zoom in and out of images. Any image or spot can have associated data files, which can link to other databases (e.g., EarthChem) or any website.

The second phase of development of StraboMicro will incorporate the comprehensive vocabulary, metadata, and workflows for microstructures determined through an iterative process with the community.

As StraboMicro is part of the StraboSpot ecosystem, data from microscale analyses is immediately tied to the field context from which a natural sample originated, including georeferencing, scale, and orientation. StraboExperimental, an app currently in development,

will allow microstructural data from a sample derived from rock deformation experiments to be tied to its experimental context. By working within one data system, images from nature and experiment can be simultaneously searched and compared.

Squishing and twisting experimental rock deformation data into the Strabo data system: Progress in developing the StraboExperimental application

Submitting Author: **Noah Phillips, Texas A&M University**
Authors: Phillips, Noah ; Mok, Ulrich ; Ash, Jason ; Kronenberg, Andreas ; Pec, Matej ; Cunningham, Hannah ; Newman, Julie ; Tikoff, Basil ; Walker, Doug

StraboExperimental is a web-based application in development which facilitates storing, sharing, visualizing, and querying of data produced through rock deformation experiments. Like other Strabo applications, StraboExperimental uses a systematic vocabulary developed by the experimental community which facilitates querying data. We have developed an apparatus repository where metadata for individual deformation apparatuses is housed and can be quickly accessed to greatly increase the efficiency of uploading data from individual experiments into the application. The repository records apparatus capabilities, schematics of the apparatus, types of sensors and their locations, calibrations for sensors (with the ability to upload calibration files), and a list of test types that each apparatus can perform. Large time-series data files produced by individual experiments can be quickly input using prescribed headers in the data files. Future development will allow users to query and plot data from various sources within the application. StraboExperimental will be seamlessly linked to projects in StraboMicro when microstructural observations of experimental run products are made, and with StraboSpot when deformation experiments are performed on natural samples collected in the field. StraboExperimental is being developed in tandem with LAPS (Laboratory Acquisition Protocols and Standards), a workflow for lab managers to produce data which can be easily uploaded into data repositories. Use of StraboExperimental and LAPS by the community will make experimental rock deformation data increasingly FAIR (findable, accessible, interoperable, and reusable).

Visualization and Reuse Competition for Digital Rocks Portal Datasets

Submitting Author: **Masa Prodanovic, The University of Texas at Austin**
Authors: Masa Prodanovic, James McClure

We present the porous media visualization challenge: the organization, short course and related Jupyter notebooks and the winning entries for our Porous Media Visualization and Data Reuse Challenge organized September 2020-January 2021. The task was to reuse any 3D dataset from the Digital Rocks Portal to create a static image, video, or 3D printed visualization. Porous media are challenging to visualize due to complexity of pore/grain/fluid surfaces and interfaces, and this creates a resource for those who wish to learned advanced 3D visualization.

An extensible interface for geochemical data sharing atop the Sparrow laboratory management system

Submitting Author: **Daven Quinn, University of Wisconsin – Madison**
Authors: QUINN, Daven P.1, LINZMEIER, Benjamin J.1, SUNDELL, Kurt2, GEHRELS, George E.2, GORING, Simon3, MARCOTT, Shaun A.1, MEYERS, Stephen R.1, PETERS, Shanan E.1, ROSS, Jake4, SCHMITZ, Mark D.5, SINGER, Brad S.1 and WILLIAMS, John W.3
(1)Department of Geoscience, University of Wisconsin – Madison, Madison, WI, (2)Department of Geosciences, University of Arizona, Tucson, AZ, (3)Department of Geography, University of Wisconsin – Madison, Madison, WI, (4)New Mexico Geochronology Research Laboratory, New Mexico Bureau of Geology and Mineral Resources, Socorro, NM 87801, (5)Department of Geosciences, Boise State University, Boise, ID

Geochronology and geochemistry laboratories maintain archives of analytical data that quantify the Earth's rock record. En masse assessment of these archival datasets is critical for building repeatable, comparable, and robust workflows. Moreover, a new generation of geoscience databases (e.g., Neotoma, the Paleobiology Database) and digital Earth models (e.g., Macrostrat) aim to integrate geochemistry and geochronology into assessments of global change. This requires large sets of contextualized and interpreted digital data.

Sparrow (https://sparrow-data.org) is an information management system that augments existing lab analytical workflows with tools to harmonize data structures and manage metadata (e.g., projects, sample context, and embargo). The system is now being used or trialed as a management layer by ten labs across geochemical and geochronological domains including detrital U-Pb, 40Ar/39Ar, SIMS, cosmogenic nuclides, and optically stimulated luminescence.

