G.E.M.S™: An Innovative Agroinformatics Data Discovery and Analysis Platform

Graham Allan, Jesse Erdmann, Andrew Gustafson, Alison Joglekar, Michael Milligan, Getiria Onsongo, Keerthi Pamulaparthy, Philip Pardey, Tom Prather, Senait Senay, Kevin Silverstein, James Wilgenbusch, Ying Zhang, and Peng Zhou[1]

Overview

The College of Food Agriculture and Natural Resource Sciences (CFANS) and the Minnesota Supercomputing Institute (MSI) at the University of Minnesota have merged domain expertise in the food and agricultural sciences with HPC and bioinformatics expertise to drive the development of a next-generation agroinformatics data discovery and analysis platform dubbed G.E.M.S™. The G.E.M.S™ platform underpins the International Agro*informatics* Alliance (IAA), a strategic partnership among public and private entities worldwide. Formed in September 2015, IAA is driven to improve agricultural productivity and sustainability by forming an agile and engaged community of research practice and by minimizing obstacles to combining and analyzing large and diverse agricultural related data sets via the use of advanced curation and analysis techniques. Since its inception, the IAA has been actively involved in two distinct, but critically important, development efforts to advance its goals. The first is to create a partnership in which its member strengths create synergies greater than the sum of their individual parts. The second is to help facilitate the development of a technical data management (GEMShare™) and analysis (GEMSTools™) platform that minimizes barriers to collaboration and discovery by giving data owners fine-grained control over what data and metadata they share, when they share these data, and with whom.

Applications at All Scales

The availability of newer and larger sources of interoperable **G**enetic, **E**nvironmental, **M**anagement, and **S**ocioeconomic data (collectively referred to as G.E.M.S) related to agriculture allows researchers to solve problems at multiple functional, temporal and spatial scales related to crop sustainability and food production. For example, production of maize (corn), a staple food crop in tropical sub-Saharan Africa, is strongly affected by the interaction of the crop variety's genetics (e.g., pathogen resistance factors, drought susceptibility factors) and the environment (e.g., rainfall, temperature, pathogen exposure), but also field management decisions (e.g., irrigation, crop rotations with legumes) and socioeconomic factors (e.g., distance from farm to a viable market, market acceptance of white vs yellow corn). Breeding decisions, varietal deployment, and even government policies can markedly benefit from taking into consideration the collective interaction of G *x* E *x* M *x* S data to elicit sustainable improvements in agricultural production. More comprehensive information derived from this multi-dimensional data matrix provides a powerful arsenal for strategic decision makers (be they technology developers, users, or regulators), pointing to a logical course of action for a whole host of agronomic and related production decisions.

While these "Grand Challenge" -type problems help to advance the field by causing us to rethink old paradigms, there are also an abundance of specific and more granular challenges that if solved will advance the goals of the Alliance and pave the practical way forward to solving larger "Grand Challenge" -type problems. Via the development and deployment of G.E.M.S™, the IAA is strategically targeting

[1] Authors listed in alphabetic order and are part of the College of Food Agriculture and Natural Resource Sciences (CFANS) and the Minnesota Supercomputing Institute (MSI) at the University of Minnesota, Twin Cities.

these discrete problems in such a way that Alliance partners can realize more immediate benefits while at the same time developing the framework and foundations needed to advance our understanding of the larger and longer term challenges. IAA's modus operandi is to target portions or discrete components of the grand matrix in a directed way to solve real-world problems, but always with an eye to the interoperability of these components. For example, consider recent converging trends in the malting, brewing and local food movements, and the consequences they have for barley breeding. Micro-breweries are "popping up" all over the country while at the same time there is increasing pressure for sustainably sourced, local foods and beverages. And yet, barley fields and malthouses are traditionally located in a limited number of places nationwide. Suddenly, there is a strong economic imperative for planting barley and locating malthouses in numerous new places near new micro-breweries. But what varieties are compatible with each location, and where should the malthouses be? These varietal breeding and deployment problems can be solved by analyzing long histories of detailed field trial data, which shows the most suitable environment types for each variety of barley, and matching that profile to the production environments that surround each microbrewery.