The common structures that Sparrow distills from analytical data provide a starting point for interfaces with community-level systems. However, each external integration requires a different subset of information (e.g., aliquot-level data for archiving with Geochron.org, or age determinations for Macrostrat). The distance in structure and specificity between internal lab datasets and community-relevant summaries presents an obstacle to data propagation.

Here, we introduce new software pipelines and user interfaces within Sparrow designed to help align lab data types to externally-recognized vocabularies and schema definitions. Schemas specify single values (e.g., an age estimate) or collections of related information (e.g., a sample's trace-element profile); they can be custom-built or derive from existing resources (e.g., USGS or EarthChem vocabularies). Schema-controlled output conforms to standard protocols (e.g., JSON-LD) and can be accessed using Sparrow's external data interfaces.

Next, we will use these tools to support operational flows of geochemical and geochronology data from labs to synthesis databases. This will be the first step towards an envisioned ecosystem of data integrations throughout geoscience communities.

iSamples (Internet of Samples): Cyberinfrastructure to support transdisciplinary use of material samples

Submitting Author: **Sarah Ramdeen, Columbia University**

Authors: Kerstin Lehnert, Columbia University 0000-0001-7036-1977
Sarah Ramdeen, Columbia University, 0000-0003-1135-5942
Neil Davies, University of California, Berkeley, 0000-0001-8085-5014
John Deck, University of California, Berkeley, 0000-0002-5905-1617
Eric Kansa, Open Context, The Alexandria Archive Institute, 0000-0001-5620-4764
Sarah Kansa, Open Context, The Alexandria Archive Institute, 0000-0001-7920-5321
John Kunze, Technology Consultant, California Digital Library, 0000-0001-7604-8041
Christopher Meyer, Smithsonian Institution, 0000-0003-2501-7952
Thomas Orrell, Smithsonian Institution, 0000-0003-1038-3028
Steven Richards, Independent contractor
Rebecca Snyder, Smithsonian Institution, 0000-0002-0028-6139
Ramona Walls, University of Arizona 0000-0001-8815-0078
David Vieglais, University of Kansas Biodiversity Institute, 0000-0002-6513-4996

Research in many Earth Science domains depends on material samples collected in the field, generated in experiments, or retrieved from space. There are often great costs associated with the capture of samples, which might be rare, unique, or irreplaceable. In 2014, EarthCube funded iSAmplEs: The Internet of Samples in the Earth Sciences. This project "promoted best practices and standards for sample identification, documentation, citation, curation, and sharing ... as a foundation to building a shared cyber-infrastructure". Based on the outcomes, a multidisciplinary research collaboration emerged to address these issues across the broader Natural Sciences including biology and anthropology. The resulting "Internet of Samples (iSamples)" is funded to develop a national cyberinfrastructure for samples. This cyberinfrastructure will provide services to uniquely, consistently, and conveniently identify material samples, record metadata, and link them to other samples, derived data, and research results published in the scientific literature. It aims to serve all research involving material samples from natural and built environments.

iSamples has two components: iSamples-in-a-Box (iSB), a standalone system that enables creation of identifiers and their associated metadata, metadata management, sample identifier resolution, and discovery of samples; and iSamples Central (iSC), a permanent Internet service that preserves and indexes sample metadata to ensure reliable discovery and retrieval, and that provides a gateway between iSB instances and identifier authorities to ensure that remote iSB content is fully synchronized with the relevant authorities (e.g., IGSNs generated on iSB are synchronized with iSC and the IGSN central authority). By providing these services that augment existing identifier authority capabilities, iSC enables support of other identifier types beyond IGSN (e.g., ARKs, DOIs).

iSamples will coordinate with the newly funded Sampling Nature RCN, which aims to build a community of collaborative researchers to exploit the potential of an accessible, integrated corpus of material sample data.