Beyond the Big Data—Open Data Paradigm

The 2012 White House Big Data Initiative[2] focused the Nation's attention on the problems and transformative benefits of new and fast-growing digital data sources and repositories. Many of the initial efforts to reap the benefits of these new digital data focused on "mining" open access data archives; thereby, sidestepping the more challenging issues of accessing private data, where a myriad of intellectual property rights and data use agreement matters greatly encumber progress. Agricultural R&D is increasingly performed by private entities, who now account for two of every three dollars spent on food and agricultural R&D in the United States (Pardey et al. 2016). The same shift towards more private participation is also occurring in other larger agricultural economies elsewhere in the world. Researchers in public institutions are often also skittish about sharing data, especially in the early (perhaps pre-publication) stages of their research. GEMShare™ is designed to explicitly confront the difficult challenge of data sharing by enabling the data provider to determine which data gets accessed by who, when. GEMShare™ addresses these hitherto, often intractable, data sharing issues related to intellectual property by providing practical technical solutions that open up new ways for research to realize the benefits of the "Big Data" revolution in agriculture.

The "international" dimension of IAA is also a critical design feature of the Alliance. Over the past 50 years the United States has lost major market share in public agricultural research carried out by the USDA and the state land grant universities: 20% of the public global agricultural R&D spend in 1961, to just 11% in 2011. Moreover, solving the challenges in securing sustainable food supplies when most of the 2 million additional mouths to be fed by 2050 will reside in sub-Saharan Africa and south Asia is a global challenge requiring focused international collaboration. IAA is already attracting key partners with global reach and a desire to tackle the local *and* global problems confronting food and agricultural systems.

[2]https://obamawhitehouse.archives.gov/the-press-office/2015/11/19/release-obama-administration-unveils-big-data-initiative-announces-200 .

The Right Data, the Right Tech, the Right Time

Transformative approaches to research depend on the juxtaposition of new data sources and new advances in supporting technologies that when taken together create an enabling digital ecosystem. Right now is that transformative moment for informatics pertaining to food and agriculture (that we dub, agroinformatics). For example, international agencies such as the CGIAR (a consortium of 15 international agricultural research centers located throughout the world) hold more than one million varietal accessions of the world's food crops (corn, cassava, wheat, rice, potatoes and more) in their gene banks. These CGIAR centers have begun sequencing the collections in their gene banks to complement measurements of yield and quality traits on those same varieties for crop trials conducted across the world. Such trials may record explicit management decisions (e.g., managed drought, low nitrogen, crop rotations) that directly affect those yields and quality traits. Concordantly, vast arrays of geospatially recorded environmental data that also influences these measurements is being collected (e.g., remote sensing data from MODIS, Landsat and Sentinel A&B, plus soil quality measures). What's more, geospatial maps of socioeconomic factors (e.g., effective time to market for any pixel on the planet to the nearest market of 20,000 people, fertilizer and irrigation adoption, spatial crop distribution maps) serve as a market-based reality check on the deployment and (environmental and economic) stewardship of new crop varieties and associate agronomic technologies. The informatic products that are enabled by integrating all these G *x* E *x* M *x* S data of varying temporal and spatial scales can substantially expand and accelerate the movement of new agricultural innovations along the entire technology value chain from experimental to farmer fields and then on to markets. The G.E.M.S™ platform is also able to leverage a number of relatively recent technological innovations. A description of these technologies and independent development efforts being used to develop these data products is given in the following section.

Implementation

A rich set of mature open source technologies now exists giving us a sustainable way to aggregate, secure, curate, federate, selectively share, and analyze these large and diverse data without having to build the entire ecosystem of required tools and technologies from scratch (Fig. 1). Also, by leveraging the work of other active developer communities in areas related to user authentication, secure data transfers, federated data access controls, big data analysis, analysis frameworks, and large scale data storage, we're not only able to accelerate IAA's development goals, but this approach will also help us to sustain this effort and extend this platform beyond this particular alliance. What follows is a brief description of the key function components of the platform.

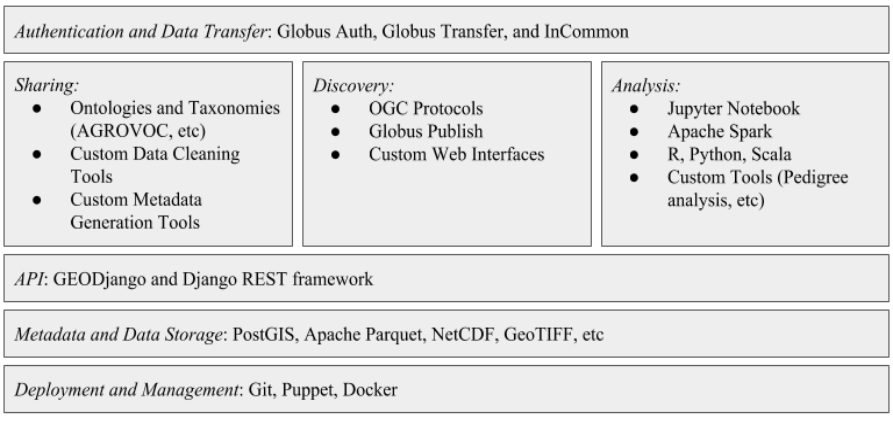

Figure 1: High-level schematic of the G.E.M.S™ architecture.

Authentication and Data Transfer: GEMShare™ uses a flexible authentication scheme in which users from many different institutions are able to easily login and transfer data. To this end, the Globus suite of technologies is leveraged for both authentication and data transfer. Globus provides a web-based federated authentication mechanism, in which users are able to login using their native institution credentials. Globus also provides a fast and reliable data transfer service, allowing users to perform scheduled high speed data transfers.

Sharing: The IAA is sensitive to the privacy concerns of its partners, and seeks to provide granular data permissions, so that collaborators may choose precisely which portions of their data sets are visible to which of their collaborators. GEMShare™ provides users with a framework in which they may choose to share particular data subsets with other users, while keeping other data sets private. Data interoperability is also key to GEMShare™, therefore we are providing custom data cleaning and metadata generation tools in conjunction with community of practice defined ontologies to standardize data as it is ingested.

Discovery: Data exploration is an important component of modern scientific research. GEMShare™ provides users with effective data exploration tools, including an intuitive exploration interface, which will allow users to discover correlations or research questions that they may not previously have expected. Providing OGC compatible catalog and map services allow existing tools such as QGIS and ArcGIS to explore the available data in familiar tools. Globus Publish allows users to share their data beyond users of the federated system with metadata to help them discover it.

Analysis: GEMShare™ supports GEMSTools™, an expanding set of modularized tools allowing both advanced and novice users to effectively analyze data sets to advance their research. The center of the GEMSTools™ analysis platform is a web based Jupyter analysis platform that allows users to run Python, Scala, and R analysis pipelines with Apache Spark, and save such pipelines in analysis notebooks. They will be able to leverage Spark's big data features to scale their pipelines. As an example, Spark-SQL enables seamless integration of programs with SQL queries. Preliminary results show Spark is significantly faster and more efficient for retrieving portions of genotype data compared to a relational database such as PostgreSQL, even with advanced database optimization techniques such as clustered indexes. Additionally, Spark can reduce storage requirement by up to a factor of three when Apache Parquet, a big data columnar storage format, is used to store genotype data. Advanced users are able to design their own pipelines, while novice users may employ some already existing pipelines created by IAA data scientists.

API: Modern research platforms provide an "Application Programming Interface" (API) layer as a powerful component to allow for external computer interaction. The GEMShare™ platform seeks to provide an accessible API interface. Django is a Python web framework with GIS and REST extensions which GEMShare™ uses to provide an API interface layer. All GEMShare™ applications use the API layer to ensure consistent operational behavior and to facilitate enforcement of authorizations. This approach also empowers third-party entities to create their own interfaces to extend the platform to meet their specific needs.

Metadata and Data Storage: Metadata search tools allow researchers to quickly locate data sets of interest. GEMShare™ uses PostGIS to store metadata, making it easy to query with standard SQL and GIS extensions. The data is stored on disk in a format fitting the type of data with an eye toward compact and performant storage. Once a data set is discovered via the metadata it can be read from disk for analysis, or transferred to another location. Where only metadata are shared, work can begin to develop appropriate data use agreements.