StraboSpot2: A new, intuitive iPad application for the collection of geologic field data

Submitting Author: **Nicolas Roberts, University of Wisconsin–Madison**
Authors: Nicolas M. Roberts, Doug Walker, Jessica Novak, Nathan Novak, Sai Damaraju, Basil Tikoff, Julie Newman, Alexander Lusk

The StraboSpot database allows users to permanently store and query geologic field data directly from a mobile application or through a web application at strabospot.org. The first version of the mobile application, StraboSpot, has many of the features needed to collect data in the field, including offline basemaps, a developed vocabulary, and a built-in compass. However, it was intended for use on small smartphones, and users found the form-style user interface inefficient and cumbersome to use. The newly developed StraboSpot2 packages the same core functionality of the original Strabospot into an intuitive, simple design that greatly increases data collection efficiency. Because the target device in an iPad (of any size), we were able to take a map-based approach that fits multiple different types of field workflows and expands the ability for the user to view and explore datasets. New core features include: 1) Color-coding of points and polygons by customizable geological unit tags or conceptual tags; 2) Viewing all data collected at a spot in a "notebook view"; 3) Making sketches or annotate photographs; 4) Filtering spots, images, or tags by map extent; and 5) Facilitating the back up of datasets using integration with the iPadOS file system. StraboSpot2 is developed in React Native, a cross-platform development environment. This development environment will allow future versions to possibly support iPhone (iOS) and Android devices.

Seamless Access to Long-Tail and Big Data in Earth and Space Science via the EarthCube Brokering Cyberinfrastructure BALTO

Submitting Author: **D. Sarah Stamps, Virginia Tech**
Authors: D. Sarah Stamps, James Gallagher, Scott Peckham, Anne Sheehan, Nathan Potter, Kodi Neumiller, Emmanuel Njinju, Maria Stoica, Zachary M. Easton, Daniel R. Fuka, Dave Fulker, Hongda Wang

The EarthCube brokering cyberinfrastructure BALTO (Brokered Alignment of Long-Tail Observations) provides streamlined access to both long-tail and big data using Web Services. BALTO consists of several distinct brokering mechanisms that we have developed since the project's inception in 2016. First, we created an extension for the OPeNDAP framework Hyrax, which is software that serves big data from USGS, NASA, and numerous other sources. The BALTO-Hyrax extension makes the big data findable and searchable through Google Dataset Search and EarthCube GeoCODES by automatically tagging dataset landing pages with JSON-LD encoding. Second, we implemented a Web Service brokering capability into the NSF mantle convection code ASPECT such that geodynamic modelers can remotely access initial condition datasets with URLs. Third, we developed methods to allow any career or citizen scientist to make their in-situ IoT based sensor data collection efforts available to the world, and extend the access of these data for IA/ML analysis. Finally, we developed a Jupyter Notebook with a GUI that allows for users to search Hyrax servers for big datasets and long-tail data. These developments comprise the entire EarthCube brokering cyberinfrastructure BALTO. We recently completed a series of online workshops to train individuals to use BALTO cyberinfrastructure. All workshop materials are available on our website at https://sites.google.com/vt.edu/balto/home. In this presentation, we provide an overview of each component of the BALTO cyberinfrastructure.

StraboSpot Workshops and Community Engagement During Covid

Submitting Author: **Basil Tikoff, University of Wisconsin - Madison**
Authors: Basil Tikoff, Julie Newman, Doug Walker, Randy Williams, Nick Roberts, Alex Lusk, Noah Phillips, Marjorie A. Chan, Casey Duncan, Liz Hajek, Diane Kamola, Stacia Gordon, Frank Spear, Michael Williams

StraboSpot (StraboSpot.org) is a digital data system for collection, management and sharing of geologic data from the macro- to the microscale. Due to the pandemic, it was not possible to organize field-based workshops and fieldtrips in Summer 2020 for demonstrating and eliciting community feedback on the StraboSpot data system. Rather, we decided to run a series of week-long virtual workshops to familiarize practioners with the StraboSpot data system. Workshops were organized by discipline (sedimentology, structural geology, and petrology), each with approximately 100 participants, including U.S.-based and foreign scientists. The daily structure of the workshop was a 2-3 hour block with presentations from several StraboSpot team members, followed by an afternoon help desk. We reassigned participant support costs (originally intended for field-based meetings) to provide a small ($200-250) stipend to U.S.-based participants if they agreed to fully input a data set into StraboSpot and make it publicly available. This approach to increasing engagement – which was not part of original dissemination plan – was extremely successful. Workshop participants who signed up for the stipend were generally very engaged in the data system, provided important feedback on the design and vocabulary, and are highly likely to use StraboSpot in the future. Further, it helped to populate the database. People who received the stipend also presented their StraboSpot projects to the larger group on the last day of the workshop, often with an impressive data set. Because of Covid restrictions on fieldwork in summer 2021, we intend to continue providing stipends for young professionals (graduate students, post-doctoral fellows) to use the StraboSpot system for their thesis or project-based fieldwork. In addition to getting detailed feedback, we hope these testers will become a cohort of motivated users that will teach and encourage others to use the StraboSpot data system.