Deployment and Management: GEMShare™ relies on DevOps best practices for continuous delivery. As developers commit code to Git repositories automated tests are executed. Successful tests result in the build of new Docker images for development, integration, and deployment. Puppet allows for modular configuration of deployed containers, virtual machines, and hardware to ensure expected behavior, and facilitate platform portability and federation efforts.

IAA Membership, Governance and Operation

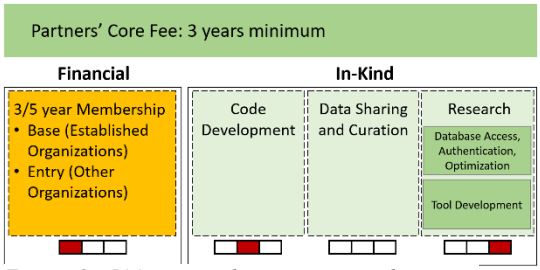

Figure 2: IAA partnership requires a three year partner's fee to cover general operating costs. In addition, some combination of financial and/or in-kind contributions must reach a level in order for a prospective partner's application to be considered by the Partner Committee (PC)..

G.E.M.S™ is designed to support an expanding portfolio of clients and partnerships. IAA has a portfolio of strategic public and private partners who make multi-year in-cash and in-kind commitments to a collectively governed Alliance (Fig.2). Alliance partners with domestic and international interests bring different, and by intent, complementary sets of informatics and domain expertise to the Allaince, the integration of which is key to the success of the IAA and the G.E.M.S™ platform they are helping to shape and focus (Fig. 3).

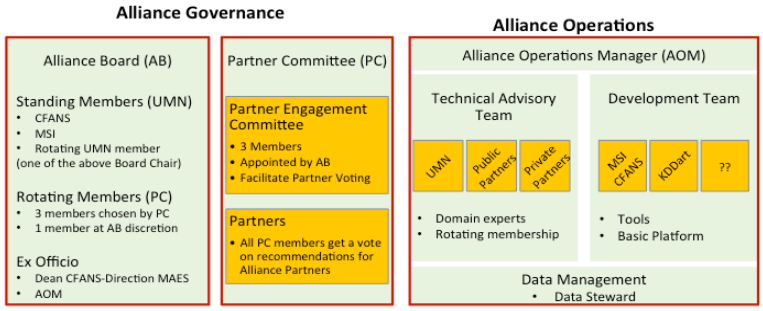

Figure 3: An overview of the key elements of the IAA governance and operational structure.

Future Fields

Over the coming months, partnership agreements for founding IAA members are being finalized, along with work agreements that clarify specific data sharing, joint code development and research activities. The G.E.M.S™ platform is on track for an open-beta release later this year followed by a version 1.0 release in Spring of 2018. Beginning over two years ago, the University of Minnesota made a long-term, strategic commitment to this agroinformatics area, including providing resources for the core G.E.M.S™ development team and finalizing hires for multiple new faculty positions with an agroinformatics focus. While this paper used the IAA to demonstrate the potential value of the G.E.M.S™ platform, there are boundless opportunities to develop additional functionality into the ever-growing suite of GEMSTools™ and to further refine and develop the GEMShare™. For example, we are working with researchers in the health informatics field to deploy the G.E.M.S™ platform and we envision other application areas that will also be able to use this platform without the requirement of major modifications.

Remaining agile, innovative and responsive to partner and market demands is part of our development DNA. Particular emphasis is being given to developing a federated network of IAA partners and G.E.M.S™ clients throughout the world. Our focus of attention over the years, and hopefully decades, ahead will be to secure additional, ideally more strategic programmatic funding (distinct from shorter-term project funding) to support new communities of research interests spanning public and private agencies, underpinned by the novel and ever-evolving data sharing and analytic platform that G.E.M.S™ represents.

References

Pardey, P.G., C. Chan-Kang, S.P Dehmer and J.M. Beddow. "Agricultural R&D is on the Move." *Nature* 15 (537)(September 2016): 301-303. Available at http://www.nature.com/news/agricultural-rd-is-on-the-move-1.20571.

Pardey, P.G. and J.M. Beddow. *Revitalizing Agricultural Research and Development to Sustain US Competitiveness.* Policy Brief, Philadelphia, PA: Farm Journal Foundation, February 28, 2017. Available at http://www.farmersfeedingtheworld.org/assets/7/6/revitalizingagresearch_print.pdf.